CRANFIELD UNIVERSITY

Hessa Abdulaziz Alsanousi

Enhancing bank's profitability by applying machine learning techniques on financial data to intelligently predict customer behaviour towards the use of electronic channels

School of Aerospace, Computational Engineering Sciences
PhD Thesis

PhD
Academic Year: 2018 - 2021

Supervisor: Dr. Irene Moulitsas
Associate Supervisor: Dr. Salvatore Fillippone
September 2021

CRANFIELD UNIVERSITY

School of Aerospace, Computational Engineering Sciences

PhD Thesis

Academic Year 2018 - 2021

Hessa Abdulaziz Alsanousi

Enhancing bank's profitability by applying machine learning techniques on financial data to intelligently predict customer behaviour towards the use of electronic channels

Supervisor: Dr. Irene Moulitsas
Associate Supervisor: Dr. Salvatore Filippone
September 2021

This thesis is submitted in partial fulfilment of the requirements for the degree of PhD

© Cranfield University 2021. All rights reserved. No part of this publication may be reproduced without the written permission of the copyright owner.

ABSTRACT

Technology is evolving rapidly, and this represents a huge prospect for investment in business. In the banking industry, new technologies have led to the creation of new electronic communication channels for customers. Therefore, banks need to do consider how to make the most of these channels. This will help them create prediction models to implement better strategies and improve their decision making. In the long run, this will help them to decrease costs and increase revenues. The aim of this study was to determine the appropriate electronic channels for specific customers based on their information. It used a big data prediction model to predict the best online channel for banking customers. The point of the model was to help banks understand their customers' preferences, thereby increasing customer satisfaction, productivity, and profits.

I obtained a substantial amount of financial data from a minimum of 100,000 customers from ten local banks in Kuwait. The data covered a period of ten years. The independent variables I studied were age, gender, number of current accounts, number of savings accounts, number of deposit accounts, income, number of consumer loans, number of instalment loans, credit card limit, outstanding credit card balance, length of relationship with the bank, continent, and nationality. The dependent variables were call centres, websites, and mobile applications. Given the size and type of data, I used machine learning. I used the Statistics and Machine Learning Toolbox, Financial Toolbox, and other functions in MATLAB to run a multinomial logistic regression analysis. I examined four different methods for analyzing the data and chose the one that was most appropriate: multinomial logistic regression. I also considered whether there was any correlation between the independent variables. I discovered one significant correlation between credit card limit and outstanding credit card balance. I explored this finding further by considering the time of execution and the key performance measures. Consequently, I decided to keep both variables in my study. I also studied the relation between the volume of training data and the accuracy of the model to see how sensitive the models were to variations in data size. The results confirmed that my method performed stably across the different

sample sizes. I addressed the issue of overfitting by running a sub-sample regression and comparing it to the original model.my results indicated that the model was not overfitted.

I differentiated between conventional banks and Islamic banks. This distinction had not been made by previous studies, and my outcomes provided more information about the difference between Islamic banks and conventional banks. I found that there were very few differences between the customers of conventional banks and those of Islamic banks. However, I found that male customers of Islamic banks tended to use internet more often than female customers, who tended to use the mobile applications. I also found that older customers tended to use call centres as their primary way of communicating with their bank. The results showed that clients with more current accounts and savings accounts were more likely to use mobile applications. However, one unexpected finding was that clients with more deposit accounts were more likely to use the internet or call centres rather than mobile applications. The findings also demonstrated that clients with higher incomes were more likely to use mobile applications than other communication platforms. In a couple of banks, customers with more loans tended to use call centres as their primary means of communication. The type of loan had no significant impact on their choices. Also, I found that customers who had been with their bank for longer were more likely to use call centres as their primary communication source. Finally, in a few banks, the results showed that a client's continent or nationality had no significant impact on their preference for a particular communication channel. These are very important findings that could change how banks operate. They will have a positive influence on decision-making and strategies in the banking industry.

Keywords:

MATLAB, Banking data, Banking prediction model, Mobile banking, Online banking, Multinomial logistic regression.

ACKNOWLEDGEMENTS

First, thank you God for granting me the blessing of being accepted to Cranfield University as a PhD student, for helping me to overcome all the difficulties I faced, and for giving me the power to finish writing this thesis.

I would like to thank the Central Bank of Kuwait and especially the Governor, Dr. Mohammed Alhashel, for their support. They granted me a fully funded scholarship and helped me to collect financial data from local banks in Kuwait. They were always available and provided invaluable help, support, and encouragement in all circumstances. I would also like to express my sincere gratitude to my advisor Dr. Irene Molitsas, who offered continuous support throughout my PhD studies and my related research. I would like to thank her for her motivation and for sharing her considerable knowledge. Her weekly meetings and constant availability helped me throughout my research and writing. I would also like to thank my supervisor Dr. Salvatore Fillippone for his support, insightful comments, and encouragement, and for his hard questions which inspired me to widen my research. My sincere thanks also go to my best friend, my central bank co-worker and PhD colleague, Najla Albarrak. Without her help and support, this work could not have happened. I would also like to thank the Head of the Centre for Computational Engineering Science Prof. Karl Jenkins for his eagerness to provide a healthy educational environment and for coordinating a monthly forum in which students and doctors could exchange experiences and information. I want to thank my university friends in the office I share for all their help and support. I also want to take this opportunity to express my sincere gratitude to the members of my doctoral committee for critically reviewing my work every year and providing me with very helpful comments.

On a personal note, a big thank you to my father, who has always encouraged and supported me. I want to tell him: thank you for believing in me; I hope I can make you proud, just like I feel proud of being your daughter. Another big thank you to my wonderful mother for enduring my complaints and responding to my many calls during the PhD. Thank you for your constant prayers and support. A big thank you goes to my husband for his constant support, encouragement, and

patience, which helped me emotionally and improved the quality of this work. Finally, thanks to my whole family for being proud of me and for their constant interest in the progress of my work.

TABLE OF CONTENTS

ABSTRACT i
ACKNOWLEDGEMENTS iii
LIST OF FIGURES x
LIST OF TABLES xii
1 INTRODUCTION 1
1.1 The concept of this research project 1
1.2 Research outline 5
1.2.1 Introduction 5
1.2.2 Preview of Literature review 6
1.2.3 Preview of methodology and results 8
1.3 Aims and objectives 14
1.4 Publications and Conferences 16
1.5 Summary 17
2 LITERATURE REVIEW 18
2.1 Introduction 18
2.2 Defining big data 18
2.3 The importance of big data in business 23
2.4 Big data techniques and software 33
2.5 Machine learning 37
2.6 MATLAB (matrix laboratory) software 43
2.7 Online banking 46
2.8 Summary 64
3 METHODOLOGY 65
3.1 Introduction 65
3.2 Data selection and collection 65
3.3 Data limitations 69
3.4 Data type and size 71
3.5 Code documentation 103
3.5.1 Data organization 103
3.5.2 Banking type prediction modelling 107
3.5.3 Banking type prediction model comparison 109
3.6 Methods 111
3.6.1 Linear regression formula 114
3.6.2 Feature scaling 118
3.6.3 Learning rate (alpha) 119
3.6.4 Multinomial logistic regression 127
3.6.5 Binary decision tree for regression 133
3.6.6 Support vector machine regression 136
3.7 Summary 140

4 RESULTS AND DISCUSSION....143
4.1 Introduction....143
4.2 Correlation matrix....143
4.3 Data training and testing....154
4.4 Sensitivity analysis....157
4.5 Multinomial logistic regression....159
4.6 Independent variable 1: age....160
4.6.1 Bank 1....161
4.6.2 Bank 2....161
4.6.3 Bank 3....161
4.6.4 Bank 4....162
4.6.5 Bank 5....162
4.6.6 Bank 6....162
4.6.7 Results summary for independent variable 1: age....163
4.6.8 Overall analysis of independent variable 1: age....164
4.7 Independent variable 2: gender....165
4.7.1 Bank 1....165
4.7.2 Bank 2....166
4.7.3 Bank 3....166
4.7.4 Bank 4....166
4.7.5 Bank 5....167
4.7.6 Bank 6....167
4.7.7 Results summary for independent variable 2: gender....168
4.7.8 Overall analysis of independent variable 2: gender....168
4.8 Independent variable 3: number of current accounts....170
4.8.1 Bank 1....170
4.8.2 Bank 2....171
4.8.3 Bank 3....171
4.8.4 Bank 4....171
4.8.5 Bank 5....172
4.8.6 Bank 6....172
4.8.7 Results summary for independent variable 3: number of current accounts....173
4.8.8 Overall analysis of independent variable 3: number of current accounts....174
4.9 Independent variable 4: number of savings accounts....174
4.9.1 Bank 1....174
4.9.2 Bank 2....174
4.9.3 Bank 3....175
4.9.4 Bank 4....175
4.9.5 Bank 5....175
4.9.6 Bank 6....176

4.9.7 Results summary for independent variable 4: number of savings accounts 176
4.9.8 Overall analysis of independent variable 4: number of savings accounts 177
4.9.9 Comparative analysis of current accounts and savings accounts 178
4.10 Independent variable 5: number of deposit accounts 179
4.10.1 Bank 1 179
4.10.2 Bank 2 180
4.10.3 Bank 3 180
4.10.4 Bank 4 180
4.10.5 Bank 5 181
4.10.6 Bank 6 181
4.10.7 Results summary for independent variable 5: number of deposit accounts 181
4.10.8 Overall analysis of independent variable 5: number of deposit accounts 182
4.11 Independent variable 6: income 183
4.11.1 Bank 1 183
4.11.2 Bank 2 184
4.11.3 Bank 3 184
4.11.4 Bank 4 185
4.11.5 Bank 5 185
4.11.6 Bank 6 185
4.11.7 Results summary for independent variable 6: income 186
4.11.8 Overall analysis for independent variable 6: income 187
4.12 Independent variable 7: number of consumer loans 188
4.12.1 Bank 1 188
4.12.2 Bank 2 189
4.12.3 Bank 3 189
4.12.4 Bank 4 189
4.12.5 Bank 5 190
4.12.6 Bank 6 190
4.12.7 Results summary for independent variable 7: number of consumer loans 191
4.12.8 Overall analysis for independent variable 7: number of consumer loans 192
4.13 Independent variable 8: number of instalment loans 192
4.13.1 Bank 1 192
4.13.2 Bank 2 193
4.13.3 Bank 3 193
4.13.4 Bank 4 193
4.13.5 Bank 5 194

4.13.6 Bank 6 194
4.13.7 Results summary for independent variable 8: instalment loans 194
4.13.8 Overall analysis for independent variable 8: number of instalment loans 195
4.13.9 Comparative analysis of number of consumer loans and instalment loans 196
4.14 Independent variable 9: credit card limit 197
4.14.1 Bank 1 197
4.14.2 Bank 2 198
4.14.3 Bank 3 198
4.14.4 Bank 4 198
4.14.5 Bank 5 199
4.14.6 Bank 6 199
4.14.7 Results summary for independent variable 9: credit card limit 199
4.14.8 Overall analysis for independent variable 9: credit card limit 200
4.15 Independent variable 10: outstanding credit card balance 201
4.15.1 Bank 1 201
4.15.2 Bank 2 202
4.15.3 Bank 3 202
4.15.4 Bank 4 202
4.15.5 Bank 5 203
4.15.6 Bank 6 203
4.15.7 Results summary for independent variable 10: outstanding credit card balance 203
4.15.8 Overall analysis for independent variable 10: outstanding credit card balance 204
4.16 Independent variable 11: duration of relationship with the bank 205
4.16.1 Bank 1 205
4.16.2 Bank 2 205
4.16.3 Bank 3 205
4.16.4 Bank 4 206
4.16.5 Bank 5 206
4.16.6 Bank 6 206
4.16.7 Results summary for independent variable 11: duration of relationship with the bank 207
4.16.8 Overall analysis for independent variable 11: duration of relationship with the bank 208
4.17 Prediction model 208
4.18 Sample results of the prediction model 209
4.19 Results for people from different continents and countries 210
4.20 Further analysis of mobile applications 211
4.21 Overfitting 213

4.22 Summary 215
5 CONCLUSION 221
5.1 Summary of the research topic, rationale, and findings 221
5.2 Contributions and achievements 223
5.3 Limitations 224
5.4 Future work 225
REFERENCES 227
APPENDICES 239
Appendix A Multinomial logistic regression results 239

LIST OF FIGURES

Figure 1-1: Digital banking use in the United Kingdom from 2007 to 2020 (source: Statista, 2020) 3

Figure 2-1: The four V's (source: Ernst & Young, 2014) 19

Figure 2-2: The definition of big data (source: Simons et al., 2013) 20

Figure 2-3: Progressive big data expansion (source: Vasarhelyi et al., 2015) .. 26

Figure 2-4: New accounting and finance professional hybrids (source: Chua, 2013) 28

Figure 2-5: Hadoop framework (source: Honest & Patel, 2016) 34

Figure 2-6: Machine learning techniques (source: Mathworks, 2016) 38

Figure 2-7: Improvement of models (source:Mathworks, 2016a) 40

Figure 2-8: Workflow at a glance (source: Mathworks, 2016c) 41

Figure 2-9: Changes in the global cloud computing marked in USD billions (source: Bataev, 2018) 42

Figure 2-10: Examples of interview questions 48

Figure 2-11: Factors affecting the use of electronic banking 49

Figure 2-12: The three most popular approaches to control for self-selection effects 55

Figure 2-13: Conceptual framework (source: Gensler et al., 2012) 56

Figure 4-1: Different methods used in the research 155

Figure 4-2: Key performance measures and validations 156

Figure 4-3: Key performance measures and validation graph 156

Figure 4-4: Sensitivity analysis using Bank 6 sample data (Islamic bank) 157

Figure 4-5: Sensitivity analysis using Bank 4 sample data (conventional bank) 158

Figure 4-6: Probability of using different electronic banking channels dependent on age 165

Figure 4-7: Probability of using different electronic banking channels dependent on gender and income 170

Figure 4-8: Probability of using different electronic banking channels dependent on number of current accounts and savings accounts 178

Figure 4-9: Probability of using different electronic banking channels dependent on number of current accounts and savings accounts in one Islamic bank 179

Figure 4-10: Probability of using different electronic banking channels dependent on deposit accounts..... 183

Figure 4-11: Probability of using different electronic banking channels dependent on income 188

Figure 4-12: Probability of using different electronic banking channels dependent on the number of consumer loans and instalment loans 197

Figure 4-13: Probability of using different electronic banking channels dependent on the number of consumer loans 201

Figure 4-14: Probability of using different electronic banking channels dependent on the duration of the customer's relationship with the bank 208

Figure 4-15: Mobile application services (conventional bank) 211

Figure 4-16: Mobile application services (Islamic bank) 212

Figure 4-17: Full sample vs sub-sample..... 215

LIST OF TABLES

Table 2-1: Structured vs unstructured data 20

Table 2-2: Unstructured data 21

Table 2-3: Data analysis pre- and post-big data (source: Richins et al., 2017) . 23

Table 2-4: Traditional vs technologically enhanced accounting measurements (source: Vasarhelyi et al., 2015) 25

Table 2-5: Illustrative uses of extended audit evidence (source: Vasarhelyi et al., 2015) 27

Table 2-6: Opportunities and challenges presented by big data in the finance profession (source: Chua, 2013) 27

Table 2-7: Big data topics that could be included in existing courses (source: Gamage, 2016) 29

Table 2-8: Technologies used to handle big data 33

Table 2-9: Big data as a service (source: Elshawi et al., 2018b) 35

Table 2-10: Summary of big data analysis framework (source: Elshawi et al., 2018b) 36

Table 2-11: Big data categories and example techniques 36

Table 2-12: Common hard and soft clustering techniques 39

Table 2-13: Common algorithms of classification and regression technique 39

Table 2-14: Descriptive statistics from 2003 to 2013 (source: Bloomberg 2014) 47

Table 2-15: Summary of all variables (source: Yang et al., 2018) 47

Table 2-16: Reliability coefficients for variables (source: Danyali, 2018) 50

Table 2-17: Hypothesis testing (source: Danyali, 2018) 51

Table 2-18: Summary of variables used in this study 52

Table 2-19: Regression results (source: Mohammad & Saad, 2011) 54

Table 2-20: Parametres of customer characteristics in propensity score model (source: Gensler et al., 2012) 59

Table 2-21: distribution of transactions by channel during June 2006 (source: Campbell & Frei, 2010) 60

Table 2-22: Six hypotheses 60

Table 3-1: Data available for electronic channels for a sample bank 70

Table 3-2. Bank 1, conventional, raw data 71
Table 3-3. Bank 2, conventional, raw data 76
Table 3-4. Bank 3, Islamic, raw data 78
Table 3-5. Bank 4, conventional, raw data 80
Table 3-6. Bank 5, Islamic, raw data 83
Table 3-7. Bank 6, Islamic, raw data 85
Table 3-8. Bank 1, conventional, data transferred from CSV to MATLAB 87
Table 3-9. Bank 2, conventional, data transferred from CSV to MATLAB 90
Table 3-10. Bank 3, Islamic, data transferred from CSV to MATLAB 93
Table 3-11. Bank 4, conventional, data transferred from CSV to MATLAB 95
Table 3-12. Bank 5, Islamic, data transferred from CSV to MATLAB 97
Table 3-13. Bank 6, Islamic, data transferred from CSV to MATLAB 99
Table 3-14. Linear regression notation table (univariate) 112
Table 3-15: Linear regression notation table (multivariate) 116
Table 3-16: Comparison of gradient descent and normal equation 123
Table 3-17: Bank 1 code execution time 131
Table 3-18: Bank 2 code execution time 131
Table 3-19: Bank 3 code execution time 132
Table 3-20: Bank 4 code execution time 132
Table 3-21: Bank 5 code execution time 133
Table 3-22: Bank 6 code execution time 133
Table 3-23. Example of results of the binary decision tree for Bank 7 135
Table 3-24: Description of nodes in decision tree regression 136
Table 3-25: Bank 7, ε results 140
Table 4-1: Bank 1 correlation matrix 145
Table 4-2: Bank 2 correlation matrix 146
Table 4-3: Bank 3 correlation matrix 147
Table 4-4: Bank 4 correlation matrix 148
Table 4-5: Bank 5 correlation matrix 149

Table 4-6: Bank 6 correlation matrix 149

Table 4-7: Multinomial logistic regression with and without the variable "outstanding credit card balance" 151

Table 4-8: Bank 2 and Bank 3 coefficients including and excluding the variable "outstanding credit card balance" 152

Table 4-9: Bank 2 and Bank 3 p-values including and excluding the variable "outstanding credit card balance" 152

Table 4-10: Bank 2 and Bank 3 standard errors, including and excluding the variable "outstanding credit card balance" 153

Table 4-11: Multinomial logistic regression results for independent variable 1: age 163

Table 4-12: Multinomial logistic regression results for independent variable 2: gender 168

Table 4-13: Multinomial logistic regression results for independent variable 3: number of current accounts 173

Table 4-14: Multinomial logistic regression results for independent variable 4: number of savings accounts 176

Table 4-15: Multinomial logistic regression results for independent variable 5: number of deposit accounts 181

Table 4-16: Multinomial logistic regression results for independent variable 6: income 186

Table 4-17: Multinomial logistic regression results for independent variable 7: number of consumer loans 191

Table 4-18: Multinomial logistic regression results for independent variable 8: number of instalment loans 195

Table 4-19: Multinomial logistic regression results for independent variable 9: credit card limit 199

Table 4-20: Multinomial logistic regression results for independent variable 10: outstanding credit card balance 203

Table 4-21: Multinomial logistic regression results for independent variable 11: duration of relationship with the bank 207

Table 4-22: Three sample results from the prediction model 209

Table 4-23: Use of electronic banking channels by people from different continents 210

Table 4-24: Use of electronic banking channels by people of different nationalities 210

Table A-1: Brief description of common big data techniques (source: Gepp et al., 2018) ..... **Error! Bookmark not defined.**

Table B-1: Multinomial logistic regression (coefficients-relative log odds)..... 239

Table B-2: Multinomial logistic regression (coefficients-relative risk exp)..... 240

Table B-3: Multinomial logistic regression (p-value)..... 241

Table B-4: Multinomial logistic regression (standard error) 242

Table B-5: Multinomial logistic regression (95% confidence lower and upper limits) 243

1 INTRODUCTION

1.1 The concept of this research project

I live in a time of rapid and unpredictable technological progress. The development of technology holds significant investment potential for investors around the world. The technology sector has become the most dominant market sector, surpassing both the industrial and financial sectors. Furthermore, technology itself also filters into these other sectors. Results of a 2019 study from Statista show that $3,360 billion has been spent on technology worldwide, accounting for almost 58% of the total global population. The growth of the technology sector is fundamentally linked to the Internet, the core pillar of the digital age, currently boasting 4.4 billion active users around the world (Industry Today, 2019). The current study makes a contribution to research that investigates how the financial industry is being affected by technology.

In the financial sector, technological advances are changing how banking is done, introducing new digital tools like call centres, websites, mobile banking applications, digital currencies, online investment platforms, and new payment systems. These technological developments have also led to a significant increase in the amount of financial data available. Financial institutions have the opportunity to use these data in their predictions (for prices, loans, and targets, for example) as well as in their marketing operations to advertise their products and services (Frew, 2016).

In general, the use of technology in financial services is referred to by the phrase ‘FinTech.’ FinTech companies aim to keep up with the most recent developments in technology to provide the best financial services for their customers (Nick, 2018). A report compiled by PwC shows that 77% of financial institutions are planning to increase their internal efforts to innovate, with many of them focusing on the disruptive environment of FinTech (Pwc, 2017). More specifically, certain important financial activities are seeing increased uses of technology, namely customer services, online banking, and fraud detection.

In the banking and finance sectors, customer service is very important. Since these sectors involve money, staff must be professional and well trained to deal with clients. Staff must also be skilled, capable of handling any problem that arises. FinTech must not be a distraction for staff but must help them improve their customer service by combining technology with a human touch.

Traditionally, banks used to consist of non-virtual branches where staff were employed. The clients of banks would go to a branch to make withdrawals, deposits, transfers, financial inquiries, and financial requests. They would interact with the staff members while performing these transactions. Nowadays, however, due to the evolution of technology, banks have changed their strategies; they have reduced the number of branches and staff, substituting them for online banking services. One of the most important reasons for this is cost control. Another important reason, though, is the ease of use. Customers can now do all their banking transactions virtually simply by pushing a button. They can access their bank account via a website, a call centre, or a mobile application.

Another major use of technology in banking is in the detection and prevention of fraud. Digital fraud protection systems make it easier to track transactions that might be fraudulent. Digital tools can also be used to track the records of a victim and a perpetrator, predicting the likelihood of fraud based on past information.

These examples (customer service, online banking, and fraud detection) are just some of the many ways that technology is being used to change how banks and financial institutions operate (Nick, 2018).

The inescapable departure of traditional banking merged with the establishment of more complicated channels of electronic banking services, it has its effect to shape customer opinions and expectations of what their banks can and must provide.

One of the most challenging tasks for banks and financial institutions, given the rapid developments in FinTech, is to provide satisfactory customer services. It is difficult for traditional banks to keep up with the pace of change when it comes to customer opinions and demands. These banks must update their services

consistently, paying attention to their competitors. Nevertheless, research has shown that one way in which traditional banks can compete with FinTech firms is to improve their customer experience (Mistry, 2021).

Banking and financial services businesses were among the first companies to adopt technological advances. Their use of digital tools extends back at least as far as the introduction of automated teller machines (ATMs) and call centres. Subsequently, secure websites began to be used for online financial transactions. Nowadays, mobile applications are also a very important channel through which customers engage in digital banking.

In the United Kingdom, the use of electronic banking increased from 2007 to 2020 (Statista, 2020). Figure 1-1 shows that in 2007, 30% of respondents stated that they had accessed their bank account virtual in the three months before the survey. This percentage increased to 76% in 2020. The research behind this report also shows that this increase was mainly due to the increased use of mobile applications (Statista, 2020).

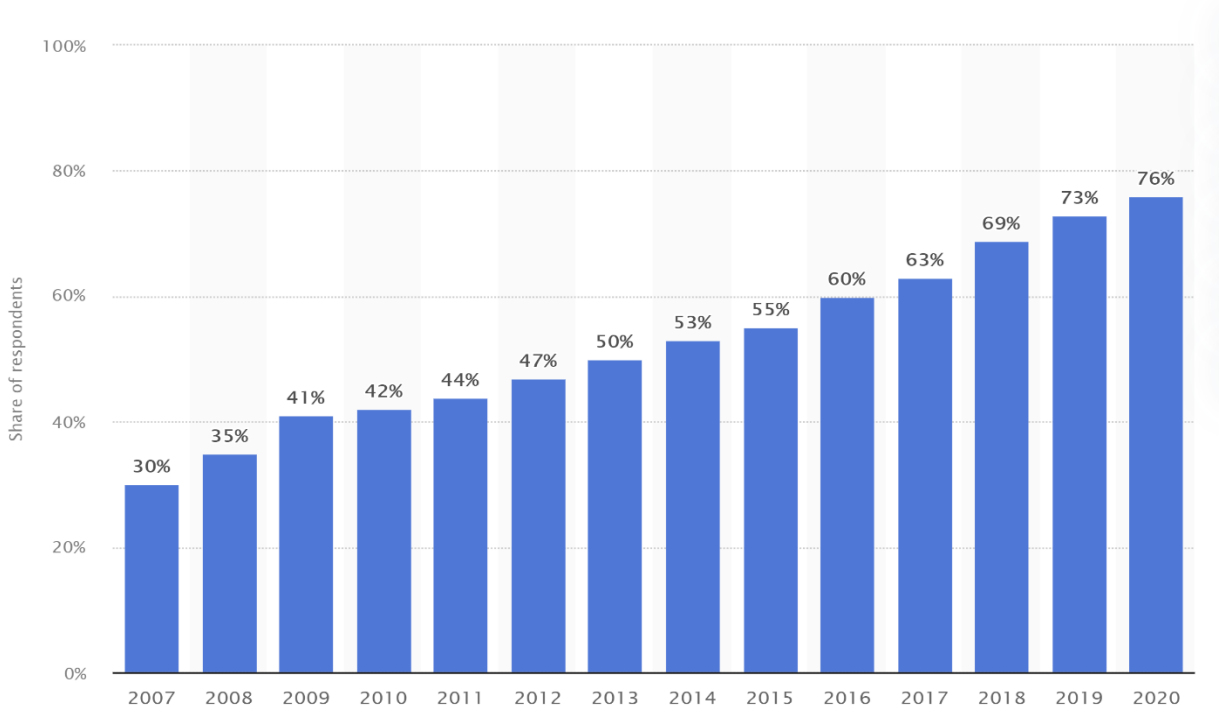


Figure 1-1: Digital banking use in the United Kingdom from 2007 to 2020 (source: Statista, 2020)

FinTech provides local banking services. For local financial institutions in Kuwait, the digital revolution has been a key part of productive growth within the New Kuwait 2035 long-term development strategy. The executive director of supervision at the Central Bank of Kuwait stated: “It is clear that I will need to embrace the digital banking transformation... FinTech will shape financial services, and banks must keep up with the rapid rate of change in the global banking environment” (Oxford Business Group, 2019).

The evolution of technology has made it incumbent upon banks to keep up with the pace of change. This is also necessary so that they can make use of a large amount of financial data to provide better service to their customers. The data can be used to build models for predicting customer behaviour, leading to better decision-making.

The current study focuses on electronic banking, which is an essential part of the banking sector. It aims to improve the efficiency of banks, leading to increased revenue and decreased costs. More specifically, it aims to provide banks with information about which online channel would be most suited to a particular customer based on their characteristics and their portfolio. The hope is that this will help banks to craft longer and more productive relationships with their customers, thereby encouraging those customers to use more products and services. As mentioned, this will increase revenue, but it will also decrease the costs of advertising and other expenses by allowing banks to determine the appropriate electronic channel for their customers. By providing banks with in-depth knowledge about whether their customers prefer to visit a branch or use digital services, the study hopes to reduce the amount of money spent on advertising services to customers who are more likely to other ones. This would also reduce the amount spent on employees and the maintenance of branches.

The study makes an intervention into the current research on specific forms of electronic banking (such as call centres, online banking, and mobile applications). This is important because electronic banking has now become an essential mechanism for carrying out transactions and exchanging financial data. This is

because it is generally faster, easier, and safer, particularly under the current circumstances created by the COVID-19 pandemic.

To make the most of the opportunities presented by technology, banks must realize that all customers need to embrace the new technologies. Banks need to be aware of the differences between customers and target their specific needs, rather than just following the latest FinTech trends.

In this study, I distinguish between two types of banks: Islamic banks and conventional banks. Islamic banks are interest-free, making them different from conventional banks. Islamic banks work under different rules and have different risk profiles compared to conventional banks. Islamic banks are regulated through a combination of centrally run, governmental banks and the Shariah Supervisory Board. These bodies authorize the products and services provided by Islamic banks and ensure that they follow specific rules. They also set the principles for Islamic banks. For example, Islamic banks have to pay registration fees because they use asset-based banking. Furthermore, Islamic banks must own the goods that they sell, thereby increasing the cost (Salman & Nawaz, 2018).

The main aim of this study is to create a prediction model, based on customer information, to predict the electronic channels that customers will use. This will show banks what to focus their investments on and which channels they should advertise to which customers.

1.2 Research outline

1.2.1 Introduction

I began this research by reading previous studies, focusing particularly on the use of big data in the business and financial sectors. I studied the techniques and software used to deal with big data in electronic banking, most notably machine learning. After deciding on my methodology, I investigated the financial data for six banks, which led to some significant results.

1.2.2 Existing Literature

The majority of the papers that I examined stated that the concept of big data can be defined using what is known as the “four V’s”: velocity, volume, veracity and variety (Zhang, Yang, & Appelbaum, 2015). Other studies argued that there are three overlapping kinds of big data: data related to an enterprise, financial data, and other external, unstructured data. I also found that as the difficulty and complexity of data increase, the intensity of the analysis increases as well. Meanwhile, more detailed analysis produces better outcomes (Simons, Masamvu, & Parks, 2013).

Big data can be structured or unstructured. Structured data is data that has been created by the systems of an organization. This type of data is managed, organized, and easy to deal with. Unstructured data, on the other hand, normally comes from outside an organization (Richins et al., 2017). Both types of data (structured and unstructured) must be used effectively and efficiently to help businesses gain a competitive advantage (Md Mujeeb, Praveen Sam, & Madhavi, 2021). As such, companies across all sorts of different industries must know how to obtain and use big data to succeed.

Experts in the financial sector are in charge of processing, assessing, reporting, and evaluating big data sets. In general, however, people in the banking sector do not have a background in computer science. Also, banks must invest in the technology required to acquire, collect, store, and analyse big data. Banking experts need to use their skills to identify the relevant information, allowing the technology to do the rest. Over time, this will improve the decision-making in financial businesses (Gamage, 2016). Despite the advantages of big data, it may also harm businesses. For example, if people are unsure about recent developments in data regulation and protection, then this might hamper the progress of businesses (Chua, 2013). Gamage (2016) has claimed that the lack of data scientists around the world is a major problem. In response, individuals across all sectors will need to undergo training in data science so that they can make the most of advanced technologies.

The mechanisms used to handle big data differ from traditional computing techniques. This is due to the need to mine, process, and store much larger amounts of data. The major difficulty when it comes to processing big data is ensuring the speed, quality, and security of the process (Honest & Patel, 2016). I wanted to understand what big data is and why it is so important for businesses, so I researched how best to obtain, process, store, and analyse big data.my research showed that statistics, computer science, machine learning, and many other tools can help researchers to handle big data sets (Elshawi, Sakr, Talia, & Trunfio, 2018a). The choice of platform for managing the data set depends on the size and type of data, as well as the desired outcome.

There are different types of technologies that can be used to handle big data. They vary in both style and cost. Given my background in accounting and financial analysts, I decided to use machine learning to produce meaningful outputs that would help in forecasting and decision-making. There are two types of machine learning: supervised and unsupervised. I found that supervised machine learning was best suited to my needs. Supervised machine learning involves models that use actual inputs and outputs to predict similar patterns (Mathworks, 2016c). Supervised machine learning uses two techniques: classification and regression (Mathworks, 2016a).

For my analysis, I found that MATLAB was the best software because it was easy for beginners. MATLAB stands for ‘Matrix Laboratory’. The toolboxes available in MATLAB were one why it was so attractive to us; they allowed us to solve several problems. Also, the software is powerful, comprehensive, and intuitive, even for people who are new to programming (Radi & ELHami, 2018). In my study, I used the Statistical and Machine Learning Toolbox, the Financial Toolbox, and several of the other available functions. I also read about other studies that had used MATLAB to gain insight into how to make the most of the software.

Several papers have examined the effect of electronic banking on the bank’s performance. For example, one paper has looked at this effect on banks in the Chinese market. The paper considered two stages in the development of electronic banking channels. It looked at the financial reports of five banks and

used some non-textual data like video records. The researchers tested five variables (financial ratios) at both stages of the investigation. They found that four of the financial ratios were improved when electronic banking was at the developed stage compared to when it was at the developing stage. One of the ratios was improved at the developing stage, but the difference was minimal (S. Yang, Li, Ma, & Chen, 2018). Similarly, Gautam (2012) has studied the effect of electronic banking on Indian banks. He concluded that electronic banking has a significant impact on customer satisfaction, leading to increased profits. A recent study by Rawwash et al. (2020) examined the factors that affect electronic banking in Jordan. The researchers used a sample of 300 Jordanian bank customers. They used questionnaires to assess the relationship between the customers' use of electronic banking and their perception of the trustworthiness, usefulness, security, intuitiveness, and privacy of the service. They found that there was a positive relationship between customers' use of electronic banking and the perceived usefulness, trustworthiness, intuitiveness, security, and privacy. They found that there was no relationship between the convenience of the service and the use of electronic baking. A further study conducted by Ikhsanova and Protsko (2016) examined the effect of certain variables on the total volume of payments made using electronic banking in Russian banks. The researchers found that a single unit increase in the average number of electronic banking services increases the total volume of payments by 44.38 million rubles.

In the literature review, I discuss several other studies that have examined the use of electronic channels in banking. From the overall review, I concluded that most of the studies I looked at were limited in the data they used, probably due to the need for confidentiality. I also noticed that some studies used a limited number of variables in their analysis. Finally, I found no studies of Kuwait, which is the focus of my research. In the subsequent sections and chapters, I outline how I tried to overcome the shortcomings of these previous studies.

1.2.3 Preview of methodology and results

The main purpose of this study is to determine the appropriate electronic channels for different customers based on their characteristics and their financial product

portfolio. I achieved this by building a model to predict customer preferences based on their personal information. I used former studies to discover the most relevant variables to study and selected those variables based on my 15 years of experience in the banking sector. During my research, I also detected some gaps in the literature, which I tried to overcome in my study. These gaps included limited data sets, limited numbers of variables, and a lack of models for predicting customer preferences based on their information. These are discussed in more detail in Chapter 3.

The study was sponsored by the Central Bank of Kuwait, which helped us to obtain financial data from local banks in Kuwait. During the preliminary stages, I signed non-disclosure agreements to ensure that all the data would be processed and stored securely. I obtained data concerning the entire customer base of ten local banks. Given the developments in technology and the shift towards electronic channels discussed above, I concentrated on different types of electronic channels rather than comparing customers' use of branches to their use of electronic channels (as had been the case in previous studies). I also used variables that had not previously been studied: the number of loans that customers had taken out; their credit card limit; their outstanding credit card balance (one previous study had used outstanding credit card balance as an independent variable but not credit card limit); income; continent; and nationality. I also used some of the variables that had been identified as important by previous studies, depending on the type and size of the data set (for example, statistical and historical data or questionnaires). I created a model to predict the data for each bank. After determining the dependent and independent variables, I obtained panel data from ten local banks in Kuwait for the period from 2008 to 2018. The data requested was on retail customers. Some of the data were obtained monthly; the rest was obtained daily. To obtain the data, I arranged appointments with all the local banks. Acquiring the data was not easy because the banks wanted to ensure that the data would be used confidentially. However, once I guaranteed this and explained how the study would positively affect their decision-making, the banks agreed to provide us with all the necessary data.

The Central Bank of Kuwait helped to ensure the cooperation of all the local banks. Nevertheless, there were some problems with the data obtained. For example, some of the banks provided us with outdated information, and one of the banks provided us with data from a shorter period than was necessary. These and other limitations are discussed in Chapter 3. As a result of these issues, I was unfortunately forced to exclude some of the banks from the study. I tried to make the most of the data available without sacrificing the quality of the research. The type and size of the data sets I used are also detailed in Chapter 3, as is the code documentation that I used.

Once I had acquired the final data sets, I began to investigate whether there was any correlation between the independent variables. Due to the large amount of data, the model would have been overly sensitive to highly correlated variables. Small variations in the data would have resulted in large residual errors. I analysed the correlations between the variables for each of the banks I studied. I noticed that many of the variables were weakly correlated, and a few were moderately correlated. But two variables (credit card limit and outstanding credit card balance) were moderately and highly correlated, respectively (Ratner, 2009). More specifically, the relation between these two variables was high in two of the banks, moderate in three of the banks, and low in one of the banks. The results were the same for Islamic and conventional banks. These variables were significant, so I decided to include them. Some customers had credit card limits that they did not use or paid the balances of their credits card immediately. I only decided to include these variables after first investigating their effect on the overall model; I found that they did not have a significant enough impact on the model to warrant excluding them from the analysis.

Based on the correlation analysis, I defined the variables that would be part of my study. I used four different methods to test them: multinomial logistic regression, multivariate linear regression, binary decision tree for regression, and support vector machine regression. These methods are discussed in more detail in Chapter 3).

To select the most suitable methods and ensure my results were valid, I used out-of-sample prediction accuracy as my key performance measure. I split the data into a training sample data set (2/3 of the total data) and out-of-sample data (1/3 of the total data). I then used each model on the training sample and predicted the outputs for the out-of-sample data. The deviation between the actual output and the predicted output was my key performance indicator in the study. The smaller the deviation, the better the performance of the model. The results showed that multinomial logistic regression was strong, but some alternatives could be used, such as multivariate linear regression and support vector machine regression.

I also assessed the relationship between the training data and the accuracy of the model, so that I could determine how sensitive the model was to variations in the size of the data set. I used data sets of different sizes, ranging from 55% of the total data to 80% of the total data, increasing the size of the data set by 5% at a time. For each sample data set, I calculated the accuracy of the out-of-sample prediction (the difference between the actual and predicted outputs). The results showed that both multinomial logistic regression and multivariate linear regression were stable across different sample sizes. Based on these tests and my sensitivity analysis, I decided to use multinomial logistic regression to test the effect of the independent variables on the dependent variables.

To run the multinomial regression analysis, I used the Statistics and Machine Learning Toolbox, the Financial Toolbox, and several other functions in MATLAB. The aim was to predict the rates at which customers would use specific electronic banking channels based on their age, gender, number of current accounts, number of savings accounts, number of deposit accounts, income, number of consumer loans, number of instalment loans, credit card limit, outstanding credit card balance, and length of their relationship with their bank.my key performance parameters were coefficients, p-values, standard errors, and the 95% confidence limits. I also tested the continent and nationality of the customers.

The results of my analysis indicated that as customers get older, they are more likely to use call centres as their primary means of communicating with their bank.

This result was anticipated, given my experience and my review of previous studies. I was able to infer from this result that banks should use their marketing and customer service tools to make encourage older people to use internet banking and mobile banking applications. To that end, I suggest that banks improve their explanatory resources, using video recordings to explain how to use their websites or mobile applications, for example.

The results also provided information about the difference between Islamic banks and conventional banks. I decided to study the difference between these banks because 50% of banks in Kuwait are now Islamic, and more than 42% of the total assets in Kuwaiti banks are in Islamic banks (Central Bank of Kuwait, 2019). Moreover, the number of Islamic banks is increasing in Kuwait and around the world. Currently, however, there is very little research into the difference between Islamic and conventional banks. The results of my research indicate that there are not many differences between the customers of conventional banks and those of Islamic banks. For all variables except gender, I found that customers in both banks had similar preferences when it came to banking channels. When it came to gender, I found that male customers at Islamic banks tended to use internet more frequently than women, whereas women tended to use the Mobile applications. A previous study by Riquelme & Rios (2010) had found that women were more interested in social norms, whereas men were more interested in the particular advantages provided by a communication channel. This implies that men at Islamic banks have yet to be convinced of the advantage of using mobile applications for banking. Islamic banks should be aware of this and encourage men to use their mobile applications.

I also expected that customers with more accounts (regardless of account type) would be more likely to use mobile applications. This was true for customers who had more current accounts and savings accounts; they were more likely to use mobile applications for their banking transactions. However, when it came to deposit accounts, I found that clients with more deposit accounts were more likely to use the internet or a call centre than a mobile application.

The results also showed that customers with higher incomes were more likely to use mobile applications than call centres or the internet. One possible explanation for this is that customers with higher incomes can afford to pay for mobile devices with more advanced networks and security. Therefore, they are likely to have more confidence in mobile banking. Another possible reason is that people with higher incomes be more likely to make more transactions. As such, they are more likely to appreciate the ease and efficiency of mobile applications.

In some banks, I found that customers with more loans preferred to use call centres as their primary source of communication. This was true for customers who had more consumer loans and those who had more instalment loans.

In general, my findings showed that customers do not necessarily prefer using a particular service just because it is the most technologically advanced. Customers with deposit accounts and loans, for example, do not need to make that many daily transactions, so they are less likely to require the most advanced digital technology. However, customers with current accounts tend to make many more daily transactions. Therefore, they are more likely to prefer mobile applications because they are easy to use. The same is true for customers with higher incomes and higher credit card limits.

Our findings also indicated that customers who had been with their banks for a long time were more likely to use call centres as their primary means of communication. This is probably because long-term customers tend to have a stronger relationship with their bank and its staff. Banks should bear this in mind and consider improving their digital services for long-term customers. For example, they could provide those customers with an electronic customer service agent to make them feel more comfortable and satisfy their needs.

The findings showed that a client's home continent or nationality had no significant effect on their preferred communication channel. I did not expect this. Since the majority of foreign workers in Kuwait (around 65%) have low incomes and low-to-moderate levels of education, I expected their banking preferences to vary based on their nationality (Statista Research Department, 2020). The results did not

back this up, so I expanded my research to look at continents, but this also did not provide any significant results.

Given the large sample size, I ran a sub-sample regression to ease my concerns about overfitting. By comparing the results to the full sample, I concluded that the model was not overfitted.

The results of the study showed that there is an ongoing trend towards the use of mobile banking applications. Therefore, I also investigated how customers tend to use their mobile applications. I found that more than 80% of the usage of mobile banking applications is connected to bank transfers. I proposed some solutions for how to encourage the use of mobile applications for payments. The details of this are provided in Chapter 4.

1.3 Aims and objectives

This study aimed to determine the appropriate electronic channel for the customers of several banks. The study used big data to build a model to predict which online channel a customer would be most likely to prefer, based on their information. The purpose of this was to help banks to understand their customers' preferences better, leading to greater customer satisfaction, as well as increased productivity and profits.

Following my review of the literature, I decided that electronic banking (and mobile banking in particular) were the most important topics in the banking sector. I identified the achievements of previous investigations and determined the most useful variables for my study. I introduced new variables (number of loans, credit card limits, income, continent and nationality) and used other variables that had been studied by other researchers.

Our main objectives were as follows. First, I defined my research question based on my experience in the banking sector. After the literature review, I came up with a focused question by the end of the first year of my PhD. The research question was: "what are the factors that influence customer behaviour when it comes to electronic banking channels?" Second, by the end of December 2019, I identified

the hypotheses that I wanted to test. Again, this was based on my experience and my review of the literature. The following hypotheses were proposed:

- Younger customers prefer using mobile applications more than older customers.
- There are no differences between men and women in their preferences for different electronic channels.
- A customer with more accounts (current accounts, savings accounts, deposit accounts) will be more likely to use mobile applications.
- A customer with more loans will be more likely to use mobile applications than other electronic channels.
- A customer with a higher credit card limit or a larger outstanding balance on their credit card will be more likely to use a mobile application.
- A customer who has been with their bank for longer will be more likely to use a call centre than a mobile application.
- A customer with a larger income will be more likely to use a mobile application.
- The type of bank (Islamic or conventional) will affect customer behaviour.

Third, I determined the type and size of the data that I required. I asked the banks to send us the data by 31 December 2019. I contacted ten local banks in Kuwait and asked them for financial data for their customers. I liaised with the Central Bank of Kuwait, who helped us with my request. Then, I travelled to Kuwait to gather the data and meet with the banks several times.my objectives for the data collection were as follows:

- Request panel data for 11 years (from 2008 to 2018) from ten Kuwaiti banks.
- Request data for a minimum of 100,000 customers per bank.
- Request the following information about customers: age, gender, length of relationship with the bank, number of current accounts, number of savings accounts, number of deposit accounts, number of consumer loans, number of instalment loans, credit card limit, outstanding credit card balance, continent, nationality, and income.

- Collect the initial data and work with the banks' contracts departments to handle any legal issues.
- Create several copies of the data and store them in safe places.
- Find an automated way (using statistics and the Machine Learning Toolbox in MATLAB) to read, understand, and clean the data.
- Limit the data to six banks following my assessment of the limitations.

Following the literature review, I applied the relevant method of analysis (having chosen between multinomial logistic regression, multivariate linear regression, binary decision tree for regression, and support vector machine regression). I also assessed my key performance measurement (using out-of-sample prediction). I decided on my methodology in June 2020. Subsequently, I analysed the results of the multinomial logistic regression. Finally, I built a prediction model for each of the banks, based on their customers and their historical data. This allowed each bank to predict which online channel their customers would choose based on their information. I met all my objectives for the study and, in some areas, even exceeded them. For example, during the data collection, I obtained data about the banks' entire customer bases, even though I had only intended to obtain information about 100,000 customers per bank.

1.4 Dissemination through Publications and Conferences

Following research articles were published as part of this PhD work:

1. Alsanousi H, Albarrak N, Moulitsas I & Filippone S (2021) Understanding customer behaviours toward the use of electronic banking given customer characteristics and financial portfolios, International Journal of Business and Social Science, 12 (1, January 2021) 54-64.
2. Alsanousi H., Albarrak N., Moulitsas I., Filippone S. (2022) Using client's Characteristics and Their Financial Products to Predict Their Usage of Banking Electronic Channels. In: Yang XS., Sherratt S., Dey N., Joshi A. (eds) Proceedings of Sixth International Congress on Information and Communication Technology. Lecture Notes in Networks and Systems, vol 217. Springer, Singapore.

The following conferences were attended in relate to the work done for this PhD:

1. Central Bank of Kuwait International Banking Conference: Shaping the Future, Kuwait.
2. The Financial Stability Institute (FSI) 20th conference conducted by the Bank of International Settlement-Basel, Switzerland.
3. Sixth International Congress on Information and Communication Technology-London, United Kingdom.

1.5 Summary

Technology is progressing rapidly, and has become a major market in today's world, moreover, it has penetrated every single sector.

The Financial sector and more specifically the banking sector need to keep pace with the fast growth in technology and include it in its operations. The use of new and ever-growing technology will help them to better understand and meet their customer's needs and demands.

Banks have large amounts of financial data available, and by implementing this new technology, they will eventually improve customer satisfaction and overall bank performance.

In this study, I used a massive amount of financial confidential data from 6 Kuwaiti banks to build a prediction model for each bank. This prediction model will use customer information to predict the preferred electronic channel of each customer. Knowing this information will help banks to focus their investments on the right channels.

I started this research by reading previous literatures. In the literature review, I examined many studies that have investigated the use of electronic channels in banks. I realized the limitations in the use of data in other studies and also noticed that several studies used a limited number of variables. Also, there is no such prediction model found in previous studies. I have attempted to overcome these previous limitations in this research by studying four models and comparing between them through the key performance factor which is the out of sample

prediction accuracy. Multinomial logistic regression shows the best results compared to the others. After that, I have determined the parameters from the regression outcomes for my analysis. From examining the parameters, I started to introduce some results that will change how banks will operate and eventually help to increase their profit. In this chapter I also mentioned the overall aim of this study and its objectives. The aim is to determine the appropriate electronic channel for the customers of local Kuwaiti banks. This is by building a model using big data to predict which online channel a customer prefers, based on their information. This is to help banks to know their customers' favorites, leading to better customer satisfaction and increased productivity and profits.

2 LITERATURE REVIEW

2.1 Introduction

This chapter provides an overview of previous research into big data and its effect on business, focusing particularly on the fields of accounting and finance. It examines the data sources, variables, and methodologies employed in other studies of prediction models in the banking sector. It illustrates the importance of machine learning for dealing with evolving financial data sets.

2.2 Defining big data

Big data is distinguished from "normal" data by its volume (larger data set), velocity (generated very fast), and variety (different data formats). Big data is used by many institutions in different fields, such as by social media companies like Facebook and Twitter, or by hospitals and care providers. The mechanisms for handling big data differ from traditional computing techniques due to the need to mine, process, and store large amounts of data. Some of the major issues involved in processing big data include the speed, quality, and security of the process (Honest & Patel, 2016). According to Ernst & Young (2014), big data has four characteristics: volume, velocity, variety, and veracity (see Figure 2-1). Big data sets can contain structured, semi-structured, or unstructured data. It increases in size and changes much more rapidly than other forms of data. As a result, it is not trusted by all companies and institutions.

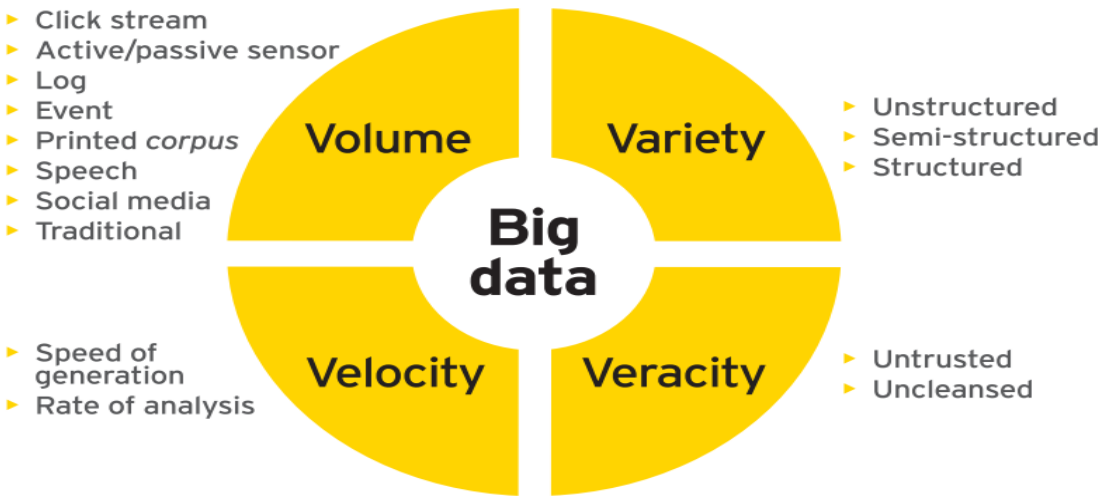


Figure 2-1: The four V's (source: Ernst & Young, 2014)

Another paper also stated that big data can be defined by the “four V’s”: volume, velocity, variety and veracity (Zhang et al., 2015). Gepp et al. (2018) defined these V’s in more detail: variety refers to the fact that big data can be structured or unstructured; velocity refers to the speed at which data is made available and updated; veracity refers to the accuracy and reliability of the data; and volume covers the size of the data.

Another study breaks big data down into three overlapping sets: financial data; other data related to the enterprise, including operational data; and other external unstructured data. The more complex the data, the higher the level of analysis. As shown by Figure 2-2, analysis that includes big data is likely to provide better results (Simons et al., 2013).

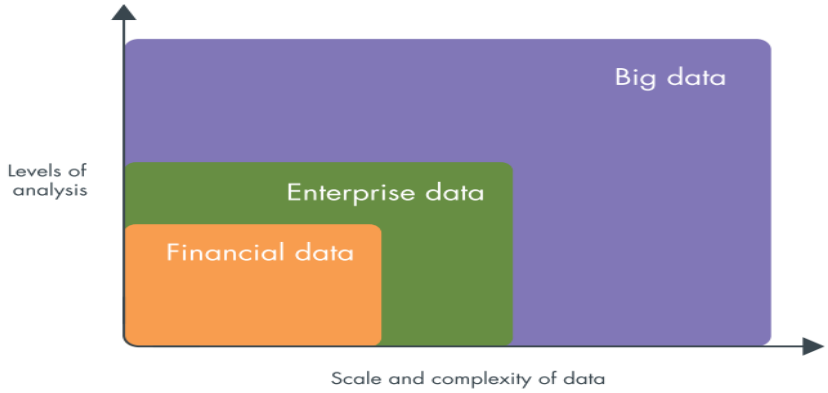


Figure 2-2: The definition of big data (source: Simons et al., 2013)

Big data includes structured and unstructured data (soft information). Structured data is data that has been produced by a firm's systems (such as its systems for processing transactions). Structured data is usually well organized and easy to process. Unstructured data comes from a range of other sources outside an organization (see Table 2-1; Richins et al., 2017).

Table 2-1: Structured vs unstructured data

	Structured	**Unstructured**
Definition	refers to data that are typically generated through the firm's transaction processing systems highly organized so that they can easily be included in a traditional relational database	data that originate from a wide variety of sources outside the firm data that appear in different formats, such as text, audio, and video lacking organization highly technical in nature

	Structured	Unstructured
Examples	point-of-sales systems (POS) inventory management system customer/supplier relationship management system (CRM and SCM)	Facebook Twitter YouTube

Computer algorithms are used to interpret different types of unstructured data, thereby providing the researcher with final figures that can be used in business operations (Warren, Moffitt, & Byrnes, 2015). Examples of unstructured data include images and videos, audio, and text (see Table 2-2):

Table 2-2: Unstructured data

	Unstructured Data		
Type	Images and Videos	Audio	Text
Examples	Recorded videos for monitoring employee productivity Videos for monitoring the condition of fixed assets	Internal workers' calls Customer phone calls Surveillance microphones	Text mining of email massages Social media SEC filling

	Unstructured Data		
Type	Images and Videos	Audio	Text
	Videos of interviews for monitoring emotions and nonverbal elements	Board of director and shareholder meetings	

Big data needs to be used productively to help businesses gain competitive advantage (Md Mujeeb et al., 2021). As Elshawi et al. (2018a) argue, one of the most important things for big data is that it must be valuable. Value is created by using the right models and techniques to identify useful trends and information within the data set. As Vasarhelyi et al. (2015) note, big data is now a common term that is widely used in different fields. This is due to recent revolutions in computing technology and telecommunications, leading to more data being produced at greater speeds. What counts as big data differs from one firm to another. It depends on the firm's information system; when the data is too big to be handled by its system, it counts as big data. To use this data as effectively as possible, a summary table of the different techniques for processing big data has been produced (see Table 2-3). This uses a conceptual framework for analysing data based on the data type and the analytical approach The model uses structured and unstructured data (Richins et al., 2017).

Table 2-3: Data analysis pre- and post-big data (source: Richins et al., 2017)

		Data: Structured	Data: Unstructured
Analysis Approach	Problem-Driven Analysis (Supervised Machine Learning)	Problem-driven analysis on structured data *(pre-Big Data)*	Problem-driven analysis on unstructured data *(post-Big Data)*
	Exploratory Analysis (Unsupervised Machine Learning)	Exploratory analysis on structured data *(post-Big Data)*	Exploratory analysis on unstructured data *(post-Big Data)*

2.3 The importance of big data in business

Experts in all areas need to be aware of importance of big data, which is set to transform the business world over the coming years. Organizations will rely heavily on dealing with big data for their success. In the financial sector, accountants will be particularly pivotal in this area, as they will be responsible for processing, measuring, reporting, and analysing big data. For example, one way in which big data will be influential is that it will allow auditors to test the entire population rather than using sample tests. Given the significance of big data, accounting courses should include training in big data and data science. This should include training in how to manage information to ensure positive decision-making. Therefore, the focus should not only be on the information systems used in accounting but also the methods for interpreting big data (Gamage, 2016). According to Bhimani & Willcocks (2014), when it comes to the financial sector, data scientists are useful because they have an understanding of the statistical methods used to handle big data, but accountants and financial experts are also fundamental because of their understanding of the business.

Accountants need to focus on the emergence of big data and its impact on accounting practices. Specifically, they need to consider how to accumulate, record, and report big data to improve performance. Accountants need to acknowledge new kinds of data and combine it with the existing data. Two of the main areas in which accountants can use big data is in managerial and financial accounting. Managerial accounting is the use of accounting data to help managers fulfil the goals and objectives of an organization. Accountants in this field are responsible for keeping the behaviour of employees and managers in line with company goals. The more data available to them, the more effective they can be in their managerial accounting. The simplest way to define financial accounting is that it is the preparation of financial statements in accordance with accounting standards, such as the United States Generally Accepted Accounting Principles (Flood, 2020) and the International Financial Reporting Standards (Burton & Jermakowicz, 2015). New data forms can be combined with traditional financial data sets for improved accounting calculations. For example, in reporting fixed assets, accountants could enhance their financials by attaching a video record of the assets, allowing users to see the condition of the assets. Moreover, big data could be used to improve reports for items not on balance sheets (such as company branding, product quality, and customer base), which are often difficult to evaluate in ways that are no subjective (Warren et al., 2015).

Studies have shown that big data affects accountants in three functional areas (management, finance, and auditing). The changes that can be expected following the introduction of big data are an area of concern. Using both structured and unstructured data, organizations can gain more information than they can by using structured data on its own. This will help management accountants to implement and monitor strategies actively, rather than simply formulating passive strategies. Big data also means that financial accountants must use new tools rather than traditional ones. New fair value analysis techniques are a good example of the benefits of big data when it comes to financial accounting. Auditors can use big data to enhance their auditing techniques by testing an entire population rather than depending on sample tests. Accountants play just as important a role as data scientists in dealing with big data due to their

understanding of businesses and their professional judgment. To improve their skills, accountants should familiarize themselves with the principles of programming (for example, by taking introductory courses to programming languages like Python). This will help them to develop their skills in using new technologies, allowing them to work side-by-side with data scientists (Richins et al., 2017).

Big data cannot be stored using traditional methods like conventional hard drives or solid-state drives. It requires much larger data storage systems. In the financial and auditing sectors, big data involves very large numbers of automated transactions, so companies require advanced processors to handle the data (Vasarhelyi et al., 2015). Accountants should use the most up-to-date technology to help them generate accurate financial statements and record real-time calculations (see Table 2-4; Vasarhelyi et al., 2015).

Table 2-4: Traditional vs technologically enhanced accounting measurements (source: Vasarhelyi et al., 2015)

Traditional Measurement	Updated Measurement	Technology Used	Accounting Concept Needed	Big Data Extensions
LIFO or FIFO inventory valuation	Current value of held inventory	Values of items in electronic markets, sales lists, infomediary price lists	Inventory revaluation adjustment to equity	RFID
Historical cost of PP&E	Current value[a] of PP&E	Search at different levels of aggregation (e.g. factory, machine part) market values	Partial valuation, geographic valuation differences, etc. Separation of different types of increase in organization net value	Peer benchmarks, extractions of histories of pricing and activity in peer markets
Depreciation	Current value comparisons across time	Values of items in electronic markets, sales lists, infomediary price lists	Partial valuation, geographic valuation differences, appreciation, etc.	
Intangibles	Reports on human resources, intellectual property, supply chain, customer reactions, service support obligations	Utilization of the ERP applications as HR, IP, SC	Additional information that may or may not link to financial statements; may or may not be dollar denominated	Social media, news pieces, labor market extracts with salaries, etc.

[a] Although there is a rich literature of price-level accounting and different bases of measurement (SFAS No. 33, replacement cost, current value, etc.), technology at that stage was not cost efficient to enable these measurements to be practical.

Different sources of data increase the overall volume of data. For this reason, technology is important for capturing, gathering, storing, and interpreting that data. Accountants should be able to distinguish the relevant information. With the help of big data, accountants and financial experts can use their analytical and modelling skills to improve decision-making in businesses (Gamage, 2016).

Vasarhelyi, Kogan and Tuttle (2015) have discussed the expanding ecosystem of corporate data (see Figure 2-3).

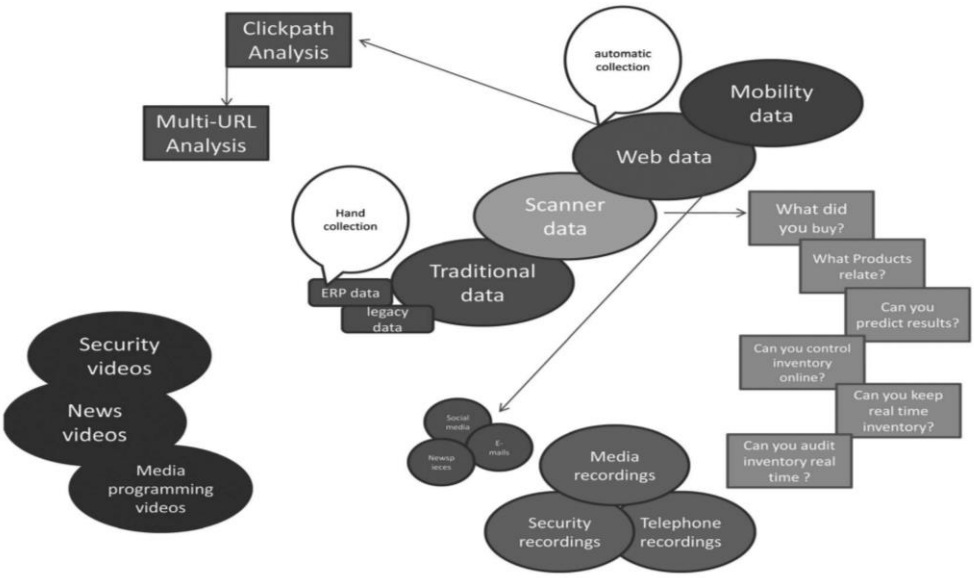


Figure 2-3: Progressive big data expansion (source: Vasarhelyi et al., 2015)

Traditional data (such as ERP) used to be developed manually by recording transactions. However, with advanced technologies such as UPS scanners, web data collection and mobile data collection can be automated. The types of data that are being recorded are becoming more diverse. Nowadays, textual, audio, and video data are being added to traditional accounting data. This process needs data bridges (such as automated facial recognition and data mining) to extract the desired data from the applications being used. Auditors can benefit from big data because it provides them with more evidence (see Table 2-5). This helps them move towards automated auditing software (Vasarhelyi et al., 2015).

Table 2-5: Illustrative uses of extended audit evidence (source: Vasarhelyi et al., 2015)

Data	Use 1	Use 2	Use 3
Security videos	Confirm receipts and exit of materials	Confirm shipping costs	Observe/deter fraud
News videos	Help in marketing	Warn of management or employees activities	Observe/deter fraud
Social networks	Evaluate consumer satisfaction and product defects	Support clients with technical problems and confirm technical support costs	Observe/deter fraud
RFID	Confirm inventory	Confirm/measure shipping	Link to price databases and estimate cost of goods sold
Web hits	Predict purchases	Predict revenue	Predict sales, cost of sales, geographic sales, product mix
Inbound mobile calls	Predict service costs	Monitor employees and estimate employee costs	Observe/deter fraud

Despite the advantages of big data, there are some threats that it poses. Chua (2013) has discussed the advantages and disadvantages of big data, as shown in Table 2-6 (Chua, 2013).

Table 2-6: Opportunities and challenges presented by big data in the finance profession (source: Chua, 2013)

Area	Opportunity	Challenge
1. Valuation of data assets	Helping companies value their data assets through the development of robust valuation methodologies Increasing the value of data through stewardship and quality control	Big data can quickly 'decay' in value as new data becomes available The value of data varies according to its use Uncertainty about future developments in regulation, global governance and privacy rights and what they might mean for data value
2. Use of big data in decision making	Using big data to offer more specialised decision-making support in real time Working in partnership with other departments to calculate the points at which big data can most usefully be shared with internal and external stakeholders	Self-service and automation could erode the need for standard internal reporting Cultural barriers might obstruct data sharing between silos and across organisational boundaries
3. Use of big data in the management of risk	Expanding the data resources used in risk forecasting to see the bigger picture Identifying risks in real-time for fraud detection and forensic accounting Using predictive analytics to test the risk of longer-term investment opportunities in new markets and products	Ensuring that correlation is not confused with causation when using diverse data sources and big data analytics to identify risks Predictive analytic techniques will mean changes to budgeting and return on investment calculations Finding ways to factor failure-based learning from rapid experimentation techniques into processes, budgets and capital allocation

In her analysis of the opportunities and threats related to big data, Chua (2013) realized that there will be an overlap between technology, finance, and

information in future models of accountancy (see Figure 2-4). Accountants will work with IT experts, data scientists, and managers, who will combine their insights for better results.

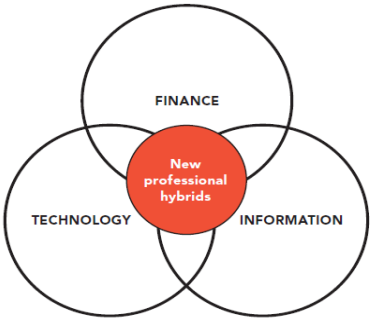


Figure 2-4: New accounting and finance professional hybrids (source: Chua, 2013)

One of the limitations of big data is that it can come from untrusted sources. Also, some organizations do not have the capabilities or the expertise to extract the relevant information. There is also the risk that organizations will not abide by ethical standards when it comes to data protection (Warren et al., 2015).

Gamage (2016) has argued that the lack of big data experts around the world is a significant problem. To overcome this, people across all professions must receive training in big data, and data science courses must be incorporated into university degree programmes (see Table 2-7). Gamage (2016) states that "universities must move quickly in partnership with industry to ensure that the graduates they produce have the required skills for the age of Big Data."

Table 2-7: Big data topics that could be included in existing courses (source: Gamage, 2016)

Course	Topics
Business Statistics	Data gathering techniques, Data exploration, Data summarisation, Data analysis, Data visualization, Communication of analytical findings
Business Information Systems	Advanced Databases, Information Retrieval, Advanced Data Mining Applications, Predictive Analytics for Decision Making, Big Data information management
Management Accounting	Application of Big Data to competitor analysis, Big Data as a strategic resource
Accounting Information Systems	Business intelligence, Enterprise analytics Information search and retrieval, Data mining, familiarity with languages such as XBRL, specialized software/reporting systems with decision support, ERP systems, Cybercrime, Data management issues
Business Finance	Financial analytics, modelling and computation of financial risks, Information Risk Management
Auditing and Assurance	Data Analytics in auditing , Mine new sources of data, Data integrity , Privacy, Safeguards, Cybersecurity, Design and evaluate IS controls, Manage IS risks and compliance, Overseeing fraud risk assessment
Forensic Accounting	Big Data, Benford's Law, Financial Analytics, Data Analytics for Fraud, Anomaly Detection in Forensics and Security
Taxation	Indirect tax and Big Data, tax value and non- tax value form data that is collected in the tax function, Visualize accounting data

Gepp et al. (2018) claim that auditors should familiarize themselves with big data techniques so that they can use them as part of their audit process. For a summary of these techniques, seeTable 2-8

Table 2-8: Brief description of common big data techniques (source: Gepp et al., 2018)

Regularized Regression (also known as shrinkage)	Aims to prevent over-fitting by shrinking variable coefficients towards zero. This shrinkage reduces the variance of the coefficient estimates that can adversely affect prediction accuracy, particularly with “highly correlated, large p” problems. It can also be used solve ill-formed problems. Further reading: James, Witten, Hastie, and Tibshirani (2013, pp. 214–228) provide further detail in Chapter 6, particularly Section 6.2.
Ridge Regression	Uses an L2 penalty based on the sum of squared coefficients, which performs well when all variables are likely to be important in relatively similar magnitudes.
Lasso/LARS	Uses an L1 penalty based on the sum of the absolute value of coefficients. The important advantage of this penalty is that it is effective at variable selection and so results in simpler models that are often desirable for their improved interpretability.

	Uses a weighted average of the L1 and L2 penalty. The weighting can be automatically chosen based on the data using a process called cross-validation. This weighted average can result in substantially improved model accuracy in some cases.
Elastic-Net	Comprise single tree model or an ensemble of them. Tree models are non-parametric models that are built in a recursive process of splitting the data into homogenous groups (usually two). Further reading: Rokach and Maimon (2014) cover single trees in detail, while Sutton (2005) also covers ensembles.
Tree-based Methods	The advantages of a single tree are that they: are resistant to outliers and irrelevant variables, automatically model interactions between variables, and do not require any variable transformations. Relatively small models are also easy to interpret and display visually. However, single trees are very sensitive to changes in the data (as are some neural networks) and so have high variance.
Single Classification and Regression Trees (CART) also known as decision trees	Ensembles of decision trees that are combined through an averaging process (Random Forests) or iterative improvement process (MART). This reduces the high variance of individual trees and usually results in increased accuracy. Random Forests are particularly good at “large p, small n” problems. Ensemble models are inherently more difficult to interpret, but there are procedures to extract information in interpretable ways.
Ensembles of decision trees including Random Forests (enhanced bagging) and Multiple Additive Regression Trees (MART or gradient boosting)	Splines involve dividing the range of independent variables into sections and fitting separate polynomials to each section. This is particularly useful when there are known breakpoints that separate different distributions. For example, the distribution for retail sales is different during holiday periods. Alternatively, MARS is one particular spline technique that automatically chooses the number of sections and where to place the breakpoints (and then fits linear models to each section). Other types of splines include natural regression splines and smoothing splines. Local regression is a popular alternative to splines. Further reading: James et al. (2013, pp. 271–282) cover splines and local regression in Chapter 7, particularly Sections 7.4–7.6. MARS is more complex and only covered in a more technical book by Hastie, Friedman, and Tibshirani (2009, pp. 321–329) in Section 9.4.
Splines Multivariate Adaptive Regression Splines (MARS)	SVMs are popular for classification problems, but are not applicable to regression problems. SVMs place hyper planes in the data to attempt to separate it into the desired groups. Kernel SVMs offer non-linear extensions. Major drawbacks include no variable

Support Vector Machines (SVMs)	selection and no easy way to calculate the associated probabilities of classification. Logistic regression with an L1 or L2 penalty is an alternative to binary classification that overcomes these drawbacks. Further reading: Provost and Fawcett (2013, pp. 89–94) briefly introduce SVMs, starting with a comparison to standard logistic regression. A simple model that assumes the variables (or features) are (conditionally) independent. This assumption is almost always violated, but it can still perform well in some circumstances, because of the low variance associated with the simple assumption. It also easily handles “large p” problems. Bayesian belief networks are generalizations of Naïve Bayes that relax some of the independence assumptions by defining a network of conditional dependencies between variables. Further reading: Provost and Fawcett (2013, pp. 233–244) introduce the basic concepts of Naïve Bayes and Alston, Mengersen, and Pettitt (2012, pp. 348–360) cover Bayesian Networks in Chapter 20.
Naïve Bayes and Bayesian (Belief) Networks	Types of evolutionary algorithms that are heavily based on Darwin’s survival of the fittest principle to evolve better solutions to a problem. They are non-parametric, and able to handle missing values and model interactions, but there are a large number of model parameters to set based on user expertise. GAs can be used for both supervised learning and unsupervised learning, and often to optimize the parameters of other models. Further reading: Negnevitsky (2011, pp. 219–257) cover evolutionary algorithms in Chapter 7. ANNs are non-parametric models designed on the inner processes of the human brain, primarily with respect to pattern learning. There are many different types of ANNs and they can be trained using supervised or unsupervised methods (such as self-organizing maps). Procedures (such as genetic algorithms) are available to automate the numerous model parameters. Advantages include their ability to model non-linear relationships and handle highly correlated variables and outliers. However, the black-box nature makes interpretation difficult, although techniques are available to extract some information. Further reading: Negnevitsky (2011, pp. 165–217) provide more detail in Chapter 6. Unsupervised learning approaches that attempt to find simple rules to describe frequently occurring patterns. For example, analyzing a department store database might reveal that customers who buy jeans also often buy music. Further reading: Zhang and Wu (2011) provide an overview of association rules.

Genetic Algorithms (GAs) including Genetic Programming (GP)	A large collection of unsupervised learning techniques designed to find sub-groups within the data, such that the data is more homogenous within each sub-group. Further reading: Provost and Fawcett (2013, pp. 163–183) provide more detail in Chapter 6, particularly in the section titled “Clustering”.
Artificial Neural Networks (ANNs), sometimes called Neural Networks or Neural Nets	A class of models that assumes there are one or more influential quantities that is hidden and unobservable. Popular examples include principal components analysis, principal curves, item response theory and multidimensional scaling, which attempt to model the complete set of data with a smaller set of latent variables. Such methods can also be used as a first step that feeds into a second supervised learning step. Further reading: Finch and French (2015) provide information on a variety of latent variable models.
Association Rules	
Clustering or Data Segmentation	Many of contemporary techniques, including those listed above; combine the results of multiple underlying models. Other techniques to combine multiple models include averaging outputs, a majority vote decision, a hierarchical approach, and more sophisticated processes such as stacking, bagging, and boosting. Ensemble models often outperform individual models in terms of accuracy, but they are inherently more complex to interpret. Further reading: Sutton (2005) introduces bagging and boosting (in Sections 1.2, 5 and 6), two popular methods to create ensemble models.
Latent Variable Models	

Previous studies have also discussed the risks of automation in accounting, as this may affect the future careers of accountancy students. The introduction of big data has been challenging for the accountancy profession. It has many risks, but it also offers opportunities for the future. Accountants need to understand big data so that they can make use of it to improve their services. Educators should develop students’ skills in analyzing big data, while also recognizing that big data may lead to the automation of some aspects of accounting. Even if automation increases, the professional judgment of accountants will continue to be necessary (Huerta & Jensen, 2017).

2.4 Big data techniques and software

Having defined big data and identified its importance in the business sector, I will now examine research that looks at how to collect, save, process, and analyse big data efficiently and profitably. Data science use statistics, computer science, machine learning, and many other tools to help draw meaningful outputs from big data that can help in decision-making (Elshawi et al., 2018a).

Choosing the correct platform for handling big data will depend on the structure of the data, the requirements for data transfer, and the desired final outcome. Cost must be taken into account when consdering these factors.

There are many technologies that can be used to handle big data. Honest and Patel (2016) provide a summary of these. They include the NosQL database, Massively Parallel Processing (MPP), and Hadoop. Table 2-9 summarises how these and other tools capture and store data.

Table 2-9: Technologies used to handle big data

NoSQL database	**Capturing and storing data**
Massively Parallel Processing (MPP)	Processing
MapReduce	Processing: Breaking down large amounts of data into smaller batches and processing them separately Computing data as a map and reducing functions
Amazon Simple Storage Service (Amazon S3)	Storing data
Hadoop Distributed File System (HDFS)	Storing data

I will now consider Hadoop in more detail. Hadoop is a framework written in java, which enables the user to store and process large data sets in a parallel and distributed format. It processes large amounts of data through clusters of computers using programming models (see Figure 2-5).

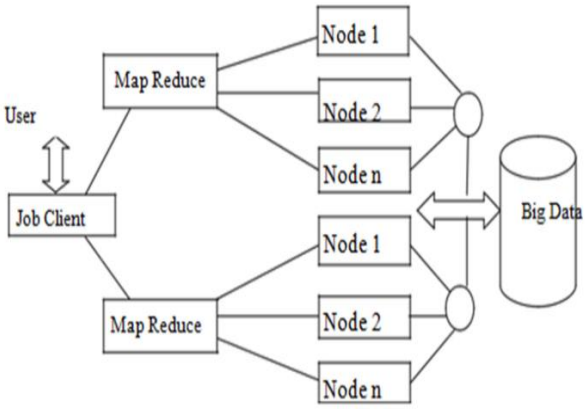


Figure 2-5: Hadoop framework (source: Honest & Patel, 2016)

Hadoop runs applications using the MapReduce algorithm, which splits the input into small parts and then allocates the parts to several computers connected to the network. Then, the results are collected to provide the final outcomes. Hadoop is capable of developing applications that run on clusters of computers, providing statistical analysis for large data sets. Hadoop consists of the HFS (the storage unit), Yane (the processing unit), MapReduce (the software framework), and the java libraries. Hadoop can be used on all computer platforms (Honest & Patel, 2016).

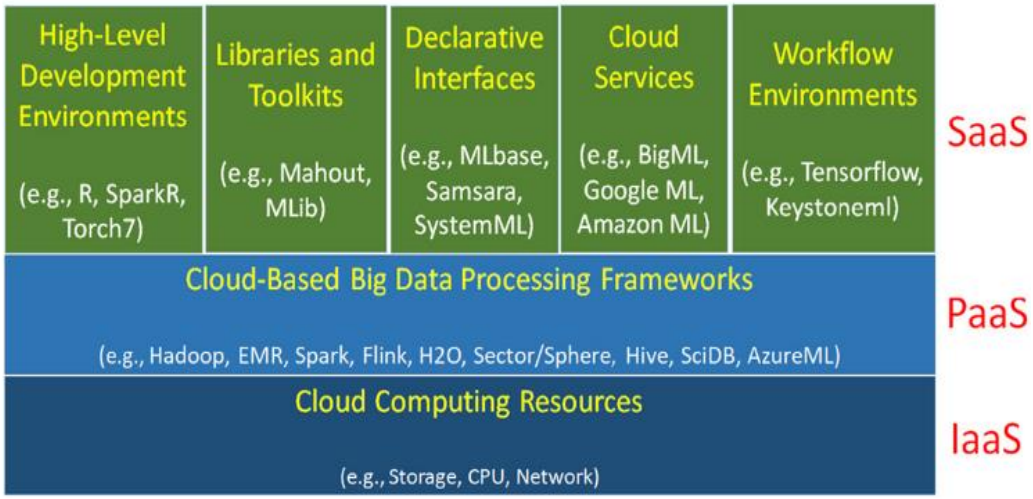
Previous studies have shown that cloud computing changes the way that companies, organizations, and individuals handle big data. Cloud computing lowers the cost of creating and managing software, leading to more meaningful insights data. These previous studies have identified three types of cloud services:

1. Infrastructure as a Service (IaaS) model.
2. Platform as a Service (PaaS) model.

3. Software as a Service (SaaS) model.

Big data can be used with each of these three types. Combining cloud computing and big data makes it easier for data scientists to collect, store, and analyze data. According to Elshawi et al. (2018) , the main requirements for using big data with cloud computing are standardized access, scalability and efficiency (Elshawi et al., 2018a). Table 2-10 provides an overview of big data software services.

Table 2-10: Big data as a service (source: Elshawi et al., 2018b)



Elshawi et al. (2018) also provided a summary of big data analysis frameworks (see Table 2-11).

Table 2-11: Summary of big data analysis framework (source: Elshawi et al., 2018b)

Framework	Model	Abstraction	Supported languages	Underlying engine	Availability
Hadoop	PaaS	MapReduce	Java, Python	Hadoop	Open source
Spark	PaaS	RDD	Scala, Python, Java, R	Spark, Yarn	Open source
H2O	PaaS	Procedural + Library	REST, R, Python	H2O	Open source
SciDB	PaaS	Declarative	SQL	SciDB	Open source
AzureML	SaaS/PaaS	Visual User Interface	REST	Microsoft Azure	Proprietary/Microsoft
R	SaaS/PaaS	Procedural	R	R	Open source
SparkR	SaaS/PaaS	Procedural	R	Spark	Open source
Mahout	SaaS/PaaS	Toolkit	Java, Scala	Hadoop, Spark	Open source
Spark MLib	SaaS/PaaS	Library	Scala, Python	Spark	Open Source
Samsara	SaaS/PaaS	Declarative	Java, Scala	Spark, Flink, H2O	Open Source
Apache SystemML	SaaS/PaaS	Declarative	R, Python	Spark	Open Source
Google ML	SaaS	Visual User Interface	Python	Google Cloud Dataflow	Proprietary/Google
Amazon ML	SaaS	Visual User Interface	N/A	Apache MXNet, TensorFlow, PyTorch	Proprietary/Amazon
BigML	SaaS/Paas	Visual User Interface	Python	BigML PredictServer	Proprietary/BigML
Tensorflow	SaaS	Visual User Interface/Library	Python, Haskell, Java, Go Julia, R, Scala	CUDA, TPU	Open Source
KeystoneML	SaaS/PaaS	Procedural + Library	Scala	Spark	Open Source

Gepp et al. (2018) simplified the process of determining which big data technique to use. They categorized big data in to three groups: “large p, small n”, “large p, large n”, and “small p, large n”, where *n* is the number of responses and *p* is the number of variables measured for each response (see Table 2-12).

Table 2-12: Big data categories and example techniques

Big data groups	**Small p, Large n**	**Large p, Small n**	**Large p, Large n**
Techniques	Existing statistical techniques	Random forests	Regularized regression (also known as shrinkage)
Examples	_	Genomics	Genomics Climate change research

In finance and accountancy, there are four main areas in which big data is used:

1. Financial distress modeling
2. Financial fraud modeling
3. Stock market prediction and quantitative modeling
4. Auditing

Accountants have used data mining strategies to predict possible failures in their future financials; auditors have used the same techniques to evaluate similar concerns. The International Standard of Accounting in the United Kingdom (ISA UK) released a revised standards for auditing in 2016 which stated that the “ISA UK requires the auditor to obtain reasonable assurance about whether the financial statements as a whole are free from material misstatement, whether due to fraud or error” (Council, 2016). Auditors can use big data techniques to attain the assurances they need.

In reviewing the literature related to each of the four areas mentioned above (financial distress modeling, financial fraud modeling, stock market prediction and qualitative modeling, and auditing), it seems that the auditing profession is behind many other professions in its use of big data techniques. Gepp et al. (2018) suggest that auditors can benefit from using big data because it provides increase transparency and accuracy.

The literature review also revealed that the most important area of big data for accountants and bankers is machine learning because banking data is often unpredictable and grows rapidly. Machine leaning can help to automate processes, allowing those with relatively little knowledge of computer science to benefit from it.

2.5 Machine learning

Machine learning uses algorithms to provide meaningful outputs that aid with decision-making and forecasting. Previous studies have claimed that there are two main types of machine learning: supervised and unsupervised. Supervised machine learning creates models that use inputs and outputs to predict similar situations. Unsupervised machine learning uses only inputs to try and find patterns or useful information in the data (Figure 2-6).

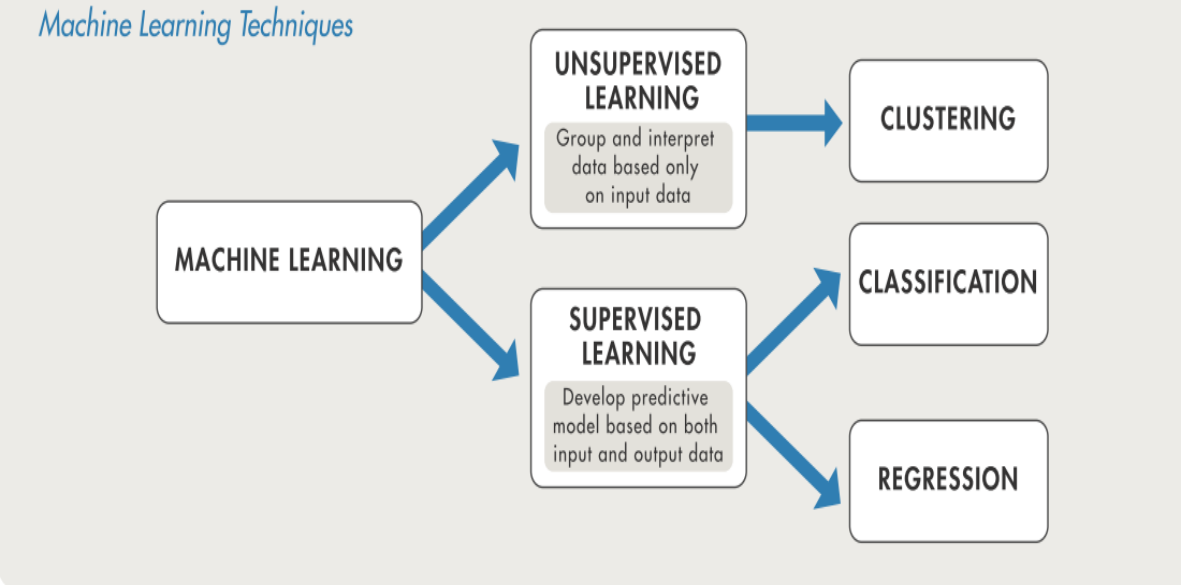


Figure 2-6: Machine learning techniques (source: Mathworks, 2016)

Data scientists must often use trial and error to find the right algorithm to use. Machine learning is the best solution when dealing with big, complex, and rapidly growing data sets (Mathworks, 2016d). Supervised machine learning is useful when there is not a clear path from the beginning to the end of the process; users are required to use multiple iterations to discover the most effective technique and algorithm (Mathworks, 2016c). Unsupervised machine learning is useful when the users have a very large data set with no expected results expected or when they want to find out whether a data set has any valuable information. Unsupervised machine learning can also help to make the data more accessible, so that it can be analyzed using supervised machine learning (Mathworks, 2016b). In unsupervised machine learning, there are two clustering techniques that are used (see Table 2-13):

- Hard clustering: this algorithm detects the data that belongs to a single cluster
- Soft clustering: this algorithm finds the data that can belong to more than one cluster

Table 2-13: Common hard and soft clustering techniques

Hard clustering algorithms	Soft clustering algorithms
k-Means, k-Medoids, Hierarchical Clustering, Self-Organizing Map	Fuzzy c-Means, Gaussian Mixture Model

Mathworks (2016b) also notes that there are three popular techniques for reducing dimensionality through unsupervised learning:

1. Principal component analysis (PCA)
2. Factor analysis
3. Nonnegative matrix factorization

Supervised machine learning is useful when users have clearly defined inputs and outputs for their data sets. There are two types of supervised learning techniques: classification and regression. In classification, the users seek to predict discrete outputs (responses). In regression, users need ongoing responses. Users must employ the correct algorithm based on their priorities, such as what characteristics they are examining and how quickly they would like the results. Table 2-14 outlines the different algorithms for classification and regression (Mathworks, 2016a).

Table 2-14: Common algorithms of classification and regression technique

Common Classification Algorithms	Common Regression Algorithms
Logistic Regression	Linear Regression
k Nearest Neighbour (kNN)	Nonlinear Regression
Support Vector Machine (SVM)	Gaussian Process Regression Model
Neural Network	SVM Regression
Naïve Bayes	Generalized Linear Model
Discriminant Analysis	Regression Tree
Decision Tree	

Bagged and Boosted Decision Trees	

Mathworks (2016a) also notes that there are some techniques for improving the accuracy and predictive powers of the algorithms (see Figure 2-7). These techniques are as follows:

- Feature selection
- Feature transformation
- Hyperparameter tuning

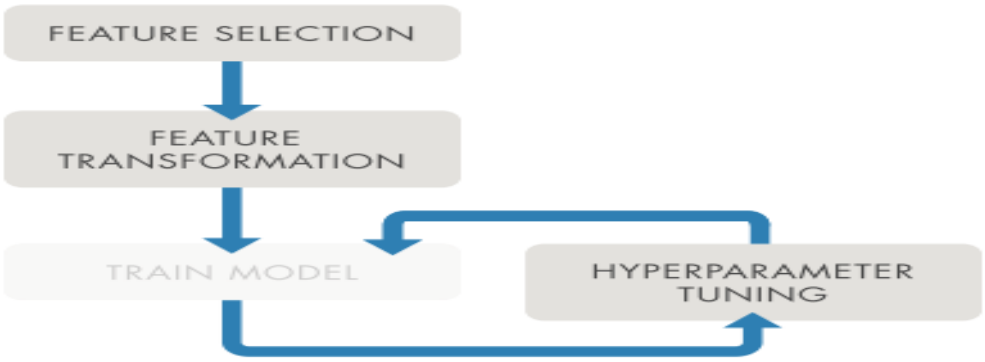


Figure 2-7: Improvement of models (source: Mathworks, 2016a)

There are some challenges when it comes to machine learning, including the complexity of the data, the variety of data formats, the need for specialized knowledge, and the time required to develop the final algorithm (Mathworks, 2016c). There are three important questions that the machine learning users should ask themselves before starting the machine learning workflow:

1. What is the type of data?
2. What do you want to know from it (insights)?
3. Where and how might these insights be applied?

Mathworks (2016c) provide a machine learning workflow example for a health monitoring app. It outlines all of the steps that must be followed to achieve the desired results (see Figure 2-8).

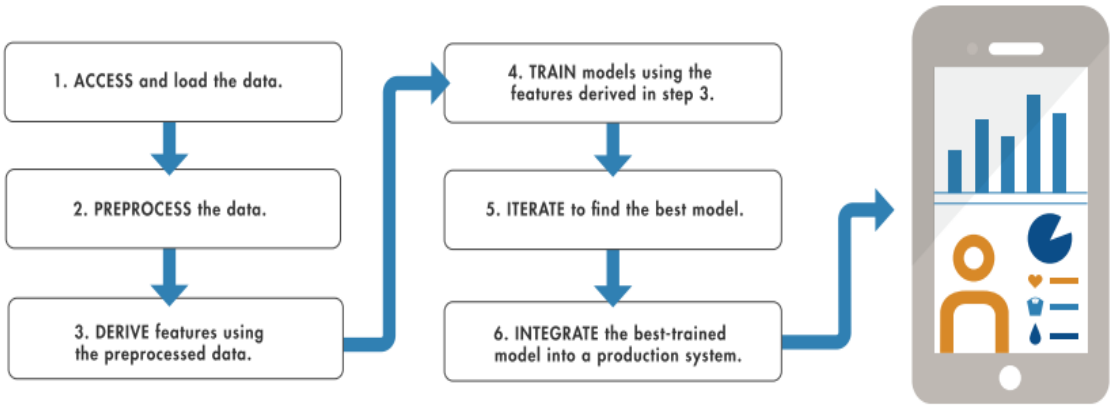


Figure 2-8: Workflow at a glance (source: Mathworks, 2016c)

Bataev (2018) has argued that new technologies are changing the way that companies and organizations are conducting their business. They are using information technology to reduce their expenses and increase the quality of their services. According to Bataev (2018), cloud computing and big data are two of the most important innovations in the areas of information and communication technologies. The total spending on cloud computing has increased significantly in comparison to the rest of the IT market. As shown by Figure 2-9, in 2018, the total spending on cloud computing was expected to have reached 240 billion USD by 2020.

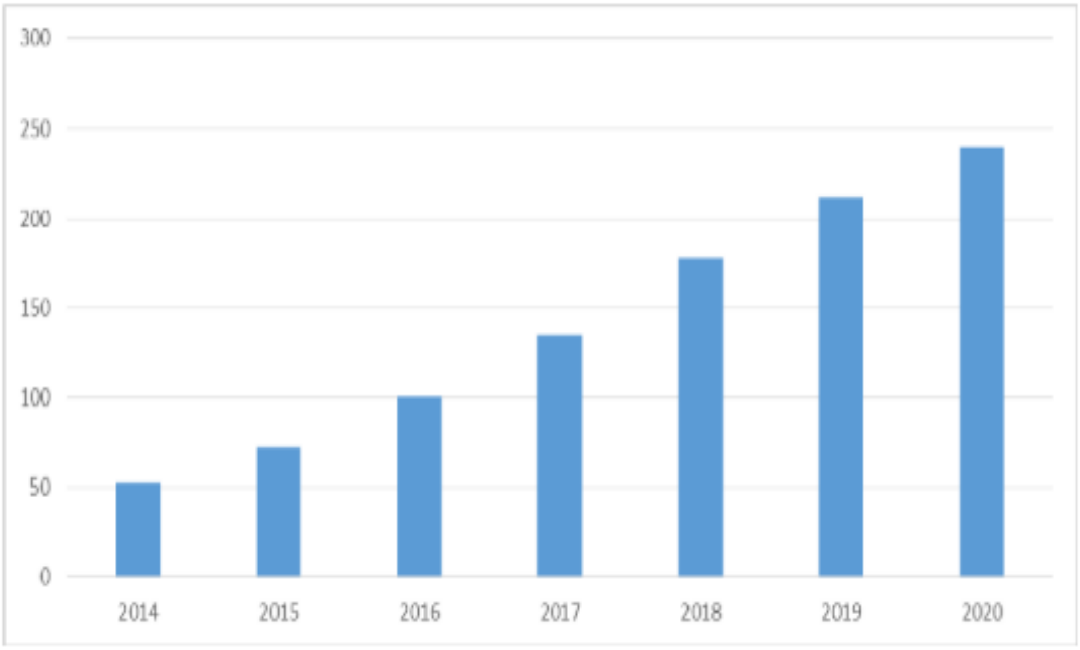


Figure 2-9: Changes in the global cloud computing marked in USD billions (source: Bataev, 2018)

the paper examined the adoption of cloud computing technologies by Russian financial institutions in comparison with institutions in other regions. The study found that the United States spent the most on adopting these technologies, with 53% of its IT spending being directed towards cloud-based business applications, equivalent to 75% of the total cost of all banking applications. On the other hand, Russia spent the least on cloud computing. The reasons for this are that transferring big data is difficult, poses security risks, and requires specialists. Some of the main advantages of shifting the business, only if the associated difficulties can be solved, towards cloud computing in the financial sector are:

- Reducing costs
- Reducing work for the IT department, allowing them to focus on other things like strategic development
- Storing data
- Using business applications

- Creating automated, cloud-based banking systems

2.6 MATLAB (matrix laboratory) software

Given that I are accountants with experience of financial analysis and not Computer Science specialists, I found MATLAB very useful and easy to use. In particular it was able to handle the quantity and type of data that I was using.

MATLAB stands for matrix laboratory; It was initially written in Fortran by Cleve Moler. MATLAB aims to make it easier to gain access to matrix software created by the LINPACK and EISPACK projects. More recent versions of MATLAB have been written in C and published by MathWorks Inc. MATLAB allows its users to exchange data with other applications and use the built-in toolboxes to process and analyse the data. The availability of toolboxes was one of the features that most attracted us to MATLAB. MATLAB and its toolboxes include many functions that allow the software to deal with various problems. It also contains links to other programming environments (like C or Fortran). The software is compelling, comprehensive, and easy to use, even for people who are not familiar with programming (Radi & ELHami, 2018). The packages in MATLAB are updated twice a year, introducing new applications, examples, exercises, and solutions (Lynch, 2014).

Until recently, most MATLAB users already had prior knowledge of programming languages (like Fortran and C) and moved to MATLAB when the software became more powerful. As such, much of the literature on MATLAB has tended to assume that the user will understand programming. Books about MATLAB have also usually focused on sophisticated topics and applications. Nowadays, however, MATLAB is being introduced to students as the first computer programming language that they learn. This was the same for us as my experience was in processing and analyzing data. There is a need for more books that teach the basics of MATLAB without assuming any previous experience in computer programming. These books should also show how to use the toolboxes for specific fields (Gilat, 2013). MATLAB offers variety of built-in toolboxes. A toolbox is a bundle of functions and tools for specific topics like Statistics and Machine

Learning Toolbox, Curve Fitting Toolbox, Control System Toolbox, Signal Processing Toolbox, Mapping Toolbox, System Identification Toolbox, Deep Learning Toolbox, DSP System Toolbox, Datafeed Toolbox, Financial Toolbox, Image Processing Toolbox, Text Analytics Toolbox and Predictive Maintenance Toolbox (Mathworks, 2021b). Mentioned toolboxes provide users with tools to process, model, visualize and analyze data. Some toolboxes require license while others can be used for free.

In the current study, I used the Statistical and Machine Learning Toolbox. Statistics and Machine Learning Toolbox supply applications and functions to illustrate, examine, and model data. Users can use expressive statistics, visualizations, and clustering for exploratory data analysis. It fit probability distributions to data; generate arbitrary numbers for Monte Carlo simulations and execute hypothesis tests. Regression and classification algorithms allow users to draw inferences from data and build predictive models. The users can operate classification and regression learner applications interactively. They also can use it programmatically through utilizing Auto Machine Learning (Mathworks, 2021o). This required us to gain a basic understanding of how to use the toolboxes. Operating with matrices is an essential part of MATLAB. Another important aspect is the use of statements, which are operators, functions, and variables that produce a matrix to be used in processing data. MATLAB also offers structures for functions and conditions; a function branches out based on specific conditions. There are many other basic concepts that need to be studied by first-time users, such as loops, relations, scalar functions, vector functions, matrix functions, submatrix, logical indexing, M-files, scripts and functions, graphics, and linear algebra (Ferreira, 2009).

In what follows, I examine several papers that have used MATLAB software in research into banking.

One study by Wei et al. (2015) (Burton & Jermakowicz, 2015) investigated the minimization of the marketing cost and the maximization of the user's perceived utility of a mobile banking system. They collected their data from responses to questionnaires by banks. They analyzed the data statistically using SPSS to

examine interactions among adoption drivers. After that, they produced a qualitative simulation to encourage interaction between these drivers. They created an optimization model based on the empirical results of the relations. They used a prototype system based on MATLAB to assess the feasibility of the framework. Their results showed that the framework could contribute to real time optimization supporting decision-making in the marketing of mobile banking systems.

Another study (Tian, Zhuang, & Yu, 2021) examined the cooperation between banks and business-to-business (B2B) platforms with regard to the design of incentives and supervision. The study focused on an electronic warehouse receipt pledge financing model. A principal-agent model for cooperation was created between the banks and the B2B platforms. The aims were to expand and compare the models by adding supervision variables and to analyse the effects of risk aversion coefficients on levels of effort level, incentive coefficients, fixed payments, and impact bank income. The researchers examined the banking system's incentives and supervision structures through big data analysis. They used MATLAB for numerical analysis. They found that banks' expected benefits when cooperating with risk-neutral B2B platforms were larger than their expected benefits when cooperating with risk-averse B2B platforms. Their results also showed that accounting for supervisory measures will positively impact profits.

A further study by Bekamiri et al. (2021) used MATLAB to discover the most valuable variables in customer behaviour analysis. The study examined the behaviour and online banking transactions of customers over a year. They collected quantitative data from interviews with experts in banking and developed a new algorithm to enhance internet security and identify the limits of services that should be available in internet banking, given the risk of data exchange. Their results imply that the potential of the given structure to understand the rate of trust in customer's behaviour in relation to online banking is at a reasonable level for experts in this area.

Recently, many banking studies have used MATLAB to deal with their data and produce results. This is due to its benefits, mentioned earlier. Given my banking

experience, I decided to use MATLAB in my research to take advantage of these benefits.

2.7 Online banking

This study focused specifically on the relevance of big data to online banking. In reviewing the literature, my aim was to identify how big data could be used to achieve maximum benefits. The definition of online or electronic banking includes banking via call centers, online banking through websites, and mobile banking through mobile applications.

Several studies have previously examined the effect of electronic banking on the performance of banks. For example, Yang et al. (2018) have argued that China is undergoing rapid grow in the e-commerce sector, leading to increased use of electronic payments. The purpose of this study was to determine the relationship between the implementation of electronic banking and the performance of banks in China, taking into account the different stages of development of online banking in the country. Secondary data were used in this study, including textual data (such as reports) and non-textual data (such as voice and video recordings). The authors selected five banks based on their major shares in the electronic banking industry, the availability of their financial data, and the inclusion of an electronic banking statement in their financial reports. The sample covered the period from 2003 to 2013. The data was analysed using SPSS and Excel to examine the profitability and efficiency of the banks. One efficiency ratio was used: non-interest expenses divided by revenue. Four profitability ratios were used: net income after taxes divided by total assets; net income divided by shareholder's equity; net interest income divided by total assets; and net income divided by revenue. The study examined the five variables in banks in the developing stage and developed stage of electronic banking using a ratio analysis (T-test). The result indicated that the developing stage and developed stage of electronic banking were noticeably different when it came to returns on assets, returns on equity, and operating margins. Electronic banking performance was much better in the developed stage than in the developing stage. In term of operating margins, there was a small difference between the two stages (electronic banking in the developed stage was

more profitable than in developing stage). The only variable for which the results were reversed was net interest margin, where the performance in the developing stage were better than that in the developed stage. However, the differences were not significant (see TablesTable 2-15 and Table 2-16).

Table 2-15: Descriptive statistics from 2003 to 2013 (source: Bloomberg 2014)

	N	Minimum	Maximum	Mean	Std. Deviation
Return on Assets	52	.04	1.39	.8560	.38000
Return on Equity	52	1.24	35.77	16.4794	6.63012
Operating Margin	52	5.46	56.75	41.2310	12.82508
Efficiency Ratio	49	25.01	62.43	44.2239	7.39484
Net Interest Margin	51	2.07	3.88	3.0488	.35365
Valid N (listwise)	49				

Table 2-16: Summary of all variables (source: Yang et al., 2018)

	ROA		ROE		OM		ER		NIM	
Groups	1	2	1	2	1	2	1	2	1	2
Mean	.7106	1.1553	15.0609	19.4000	36.6817	50.5971	44.9934	42.3000	3.0557	3.0338
Mean Difference	-.44472		-4.33914		-13.91534		2.69343		.02196	
F	13.136		10.437		7.306		.017		4.477	
Sig.	.001		.002		.009		.898		.039	
t	-5.981		-3.128		-4.236		1.156		.243	
P(Sig.)	.000		.003		.000		.254		.809	
Conclusion	Positive		Positive		Positive		Negative		Negative	

Another study by Gautam (2012) examined the effect of electronic banking on the performance of Indian banks. It showed that electronic banking has a significant impact on customer satisfaction, leading to increased profits. The main objective of the study was to examine the quality of online banking services, as well as the efficiency of the operation, and the profitability of the bank. The study was qualitative in nature, involving a field survey with open-ended questions (see Figure 2-10). Interviews were conducted with the bank managers of 16 Indian banks, namely Syndicate Bank, ICICI Bank, the Bank of Baroda, the Indian Overseas Bank, the State Bank of India, the Union Bank of India, the Standard Chartered Bank, the Punjab National Bank, UCO bank, the United Bank of India, City Bank, HDFC Bank, IDBI Bank, the Oriental Bank of Commerce, AXIS bank,

and HSBC bank. These banks all provided similar electronic services such as mobile banking, money transfers, debit and credit card payments, automated teller machine (ATM) services, and more.

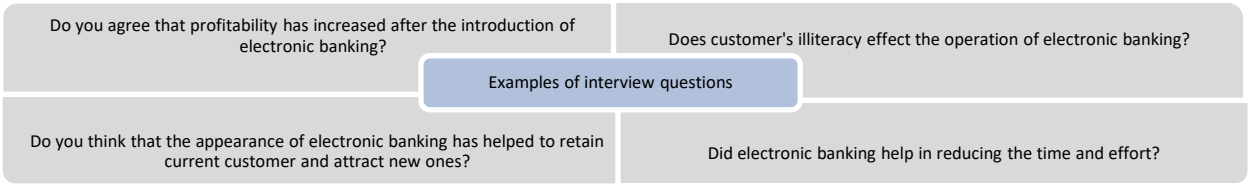


Figure 2-10: Examples of interview questions

The findings of this study illustrated that the managers were positive regarding the impact of electronic banking on the profitability of their banks. They said that electronic banking had helped to speed up operations and increase customer satisfaction, thereby leading to more profits. The managers also confirmed that their employees had been well trained to deal with customers who were not digitally literate. They felt that their banks had become more efficient by reducing the need for manual maintenance and the cost of labour.

Electronic banking also increases client satisfaction by allowing them to manage their finances while at home, in the office, or anywhere else (Yee & Faziharudean, 2010).

Rawwash et al. (2020) have studied the factors that affect Jordanian electronic banking services. These factors can be categorized into three groups: elements related to technological characteristics; elements related to individual characteristics; and external elements related to the social environment (Moga et al., 2012; see Figure 2-11).

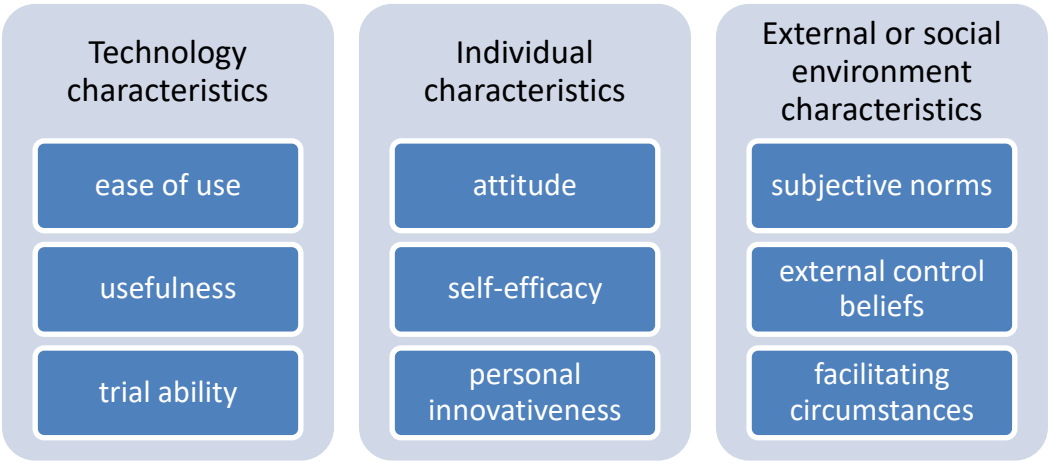


Figure 2-11: Factors affecting the use of electronic banking

Rawwash et al. (2020) used a sample of 300 bank clients that had bank accounts in local Jordanian banks. They were given questionnaires to examine the relationship between their use of electronic banking and their perception of the usefulness, trustworthiness, ease, security, and privacy of the service. The results showed a positive relationship between customer's use of electronic banking and the perceived usefulness, ease of use, trustworthiness, security, and privacy of the service. On the other hand, there was no relationship between convenience and the use of electronic baking.

Danyali (2018) studied the factors that affect the behaviour of customers when using mobile banking. A quantitative method was applied in this study. A random sample of 384 customers of Tehran Province Tejarat Bank was chosen. This sample was selected from a population of 400 customers who were using online and mobile banking. These customers were asked to respond to 39 closed questions on questionnaire and provide information about their gender, education, age, and monthly income. Cronbach's α coefficient was used to assess the reliability of the questions (see Table 2-17). To analyze the data, a T-test was used (see Table 2-18). The results showed that perceived advantage and compatibility had a positive relationship with customer attitudes. Complexity was negatively correlated with customer attitudes. There was a positive correlation between subjective norms withing peer groups and superior groups. The results

showed a positive relationship between self-efficacy and perceived behavioural control. There were also positive relationships between the source facilitator conditions and the perceived behavioural control. as well as between the technology facilitator conditions and the perceived behavioural control. The results showed a significant correlation between bank customers' tendency to use online banking and their attitudes, subjective norms, and perceived behavioural control. Finally, there was an important correlation between behavioural tendencies and actual behaviour with regard to mobile banking.

Table 2-17: Reliability coefficients for variables (source: Danyali, 2018)

Variable	α
Perceived advantage	0.82
Complexity	0.92
Compatibility	0.76
Peer groups	0.81
Superior groups	0.79
Self-efficacy	0.71
Source facilitator conditions	0.73
Technology facilitator conditions	0.75
Attitude	0.89
Subjective norms	0.87
Perceived behavioral control	0.84
Behavioral tendency of using	0.89
Actual behavior of using	0.82

Table 2-18: Hypothesis testing (source: Danyali, 2018)

Hypothesis	Standardized coefficient	t-value	p-value	Result
Relationship between perceived advantage and bank customers	0.54	9.66	< 0.01	+
Relationship between complexity of using mobile banking and customers' attitude	0.36	−6.55	< 0.01	+
Relationship between compatibility of using mobile banking and customers' attitude	0.14	2.89	< 0.01	+
Relationship between peer groups and subjective norms	0.44	8.12	< 0.01	+
Relationship between superior groups and subjective norms	0.39	7.10	< 0.01	+
Relationship between self-efficacy and perceived behavioral control	0.09	1.89	< 0.01	−
Relationship between source facilitator conditions and perceived behavioral control	0.41	7.67	< 0.01	+
Relationship between technology facilitator conditions and perceived behavioral control	0.41	7.66	< 0.01	+
Relationship between bank customers' attitude and behavioral tendency to use	0.60	10.97	< 0.01	+
Relationship between subjective norms and behavioral tendency to use	0.44	8.22	< 0.01	+
Relationship between perceived behavioral control and behavioral tendency to use	0.34	6.42	< 0.01	+
Relationship between and behavioral tendency to use mobile banking and customers' actual behavior	0.67	12.35	< 0.01	+

Another study by Ikhsanova and Protsko (2016) used economic and mathematical modeling to study the effect of their determined variables on the total volume of payments made through electronic banking in Russian banks. They performed a correlation analysis using a correlation matrix. They used the GRETL program to run a simulation. A Fisher Test was used to analyze the significance of the economic and mathematical models. By applying regression analysis, they reached the conclusion that the following equation was the best fit for their data set:

$$Y = 712.67 + (42.9501)(X2) + (0.00154757)(X3) + (44.3815)(X9) \quad (2\text{-}1)$$

Where:

Y = total volume of electronic banking payments.

x2 = number of banks implement internet banking system.

x3 = number of Internet users.

x9 = average number of services provided by banks through electronic banking channels.

The results of the study showed that if the number of banks that provided online banking were to increase by one, the total volume of payments made via electronic banking would increase by 42.95 million rubles. If the number of users increased by one, then the total volume of payments would increase by 1500 rubles. A single unit increase in the average number of electronic banking services would increase the total volume of payments by 44,38 million rubles. The results also predicted that there would be an increase in the use of electronic banking in the future. Russian banks are currently in the process of building new "client-bank" relationships by developing their electronic banking services.

Mohammad and Saad (2011) have argued that banks need to focus their attention on electronic banking and adopt new technologies that will help them meet the standards expected by customers. According to the Basel Committee Report on Banking Supervision (2003), electronic banking has significant value for both retail and small value banking operations. Mohammad and Saad (2011) conducted an empirical investigation into data collected from 15 Jordanian banks during the period from 2000 to 2010 (a major period of growth for electronic banking). The financial reports of the banks and the annual reports of the Central Bank of Jordan were the main sources of data of this study. This accounting data was combined with macroeconomic data to measure the banks' performance. Pooled Ordinary Least Squares (OLS) regression analysis was used to observe the effect of electronic banking on the performance of the banks. The variables used in the study are summarized in Table 2-19.

Table 2-19: Summary of variables used in this study

	Dependent variable	**Independent variables**	**Control variables**	
	Bank's performance	Dummy variable	Bank's specific	Macroeconomic determinants

Variables names	ROE: net income/equity	EBANK	SIZE: total assets (proxy) CAPITAL: equity capital/assets CREDIT RISK: non-performing loan /total loans EXPENSES MANAGEMENT: operating cost/total assets LIQUIDITY: total loans/total deposits	INFLATION: rate of change of CPI ECONOMIC GROWTH: GDP growth
Comments	Common measure of bank performance	This variable is created and takes a value of 1 if the bank has implemented e- banking; otherwise, it takes a value of 0	Possible determinants of banks' performance	Possible determinants of banks' performance

Regression analysis was conducted and mathematically expressed as:

(2-2)

$$ROE_{it} = \beta_0 + \beta_1 EBAN_{it} + \beta_2 SZE_{it} + \beta_3 CAP_{it} + \beta_4 CRR_{it} + \beta_5 EXPM_{it} + \beta_6 LIQ_{it} + \beta_7 GDPG_{it} + \beta_8 INF_{it} + \varepsilon.$$

Where β0 refers to the intercept and ε refers to random error. The results of the regression are shown in Table 2-20.

Table 2-20: Regression results (source: Mohammad & Saad, 2011)

Variable	Coefficient	t-value	p-value
Constant	0.235	2.668	0.008
EBANK	-0.323*	-1.727	0.085
SZE	0.721***	12.215	0.009
CAP	0.017	0.821	0.412
CRR	0.013***	2.949	0.004
EXPM	-8.363	-0.492	0.624
LIQ	-0.573**	-2.323	0.022
GDPG	0.173***	7.004	0.000
INF	-0.040	-1.600	0.110
R. squared	0.431		
F- statistics	4.272		
Prob.(F-statistics)	0.000		

Note: (EBANK) a dummy variable for e-banking, (SZE) is the bank size, (CAP) is the capital ratio, (CRR) is the credit risk, (EXPM) is the expanses management, (LIQ) is the liquidity ratio, (GDPG) is the economic growth, and (INF) is the inflation. The results were corrected for heteroscedsticty using White matrix. ADF test confirmed that the variables are stationary.*p<.10. **p<.05. ***p<.01

The results of the regression analysis show that the electronic banking variable (EBANK) had a significant negative effect on the banks' performance. The authors explained this by pointing to the cost of electronic banking, which involves infrastructural costs, maintenance costs, and training programmes for employees. They considered that this cost was possibly higher than the revenue from electronic banking. The size of the banks showed a positive correlation with performance: big banks generated more profit than small banks. Capital was slightly related to performance measures. However, the higher the credit risk, the lower the performance of the bank. Therefore, managers had to work hard on credit policies to deal with credit risks. High expenses management and liquidity ratios led to lower profits. Therefore, the banks had to manage their costs and liquidity in a prudent way to avoid higher costs and excess liquidity. The results also showed that economic growth positively affected the performance of the

banks. There was an insignificant negative correlation with inflation. This indicates that good economic environments have a positive impact on bank performance.

According to Gensler, Leeflang, and Skiera (2012), online banking is vital because it enhances profits by increasing revenue from customers while reducing the costs required to serve those customers. They noted that it is not sufficient simply to compare online and offline channels as this will not give accurate results. They argued that three factors must be taken into consideration when comparing online and offline channels:

1. The effects of self-selection: customers differ in their characteristics (see Figure 2-12).
2. The need to distinguish between the revenue firms acquire from online customers and the cost to serve those customers.
3. The product portfolio: different products might affect online use.

The authors studied the effects of customers' online use on revenue and cost by considering the effects of self-selection effects and the banks' product portfolios.

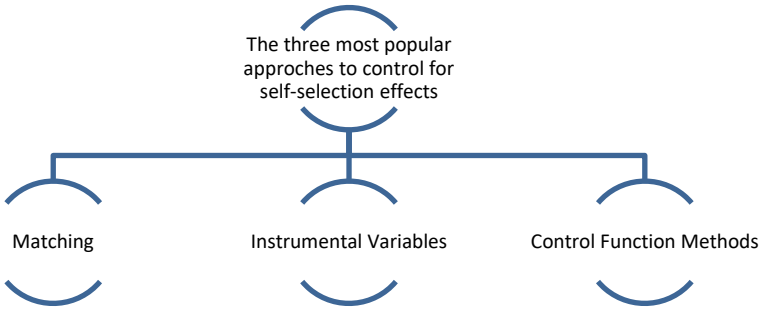


Figure 2-12: The three most popular approaches to control for self-selection effects

As Degeratu, Rangaswamy, and Wu (2000) explain, customer revenue from online use will increase and the cost to serve that customer will decrease only if the online networks encourage the customer to change his or her behaviour. In the conceptual framework of Gensler, Leeflang and Skiera, (2012), customer

behaviour is reflected through two elements: the customers' demand for the products and the number of transactions by customers in each online channel (see Figure 2-13).

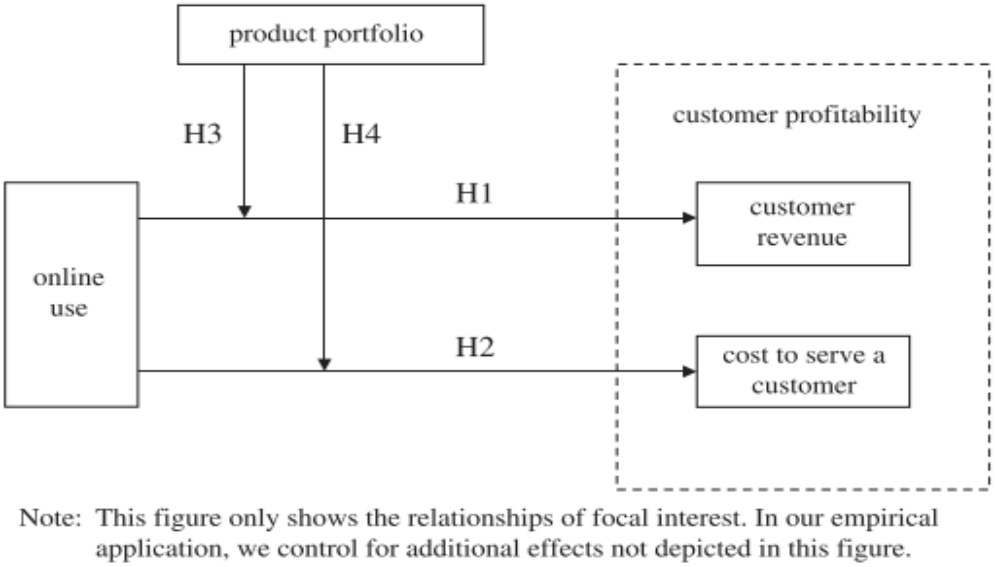


Figure 2-13: Conceptual framework (source: Gensler et al., 2012)

This study noted three ways in which online channels can increase customer demand for a product: by providing more access to information about the product, thereby influencing decision-making; by offering interactive tools to help customers make decisions; by making it possible for customers to make decisions at any time, thereby increasing convenience. Thus, customer demand will increase following the introduction of online channels, leading to increased revenue for firms. Firms should invest in online infrastructure so they can lower the cost to serve per transaction compared to offline channels.

As Danaher, Wilson, and Davis (2003) note, products that are frequently used benefit more from digitization than products that are rarely used. As such, product portfolios play a crucial role in the effect of electronic banking on customer habits

and profits. This study mentioned that profitability in the banking sector is related to customer revenue, cost to serve, and risks by the following equation:

Customer profitability = customer revenue – customer cost to serve – risk cost (2-1)

This empirical study used transaction data for approximately 87,000 customers from a European retail bank for a period of three months. Approximately 38% of the customers were registered for online banking, but only 2% used it actively. The data that was collected for this study were:

- Number of transactions per customer (monthly).
- Number of transactions per customer in each channel.
- Number and balance of checking accounts per customer.
- Number and balance of saving accounts per customer.
- Number and balance of brokerage account per customer.
- Number and balance of credit cards per customer (credit card balance or checking account balance).
- Customer age.
- Length of the customers' relationship with the bank.
- Number of joint checking accounts per customer

The researchers also obtained in-house information about the net contribution margins and the cost to serve, allowing them to calculate the profitability of each customer. They gathered data about the customers at two time points: six months prior to the study and three months after the start of the study. They found that

108 customers were enrolled online but did not use any online channels; these customers were treated as a “holdout sample”. The study used a hybrid matching method that estimated a customer’s likelihood of using an online channel based on the following factors:

- Customer age (AGE).
- Length of the customer's relationship with the bank (LOR).
- Customer's checking accounts (CHECK).
- Customer's savings accounts (SAV).
- Customer's brokerage accounts (BROK).
- Customer's credit cards (CREDIT).
- Customer's joint accounts (JOINT).

The effects of those factors on each customer’s propensity to use a given online channel are shown in Table 2-21. The results demonstrated that young customers used online banking more frequently than old customers; customers with who had been with their bank for longer used online channels less than newer customers; customers with checking accounts used online banking more than customers without checking accounts; customers with joint accounts or credit cards used online banking more frequently; customers with savings accounts or brokerage accounts used online banking less frequently. In summary, online banking has a positive effect on customer profitability, especially revenue. The percentage revenue increase is greater than the percentage decrease in the cost to serve. The effects of online banking also differ depending on the product portfolio.

Table 2-21: Parametres of customer characteristics in propensity score model (source: Gensler et al., 2012)

Customer characteristic	Parameter	*t*-Value	Reduction in bias (%)[a]
Age (years)	**−.04**	−8.51	55.36
Length of relationship (month)	**−.02**	−19.94	43.94
Joint account	**.32**	3.31	−82.14
Checking account ownership	**5.35**	5.34	74.20
Credit card ownership	**.48**	5.74	34.59
Brokerage account ownership	−.00	−.02	6.94
Savings account ownership	−.03	−.59	35.14
Intercept	**−7.22**	−7.16	
Sample size = 86,754; log-likelihood = −7023.04, pseudo R^2 = .131			

Bold values are significant at the 5% level.
[a] Computation based on Eq. (3).

According to Campbell and Frei (2010), firms are investing in self-service technologies to alter the behaviour of customers. This will result in lower costs and higher profitability. It will also improve the quality of services and increase customer retention. Campbell and Frei (2010) study focused on one self-service channel (online banking). They assessed its impact on customer behaviour by focusing on several variables: the number of transactions, profitability, cost, retention, and market share. Self-service technologies differ between industries, but the authors found that studies of the banking industry should focus on the relation between self-service channels and a given firm's performance. Their reasons for this are as follows. First, in retail banking (service firms), customer behaviour is the main reason why firms change any of their services. Second, it is in the interest of banks to lower their costs by introducing new self-service technologies. The data obtained by Campbell and Frei (2010) came from one of the biggest banks in the United States. The random sample included 100,000 customers who were enrolled in online banking in 2006. The monthly data set ran from the end of 2004 to May 2007. The following information was recorded about monthly transactions: the channel, type of product (checking account, loan, or investment), the number of accounts for each product, balance by product type, customer profitability, the relationship with the bank (tenure), and customer age. The researchers also extracted another data set from bank to examine the relation between online banking and customer retention. This was based on a random sample of 100,000 online and offline customers from the end of 2003 until the end of 2006. The same elements were recorded, along with each customer's relationship status at the end of 2004, 2005, and 2006 (to examine customer

retention over the years). The researchers collected information about the total number of monthly transactions across all channels (online banking, in branch, ATMs, call centres, and automated call centres). The spread of transactions across each channel is detailed in Table 2-22.

Table 2-22: distribution of transactions by channel during June 2006 (source: Campbell & Frei, 2010)

	Transactions (%)	Cumulative (%)
Branch		
Deposit	58.80	58.80
Cash check	27.25	86.05
Verify funds	6.71	92.76
Payment	4.24	97.01
Purchase	2.85	99.85
Withdrawal	0.12	99.97
Miscellaneous	0.03	100.00
ATM		
Withdrawal	65.21	65.21
Inquiry	18.04	83.25
Deposit	14.80	98.04
Transfer	1.30	99.34
Purchase	0.58	99.92
Payment	0.08	100.00
Online		
Query history	51.03	51.03
View basic account information	38.83	89.85
Payment	5.48	95.33
Transfer	4.06	99.40
Download financial information	0.60	100.00

Notes. This table shows the distribution of transaction types for each of the branch, ATM, and online channels. The percentages reported represent all transactions performed in each channel at National Bank during June 2006. Because of data limitations, similar breakdowns cannot be provided for either of the phone channels (VRU or call center).

The study calculated the profitability of customers based on the following equation: net interest income + fees & service charges – product unit cost – provision for loan loss. This provided information on the unit cost for each product type in each channel (determined by the bank's activity-based costing system), while also taking all overhead costs into account. The researchers also identified the fee revenue for each product in each channel.

The data were used to test the authors six hypotheses, which are summarized in Table 2-23.

Table 2-23: Six hypotheses

Hypothesis 1	Transactions in offline self-service channels decrease following the adoption of the online channel ("Substitution Effect").

Hypothesis 2	Transactions in offline assisted-service channels change following the adoption of the online channel ("Substitution Effect" or "Augmentation Effect").
Hypothesis 3	Total transaction volume increases following the adoption of the online channel ("Volume Effect").
Hypothesis 4	There is a change in customer profitability following the adoption of the online channel.
Hypothesis 5	There is an association between customer retention and the use of online banking.
Hypothesis 6	Market share is increasing in the degree of online banking penetration among a firm's customer base.

In their methodology, the researchers created a control group. Customers in the control group were enrolled in online banking but never used it. They were named "inactive enrollers". The researchers broke their analysis down into two modules: standard difference-in-difference estimation and economic specification. The latter considered more factors and helped to identify some issues such as the fact that the fixed differences between customers were related to unobservable phenomena (such as demographic characteristics). However, they dealt with these issues by using the conditional fixed-effect Poisson model (Hausman, Hall, and Griliches 1984). To test their hypotheses, the researchers summarized the equations used by Campbell and Frei (2010) ,in the basic Poisson model, the probability of observing yit transactions in a given offline channel for customer I at time t is as follows:

$$P(y_{it}) = \frac{e^{-\lambda_{it}} \lambda_{it}^{y_{it}}}{y_{it}!}, \qquad (2\text{-}2)$$

where λ_{it}, the Poisson parameter, is the expected value of yit.

Equation (2-3) is the basic Poisson model, which estimates the probability of observing transactions in an offline channel for a specific customer at a given time. They parameters were modelled to test the hypotheses using the following equations.

To test hypotheses 1, 2, and 3:

$$E(y_{it}) = exp(\beta_0 + \beta_1 POST_{it} + (\beta_2 POST_{it})(PASSIVE_i) + (\beta_3 POST_{it})(ACTIVE_i) + \gamma EnrollMonth_{it} + X_{it}\delta + \alpha_i), \quad (2\text{-}3)$$

Where online customers are divided into two groups: passive and active. This depends on the relationship between the customer's online transactions and the median number of all online transactions over a period. If there are fewer customer transactions than the median then the customer can be considered a passive customer. If there are more transactions than the median then customer is active. Use an indicator that takes values of 0 or 1 depending on the month being tested. If it is before the online enrolment then it takes a value of 0; if it is after the online enrolment, then it takes a value of 1. If the customer enrolled in an online channel during the months being study, then the "enrollmonth" indicator will take a value of 1. The authors constructed the equation for each online channel (transactions) and then reconstructed the equation for all offline channels to test the substitution and augmentation among the channels. For hypothesis 3, they replaced the online transactions in the equation with the total volume of online and offline transactions across the channels. They also included some other indicators, such as the number of accounts by product type and their balances. The controlled for seasonality and time by adding year and month indicators.

To test hypothesis 4:

$$P_{it} = \rho P_{it-1} + \lambda_1 POST_{it} + (\lambda_2 POST_{it})(PASSIVE_i) + (\lambda_3 POST_{it})(ACTIVE)_i + \gamma EnrollMonth_{it} + v_i + \mu_{it}, \quad (2\text{-}4)$$

The profit equation accounts for the customer specific fixed effect and the random error term.

To test hypothesis 5:

$$\Pr(RETAIN_i = 1|\cdot) = f(\gamma_0 + \gamma_1 ONLINEUSE_i + \gamma_2 TENURE_i + \gamma_3 AGE_i + X_i\beta) \quad (2\text{-}5)$$

The relationship with the bank and the age of the customer are important indicators in examining customer retention.

To test hypothesis 6:

$$MarketShare_{mt+1} = \gamma_0 + \rho MarketShare_{mt} + \gamma_1 OnlineRate_{mt} + \varepsilon_{mt+1} \quad (2\text{-}6)$$

Where market share is the share of all the deposits in the market that are controlled by the bank, and online rate represents the online penetration in a specific market.

The findings of Campbell and Frei (2010) illustrate that:

1. Online banking substituted for the use of self-service delivery channels (especially the costly ones, such as ATMs).
2. The introduction of online banking increased the assisted-service delivery channels (such as branches).
3. There was a significant increase in the total number of transactions.
4. There was an increase in the average cost to serve.
5. There was a decrease in customer profitability for the short term.
6. Online banking increased customer retention.

2.8 Summary

The amount of data relating to banking is growing rapidly. Much of it is unutilized information that could make a significant difference to the banking industry. Electronic banking (involving call centers, the internet, and mobile applications) has become a dominant way of exchanging financial information and communicating with banks. Data from electronic platforms are increasing, so there is a need for models that can make better use of the data. I began this chapter by looking at previous research into big data. Subsequently, I discussed several papers that examined the different techniques used in big data science and how they apply to finance. I then focused specifically on MATLAB and considered the advantages that it provides through its toolboxes. Finally, I examined the research into electronic banking and the factors that can influence the use of electronic channels. I recognized that most of the studies were limited in their data sets, relying on questionnaires and interviews. I also realized that there were several important variables that had not previously been addressed. Most importantly, no previous studies had addressed the importance of big data and electronic banking for Kuwait, which is the focus of the current study.

3 METHODOLOGY

3.1 Introduction

In this study, it was an important and difficult task to select, collect, clean, and test the data. I visited ten banks several times, travelling back and forth between Kuwait and the UK. I collected a substantial amount of confidential data, so I had to ensure that I had followed all the requisite legal procedures, including signing non-disclosure agreements with several independent banks, the Central Bank of Kuwait, and Cranfield University. I collected approximately 1TB of financial data from ten large and medium-sized Kuwaiti banks. The data included information about customer characteristics, product portfolios, banking channels (including branches), automated teller machines, points of sale, internet, call centres, and mobiles.my aim was to build a model to predict the electronic channels (dependent variables) that a customer would choose based on their information (independent variables). Given the data I collected and the evolution of data in the banking world today, I decided to use machine learning to handle the data as this would be the most beneficial way for financial analysts to understand the data.

3.2 Data selection and collection

The study was conducted to determine the appropriate electronic channel for each customer given their characteristics and financial portfolio. I reached this assessment by building a prediction model that banks can use to predict customer preferences based on their personal information. As discussed in the literature review, I examined previous studies to identify the most relevant variables and the data they would provide. I also used my 15 years of personal experience in the banking sector, my knowledge of local banks and the Central Bank of Kuwait, and my understanding of how the banking system works to identify several other independent variables that I thought would have a significant effect on the dependent variables.

I also identified some gaps in the literature, which I tried to address in my research. These gaps were as follows:

- The research data for all the studies were limited. Most of the studies used questionnaires with limited sample sizes, resulting in small financial data sets that were not generalizable.
- Most studies focused only on customers' preferences for electronic forms of communication as opposed to in-person branches. By contrast, I focused on different types of electronic channels that provided similar services, namely the internet, call centres, and mobile applications.
- In previous studies, the number of variables was limited; my study included a more comprehensive set of variables.
- None of the previous studies provided a model for predicting customer preferences based on their information. More importantly, there is no prediction model of this kind available for the Kuwait market.

The current research was sponsored by the Central Bank of Kuwait, who served as a liaison in helping us to obtain the data. They helped to process the confidentiality agreements and guarantee the security of the customer data for each of the local banks.

As mentioned, given the rapid evolution in technology, I decided to focus on a comparison among different types of electronic channels, rather than comparing electronic channels with in-person branches. The Central Bank of Kuwait eased the process of data collection by helping us obtaining data, handling the confidentiality requirements and agreements, from all local banks, including their entire customer base. The more data is available for each variable, the more explanatory power that variable will have. In my study, I also introduced new variables that had not previously been studied in this field. These included the number of loans a customer had taken out; credit card limit; outstanding credit card balance (one previous study had used outstanding credit card balance as an independent variable but not credit card limit); income; continent; and nationality. I also used some of the variables that had been identified by previous studies, though the type and size of my data set were different. Finally, I built a prediction

model for each bank using the data I obtained; such prediction models were not previously available in Kuwait and are in high demand.

After identifying all the dependent and independent variables, I have collected a panel of data from 10 Kuwaiti banks from 2008 to 2018. I requested the same data from all the different banks. The data for some variables were collected on a monthly basis; the data for other variables was collected on a daily basis. For others, I requested a monthly sample of 100,000 customers. Thus, the size of the data set was considerable, helping to improve the reliability and effectiveness of the results. I requested the retail numbers and monthly online transactions for every customer at every bank for each online channel (ATM, POS, online banking, mobile apps, automated call centre), though some of these online channels were removed later on, for reasons discussed below. I also collected data on the specific type of transaction that was performed through each of these channels (deposits, withdrawals, purchases, payments, or transfers). Finally, I also requested information about the following:

- Customer (anonymized) identifier
- Customer gender
- Customer address (I am not gathering the exact address but rather aggregating at district level)
- Customer age
- Date at which the customer's relationship with the bank began
- Customer level of education
- Customer nationality
- Customer income on a monthly basis
- Daily customer transactions, including the date, value, and currency of each transaction, as well as the type of card used and the type of transaction (purchase, withdrawal, payment, deposit, or transfer)

- Number of current accounts, savings accounts, retail deposits, and joint accounts, along with their respective values.
- Total amount, principle paid, interest paid, and outstanding balance of consumer loans and instalment loans
- Consumer risk factor
- Credit card limits and any outstanding balances
- The number and value of each transaction for each service (withdrawals, deposits, transfers, purchases, and payments) across each online channel (ATM, POS, online banking, mobile application, and automated call centre) for each customer.

Having identified the variables that, I needed to test for, I arranged appointments with all the local banks in Kuwait. The Central Bank of Kuwait coordinated these meetings. I started attending these meetings in December 2019, travelling from the UK to Kuwait several times. I explained the kind of data and size of data set that I required, and I also explained the rationale behind the study.

Obtaining the data was not straightforward. The banks were keen to know whether sharing this confidential information would benefit them in the future. I explained that the purpose of the research was to change how banks operate, ultimately leading to increased profits. The banks were eventually convinced by this and agreed to provide us with the requested data.

Due to the amount of data requested, the banks required between two and six months to send us all of the information. During that period, I visited the banks several times to make sure that they were preparing the correct type, size, and format of data. At the same time, I also finished drafting and signing the non-disclosure agreements between ourselves and the Central Bank of Kuwait and between the Central Bank of Kuwait and Cranfield University. Once the agreements were in place, I received the data from the banks in comma separated values (CSV) format. Each of the banks delivered the data on a separate hard

drive, and I then made three back-ups of each of the hard drives and stored them in different locations.

Once I received the data, I began to investigate whether there were any clear correlations between the independent variables. I created a correlation matrix for each bank, which will be discussed in more detail in chapter 5.

3.3 Data limitations

As mentioned above, the data collection process was somewhat difficult and required the collaboration of the local banks and the Central Bank of Kuwait. As a result, there were some limitations to my final data set, which are summarized below:

- Some of the banks did not have records for their customers' level of education, and some had information that was out of date.
- Some banks did not provide us with information about their customers' nationalities.
- In response to my request for daily transaction data, some banks provided us with all of the transaction amounts in Kuwaiti Dinars, even when a transaction had been carried out in another currency.
- Only a few banks provided us with details about joint accounts.
- One of the banks provided us with data for a shorter period (from 2011 to 2018).
- Some banks did not record credit risk factors, and others provided us with a substitute type of risk assessment based on their own definition.
- Most of the local banks used a third party to handle credit cards, so it was difficult to get information about customer credit cards. One of the banks only provided us with four years of information about credit cards.
- Some banks combined online banking with mobile app transactions, which made it difficult to study these two variables in isolation.

- Some banks recorded all transactions made through a mobile application as transfers. They did not record them as purchases or payments. This made it difficult to the study of effect of each service separately.
- Some banks did not have a specific field for Governance, so it was picked up from the address fields where possible.
- Some of the banks had no historical data on risk factors.
- One of the banks had finished their 'Core Banking' upgrade in October 2017. This involved the development of their digital banking channels. As a result, the data for this bank was more granular after the upgrade, which caused some illusory fluctuations.
- The electronic channels had been introduced at different times at different banks. Occasionally, some of the information was recorded for two services under the same name. For an explanation of this, see Table 3-1, which provides information for a sample data set.

Table 3-1: Data available for electronic channels for a sample bank

Data available	Deposits	withdrawals	purchases	payments	transfers
ATM	2017-2018	2008-2018			
P.O.S			2008-2018		
Internet				2017-2018	2008-2018
Mobile				2015-2018	2017-2018
Call Center				2008-2018	2017-2018

To take an example from Table 3-1, when looking at call centres, I can see that the payments and transfers were recorded under payments until 2017. From 2017 onwards, the bank started to record payments and transfers separately.

Given these limitations, I was forced to remove four of the banks from my data set. I also had to discard some of the variables from the study. However, thanks to the contract I negotiated with the Central Bank of Kuwait, the data will remain available after the end of the PhD, so I will be able to use the data from the banks in future studies.

I also decided to focus solely on electronic channels that provided the same services: the Internet, mobile applications, and call centres. All of these channels provided both payments and transfers, whereas the other offered other services like deposits, withdrawals, and purchases. Finally, I focused on the number of transactions for each channel rather than the value of each transaction. This was because I believe that people with older ages generate more wealth and eventually engage in transactions of a higher value, negatively influencing the analysis of usage of electronic channels.

3.4 Data type and size

In this section, I created a table for each bank to summarize the type and size of raw data I obtained (Tables Table 3-2 to Table 3-7). Each table specifies the names of the files, the file elements, formatting type, and size of the data. I also produced a second other set of tables (Tables Table 3-8 toTable 3-13) to record the data once it had been transferred into MATLAB. These tables record the file name, file component, file format, file size, and number of observations.

Table 3-2. Bank 1, conventional, raw data

Name	Comprising (Column Headings)	Type	Size (mb)
CBK1_Branch	CIF_ID Transaction Date Value of Transaction Billing_Currency Type_Of_Card	Comma separated values file	245.56

Name	Comprising (Column Headings)	Type	Size (mb)
	Tranxaction_Type Year		
CBK1_Cards (2009 and above)	CIF_ID Transaction Date Value of Txn BILLING_CURRENCY TYPE_OF_CARD TXN_TYPE Year	Comma separated values file	1,993.36
CBK1_Cards_2008	CIF_ID Transaction Date Value of Txn BILLING_CURRENCY TYPE_OF_CARD TXN_TYPE Year	Comma separated values file	138.03
CBK2_Customer_Data	CIF_ID Age Gender Address Relationship date	Comma separated values file	12.99
CBK2_Customer_Data_Nationality	A_CIF_ID AGE GENDER ADDRESS Relationship Date Nationality Country description	Comma separated values file	24.25
CBK2_Education_Level	CIF_ID Age Gender Address Education level	Comma separated values file	11.08
CBK3_Income	CIF_ID		62.61

Name	Comprising (Column Headings)	Type	Size (mb)
	The income from January until December year	Comma separated values file	
CBK4_Current Account Balance	CIF_ID The balance from January until December year	Comma separated values file	20.71
CBK4_ Number of Current Account	CIF_ID The Current Account Number from January until December year	Comma separated values file	8.76
CBK4_Deposit Account Balance	CIF_ID The balance from January until December year	Comma separated values file	6.11
CBK4_Number of Deposit Account	CIF_ID The Deposit Account Number from January until December year	Comma separated values file	2.72
CBK4_Joint Account Balance	CIF_ID The balance from January until December year	Comma separated values file	0.90
CBK4_Number of Joint Account	CIF_ID The Joint Account Number from January until December year	Comma separated values file	0.38
CBK4_Saving Account Balance	CIF_ID The balance from January until December Year	Comma separated values file	84.14

Name	Comprising (Column Headings)	Type	Size (mb)
CBK4_Number of Saving Account	CIF_ID The Saving Account Number from January until December year	Comma separated values file	37.34
CBK5_Consumer_Loan_Count	CIF_ID The count from January until December Year	Comma separated values file	4.54
CBK5_Consumer_Loan_Interest	CIF_ID The Interest from January until December Year	Comma separated values file	5.56
CBK5_Consumer_Loan_Outstanding Balance	CIF_ID The Outstanding Balance from January until December Year	Comma separated values file	10.68
CBK5_Consumer_Loan_Principle_Payment	CIF_ID The Principle Payment from January until December Year	Comma separated values file	5.88
CBK5_Installment_Loan_Count	CIF_ID The count from January until December Year	Comma separated values file	6.36
CBK5_ Instalment _Loan Interest	CIF_ID The Interest from January until December Year	Comma separated values file	22.72
CBK5_ Instalment _Loan Outstanding Balance	CIF_ID The Outstanding Balance from January until December Year	Comma separated values file	19.35

Name	Comprising (Column Headings)	Type	Size (mb)
CBK5_Installment_Loan_Principle_Payment	CIF_ID The Principle Payment from January until December Year	Comma separated values file	23.75
CBK6_ Credit Cards limit	CIF_ID The Credit Card Limit from January until December Year	Comma separated values file	12.95
CBK6_Credit Cards Outstanding Balance	CIF_ID The Outstanding Credit Card Balance from January until December Year	Comma separated values file	12.36
CBK6_Risk_Factor	CIF_ID Customer risk score (1-100) Customer risk rating (low-medium-high)	Comma separated values file	5.64
CBK7_ATM	CIF_ID Year and month Monthly Withdrawals (counts & values) Monthly deposits (counts and values) Monthly transfers (counts and values) Monthly purchases (count and values) Monthly payments (counts and values)	Comma separated values file	43.85
CBK7_Call_Center	CIF_ID Year and month Monthly transfers (counts and values) Monthly purchases (count and values) Monthly payments (counts and values)	Comma separated values file	3.86

Name	Comprising (Column Headings)	Type	Size (mb)
CBK7_Internet	CIF_ID Year and month Monthly transfers (counts and values) Monthly purchases (count and values) Monthly payments (counts and values)	Comma separated values file	7.86
CBK7_Mobile	CIF_ID Year and month Monthly transfers (counts and values) Monthly purchases (count and values) Monthly payments (counts and values)	Comma separated values file	5.09
CBK7_POS	CIF_ID Year and month Monthly transfers (counts and values) Monthly purchases (count and values) Monthly payments (counts and values)	Comma separated values file	44.68
Total raw data size for bank 1			2,884.22

Table 3-3. Bank 2, conventional, raw data

Name	Comprising (Column Headings)	Type	Size (mb)
CBK1 Daily Account Transactions	CIF_ID Transaction Date Billing Currency Type of Card Transaction Type Value of Transactions	Comma separated values file	6,922
	CIF_ID Age		4

Name	Comprising (Column Headings)	Type	Size (mb)
CBK2_Customer_Data Customer General Information	Gender Address Date of relationship	Comma separated values file	
CBK3_Income Customer Income 2008-2018	CIF_ID The income from January until December for each year	Comma separated values file	111
CBK4_Accounts counts and balances 2008-2018	CIF_ID Type of account (current, saving or deposit) The Account counts from January until December for each month in 8 years	Comma separated values file	482
CBK5_Loans 2008-2018	CIF_ID Loan type Loans counts Total loan amount Principle paid Interest paid Outstanding balance From January until December for each year	Comma separated values file	298
CBK6_ Credit Cards	CIF_ID The Credit Card Limit The Outstanding Credit Card Balance From January until December for each year	Comma separated values file	697
CBK7_Electronic banking transactions (ATM, Call centre, Mobile, Internet, POS) 2008-2018	CIF_ID Year and month Monthly Withdrawals (counts & values) Monthly deposits (counts and values) Monthly transfers (counts and values) Monthly payments (counts and values)	Comma separated values file	95.2
	Customer ID		8.74

Name	Comprising (Column Headings)	Type	Size (mb)
CBK8_Customer information 2008-2018	Nationality address Gender	Comma separated values file	
Total raw data size for bank 2			8,609

Table 3-4. Bank 3, Islamic, raw data

Name	Comprising (Column Headings)	Type	Size (mb)
CBK1 Daily Account Transactions	CIF_ID Dates Business Date Number of Transactions Value of Transactions Value of Transaction in Foreign Currency Currency Transaction Type Type of Card	Comma separated values file	16,015.94
CBK2_Customer_Data Customer General Information - All Active Customers	CIF_ID Age Gender Address Education	Comma separated values file	35.95
CBK3_Income Customer Income 2008-2018 - All Active Customers	CIF_ID The income From January until December for each 11 years	Comma separated values file	137.53
CBK4_Accounts counts and balances 2008-2018 – All Customers	CIF_ID Type of account (current, saving or deposit) The Account counts from January until December for each month in 11 years The account balance from January until December for each month in 11 years	Comma separated values file	834.83

Name	Comprising (Column Headings)	Type	Size (mb)
CBK5_Loans 2008-2018 – All Active Customers	CIF_ID Loan type Loans counts Total loan amount Principle paid Interest paid Outstanding balance From January until December for each month in the 11 Years	Comma separated values file	285.54
CBK6_ Credit Cards 2008-2018 – All Active Customers	CIF_ID The Credit Card Limit The Outstanding Credit Card Balance From January until December for each month in the 11 Years	Comma separated values file	42.37
CBK 6 Credit Card - Credit Score -2008-2018 – all active Customers	CIF_ID Credit card credit scoring	Comma separated values file	118.15
CBK7_Electronic banking transactions (ATM, Call centre, Mobile Internet, POS) 2008-2018 - All Active Customers	CIF_ID Year and month Monthly Withdrawals (counts & values) Monthly deposits (counts and values) Monthly transfers (counts and values) Monthly payments (counts and values)	Comma separated values file	1,351.51
CBK8_Customer information 2008-2018 - All Active Customers	CIF_ID Education level	Comma separated values file	528.78
Total raw data size for bank 3			19,350.60

Table 3-5. Bank 4, conventional, raw data

Name	Comprising (Column Headings)	Type	Size (mb)
CBK1 Daily Account Transactions	CIF_ID Transaction Date Value of Transactions Billing Currency Type of Card Transaction Type Year	Comma separated values file	32,101
CBK2_Customer_Data Customer General Information - All Active Customers	CIF_ID Age Gender Address Relationship date	Comma separated values file	30.41
CBK2_Education_Level Customer Education 2014-2018 - All Customers	CIF_ID Monthly Education level	Comma separated values file	534.11
CBK3_Income Customer Income 2014-2018 - All Active Customers	CIF_ID The income from January until December for each month in the 5 years	Comma separated values file	286.28
CBK4_Current Account Balance Current Account - End Balance 2014-2018 - 100k Customers	CIF_ID The balance from January until December for each month in the 5 years	Comma separated values file	2.12
CBK4_ Number of Current Account Current Account - No. of Account 2014-2018 - 100k Customers	CIF_ID The Current Account Counts from January until December for each month in 5 years	Comma separated values file	0.77
CBK4_Deposit Account Balance Term Deposits - End Balance 2014-2018 - 100k Customers	CIF_ID The balance from January until December for each month in the 5 years	Comma separated values file	0.43
CBK4_Number of Deposit Account Term Deposits - No. of Account 2014-2018- 100k Customers	CIF_ID The Current Account Counts from January	Comma separated values file	0.201

Name	Comprising (Column Headings)	Type	Size (mb)
	until December for each month in 5 years		
CBK4_Saving Account Balance Savings Account - End Balance -2014-2018 - 100k Customers	CIF_ID The balance from January until December for each month in the 5 years	Comma separated values file	36.99
CBK4_Number of Saving Account Savings Account - No. of Account 2014-2018 - 100k Customers	CIF_ID The Current Account Counts from January until December for each month in 5 years	Comma separated values file	11.96
CBK5_Consumer_Loan_Count Consumer - No. of Account 2014-2018 - 100k Customers	CIF_ID The count from January until December for each month in the 5 Years	Comma separated values file	1.57
CBK5_Consumer_Loan_Interest Loans - Consumer - Interest Paid 2014-2018 - 100k Customers	CIF_ID The Interest from January until December for each month in the 5 Years	Comma separated values file	7.53
CBK5_Consumer_Loan_Outstanding Balance Loans - Consumer - Outstanding Balance 2014-2018- 100k Customers	CIF_ID The Outstanding Balance from January until December for each month in the 5 Years	Comma separated values file	5.78
CBK5_Consumer_Loan_Principle_Payment Loans - Consumer - Principal Paid 2014-2018 - 100k Customers	CIF_ID The Principle Payment from January until December for each month in the 5 Years	Comma separated values file	10.11
CBK5_Installment_Loan_Count Loans - Instalment - No. of Account 2014-2018 - 100k Customers	CIF_ID The count from January until December for each month in the 5 Year	Comma separated values file	1.92
CBK5_ Instalment _Loan Interest Loans - Instalment - Interest Paid 2014-2018 - 100k Customers	CIF_ID The Interest from January until December for each month in the 5	Comma separated values file	15.72

Name	Comprising (Column Headings)	Type	Size (mb)
	Years		
CBK5_ Instalment _Loan Outstanding Balance Loans - Instalment - Outstanding Balance 2014-2018 - 100k Customers	CIF_ID The Outstanding Balance from January until December for each month in the 5 Years	Comma separated values file	8.19
CBK5_Installment_Loan_Principle_Payment Loans - Instalment - Principal Paid 2014-2018 - 100k Customers	CIF_ID The Principle Payment from January until December for each month in the 5 Years	Comma separated values file	14.62
CBK6_ Credit Cards limit Credit Card - Total Limit 2014-2018 - 100k Customers	CIF_ID The Credit Card Limit from January until December for each month in the 5 Years	Comma separated values file	6.28
CBK6_Credit Cards Outstanding Balance Credit Card - Balance 2014-2018 - 100k Customers	CIF_ID The Outstanding Credit Card Balance from January until December for each month in the 5 Years	Comma separated values file	10.31
CBK 6 Credit Card - Credit Score -2016-2018 - 100k Customers	CIF_ID Credit card credit scoring	Comma separated values file	0.43
CBK5 Loans - Credit Score -2016-2018 - 100k Customers	CIF_ID Loan credit score	Comma separated values file	0.41
CBK7_ATM ATM Transactions 2014-2018 - All Active Customers	CIF_ID Year and month Monthly Withdrawals (counts & values) Monthly deposits (counts and values)	Comma separated values file	533.72
CBK7_Call_Center Automated Call Centre Transactions 2014-2018 - All Active Customers	CIF_ID Year and month Monthly transfers (counts and values) Monthly payments (counts and values)	Comma separated values file	32.43

Name	Comprising (Column Headings)	Type	Size (mb)
CBK7_Internet Online Banking Transactions 2014-2018 - All Active Customers	CIF_ID Year and month Monthly transfers (counts and values) Monthly payments (counts and values)	Comma separated values file	81.75
CBK7_Mobile Mobile Banking Transactions 2014-2018 - All Active Customers	CIF_ID Year and month Monthly transfers (counts and values) Monthly payments (counts and values)	Comma separated values file	65.5
CBK7_POS POS Transactions 2014-2018 - All Active Customers	CIF_ID Year and month Monthly purchases (count and values)	Comma separated values file	557.64
CBK8_Customer_Data	CIF_ID Education level	Comma separated values file	610.53
Total raw data size for bank 4			34,969

Table 3-6. Bank 5, Islamic, raw data

Name	Comprising (Column Headings)	Type	Size (mb)
CBK1 Daily Account Transactions	CIF_ID Transaction Date Billing Currency Type of Card Transaction Type Value of Transactions	Comma separated values file	488.37
CBK2_Customer_Data Customer General Information	CIF_ID Age Gender Address	Comma separated values file	13,420.98
	CIF_ID		7.6

Name	Comprising (Column Headings)	Type	Size (mb)
CBK3_Income Customer Income 2012-2018 -	The income from January until December for each year	Comma separated values file	
CBK4_Accounts counts and balances 2012-2018	CIF_ID Type of account (current, saving or deposit) The Account counts from January until December for each month in 11 years The account balance from January until December for each year	Comma separated values file	13,420.98
CBK5_Loans_ 2012-2018	CIF_ID Loan type Loans counts Total loan amount Principle paid Interest paid Outstanding balance From January until December for each year	Comma separated values file	77.33
CBK6_ Credit Cards	CIF_ID The Credit Card Limit The Outstanding Credit Card Balance From January until December for each year	Comma separated values file	6.29
CBK7_Electronic banking transactions (ATM, Call centre, Mobile, Internet, POS) 2012-2018	CIF_ID Year and month Monthly Withdrawals (counts & values) Monthly deposits (counts and values) Monthly transfers (counts and values) Monthly payments (counts and values)	Comma separated values file	69.2
	CIF_ID		1.83

Name	Comprising (Column Headings)	Type	Size (mb)
CBK8_Customer information 2012-2018	Customer nationality Date of relationship	Comma separated values file	
Total raw data size for bank 5			27,493

Table 3-7. Bank 6, Islamic, raw data

Name	Comprising (Column Headings)	Type	Size (mb)
CBK1 Daily Account Transactions	CIF_ID Transaction Date Value of Transactions Billing Currency Type of Card Transaction Type	Comma separated values file	91,399.90
CBK2_Customer_Data Customer General Information	CIF_ID Age Gender Address Date of relationship Nationality	Comma separated values file	63,823.67
CBK3_Income Customer Income 2008-2018	CIF_ID The income from January until December for each year	Comma separated values file	490.47
CBK4_Accounts counts and balances 2008-2018	CIF_ID Type of account (current, saving or deposit) The Account counts from January until December for each month in 8 years The account balance from January until December for each year	Comma separated values file	2,197.98
CBK5_Loans_ 2008-2018	CIF_ID Loan type Loans counts	Comma separated values file	4,660.12

Name	Comprising (Column Headings)	Type	Size (mb)
	Total loan amount Principle paid Interest paid Outstanding balance from January until December for each year		
CBK6_ Credit Cards	CIF_ID The Credit Card Limit The Outstanding Credit Card Balance from January until December for each year	Comma separated values file	165.93
CBK7_Electronic banking transactions (ATM, Call centre, Mobile, Internet, POS) 2008-2018	CIF_ID Year and month Monthly Withdrawals (counts & values) Monthly deposits (counts and values) Monthly transfers (counts and values) Monthly payments (counts and values)	Comma separated values file	631.49
CBK8_Customer information 2008-2018	Customer General information Customer Income Customer education history	Comma separated values file	2,500.32
Total raw data size for bank 6			165,869.88

From the raw data contained in the previous tables, I can see what types of data have been collected. The financial data set is detailed and private. Furthermore, as mentioned, the data was collected on daily and monthly bases, thereby enhancing its accuracy.

After collecting the data, I had to clean it. This is discussed in the following section. The data after cleaning is presented in Tables 3-8 to 3-13.

Table 3-8. Bank 1, conventional, data transferred from CSV to MATLAB

Name	Comprising (Column Headings)	Type	Format	Number of Observations	Size (mb)
CBK1_Branch	CIF_ID Transaction Date Value of Txn BILLING_CURRENCY TYPE_OF_CARD TXN_TYPE Year	MATLAB	Table	40,483,159	50,851
CBK1_Cards (2009 and above)	CIF_ID Transaction Date Value of Txn BILLING_CURRENCY TYPE_OF_CARD TXN_TYPE Year				
CBK2_Customer_Data	A_CIF_ID AGE GENDER ADDRESS	MATLAB	Table	295,801	249.45

Name	Comprising (Column Headings)				Type	Format	Number of Observations				Size (mb)	
	Relationship Date Nationality											
CBK3_Income	CIF_ID The Monthly income From January until December for each customer dates				MATLAB	Table	8,345,352				17.37	
CBK 4_Accounts	TABLE 1 CIF_ID Dates CA. counts DA counts JA counts SA counts		TABLE 2 CIF_ID Dates CA. balances DA balances JA balances SA balances		MATLAB	Structure (table 1: counts, table 2: balances	TABLE 1 12,552,552		TABLE 2 12,552,552		49.44	
CBK5_Loans	Table 1 CIF_ID Dates Consumer counts	Table2 CIF_ID Dates Consumer interest	Table3 CIF_ID Dates Consumer outstand.	Table4 CIF_ID Dates Consumer principle	MATLAB	Structure (Table 1: counts, Table 2: interest, Table 3: outstanding, Table 4: principle)	Table1 2,639,784	Table2 2,198,508	Table3 2,639,784	Table4 2,147,100	33.57	

Name	Comprising (Column Headings)				Type	Format	Number of Observations			Size (mb)	
	Instalments counts	Instalment interest	Instalment outstand.	Instalment principle							
CBK6_Credit cards and Risk factor	CIF_ID Dates Credit card limit Credit card outstanding balances Risk score Risk rating				MATLAB	Table	1,659,112			686.36	
CBK7_Electronic channels	Table 1 Table 2 Table 3 Table 4 Table 5 For each table: CIF_ID Year and month Monthly Withdrawals (counts & values) Monthly deposits (counts and values) Monthly transfers (counts and values) Monthly purchases (count and values) Monthly payments (counts and values)				MATLAB	Structure (Table 1: ATM, Table 2: call centre, Table 3: internet, Table 4: mobile, Table 5: P.O.S)	Table1: 735,356 Table2: 63,792 Table3: 129,215 Table4: 83,597 Table5: 751,168			11.85	
CBK8_Education_Level	CIF_ID Age Gender Address				MATLAB	Table	236,927			295	

Name	Comprising (Column Headings)	Type	Format	Number of Observations	Size (mb)
	Education level				
Total MATLAB data					52,194

Table 3-9. Bank 2, conventional, data transferred from CSV to MATLAB

Name	Comprising (Column Headings)		Type	Format	Number of Observations		Size (mb)
CBK1 Monthly Average Account Transactions	CIF_ID Dates Number of Transactions Value of Transactions		MATLAB	Table	7,853,183		160
CBK2_Customer_ Data	CIF_ID Age Gender Address Date of relationship		MATLAB	Table	95,429		0.92
CBK3_Income	CIF_ID The Monthly income From January until December for each customer Dates		MATLAB	Table	3,357,771		10.26
CBK 4_Accounts	Table 1 CIF_ID	Table 2 CIF_ID	MATLAB	Structure (table 1:	Table 1 4,103,174	Table 2 4,110,543	18.68

Name	Comprising (Column Headings)				Type	Format	Number of Observations				Size (mb)
	Dates CA. counts JA. Count		Dates CA. balances JA balances			counts, table 2: balances)					
CBK5_Loans	Table 1 CIF_ID Dates Consumer counts Instalments counts	Table2 CIF_ID Dates Consumer interest Instalment interest	Table3 CIF_ID Dates Consumer outstand. Instalment outstand	Table4 CIF_ID Dates Consumer principle Instalment principle	MATLAB	Structure (table 1: counts, table 2: interest, table 3: outstanding, table 4: principle)	Table1 976,156	Table2 882,911	Table3 950,305	Table4 875,343	13.63
CBK6_Credit cards	CIF_ID Dates Credit card limit Credit card outstanding balances Credit Score				MATLAB	Table	1,213,261				2.46
CBK7_Electronic channels	Table 1 Table 2 Table 3 Table 4 Table 5 For each table: CIF_ID Year and month Monthly Withdrawals (counts & values)				MATLAB	Structure (table 1: ATM, table 2: call centre, table 3: internet, table 4: mobile, table 5: P.O.S)	Table1: 1,431,388 Table2: 33,048 Table3: 120,989				7.45

Name	Comprising (Column Headings)	Type	Format	Number of Observations	Size (mb)
	Monthly deposits (counts and values) Monthly transfers (counts and values) Monthly purchases (count and values) Monthly payments (counts and values)			Table4: 2,017 Table5: 101,190	
CBK8_Customer_ Data	CIF_ID Nationality Address Gender	MATLAB	Table	240,993	1.23
Total MATLAB Data					214

Table 3-10. Bank 3, Islamic, data transferred from CSV to MATLAB

Name	Comprising (Column Headings)		Type	Format	Number of Observations		Size (mb)
CBK1 Monthly Average Account Transactions	CIF_ID Dates Number of Transactions Value of Transactions		MATLAB	Table	3,796,794		21.02
CBK2_Customer_Data	CIF_ID Age Gender Address Relationship date Education		MATLAB	Table	392,995		3.88
CBK3_Income	CIF_ID The Monthly income From January until December for each customer Dates		MATLAB	Table	4,950,996		24.2
CBK 4_Accounts	Table 1 CIF_ID Dates CA. counts	Table 2 CIF_ID Dates CA. balances	MATLAB	Structure (table 1: counts, table 2: balances)	Table 1 15,201,378	Table 2 15,201,378	71.69

Name	Comprising (Column Headings)				Type	Format	Number of Observations				Size (mb)
CBK5_Loans	Table 1 CIF_ID Dates Consumer counts Instalments counts	Table2 CIF_ID Dates Consumer interest Instalment interest	Table3 CIF_ID Dates Consumer outstand. Instalment outstand.	Table4 CIF_ID Dates Consumer principle Instalment principle	MATLAB	Structure (table 1: counts, table 2: interest, table 3: outstanding, table 4: principle)	Table1 2,878,089	Table2 2,878,089	Table3 2,878,089	Table4 2,878,089	62.21
CBK6_Credit cards	CIF_ID Dates Credit card limit Credit card outstanding balances				MATLAB	Table	1,079,603				3.48
CBK7_Electronic channels	Table 1 Table 2 Table 3 Table 4 Table 5 For each table: CIF_ID Year and month Monthly Withdrawals (counts & values) Monthly deposits (counts and values) Monthly transfers (counts and values) Monthly purchases (count and values) Monthly payments (counts and values)				MATLAB	Structure (table1: ATM, table 2: call centre, table 3: internet, table 4: mobile, table 5: P.O.S)	Table1: 7,734,642 Table2: 22,450 Table3: 1,072,257 Table4: 813,364 Table5: 7,547,495				250.56
CBK8_Customer_Data	CIF_ID Education level				MATLAB	Table	392,995				0.73

Name	Comprising (Column Headings)	Type	Format	Number of Observations	Size (mb)
Total MATLAB Data					437.77

Table 3-11. Bank 4, conventional, data transferred from CSV to MATLAB

Name	Comprising (Column Headings)		Type	Format	Number of Observations		Size (mb)
CBK1 Monthly Average Account Transactions	CIF_ID Dates Number of Transactions Value of Transactions		MATLAB	Table	400,781,903		12,825
CBK2_Customer_Data	CIF_ID Age Gender Address Relationship date		MATLAB	Table	875,191		1,146
CBK3_Income	CIF_ID The Monthly income From January until December for each customer Dates		MATLAB	Table	52,511,520		65.15
CBK 4_Accounts	Table 1 CIF_ID	Table 2 CIF_ID	MATLAB	Structure (table 1:	Table 1 5,370,264	Table 2 5,370,264	19.15

Name	Comprising (Column Headings)				Type	Format	Number of Observations				Size (mb)
	Dates CA. counts		Dates CA. balances			counts, table 2: balances)					
CBK5_Loans	Table 1 CIF_ID Dates Consumer counts Instalments counts	Table2 CIF_ID Dates Consumer interest Instalment interest	Table3 CIF_ID Dates Consumer outstand. Instalment outstand.	Table4 CIF_ID Dates Consumer principle Instalment principle	MATLAB	Structure (table 1: counts, table 2: interest, table 3: outstanding, table 4: principle)	Table1 1238088	Table2 1207296	Table3 1238088	Table4 1211628	16.72
CBK6_Credit cards and Credit Card - Credit Score	CIF_ID Dates Credit card limit Credit card outstanding balances Credit cards credit score				MATLAB	Table	2,140,200				6.1
CBK7_Electronic channels	Table 1 Table 2 Table 3 Table 4 Table 5 For each table: CIF_ID Year and month Monthly Withdrawals (counts & values)				MATLAB	Structure (table1: ATM, table 2: call centre, table 3: internet,	Table 1: 2,682,302 Table 2: 139,701 Table 3: 327,794				33.31

Name	Comprising (Column Headings)	Type	Format	Number of Observations	Size (mb)
	Monthly deposits (counts and values) Monthly transfers (counts and values) Monthly purchases (count and values) Monthly payments (counts and values)		table 4: mobile, table 5: P.O.S)	Table 4: 263,586 Table 5: 2,512,264	
CBK8_Customer_Data	CIF_ID Education level	MATLAB	Table	1,750,384	3.73
Total MATLAB Data					14,115

Table 3-12. Bank 5, Islamic, data transferred from CSV to MATLAB

Name	Comprising (Column Headings)	Type	Format	Number of Observations	Size (mb)
CBK1 Monthly Average Account Transactions	CIF_ID Dates Number of Transactions Value of Transactions	MATLAB	Table	663,667	4.98
CBK2_Customer_Data	CIF_ID Age Gender Address	MATLAB	Table	70,501	0.92

Name	Comprising (Column Headings)				Type	Format	Number of Observations				Size (mb)
CBK3_Income	CIF_ID The Monthly income From January until December for each customer Dates				MATLAB	Table	599,028				1.1
CBK 4_Accounts	Table 1 CIF_ID Dates CA. counts JA. Count		Table 2 CIF_ID Dates CA. balances JA balances		MATLAB	Structure (table 1: counts, table 2: balances)	Table 1 1,327,392		Table 2 1,327,392		5.06
CBK5_Loans	Table 1 CIF_ID Dates Consumer counts Instalments counts	Table2 CIF_ID Dates Consumer interest Instalment interest	Table3 CIF_ID Dates Consumer outstand Instalment outstand.	Table4 CIF_ID Dates Consumer principle Instalment principle	MATLAB	Structure (table 1: counts, table 2: interest, table 3: outstanding, table 4: principle)	Table1 2,146,620	Table2 2,146,620	Table3 2,146,620	Table4 2,146,620	9.37
CBK6_Credit cards	CIF_ID Dates				MATLAB	Table	279,600				0.62

Name	Comprising (Column Headings)	Type	Format	Number of Observations	Size (mb)
	Credit card limit Credit card outstanding balances				
CBK7_Electronic channels	Table 1 Table 2 Table 3 Table 4 Table 5 For each table: CIF_ID Year and month Monthly Withdrawals (counts & values) Monthly deposits (counts and values) Monthly transfers (counts and values) Monthly purchases (count and values) Monthly payments (counts and values)	MATLAB	Structure (table1: ATM, table 2: call centre, table 3: internet, table 4: mobile, table 5: P.O.S)	Table1: 892,089 Table2: 1,100,740 Table3: 412,332 Table4: 278,556 Table5: 1,005,757	9.13
CBK8_Customer_Data	CIF_ID Nationality Date of relationship	MATLAB	Table	70,501	28.98
Total MATLAB Data					55.18

Table 3-13. Bank 6, Islamic, data transferred from CSV to MATLAB

Name	Comprising (Column Headings)		Type	Format	Number of Observations		Size (mb)
CBK1 Monthly Average Account Transactions	CIF_ID Dates Number of Transactions Value of Transactions		MATLAB	Table	74,632,245		515.37
CBK2_Customer_Data	CIF_ID Age Gender Address Date of relationship Nationality		MATLAB	Table	1,382,942		9.13
CBK3_Income	CIF_ID The Monthly income From January until December for each customer Dates		MATLAB	Table	34,566,719		93.95
CBK 4_Accounts	Table 1 CIF_ID Dates CA. counts JA. Count	Table 2 CIF_ID Dates CA. balances JA balances	MATLAB	Structure (table 1: counts, table 2: balances)	Table 1 106,701,077	Table 2 117,078,628	453.38

Name	Comprising (Column Headings)				Type	Format	Number of Observations				Size (mb)
CBK5_Loans	Table 1 CIF_ID Dates Consumer counts Instalments counts	Table2 CIF_ID Dates Consumer interest Instalment interest	Table3 CIF_ID Dates Consumer outstand. Instalment outstand.	Table4 CIF_ID Dates Consumer principle Instalment principle	MATLAB	Structure (table 1: counts, table 2: interest, table 3: outstanding, table 4: principle)	Table1 3,024,481	Table2 3,024,481	Table3 3,024,481	Table4 3,024,481	44.82
CBK6_Credit cards	CIF_ID Dates Credit card limit Credit card outstanding balances Risk factor number				MATLAB	Table	2,393,127				7.42
CBK7_Electronic channels	Table 1 Table 2 Table 3 Table 4 Table 5 For each table: CIF_ID Year and month Monthly Withdrawals (counts & values) Monthly deposits (counts and values) Monthly transfers (counts and values) Monthly purchases (count and values)				MATLAB	Structure (table1: ATM, table 2: call centre, table 3: internet, table 4: mobile, table 5: P.O.S)	Table1: 931,605 Table2: 22,822 Table3: 38,722 Table4: 381,589 Table5: 1,047,685				9.13

Name	Comprising (Column Headings)	Type	Format	Number of Observations	Size (mb)
	Monthly payments (counts and values)				
CBK8_Customer_Data	CIF_ID Age Gender Nationality Date of relationship	MATLAB	Table	1,382,942	584.08
Total MATLAB Data					1,192.78

3.5 Code documentation

3.5.1 Data organization

After going through all the process of data collection and obtaining the data, I started a new phase of data cleaning. I created similar structure for all banks; each bank has 8 tables. Listed below is a detailed explanation of the functions used, during cleaning process, in each of the 8 tables:

1. *Database 1*: The main function for this table is ‘varcollection_CBK1.m’. The original data are in the daily frequency and have 6 columns with headers of: 'CIF_ID', 'Dates', 'Value_Of_Txn', 'CURRENCY', 'TYPE_OF_CARD', and 'TXN_TYPE'. The original data in csv files are converted to MATLAB table and saved as ‘CBK1_raw’.

Table 1 is the output of the main function and has 4 columns with headers of: 'CIF_ID', 'Dates', 'Num_of_Txn', and 'Value_Of_Txn'. Table 1 is saved as ‘CBK1’. The data in Table 1 are in the monthly frequency. 'Num_of_Txn' is the monthly average number of transactions, and 'Value_Of_Txn' is the monthly average value of transactions. MATLAB built-in function ‘readtable.m’ is used to read the data from original csv files. ‘sortrows.m’ is used to sort rows in the table by dates and CIF_ID. ‘varfun.m’ is used to aggregate the daily data into monthly frequency. A user specified function ‘changedate2mthend.m’ is used to change the daily dates into month end dates (Mathworks, 2021l) (Mathworks, 2021m).

2. *Database 2*: The main function for this table is ‘varcollection_CBK2.m’. The original data have 5 columns with headers of: 'CIF_ID', 'age', 'gender', 'address', and 'DateOfRelationship'. Table 2 is saved as ‘CBK2’. If the same CIF_ID has multiple rows, only the last row is kept removing duplicates. Variable ‘gender’ has three lables: ‘F’, ‘M’, and ‘NA’. (Female, Male, and Not Available).

3. *Database 3*: The main function for this table is ‘varcollection_CBK3.m’. The original data have 14 columns with headers of: "Year", "CIF_ID", "Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov", and

"Dec". The data contain monthly income of all CIF_IDs from 2008 to 2018. The original data are pivoted into a long table with 3 columns with headers of: ‘CIF_ID’, ‘Dates’, and ‘Income’. The Dates are month end dates. The pivoting function is a user specified function: ‘convertincometable.m’. Table 3 is the output of ‘varcollection_CBK3.m’ and saved as ‘CBK3’. In ‘convertincometable.m’, if the same CIF_ID has multiple incomes within the same month, these incomes are added together into one entry. MATLAB function ‘splitapply.m’ is used to do the aggregation. More specifically, a new categorial variable combining CIF_ID and Year is added to the original table; ‘@rowsum’ is applied (through ‘splitapply’) to rows with the same category of this variable, i.e., CIF_ID+Year in the columns "Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov", and "Dec" (Mathworks, 2021n).

4. *Database 4*: The main function for this table is ‘varcollection_CBK4.m’. The original data are monthly account balance and number of account, and each csv file has 15 columns with headers of: "Year", "ACCT_TYPE", "CIF_ID", "Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov", and "Dec". ‘ACCT_TYPE’ has four categories: "'Current Accounts'", "'Deposit Accounts'", "'Joint Accounts '", and "'Saving Accounts '". These categories are abbreviated as ‘CA’, ‘DA’, ‘JA’, and ‘SA’ in table 4. Table 4 is the output of ‘varcollection_CBK4.m’ and saved as CBK4. CBK4 is a structure variable with two tables inside: count and balance, where count is the table for number of accounts and balance is the table for monthly account balance. count has 6 columns with headers of: ‘CIF_ID’, ‘Dates’, ‘CA_Count’, ‘DA_Count’, ‘JA_Count’, and ‘SA_Count’; balance has 6 columns with headers of: ‘CIF_ID’, ‘Dates’, ‘CA_Bal’, ‘DA_Bal’, ‘JA_Bal’, and ‘SA_Bal’. Like in table 3, the two tables: count and balance are from the pivoted original data. The pivoting function is also ‘convertincometable.m’. For balance, ‘splitapply.m’ and ‘@rowsum’ are applied the exactly same way in table 3; for count, ‘splitapply.m’ and ‘@rowmax’ are applied: instead of taking the sum, it takes the maximum in rows with the same category CIF_ID+Year in the columns "Jan", "Feb",

"Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov", and "Dec". MATLAB function ‘outerjoin.m’ is used to join tables in loops (Mathworks, 2021k).

5. *Database 5*: The main function for this table is ‘varcollection_CBK5.m’. The original data are monthly loan data and have 10 columns with headers of: 'YearMonth', 'LoanType', 'CIF_ID', 'count', 'LoanTotal', 'principle', 'interest', 'outstanding', 'RiskFactor', and 'RiskFactorNumber'. The raw data are read in MATLAB. Empty rows are removed using MATLAB function ‘erase.m’ and ‘cellfun.m’ with ‘@isempty’. Column ‘Dates’ created from 'YearMonth' using MATLAB functions ‘datetime.m’ and ‘lbusdate.m’ is added to the original table; then the rows in the table are sorted first by 'CIF_ID' and second by ‘Dates’ in the ascending order; the raw data table is saved as ‘CBK5_raw’. ‘LoanType’ has two categories: ‘Consumer’ and ‘Installment’.

Table 5 is the output of ‘varcollection_CBK5.m’ and saved as CBK5. CBK5 is a structure variable with four tables inside: 'count', 'interest', 'outstanding', and 'principle', where count is the table for the number of loans with 4 columns 'CIF_ID', 'Dates', ‘Consumer_Count’, and ‘Installment_Count’; interest is the table for the interests paid with 4 columns 'CIF_ID', 'Dates', ‘Consumer_Interest’, and ‘Installment_Interest’; outstanding is the table for the loan outstanding with 4 columns 'CIF_ID', 'Dates', ‘Consumer_Outstanding’, and ‘Installment_Outstanding’, principle is the table for the principle paid with 4 columns 'CIF_ID', 'Dates', ‘Consumer_Principle’, and ‘Installment_Principle’ (Mathworks, 2021d) (Mathworks, 2021h) (Mathworks, 2021e) (Mathworks, 2021c).

6. *Database 6*: The main function for this table is ‘varcollection_CBK6.m’. The original data are monthly credit card data and have 6 columns with headers of: 'YearMonth', 'CIF_ID', 'limit', 'Outstanding_Bal', 'RiskFactor', and 'RiskFactorNumber'. The original data are read into MATLAB using ‘readtable.m’. Table 6 is the output of ‘varcollection_CBK6.m’ and saved as ‘CBK6’ which has 5 columns with headers of: 'CIF_ID', 'Dates', 'limit', 'Outstanding_Bal', and 'RiskFactorNumber'. Column ‘Dates’ is created

from ‘YearMonth’ using MATLAB functions ‘datetime.m’ and ‘lbusdate.m’. The rows in the tables are sorted first by 'CIF_ID' and second by ‘Dates’ in the ascending order.

7. *Database 7*: The main function for this table is ‘varcollection_CBK7.m’. The original data are monthly banking behaviour data, for 'ATM', 'Call_Center', and 'POS' the csv files have 17 columns with headers of: "Year", "Channel_Type", "Channel_Content", "Trans_Type", "CIF_ID", "Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov", and "Dec", for 'Internet' and 'Mobile', the csv files have 17 columns with headers of: "Year", "Channel_Content", "Trans_Type", "Channel_Type", "CIF_ID", "Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov", and "Dec". ‘Trans_Type’ has four categories: 'Withdrawals', 'Deposits', 'Transfers', 'Payments'. ‘Channel_Content’ has two categories: 'Number Of Transaction', 'Value Of Transactions'. Table 7 is the output of ‘varcollection_CBK7.m’ and saved as CBK7. CBK7 is a structure variable with 5 tables inside: 'ATM', 'Call_Center', 'Internet', 'Mobile', and 'POS'. All tables have 10 columns with headers of: 'CIF_ID', 'Dates', ‘'Withdraw_no_count', 'Deposit_no_count', 'Transfer_no_count', 'Payments_no_count', 'Withdraw_Val_Trn', 'Deposit_Val_Trn', 'Transfer_Val_Trn', and 'Payments_Val_Trn', where ‘no_count’ is 'Number Of Transaction' and ‘Val_Trn’ is ‘Value Of Transactions’. Like in table 3, these five tables are from the pivoted original data. The pivoting function is a user specified function ‘convertmonthtable_simpler.m’. There are not duplicates in the original data. Therefore, ‘splitapply.m’ is not necessary. The rows in the tables are sorted first by 'CIF_ID' and second by ‘Dates’ in the ascending order. MATLAB function ‘outerjoin.m’ is used to join tables in loops.

8. *Database 8*: The main function for this table is ‘varcollection_CBK8.m’. The original data are the client information data and have 6 columns with headers of: "CIF_ID", "Age", "Gender", "Address", "RelationShipdate_str", and "Nationality". The original data are read into MATLAB using

‘readtable.m’. Rows with empty ‘RelationShipdate_str’ are removed using MATLAB function ‘erase.m’ and ‘cellfun.m’ with ‘@isempty’. Table 8 is the output of ‘varcollection_CBK8.m’ and saved as CBK8. CBK8 has 5 columns with headers of: "CIF_ID", "Age", "Gender", "RelationShipdate", "Nationality". The rows are sorted by 'CIF_ID' in the ascending order. Rows with duplicates of 'CIF_ID' are removed.

3.5.2 Banking type prediction modelling

In this section I describe in detail all the functions used to construct regression table, build banking type prediction model estimation, develop bank type prediction model for practical use and create a formula to check one variable’s effect on banking type preference.

1. *Variable collection*: the main function in this step is ‘construct_regression_table_morevars.m’.
   In‘construct_regression_table_morevars.m’, ‘CBK2’, ‘CBK3’, ‘CBK4’, ‘CBK5’, ‘CBK6’, and ‘CBK7’ are loaded for variable collection. zero and missing data in ‘CBK6.limit’ are removed before being collected. ‘Relationshipduration’ is constructed as the number of calendar days between ‘CBK2.DateOfRelationship’ and 2019-12-31 divided by 365. Non ‘F’ or ‘M’ elements in ‘CBK2.gender’ are removed before being collected. The five banking types considered are 'ATM', 'Call_Center', 'Internet', 'Mobile', and 'POS'. Total number of accounts in each of the five banking types are calculated and collected from ‘CBK7’. These numbers of accounts are used as the dependent variables in the banking type prediction analysis. The following variables are collected from ‘CBK2’ to ‘CBK6’: ‘age’, ‘gendernum’ (numeric, one for ‘M’ and zero for ‘F’), ‘CA_Count’, ‘DA_Count’, ‘SA_Count’, ‘Consumer_Count’, ‘Income’, ‘Installment_Count’, ‘limit’, ‘Outstanding_Bal’, and ‘Relationshipduration’. These variables alongside ‘ATM’, ‘Call_Center’, ‘Internet’, ‘Mobile’, and ‘POS’ are placed in a table called ‘table4regression’ and saved to the data folder. Therefore ‘table4regression’ has 2 + 14 = 16 columns with the first two headers being ‘CIF_ID’ and ‘Dates’.

2. *Banking type prediction model estimation*: the main function in this step is ‘bankingtype_estimation_morevars.m’. This function first loads ‘table4regression’ saved from ‘construct_regression_table_morevars.m’. It first collects the following variables: ‘age’, ‘gendernum’, ‘CA_Count’, ‘DA_Count’, ‘SA_Count’, ‘Consumer_Count’, ‘Income’, ‘Installment_Count’, ‘limit’, ‘Outstanding_Bal’, and ‘Relationshipduration’ into the explanatory variable container, ‘X’, then collects the following variables: ‘ATM’, ‘Call_Center’, ‘Internet’, ‘Mobile’, and ‘POS’ into the dependent variable container, ‘Y’. ‘X’ and ‘Y’ are fed into a MATLAB built-in function ‘mnrfit.m’ to estimate the full sample model (outputs ‘B’ and ‘stats’). The model used for banking type prediction is Multinomial logistic regression model. In the cross-validation, 1/3 of the sample is held out for validation purpose and only 2/3 of the sample is used to estimate the model. The full sample model parameters are saved as ‘B’ and the partial sample (2/3) model parameters are saved as ‘B_trained’. Alongside ‘B’ and ‘B_trained’, the subsample data and anther model estimate output, ‘stats’, are also saved (‘stats’ is needed for model prediction).

mnrfit (Multinomial logistic regression) (Mathworks, 2021i):

Bank type prediction for practical use: the main function in this step is ‘bankingtype_prediction_morevars.m’. This function takes the following inputs: ‘age’, ‘gendernum’, ‘CA_Count’, ‘DA_Count’, ‘SA_Count’, ‘Consumer_Count’, ‘Income’, ‘Installment_Count’, ‘limit’, ‘Outstanding_Bal’, and ‘Relationshipduration’. These inputs are collected in the explanatory variable container ‘X’. This function loads the estimated Multinomial logistic regression model ('B' and 'stats') from the data folder. ‘B’, ‘X’, and ‘stats’ are fed into a MATLAB built-in function ‘mnrval.m’ to produce the probabilities of choosing the five different types of banking, namely, ‘ATM’, ‘Call_Center’, ‘Internet’, ‘Mobile’, and ‘POS’. The conditional probabilities of choosing ‘Call_Center’, ‘Internet’, and ‘Mobile’ are calculated as well. The conditional probability measures the likelihood of choosing one of the three types assuming that the client does not choose the other two, namely, ‘ATM’ and ‘POS’. The values of the predictors and the

calculated probabilities are the outputs of the function and printed in the MATLAB command window.

3. *Check one variable's effect on banking type preference*: the main function in this step is 'bankingtype_prediction_visualisation_givenvar.m'. This function takes one variable name only. This name must be chosen from the following list: 'CA_Count', 'DA_Count', 'SA_Count', 'Consumer_Count', 'Income', 'Installment_Count', 'limit', 'Outstanding_Bal', 'age', 'Relationshipduration'. Other than this input variable, all variables are fixed at their sample average calculated using the data in 'table4regression' loaded from the data folder. If the input variable is one of 'CA_Count', 'DA_Count', 'JA_Count', 'SA_Count', and 'Consumer_Count', the range for the input variable is from zero to its sample maximum in 'table4regression'; otherwise, the range for the input variable is from its sample 10 percentile to sample 90 percentile in 'table4regression'. 100 points are chosen on this range with equal spacing. Each point of these 100 points alongside all other variables fixed at their sample averages are fed into 'bankingtype_prediction_morevars.m' to produce banking type probabilities (both conditional and unconditional ones). Once all 100 points are looped through in this computation, the resulting probabilities are plotted against the values of these 100 points to show the effect of this input variable on preference of the different banking types. These plots are produced separately for male and female.

3.5.3 Banking type prediction model comparison

In this part, I list all the functions used for the four alternative models in this research, Robustness to the size of the hold out sample, and Results comparison.

1. *alternative models*: there are four alterative models considered for the banking type prediction. 'mvregress' using the 'cwls' algorithm (in function 'bankingtype_estimation_morevars_alternative.m'), 'mvregress' using the 'ecm' algorithm (in function

'bankingtype_estimation_morevars_alternative2.m'), 'fitrtree' (in function 'bankingtype_estimation_morevars_alternative3.m'), 'fitrsvm' (in function 'bankingtype_estimation_morevars_alternative4.m'). In each of these functions, the data table 'table4regression' saved from 'construct_regression_table_morevars.m' is loaded, the following variables: 'age', 'gendernum', 'CA_Count', 'DA_Count', 'SA_Count', 'Consumer_Count', 'Income', 'Installment_Count', 'limit', 'Outstanding_Bal', and 'Relationshipduration' are collected into the explanatory variable container, 'X', and the following variables: 'ATM', 'Call_Center', 'Internet', 'Mobile', and 'POS' are collected into the dependent variable container, 'Y'. 'X' and 'Y' are fed into the corresponding model in the function. One third of the data are held out for out-of-sample test, in which the average out-of-sample prediction errors is recorded as the performance measure and saved in 'deviate_outofsample'. The formula for this calculation is: aveprob_outofsample = average pihat_outofsample, deviate_outofsample = (aveprob_outofsample/aveprob_benchmark_oos - 1), where 'aveprob_benchmark_oos' is the average observed probabilities of each banking type, and 'aveprob_outofsample' is the average model-produced probabilities of each banking type. The smaller this error the best out-of-sample performance (Mathworks, 2021j) (Mathworks, 2021g) (Mathworks, 2021f) (Mathworks, 2021a).

2. *Robustness to the size of the hold out sample*: the main function in this step is 'samplesize_and_OOS_accuracy.m'. This function calls the following subfunction within a for loop: 'bankingtype_estimation_OOS_accuracy.m' (base line model), 'bankingtype_estimation_OOS_accuracy_alternative.m', 'bankingtype_estimation_OOS_accuracy_alternative3.m', 'bankingtype_estimation_OOS_accuracy_alternative4.m' to compute the OOS accuracy with changing sample size. Each of this subfunction takes a given sample and calculates the OOS accuracy defined in step 1 of this

section. The output of this function is a plot showing the relation between the side of the reserved data for training and the OOS accuracy.

3. *Results comparison*: from the output of 'samplesize_and_OOS_accuracy.m', it is found that the baseline and alternative 1 deliver similar OOS performance better than alternative 2 with changing sample size. Other alternatives do not provide clear advantage over 'mnrfit'. But 'mnrfit' is easy to run. 'mvregress' has similar performance as mnrfit, so using 'mnrfit' or 'mvregress' would make little difference. This is true for other models as well. This is also an indication of robustness of 'mnrfit'.

3.6 Methods

Changes in technology have led to a rapid growth in data, particularly in the banking sector. Bankers need to find the best techniques to deal with big data and use it to their advantage. Therefore, they need to collaborate with data analysts. Bankers and financial analysts are experts in determining the type of data that is required, as well as the best parameters and methods for analysing the results. However, they need data scientists to help them clean and process the data. Given the amount of data I obtained (approximately 1 TB), I decided to use machine learning to clean and process my data. This helped us to develop the predictive analysis for my banks. I researched which software to use and decided on the Statistics and Machine Learning Toolbox in MATLAB (The MathWorks Inc., 2020). This software offers very useful functions for describing, examining, and modelling data.

I began by studying different machine learning techniques. Machine learning involves computer algorithms working in the same way as human beings to solve problems. In machine learning, computers watch and learn without needing to be specifically programmed; they will get better at responding to the data independently. Through automation, which emerged at the same time as big data, algorithms can be used to process data and save time. There are two main types

of machine learning: supervised and unsupervised (Mathworks, 2016c). Supervised learning involves observing an actual set of data and trying to reach a specific output from a given number of inputs. Regression and classification problems are the most common learning algorithms for predicting continuous and discrete outputs, respectively (Mathworks, 2016a). By contrast, unsupervised machine learning is used to discover what information a data set might contain (Mathworks, 2016b). Unsupervised machine learning makes it possible for the algorithm to determine any suitable structure in the given data set. The unsupervised algorithm tries to gather the data based on similar variables (clustering) or discover structure in messy data (non-clustering).

Arthur Samuel used machine learning to train computers to learn to play checkers like a human being. He began programming the 1950s but quickly found that he could not design algorithms that could beat human players. However, by 1969, he had developed a machine learning algorithm that had learned to play checkers effectively (Samuel, 1969).

In my research, I used supervised learning to observe a data set consisting of several inputs and outputs. I attempted to solve a classification problem by using multinomial logistic regression (MLR). I began by clarifying the linear regression and then moved on to study more advanced types of general regression.

The linear regression model used a single variable (supervised learning) to produce real value outputs from the inputs. The learning algorithm used the data set (known as the training set) to output the hypothesis function. The function takes the variables (features) as the input and tries to predict the output value (Coursera, 2019). Table 3-14 shows the meaning of some notation used in this model.

Table 3-14. Linear regression notation table (univariate)

Notation	Meaning
(i)	Index
x_i	Input variables

y_i	Output variables (actual)
(x_i, y_i)	Training example
m	Number of training examples
$h\ or\ h(x)$	Hypothesis, function of x to produce y
θ	Theta, parameter (θ_0, θ_1, etc.)

The hypothesis uses a linear function to predict the outcomes. The function consists of parameters (θs) and variables (Xs) as shows in Equation $h\theta(x) = \theta_0 + \theta_1 x$ (3-1).

$$h_\theta(x) = \theta_0 + \theta_1 x \qquad (3\text{-}1)$$

The main job of the hypothesis function is to determine the values of the parameters ($\theta_0\ and\ \theta_1$). In this example, I only used two parameters to predict the output ($h_\theta(x)$); this function tries to reach the minimum difference between the predicted output ($h_\theta(x)$) and the actual output (y). To test the accuracy of the hypothesis function, I used a cost function (see Equation $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}_i - y_i)^2 = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i)^2$ (3-2). The later function calculates the average (mean) difference between all the predicted outputs (function of Xs) and the actual outputs. As I can see from the cost formula, the mean is multiplied by ½. This is to make the computation more convenient.

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}_i - y_i)^2 = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i)^2 \qquad (3\text{-}2)$$

The main objective of the learning algorithm is to minimize the cost function $J(\theta_0, \theta_1)$. To reach an optimal level, the learning algorithm chooses the best values

among the parameters, finding a line that fits the training data properly (see Equation 3-3)

3.6.1 Linear regression formula

Hypothesis: $h_\theta(x) = \theta_0 + \theta_1 x$

Parameters: θ_0, θ_1

Cost Function: $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$ (3-3)

Goal: $\min_{\theta_0, \theta_1} J(\theta_0, \theta_1)$

Having defined the formulae and goal of the linear regression, I will now discuss how the algorithm works to minimize the cost function through a gradient descent algorithm. This algorithm is used in machine learning to minimize the different type of functions. To illustrate how the gradient descent algorithm works, I will use two parameters ($\theta_0 \text{ and } \theta_1$). As mentioned, the goal is to minimize the cost function of $J(\theta_0, \theta_1)$ by continuously trying new values for $\theta_0 \text{ and } \theta_1$. These are reduced until the minimum values are reached (the minimum difference between the predicted values compared to the real values). The gradient descent starts by assigning initial values for the parameters (in my example, $\theta_0 \text{ and } \theta_1$). Then, the values for each parameter are changed using the following formula (Equation $\theta j := \theta_j - \frac{\partial}{\alpha \theta_j} J(\theta_0, \theta_1) \quad (for\ j = 0\ and\ j = 1)$ (3-4):

$$\theta_j := \theta_j - \frac{\partial}{\alpha \theta_j} J(\theta_0, \theta_1) \quad (for\ j = 0\ and\ j = 1) \qquad (3\text{-}4)$$

Here, the gradient descent algorithm tries to reach the minimum point of the cost function by recognizing the best direction that the parameters should follow to reach that point. The parameters change simultaneously (see Equation Cost Function: $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$ (3-3) until the minimum point is reached. It is very important for the gradient descent to be

correctly implemented. In my example, both θ_0 and $\theta_1$must be updated at the same time; ignoring this might give different results.

In this study, I used MATLAB software to apply this learning algorithm. The notation ":=" in Equation $\theta j := \theta_j - \frac{\partial}{\alpha\theta_j} J(\theta_0, \theta_1) \quad (for\ j = 0\ and\ j = 1)$ (3-4 refer to the unequal assignment operator (Attaway, 2017). α (learning rate) in this equation is responsible for the downward steps. The derivatives ($\frac{\partial}{\alpha\theta_j} J(\theta_0, \theta_1)$) in Equation $\theta j := \theta_j - \frac{\partial}{\alpha\theta_j} J(\theta_0, \theta_1) \quad (for\ j = 0\ and\ j = 1)$ (3-4 represent the slop of the tangent line to the cost function (after selecting the initial point of the thetas). Given that

temp0 : = $\theta_0 - \alpha \frac{\partial}{\alpha\theta_j} J(\theta_0, \theta_1)$

temp1 : = $\theta_1 - \alpha \frac{\partial}{\alpha\theta_j} J(\theta_0, \theta_1)$ (3-5)

θ_0 : = temp0

θ_1 : = temp1

Alpha

Having developed the linear regression and the gradient descent, it is now possible to combine them together by applying the gradient descent as part of the linear regression problem. The cost function uses the gradient descent algorithm in its equation to alter the thetas, until it reaches the minimum point required, where the slope of the derivatives is zero. Equation $\theta 0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})$ (3-6) shows the combination of the gradient descent algorithm and the linear regression model (using only two variables).

Repeat until convergence {

$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})$ (3-6)

$$\theta_1 \coloneqq \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \, . x^{(i)}$$

}

I can see from Equation $\theta 0 \coloneqq \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})$ (3-6 (in the derivative part) that the algorithm is taking the whole sum of the training set. So, each time the gradient descent calculates the values for the new parameters, it considers all the training examples. For this reason, machine learning users refer to it as “batch gradient descent”.

Now I will take a further step and discuss the linear regression with multiple variables. In the previous one I used one variable (x), but now there are multi variables $(x_1, x_2, x_3, \dots, x_n)$ that are used in the hypothesis function to come up with expected output. Table 3-15 represents some new notations for the multivariate linear regression.

Table 3-15: Linear regression notation table (multivariate)

Notation	**Meaning**
(i)	index
x_j^i	Value of variable j in the ith training set
x_i	Input variables, the feature of the ith training set
y_i	Output variables (actual)
(x_i, y_i)	Training example
m	Number of training examples
$h \; or \; h(x)$	Hypothesis, function of x to produce y
θ	Theta, parameter (θ_0, θ_1, etc.)
n	Number

$= \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$ (3-8. The number of parameters increases depending on the number of features.

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \cdots + \theta_n x_n \quad (3\text{-}7)$$

Cost Function

$$J(\theta_0, \theta_1, \ldots, \theta_n) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 \quad (3\text{-}8)$$

To simplify the equation (especially in cases where there are many features), I can consider $J(\theta_0, \theta_1, \ldots, \theta_n) = J(\theta)$, where θ is an n+1–dimensional vector.

To use the MATLAB in matrix operation, I will assume $x_\theta(1) = 1$ for i (1…m). One of the requirements for matrix multiplication is that the number of columns of one matrix should equal the number of rows of the other matrix. Therefore, the multivariate hypothesis function can be written as follows (Equation $h\theta(x) = [\theta_0 \quad \theta_1 \quad \ldots \quad \theta_n] \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix} = \theta^T x$ (3-9):

$$h_\theta(x) = [\theta_0 \quad \theta_1 \quad \ldots \quad \theta_n] \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix} = \theta^T x \quad (3\text{-}9)$$

There is one difference in applying gradient descent for multivariate linear regression compared to univariate regression: in univariate regression, I use x_i when n = 1, but in multivariate regression, I use $x_j^{(i)}$ when n ≥ 1 (Equation (3-10).

Repeat {

$(simultaneously\ update\ \theta_j\ for\ j = 0, \ldots, n)$

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})\ x_j^{(i)} \tag{3-10}$$

}

In what follows, I describe some actions that make the gradient descent algorithm work better:

3.6.2 Feature scaling

When it comes to multivariate regression with lots of variables, I need to ensure that all variables have the same range of values. The cost function $(J(\theta))$ is a function of all thetas (parameters). Therefore, having different scales for the values of features will create a very difficult path for the gradient descent, making it harder to reach the required global minimum. In general, the range of values should be approximately -1 ≤ x_i≤ 1, but it is possible to be flexible with small changes. Below are two methods of feature scaling:

1. Dividing the feature by the maximum value of the training examples related to that feature or the range (maximum minus minimum)
2. Mean normalization: subtracting the average value of the feature (input variable) from the value of that feature. The reason for doing this is to make the new average for that variable equal to zero.

I can combine both techniques in the following equation:

$$x_i := \frac{x_i - \mu_i}{s_i} \quad (3\text{-}11)$$

Where μ_i is the average value of all the examples of variable (i), and s_i is the difference between the maximum and the minimum values (the range) or the standard deviation.

3.6.3 Learning rate (alpha)

As mentioned above, the gradient descent algorithm changes the parameters in each iteration by applying the following formula (Equation $\theta j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$ (3-12):

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \quad (3\text{-}12)$$

The goal is to reach the global minimum of the cost function $J(\theta)$. Checking whether alpha (the learning rate) has the correct value and trying to “debug” the incorrect pattern of alpha will help to improve the gradient descent results. The two methods for checking whether alpha needs to be altered are as follows:

1. Plot the cost function when the gradient descent is going through the iterations:

I plot the cost function as a function of the number of iterations, using a graph in which the x-axis represents the number of iterations and the y-axis represents the cost function. I expect the cost function to decrease for every iteration until it converges. If the results show numbers that differ from those I expect, I interfere and change the value of the learning rate (alpha).

2. Automatic convergence test:

This works by requesting the algorithm to announce the convergence if $J(\theta)$ decreases by less than the specified threshold (for example, 10^{-3}) in a single iteration.

To clarify, I can summarize the process in the following way: linear regression is a linear method for modelling the relationship between a dependent variable and one or more independent variables. Simple linear regression involves one independent variable (x_1). Multiple linear regression involves more than one independent variable ($x_1, \dots, x_n$) and one dependent variable (y). Multivariate linear regression involves multiple dependent variables ($y_1, \dots, y_n$). Thus, there are many extensions of basic linear regression which accept all assumptions and allow the basic model to be altered (Freedmen, 2009).

The simple model for multiple linear regression is as follows:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i \qquad (3\text{-}13)$$

for each observation $i = 1, \dots, n$

In the equations above, I consider n observations of one dependent variable and p independent variables. So, Y_1 is the i^{th} observation of the dependent variable, X_{ij} is the i^{th} observation of the j^{th} independent variable, and $j = 1,2, \dots, p$. The value β_j signifies the parameters to be estimated, and ϵ_j is the i^{th} independent variable identically issued normal error (Y. Yang, Sun, & Guo, 2018).

But with more general multivariate linear regression, there is one equation of the above form for each of $m > 1$ dependent variables that share the same set of explanatory variables. These variables are, therefore, estimated simultaneously with each other:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{i1} + \beta_{2j}X_{i2} + \cdots + \beta_{pj}X_{pi} + \epsilon_{ij} \quad (3\text{-}14)$$

These equations hold for all the observations of $i = 1, \ldots, n$ and for all dependent variables of $j = 1, \ldots, m.$

One of the methods I used for this study is multivariate linear regression, which entails multiple independent and dependent variables (Hidalgo & Goodman, 2013).

Multivariate linear regression expresses a *d*-dimensional response vector as a linear combination of predictor terms in addition to a vector of error terms with a multivariate normal distribution. Hence, let $Y_i = (Y_{i1}, \ldots, Y_{id})'$ represent the response vector for observation i, $i = 1, \ldots, n$ (MathWorks, 2020). In general, given the *d*-by-*K* design matrix X_i and the *K*-by-1 vector of coefficients β, the multivariate linear regression model is:

$$Y_i = X_i\beta + \varepsilon_i \quad (3\text{-}15)$$

Where the *d*-dimensional vector of error terms uses a multivariate normal distribution, the model is:

$$\varepsilon_i \sim MVN_d(0, \Sigma) \quad (3\text{-}16)$$

The model takes on independence between observations, meaning that the error variance-covariance matrix for the *n d*-dimensional response vectors is:

$$I_n \otimes \Sigma = \begin{pmatrix} \Sigma & & 0 \\ & \ddots & \\ 0 & & \Sigma \end{pmatrix} \qquad (3\text{-}17)$$

If y represents the *nd*-by-1 vector of the stacked *d*-dimensional responses, and X symbolizes the *nd*-by-*K* matrix of the stacked design matrices, then the distribution of the response vector is:

$$y \sim MVN_{nd}(X\beta, I_n \otimes \Sigma) \qquad (3\text{-}18)$$

To apply multivariate linear regression models from MATLAB Statistics and Machine Learning Toolbox, I used the function “mvregress”:

$$y_i = X_i\beta + \varepsilon_i,\ \varepsilon_i \sim MVN_d(0, \Sigma) \qquad (3\text{-}19)$$

Sometimes, the linear function does not fit the data well, so I can change the linear hypothesis function into a quadratic, cubic, square root, or any other function. I can also improve the features by, for example, combining two features to get a new feature. This is called polynomial regression. It enhances my functions and variables in different ways. It is worth mentioning that if I decide to shift from nominal to another type of equation, I need to monitor the values of the features and perform feature scaling to unite the range of values.

Depending on the problem I want to address and the number of features I have, I can decide which method to use to give faster and more accurate results. For relatively few features (e.g., from 100 to 1000), I can use a “normal equation” rather than a gradient descent for optimal results. In a normal equation, I can get the values for theta that minimize $J(\theta)$ in one step, rather than trying many iterations to reach convergence. This can be done by setting the derivative of $J(\theta)$

to zero and then solving for the values of theta that minimize the cost function $J(\theta)$. The normal equation formula is as follows (Equation 3-20):

$$\theta = (X^T X)^{-1} X^T y \qquad (3\text{-}20)$$

In MATLAB, I can enter “pinv(x`*x)*x`*y” to get the above equation. X` is the notation to compute x transpose, and pinv or inv commands compute the inverse of (x`*x). Table 3-16 summarizes some of the advantages and disadvantages of using gradient descent as opposed to a normal equation.

Table 3-16: Comparison of gradient descent and normal equation

Gradient Descent	Normal Equation
Need to choose alpha	No need to choose alpha
Needs many iterations	No need to iterate
$0(kn^2)$	$0(n^3)$, need to calculate inverse of $X^T X$
Works well when (n) is large	Slow if (n) is very large

Having described the regression, I will now consider the classification, where the values of the outputs are discrete values. A linear regression algorithm is not recommended for the classification problem because the outputs can be below or above the required values of the output (for example, 0 or 1). Therefore, the outputs can be higher than one or lower than zero. A logistic regression algorithm will always give values between zero and one (in my previous example), so this type of algorithm will work for such classification problems. Below is the new form of the hypothesis equation used in logistic regression algorithm (Equation 3-21 and Equation 3-22):

$$h_\theta(x) = g(\theta^T x)$$
$$z = \theta^T x$$

$$g(z) = \frac{1}{1+e^{-z}} \quad (3\text{-}21)$$

$$h_\theta(x) = P(y = 1|x;\theta) = 1 - P(y = 0|x;\theta)$$
$$P(y = 0|x;\theta) + P(y = 1|x;\theta) = 1 \quad (3\text{-}22)$$

This equation will give us the probability that the output is equal to one. So, the probability that the predicted output equals zero will always correspond to the probability of being equal to one. To understand the logistic regression hypothesis function better, it is also necessary to understand what is known as the “decision boundary”. A decision boundary is a line (straight or otherwise) drawn (extracted from) by the logistic regression hypothesis function to separate the two regions where y = 0 and y = 1 (in my example).

By solving the following equation (3-23), I can identify the areas where the output is equal to one or zero.

$$\theta^T x \geq 0 \Rightarrow y = 1$$
$$\theta^T x \geq 0 \Rightarrow y = 0 \quad (3\text{-}23)$$

In the logistic regression model, I cannot use the same cost function as the one used for linear regression because this will result in a non-convex function (curvy outputs with many local minimum points). So, the cost function for the logistic regression model (in my example) will be as follows:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} Cost\left(h_\theta\left(x^{(i)}\right), y^{(i)}\right)$$
$$Cost(h_\theta(x), y) = -\log\left(h_\theta(x)\right) \qquad if\, y = 1$$

$$Cost(h_\theta(x), y) = -\log\big(1 - h_\theta(x)\big) \quad if\, y = 0 \qquad (3\text{-}24)$$

Where

$$Cost(h_\theta(x), y) = 0\ if\ h_\theta(x) = y$$
$$Cost(h_\theta(x), y) \to \infty\ if\ y = 0\ and\ h_\theta(x) \to 1$$
$$Cost(h_\theta(x), y) \to \infty\ if\ y = 1\ and\ h_\theta(x) \to 0 \qquad (3\text{-}25)$$

Usually, when using the logistic regression model, the cost function is written in a different way (see Equation 3-26) to make finding the appropriate parameters (thetas) more efficient. Also, the convex shape helps to arrive at the optimal output.

$$J(\theta) = -\tfrac{1}{m}\textstyle\sum_{i=1}^{m}[y^{(i)} \log\left(h_\theta\left(x^{(i)}\right)\right) + \left(1 - y^{(i)}\right) \log\left(1 - h_\theta\left(x^{(i)}\right)\right)] \qquad (3\text{-}26)$$

Again, the objective here is to minimize the cost function. As discussed earlier in this chapter, I can use the gradient descent algorithm to predict thetas that minimize the cost function (see Equation3-27).

Gradient Descent:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)} \log h_\theta\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - h_\theta\left(x^{(i)}\right)\right)]$$

Want $min_\theta\, j(\theta)$:

Repeat {

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

$(simultaneously\ update\ all\ \theta_j)$ (3-27)

}

If I rewrite the derivative part, I can arrive at the gradient descent function.

A vectorized implementation is as follows:

Repeat {

$$\theta_j \coloneqq \theta_j - \alpha \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x_j^{(i)}$$

$(simultaneously\ update\ all\ \theta_j)$

} (3-28)

$$\theta \coloneqq \theta - \frac{\alpha}{m} X^T (g(X\theta) - \vec{y}) \quad (3\text{-}29)$$

This study has provided a detailed illustration of the gradient descent algorithm, but there are other algorithms that can be used to calculate the optimum cost functions. Examples of these more advanced algorithms are conjugate gradient BFGS and L-BFGS (Coursera, 2019). In MATLAB, there are functions that make it possible to run such algorithms without writing new code. The main advantage of using these sophisticated algorithms is that they are faster than gradient descent. Also, there is no need to calculate the learning rate (alpha) manually.

Now, let us expand my example to include more than two discrete values (y = 0 or 1). Previously, I talked about the binary classification problem, but what

happens with a multiclass classification problem, in which y can take any value from 1 to n? In these cases, I can use the same techniques as I used for binary classification by choosing a class (e.g., y = 1) and grouping all the other classes (y = 2 ... n) in a single class. In this way, I create two classes to be treated as a binary. I can then repeat the process with all the other classes by using the following formulae (Equation 3-30):

$$
\begin{aligned}
&y \in \{0, 1 \ldots n\} \\
&h_\theta^{(0)}(x) = P(y = 0|x; \theta) \\
&h_\theta^{(1)}(x) = P(y = 1|x; \theta) \\
&\ldots \\
&h_\theta^{(n)}(x) = P(y = n|x; \theta) \\
&prediction = \max_i(h_\theta^{(i)}(x))
\end{aligned}
\qquad \text{(3-30)}
$$

3.6.4 Multinomial logistic regression

As mentioned, based on my understanding of the two types of machine learning (supervised and unsupervised), I decided to use multinomial logistic regression (MLR) analysis.

MLR is a classification method that uses the same techniques as binary logistic regression but for a multiclass problem. It generalizes a binary regression into a multiclass regression. In this paper, I used MLR to predict the probabilities of different possible outcomes of a response variable, given several independent variables. When there are more than two possible outcomes, or nominal responses, MLR can be used to calculate the probabilities of each of these outcomes. It predicts the probabilities of multiple possible outcomes of a categorical variable by using a set of independent variables (McCullagh & Nelder FRS, 1990; Carpenter, 2008). MLR is used in studies that have more than one categorical dependent variable and more than one outcome (Greene, 2012). In this research, the dependent variables were the electronic channels.

MLR is used to model unordered and independent categorical response variables. I found this model useful for my analysis because it does not assume linearity, normality, or homoscedasticity. MLR allows each category to be compared to a reference category, thereby providing several MLR models. This makes it possible to make certain comparisons of the response categories. When there are *j* categories for the response variable, the model requires *j-1* logic equations to fit simultaneously (Okasha & Shehada, 2016).

MLR classifies *d*-dimensional real-valued input vectors $x \in \mathbb{R}^d$ into one of the k outcomes $c \in \{0, \dots, k-1\}$ using $k-1$ parameter vectors $\beta_0, \dots, \beta_{k-2} \in \mathbb{R}^d$. The response variable $Y \in \{1, \dots, k\}$ has k possible values/categories (Carpenter, 2008). The general form of the MLR is given by

$$p(Y = r|x) = \frac{exp\ (xT\beta r)}{\sum_{s=1}^{k} exp\ (xT\beta s)} = \frac{exp(\eta r)}{\sum_{s=1}^{k} exp\ (\eta s)} \qquad (3\text{-}31)$$

Where $\beta_r^T = (\beta_{r0}, \dots, \beta_{rp})$. (3-32)

It is noticeable that I have to specify some additional constraints since the parameters $\beta_1^T, \dots, \beta_k^T$ are not recognizable. Here, I have to select a reference category. When category k is selected, I set $\beta_k^T = (0, \dots, 0)$ producing $\eta_k = 0$. I can pick any other category as a reference, which is mobile in this study. If I fit a model using a reference category k, the corresponding equation is

$$p(Y = r|x) = \frac{exp\ (xT\beta r)}{1+\sum_{s=1}^{q} exp\ (xT\beta s)} \quad for\ r = 1, \dots, q \qquad (3\text{-}33)$$

Muller (2004) states that MLR is used if the response and dependent variable for each observation *i* is one out of more than two alternative categories. One of the

categories has to be chosen as a reference category, with no loss of generality. In this research, I used the mobile application (for example, π_1). With π_j, I expressed the probability $P(Y = j|X)$. Then, I could consider the MLR with respect to the reference category (see Equation 3-34).

$$logit(\pi_j) = \log\left(\frac{\pi_j}{\pi_1}\right) = X_j^T \beta_j \qquad (3\text{-}34)$$

The terms X_j and β_j indicate that the explanatory variables and their corresponding coefficients may depend on category J. The model can be defined by

$$P(Y = 1|X) = \frac{1}{1 + \sum_{k=2}^{J} \exp(X_k^T \beta_k)}$$

$$P(Y = j|X) = \frac{X_j^T \beta}{1 + \sum_{k=2}^{J} \exp\,(X_k^T \beta_k)} \qquad (3\text{-}35)$$

It is important to recognize that the logit model is a special case of the multinomial model for exactly two alternatives. If the alternatives (categories) are ordered in some way, then this information must also be considered (Long & Freese, 2003).

Emptage and Dobson (1992) stated that the nominal MLR model is used when there is no natural order among the response categories. One category is subjectively chosen as the reference category. In my study, this was the mobile application. If I suppose that this is the first category, then the logits for the other categories are described by

$$logit(\pi_j) = \log\left(\frac{\pi_j}{\pi_1}\right) = x_j^T \beta_j, \quad for\ j = 2, \dots, J \qquad (3\text{-}36)$$

The $J-1$ logit equations are used to estimate the parameters β_j. Once the estimates for b_j have been obtained, the linear predictors $x_j^T b_j$ can be computed using Equation (3-37):

$$\hat{\pi}j = \widehat{\pi_1}\exp\left(x_j^T b_j\right) \qquad for\ j = 2, \dots, J$$

$$But\ \widehat{\pi_1} + \widehat{\pi_2} + \cdots + \hat{\pi}j = 1, \qquad so$$

$$\widehat{\pi_1} = \frac{1}{1 + \sum_{j=2}^{J}\exp\left(x_j^T b_j\right)}$$

and

$$\hat{\pi}j = \frac{\exp\left(x_j^T b_j\right)}{1+\sum_{j=2}^{J}\exp\left(x_j^T b_j\right)} \qquad (3\text{-}37)$$

In this study, I focused on three electronic channels as dependent variables. In MLR, it is necessary to choose one dependent variable as a reference variable. In this study, “mobile application” was chosen as the reference variable. The probability of using any of the other online channels was compared to the probability of using a mobile application. Rather than running individual regression analysis for each parameter and then comparing the results, I evaluated the parameters simultaneously. This made the comparison simpler and was more efficient.

The regression evaluated parameters simultaneously. I focused on the coefficients and *p*-values to see the effect of each variable on the probability of using each electronic banking channel. The coefficients showed the statistical relationship between the independent and dependent variables. A *p*-value represents a pre-set level of significance. I considered any variable with a *p*-value of less than 0.05 to be significant, meaning that an independent variable had an important effect on the relevant dependent variable (Gibbons & Pratt, 1975). I also assessed the standard errors and confidence intervals for all the variables. This

is discussed in more detail in Chapter 4, which provides the results and discussion.

Tables Table 3-17 to Table 3-22 show the codes for cleaning, constructing tables, and running the MLR analysis. However, this does not reflect all the time taken to prepare the results. Using the toolboxes available in MATALAB and finding, changing, and constructing it required a lot of time that could not be measured. On the other hand, implementing the model in banks require preparation for manpower, computing, and how will this model fits into the bank's operations. Banks need to decide whether to invest on the existing staff by excessive training programs or to hire specialists in data scientists to deal with all the implementation requirements. Also, they must decide which software they will use depends on the effectiveness and efficiency of the software system. Implementing such model will facilitate product and service developments.

Table 3-17: Bank 1 code execution time

<table>
<tr><th></th><th>Transfer from CVS to MATLAB (cleaning)</th><th>Construct Regression Table</th><th>Fit MLR model</th><th>Predict probabilities for MLR</th></tr>
<tr><td>CBK 1</td><td rowspan="3">5:23 hrs</td><td rowspan="8">8 min</td><td rowspan="8">3 min</td><td rowspan="8">10 seconds approximately</td></tr>
<tr><td>CBK 2</td></tr>
<tr><td>CBK 3</td></tr>
<tr><td>CBK 4</td><td>47:43 hrs</td></tr>
<tr><td>CBK 5</td><td>15:20 hrs</td></tr>
<tr><td>CBK 6</td><td>15 min</td></tr>
<tr><td>CBK 7</td><td>2 min</td></tr>
<tr><td>CBK 8</td><td>4 min</td></tr>
</table>

Table 3-18: Bank 2 code execution time

	Transfer from CVS to MATLAB (cleaning)	Construct regression table	Fit MLR model (min)	Predict probabilities for MLR (sec)
CBK 1	28 min	1 min	3 min	

CBK 2	1 min			10 seconds approximately
CBK 3	1 min			
CBK 4	2 min			
CBK 5	1 min			
CBK 6	1 min			
CBK 7	1 min			
CBK 8	1 min			

Table 3-19: Bank 3 code execution time

	Transfer from CVS to MATLAB (cleaning)	**Construct regression table**	**Fit MLR model**	**Predict probabilities for MLR**
CBK 1	7 min	4 min	1 min	10 seconds approximately
CBK 2	1 min			
CBK 3	2 min			
CBK 4	3 min			
CBK 5	1 min			
CBK 6	1 min			
CBK 7	7 min			
CBK 8	2 min			

Table 3-20: Bank 4 code execution time

	Transfer from CVS to MATLAB (cleaning)	**Construct regression table**	**Fit MLR model**	**Predict probabilities for MLR**
CBK 1	3.40 hrs	4 min	5 min	10 seconds approximately
CBK 2	5 min			
CBK 3	48.48 hrs			
CBK 4	42 min			
CBK 5	10 min			
CBK 6	6 min			
CBK 7	1 min			
CBK 8	2 min			

Table 3-21: Bank 5 code execution time

	Transfer from CVS to MATLAB (cleaning)	Construct regression table	Fit MLR model	Predict probabilities for MLR
CBK 1	29 min	1 min	4 min	10 seconds approximately
CBK 2	1 min			
CBK 3	4 min			
CBK 4	10 min			
CBK 5	29 min			
CBK 6	2 min			
CBK 7	42 min			
CBK 8	1 min			

Table 3-22: Bank 6 code execution time

	Transfer from CVS to MATLAB (cleaning) minutes	Construct regression table	Fit MLR model (min)	Predict probabilities for MLR (sec)
CBK 1	6.43 hrs	3 min	10 min	10 seconds approximately
CBK 2	2 min			
CBK 3	17:58 hr			
CBK 4	97:34 hr			
CBK 5	1 min			
CBK 6	2 min			
CBK 7	1min			
CBK 8	2min			

3.6.5 Binary decision tree for regression

Another supervised learning method is the binary decision tree. This method was applied by Emrouznejad and Anouze (2010). They proposed a framework for bringing together data envelopment analysis with classification and regression, making it possible to evaluate the efficiency and productivity of decision making in banking. The result of this combined model is a set of guidelines that can be used to understand efficient and inefficient decision-making. It uses classification

and a regression tree to examine factors related to efficiency in the banking sector in countries in the Gulf Cooperation Council.

A decision tree is a tree of predictions. The overall idea behind this method is that the algorithm segments the predictor space into different regions. So, if I have several observations ($x_1 \ldots, x_n$), and I want to make a prediction based on them, I use the mean of the training set of data in the area or region to which it belongs. This method is called a “decision tree method” because it involves a series of bifurcating rules to segment the predictor space. These can be represented as a tree.

Subsequently, to provide more information than the basic decision tree method, I can predict response variables from inputs $x_1, x_2 \ldots, x_n$. I do this by drawing and expanding the binary tree. At each internal node in the tree, I apply a test to one of the inputs (say x_i). Depending on the result of the test, I follow either the left or the right branch of the tree. Finally, I reach a leaf node, where I make a prediction. This prediction aggregates or averages all the training data points that have been passed through to reach that leaf.

Decision trees can be used for both regression and classification problem. I used regression trees.my methods were uncomplicated and useful for analysis and interpretation (Steorts, 2017). To build the regression tree, I divided the predictor area (that is, the set of possible values for $x_1 \ldots, x_p$) into *J* distinct and non-overlapping regions, $R_1 \ldots, R_j$. For every observation that went into the region R_j, I performed the same prediction, which was the mean of the response values for the training observations in R_j. For example, let us presume that I obtained two regions in step one and the mean of the responses of the training observations in the first region were 10, while the response mean in the second region was 20. So, for a given observation X = x, if x ∈ R_1 , I can predict a value of 10, and if x ∈ R_2, I can predict a value of 20. Regions can have any shape, and I construct a region by choosing to divide the predictor space into high-dimensional rectangles or boxes.

The goal here is to find boxes $R_1 \ldots, R_j$ that minimize the residual sum of squares given by

$$RSS = \sum_{j=1}^{J} \sum_{i \epsilon R_j} (y_i - \hat{y}\, R_j)^2 \qquad (3\text{-}38)$$

where $\hat{y}\, R_j$ is the mean response for the training observations within the j^{th} box (Steorts, 2017).]

A study by Loh (2004) found that a regression tree is a piecewise constant or piecewise linear estimate of a regression function, constructed by recursively partitioning the data and the sample space. The AID algorithm (Morgan & Sonquist, 1963) was the first implementation of this idea. It searched all axis-orthogonal dividers and generated a piecewise constant estimate. At each stage, the binary partition that minimized the total sum of the squared errors (SSE) was selected. The splitting stopped if the fractional decrease in total SSE was less than a pre-specified value γ or if the sample size was too small.

This is a general explanation of the concept of the binary decision tree. As mentioned earlier, I used machine learning toolboxes from MATLAB to apply these supervised learning methods. For the binary decision tree, I began by determining all the Xs and Ys inmy study. Then, I used the syntax "fitrtree (x, y)" to fit this model. The algorithm created a decision tree for each response variable I had (Ys: call centre, online, and mobile application), the variables I wanted to predict, using a number of observations for each of the input variables (Xs: $x_1 \ldots x_{11}$). Below is an example of this, presented as a sample bank (Bank 7).

Table 3-23. Example of results of the binary decision tree for Bank 7

Response variable	Number of observations	Number of nodes
Call center	16,015	4,597

Response variable	Number of observations	Number of nodes
Internet	47,103	8,261
Mobile application	617,567	22,429

The algorithm applied a test for each predictor variable (x) on the tree. Depending on the outcome of the test, the algorithm would go either left or right to the next sub-branch of the tree. It would take the mean of all the observations for the given variable (x). Anything below the mean would require going to the left, and anything above the mean would require going to the right. Below is a description of some of the types of nodes used in the decision tree regression,

Table 3-24: Description of nodes in decision tree regression

Type of nodes	Description
parent	Nodes containing child nodes
child	Nodes coming from the parents
leaf	The root nodes with no child

Eventually, I arrive at a leaf node. At this point, I can make a prediction (which aggregates or averages all the training data points for each leaf node). This makes it possible to predict the number of transactions carried out by each customer in each tested channel.

3.6.6 Support vector machine regression

Support vector machine regression is the last method that I used in my study. It is also a form of supervised machine learning. Support vector machines (SVMs) are often used in classification problems, but they are rarely used in regression problems. When they are used for the latter, they form part of support vector regression (SVR) analysis.

In most linear regression models, the intention is to minimize the sum of squared errors. Take Ordinary Least Squares (OLS) regression as an example. The objective function for OLS is as follows:

$$MIN \sum_{i=1}^{n}(y_i - w_i x_i)^2 \qquad (3\text{-}39)$$

where y_i is the target, x_i is the predictor, and w_i is the coefficient. The goal of most linear regression models is to minimize the sum of squared error between the actual and the predicted data. In SVR, however, the only concern is to reduce the error to a specific degree. It does not matter how significant the errors are, as long as they fall within an acceptable limit. SVR provides flexibility, making it easy to identify how much error is reasonable in a particular model. This shows how to construct an appropriate line (hyperplane in higher dimensions) to fit the data.

SVM is an algorithm that seeks to find an optimal hyperplane that clearly classifies the data points in N-dimensional space (where N is the number of features). SVR uses the idea of a SVM to perform the regression analysis. It performs linear regression in a higher dimension and predicts continuous variables. A kernel function is used to plot lower dimensional data points into higher data points. Here I used the Gaussian kernel. Other linear regression models try to minimize the error between the predicted and the actual value, whereas SVR tries to fit the best line within a predefined or threshold error (ϵ). Furthermore, SVR also tries to categorize all the prediction lines into two types: those that pass through the error boundary and those that do not pass. In an SVM, the hyperplane is a line used to split data classes into higher dimensions than the actual dimension. The boundary lines are two parallel lines drawn on each side of the support vector with the error threshold (ϵ). They create a margin between the data points. These boundary lines try to fit as many cases in as possible without breaching the margin. The width of the boundary is monitored by the error (ϵ) threshold (Sethi, 2020).

The Statistics and Machine Learning Toolbox in MATLAB uses linear, epsilon-insensitive SVM (ϵ-SVM) regression. This is also known as L1 loss. In ε-SVM regression, the training data includes predictors and response variables. The aim is to discover a function f(x) that deviates from yn by a value of no more than ε for each training data point x. Ideally, the result will be as flat as possible.

To understand this, let us suppose that I have training data where xn is a multivariate set of N observations with observed response to the values yn.

To find the linear function:

$$f(x) = x'\beta + b, \qquad (3\text{-}40)$$

To make sure that it is as flat as possible, find $f(x)$ with the minimal norm value $(\beta'\beta)$. This is formulated as a convex optimization problem.

$$J(\beta) = \frac{1}{2}\,\beta'\beta \qquad (3\text{-}41)$$

This is subject to residuals having a value less than ε. As an equation:

$$\forall n: |y_n - (x_n'\beta + b)| \leq \varepsilon \qquad (3\text{-}42)$$

Sometimes, no function $f(x)$ exists to satisfy these constraints for all points. To deal with otherwise infeasible constraints, I can propose slack variables ξn and $\xi * n$ for each point. These slack variables allow regression errors to exist up to the value of ξn and $\xi * n$ while still satisfying the required conditions. Including

slack variables leads to the objective function, also known as the primal formula (Vapnik & N., 1995):

$$J(\beta) = \frac{1}{2}\beta'\beta + C\sum_{n=1}^{N}(\xi_n + \xi_n^*)$$

$$subject\ to:$$

$$\forall n:\ y_n - (x_n'\beta + b) \leq \varepsilon + \xi_n$$

$$\forall n: (\ x_n'\beta + b) - y_n \leq \varepsilon + \xi_n^*$$

$$\forall n: \xi_n^* \geq 0$$

$$\forall n: \xi_n \geq 0 \qquad (3\text{-}43)$$

The constant C is the box constraint, a positive numeric value that regulates the observations that lie outside the epsilon margin (ε) and avoids overfitting. This value determines the trade-off between the smoothness of x and the amount up to which deviations larger than $\boldsymbol{\varepsilon}$ are tolerated.

The linear ε-insensitive loss function disregards errors that are within the distance $\boldsymbol{\varepsilon}$ of the observed value by treating them as equal to zero. The loss is measured based on the space between the observed value y and the ε boundary (MathWorks, 2021). This is described by

$$L_\varepsilon = \begin{cases} 0 & if\,|y - f(x)| \leq \varepsilon \\ |y - f(x)| - \varepsilon & otherwise \end{cases} \qquad (3\text{-}44)$$

This is a general description of the concept of the SVR. As stated previously, I used the Statistics and Machine Learning Toolbox from MATLAB to apply these supervised learning methods. For the SVM regression, I determined all the responses and predictors, then I used the syntax "fitrsvm $(\ x, y)$" to fit this model. The model produced the support vectors and its alphas, from which it was

possible to draw a hyperplane with ε values for the variables. The values for ε for Bank 7 are listed below as an example. Ultimately, the model produced the prediction values based on the rules of the SVM regression.

Table 3-25: Bank 7, ε results

	Mobile application	**Call centre**	**Internet**
ε (epsilon)	0.45	0.1	0.1

3.7 Summary

In this study, I selected the appropriate data needed based on the previous research information along with my banking experience. Then I started the process of requesting and collecting the data from ten local banks in Kuwait. This data collection process was the most challenging part of this research. Negotiation with banks about the confidentiality of the data, what type and size of data I want to get from them, what benefit they will gain from my study, the non-disclosure agreement that need to be signed by different parties and the security of the information, all this took a lot of time and effort. After obtaining the data I dealt with the limitation of the data trying to get the maximum benefit from it with minimum scarification of the valuable financial data I collected. I also documented all the codes that have been used in this study started with data cleaning and organization and ended with the prediction models. Furthermore, I applied four different methods of analysis: Multinomial Logistic Regression (Baseline), Multivariate Linear Regression (Alt 1), Binary Decision Tree for Regression (Alt 2), and Support Vector Machine Regression (Alt 3). To choose the appropriate methods and validate my results, I took the out-of-sample forecast as the key performance measure and found that multinomial logistic regression was the

model best suited to my needs. The reason for this is presented in more detail in the results and discussion chapter.

4 RESULTS AND DISCUSSION

4.1 Introduction

The aim of this study was to focus on the theory of how computationally intelligent data is handled by algorithms in the banking sector, which relies on quick and automated decision-making. I wanted to know which electronic channels banks should use to target customers, based on their characteristics, income, and product portfolio. This would help banks to select the most suitable channels for promoting electronic banking to customers, thereby boosting the performance and profitability of the bank.

I began by identifying the independent variables: age, gender, number of current accounts, number of savings accounts, number of deposit accounts, number of consumer loans, number of instalment loans, income, credit card limit, outstanding credit card balance, duration of relationship with the bank in years, continent, and nationality. Several of these variables had been analyzed in previous studies, as described in the literature review. Others, however, had not previously been analyzed and were included here to provide additional data. The new variables were number of loans (consumer and instalment), credit card limits, income, continent, and nationality.

I then established the independent variables. These were the three electronic channels: call centre, internet, and mobile application. The study included six banks, three of which were ‘Islamic’ and three ‘Conventional’. I examined the coefficients, p-values, standard errors, and 95% confidence intervals for each of the variables. This chapter gives an account of the results for every variable across each of the six banks. It also provides an overview of the key findings for each of the different banks.

4.2 Correlation matrix

Before I began testing my models, I investigated whether there was any correlation between the independent variables. I created a correlation matrix for each bank. The correlation matrix was produced through using the “xcorr” function

in MATLAB which is an implementation of cross-correlation. Cross-correlation determines the resemblance between a vector x and shifted (lagged) versions of a vector y as a function of the lag (Mathworks, 2021p). If x is an M × N matrix, then xcorr(x) returns a (2M – 1) × N^2 matrix with the automated-correlations and cross-correlations of the columns of x. The results are shown in TablesTable 4-1 to Table 4-6. As can be seen from Table 4-1, I found that for Bank 1, there was a negative correlation of -0.44 between the number of current accounts and the number of savings accounts. Customers who had more than one current account tended to have fewer savings accounts and vice versa. The data also showed that the correlation between credit card limits and outstanding credit card balance was 0.56, which indicated that these two variables were positively correlated. Higher credit card limit was related to higher credit card balance. All other variables were either weakly or moderately correlated with one another.

Table 4-1: Bank 1 correlation matrix

	'age'	'gender'	'CA Count'	'DA Count'	'SA Count'	'Consumer Count'	'Income'	'Instalment Count'	'limit'	'Outstanding Bal'
'Age'	1.00	0.09	-0.02	0.03	0.06	0.20	0.02	-0.12	0.03	0.02
'Gender'	0.09	1.00	-0.01	-0.02	-0.02	0.03	0.04	0.00	0.06	0.05
'CA_Count'	-0.02	-0.01	1.00	0.00	-0.44	-0.19	0.12	0.03	0.09	0.06
'DA_Count'	0.03	-0.02	0.00	1.00	0.05	-0.03	0.10	0.06	0.18	0.07
'SA_Count'	0.06	-0.02	-0.44	0.05	1.00	0.10	-0.01	-0.05	-0.02	-0.03
'Consumer_Count'	0.20	0.03	-0.19	-0.03	0.10	1.00	-0.08	-0.20	-0.10	-0.06
'Income'	0.02	0.04	0.12	0.10	-0.01	-0.08	1.00	0.12	0.28	0.16
'Instalment_Count'	-0.12	0.00	0.03	0.06	-0.05	-0.20	0.12	1.00	0.07	0.09
'Limit'	0.03	0.06	0.09	0.18	-0.02	-0.10	0.28	0.07	1.00	0.56
'Outstanding_Bal'	0.02	0.05	0.06	0.07	-0.03	-0.06	0.16	0.09	0.56	1.00

Table 4-2 shows that the correlation coefficient for Bank 2 was -0.57, indicating a negative correlation between the number of consumer loans and the number of instalment loans. Customers who had more than one consumer loan tended to have fewer instalment loans and vice versa. The results also show a positive correlation between income and outstanding credit card balance, with a correlation coefficient of 0.55. There was a high correlation (0.73) between credit card limit and outstanding credit card balance. The higher the limit the higher the outstanding balance. Most variables were weakly correlated, and a few variables were moderately correlated.

Table 4-2: Bank 2 correlation matrix

	'age'	'gender'	'CA Count'	'DA Count'	'SA Count'	'Consumer Count'	'Income'	'Instalment Count'	'limit'	'Outstanding Bal'	'Relationship duration'
'age'	1.00	0.08	0.08	0.02	0.01	0.00	0.17	0.02	0.11	0.14	0.23
'gender'	0.08	1.00	0.00	0.01	0.03	-0.06	0.09	0.05	0.02	0.10	0.10
'CA_Count'	0.08	0.00	1.00	0.09	-0.34	-0.11	0.33	0.12	0.17	0.31	0.16
'DA_Count'	0.02	0.01	0.09	1.00	0.00	-0.08	0.10	0.05	0.04	0.11	0.07
'SA_Count'	0.01	0.03	-0.34	0.00	1.00	0.04	0.01	-0.03	-0.01	-0.02	0.15
'Consumer_Count'	0.00	-0.06	-0.11	-0.08	0.04	1.00	-0.22	-0.57	-0.15	-0.19	-0.23
'Income'	0.17	0.09	0.33	0.10	0.01	-0.22	1.00	0.33	0.41	0.55	0.31
'Instalment_Count'	0.02	0.05	0.12	0.05	-0.03	-0.57	0.33	1.00	0.28	0.27	0.32
'limit'	0.11	0.02	0.17	0.04	-0.01	-0.15	0.41	0.28	1.00	0.73	0.22
'Outstanding_Bal'	0.14	0.10	0.31	0.11	-0.02	-0.19	0.55	0.27	0.73	1.00	0.29
'Relationshipduration'	0.23	0.10	0.16	0.07	0.15	-0.23	0.31	0.32	0.22	0.29	1.00

Table 4-3 shows a correlation coefficient of 0.54 between credit card limit and outstanding credit card balance in Bank 3. The higher the limit the higher the outstanding balance. Most variables were weakly correlated, though a few variables were moderately correlated.

Table 4-3: Bank 3 correlation matrix

	'age'	'gender'	'CA Count'	'DA Count'	'SA Count'	'Consumer Count'	'Income'	'Instalment Count'	'limit'	'Outstanding Bal'	'Relationship duration'
'age'	1.00	-0.01	0.14	0.17	-0.02	0.03	0.23	0.01	0.23	0.12	0.22
'gender'	-0.01	1.00	0.05	-0.03	-0.01	0.03	0.18	0.10	0.11	0.04	0.16
'CA_Count'	0.14	0.05	1.00	0.04	-0.02	0.05	0.10	0.02	0.08	0.06	0.07
'DA_Count'	0.17	-0.03	0.04	1.00	0.08	0.03	0.09	-0.05	0.12	0.05	0.04
'SA_Count'	-0.02	-0.01	-0.02	0.08	1.00	0.06	0.10	-0.04	0.13	0.05	0.04
'Consumer_Count'	0.03	0.03	0.05	0.03	0.06	1.00	0.06	-0.46	0.09	0.10	0.01
'Income'	0.23	0.18	0.10	0.09	0.10	0.06	1.00	0.01	0.28	0.17	0.09
'Instalment_Count'	0.01	0.10	0.02	-0.05	-0.04	-0.46	0.01	1.00	-0.07	-0.16	0.12
'limit'	0.23	0.11	0.08	0.12	0.13	0.09	0.28	-0.07	1.00	0.54	0.10
'Outstanding_Bal'	0.12	0.04	0.06	0.05	0.05	0.10	0.17	-0.16	0.54	1.00	0.01
'Relationshipduration'	0.22	0.16	0.07	0.04	0.04	0.01	0.09	0.12	0.10	0.01	1.00

Table 4-4 shows the results for Bank 4. There was a positive correlation of 0.55 between income and credit card limit. There was also a positive correlation coefficient of 0.59 between credit card limit and outstanding credit card balance. Table 4-5 and Table 4-6 show the results for Bank 5 and Bank 6, which had a moderate and positive correlation of 0.46 and 0.85, respectively, between credit

card limit and outstanding credit card balance. Most of the other variables were weakly correlated. A few variables were moderately correlated.

Table 4-4: Bank 4 correlation matrix

	'age'	**'gender'**	**'CA Count'**	**'DA Count'**	**'SA Count'**	**'Consumer Count'**	**'Income'**	**'Instalment Count'**	**'limit'**	**'Outstanding Bal'**	**'Relationship duration'**
'age'	1.00	0.08	0.34	0.10	-0.02	0.00	0.25	-0.08	0.28	0.09	0.56
'gender'	0.08	1.00	0.12	-0.03	-0.06	0.06	0.18	0.01	0.12	0.03	0.03
'CA_Count'	0.34	0.12	1.00	0.09	0.00	-0.07	0.26	0.01	0.32	0.14	0.39
'TD_Count'	0.10	-0.03	0.09	1.00	0.08	-0.02	0.05	0.00	0.08	0.02	0.07
'SA_Count'	-0.02	-0.06	0.00	0.08	1.00	-0.08	0.12	0.11	0.09	0.07	0.02
'Consumer_Count'	0.00	0.06	-0.07	-0.02	-0.08	1.00	-0.05	-0.25	-0.03	0.04	-0.12
'Income'	0.25	0.18	0.26	0.05	0.12	-0.05	1.00	0.08	0.55	0.29	0.23
'Instalment_Count'	-0.08	0.01	0.01	0.00	0.11	-0.25	0.08	1.00	-0.04	0.06	0.10
'limit'	0.28	0.12	0.32	0.08	0.09	-0.03	0.55	-0.04	1.00	0.59	0.24
'Outstanding_Bal'	0.09	0.03	0.14	0.02	0.07	0.04	0.29	0.06	0.59	1.00	0.11
'Relationshipduration'	0.56	0.03	0.39	0.07	0.02	-0.12	0.23	0.10	0.24	0.11	1.00

Table 4-5: Bank 5 correlation matrix

	'age'	'gender'	'CA Count'	'DA Count'	'SA Count'	'Consumer Count'	'Income'	'Instalment Count'	'limit'	'Outstanding Bal'	'Relationship duration'
'age'	1.00	0.05	0.04	0.06	0.00	0.00	0.16	-0.09	0.16	0.04	-0.01
'gender'	0.05	1.00	0.06	-0.02	-0.01	-0.04	0.12	0.08	0.09	0.02	-0.02
'CA_Count'	0.04	0.06	1.00	0.02	-0.11	0.06	0.04	0.02	0.06	0.04	-0.04
'DA_Count'	0.06	-0.02	0.02	1.00	0.16	0.04	0.06	-0.06	0.07	0.00	0.00
'SA_Count'	0.00	-0.01	-0.11	0.16	1.00	0.15	0.09	0.00	0.16	0.04	0.12
'Consumer_Count'	0.00	-0.04	0.06	0.04	0.15	1.00	0.08	-0.34	0.09	0.04	-0.22
'Income'	0.16	0.12	0.04	0.06	0.09	0.08	1.00	0.04	0.23	0.06	0.09
'Instalment_Count'	-0.09	0.08	0.02	-0.06	0.00	-0.34	0.04	1.00	0.03	0.03	0.37
'limit'	0.16	0.09	0.06	0.07	0.16	0.09	0.23	0.03	1.00	0.46	0.17
'Outstanding_Bal'	0.04	0.02	0.04	0.00	0.04	0.04	0.06	0.03	0.46	1.00	-0.03
'Relationshipduration'	-0.01	-0.02	-0.04	0.00	0.12	-0.22	0.09	0.37	0.17	-0.03	1.00

Table 4-6: Bank 6 correlation matrix

	'age'	'gender'	'CA Count'	'DA Count'	'SA Count'	'Consumer Count'	'Income'	'Instalment Count'	'limit'	'Outstanding Bal'	'Relationship duration'
'age'	1.00	0.00	0.04	0.13	0.03	0.08	0.20	-0.06	0.16	0.07	0.55
'gender'	0.00	1.00	0.03	-0.06	-0.03	0.06	0.23	0.13	0.04	0.03	-0.03
'CA_Count'	0.04	0.03	1.00	-0.11	-0.22	0.00	0.02	0.01	0.03	0.02	0.00

	'age'	'gender'	'CA Count'	'DA Count'	'SA Count'	'Consumer Count'	'Income'	'Instalment Count'	'limit'	'Outstanding Bal'	'Relationship duration'
'DA_Count'	0.13	-0.06	-0.11	1.00	-0.07	0.01	0.07	0.00	0.09	0.06	0.12
'SA_Count'	0.03	-0.03	-0.22	-0.07	1.00	0.02	0.05	-0.02	0.05	0.03	0.09
'Consumer_Count'	0.08	0.06	0.00	0.01	0.02	1.00	0.14	-0.36	0.07	0.04	0.14
'Income'	0.20	0.23	0.02	0.07	0.05	0.14	1.00	0.08	0.39	0.28	0.27
'Instalment_Count'	-0.06	0.13	0.01	0.00	-0.02	-0.36	0.08	1.00	0.00	0.06	-0.01
'limit'	0.16	0.04	0.03	0.09	0.05	0.07	0.39	0.00	1.00	0.85	0.24
'Outstanding_Bal'	0.07	0.03	0.02	0.06	0.03	0.04	0.28	0.06	0.85	1.00	0.16
'Relationshipduration'	0.55	-0.03	0.00	0.12	0.09	0.14	0.27	-0.01	0.24	0.16	1.00

Due to the large amount of data, the model would become sensitive to highly correlated variables, which means that small variations in the data could cause large residual errors in the predictions. In my case, however, the only correlation coefficient that was too high was between credit card limits and outstanding credit card balance (Ratner, 2009). The relation between these two variables was considered high in two of the banks, moderate in three of the banks, and low in one of the banks.

I also noticed that the results of the Islamic and conventional banks were comparable. Most of the variables were weakly correlated, and a few were moderately correlated. However, credit card limit and outstanding credit card

balance were moderately and highly correlated in both Islamic and conventional banks. Given the importance of these two variables, I monitored the behaviour of customers who had balances due with the bank (that is, an outstanding balance). There were lots of customers who had credit card limits that they did not use or who paid off their credit card immediately. For this reason, I decided to include both variables, but only once I had completed further investigation into the effect of including both variables in the model. I ran the model twice for two samples banks: one Islamic and one conventional. The first time, I included the two variables; the second time, I excluded the outstanding balance variables. I compared the two models based on the time of execution and the key performance parameters (see Table 4-7). I did not find any noticeable differences between the two models; therefore, I decided to include all the independent variables in my model. Table 4-8, Table 4-9, and Table 4-10 show the outcomes of the coefficients, p-values, and standard errors from running the two multinomial logistic regressions (including and excluding the outstanding credit card outstanding balance).

Table 4-7: Multinomial logistic regression with and without the variable "outstanding credit card balance"

	Sample 1: Bank 2 (conventional)		**Sample 2: Bank 3 (Islamic)**	
	Before excluding	After excluding	Before excluding	After excluding
Model running time	3 min	5 min	1 min	2 min
Coefficients	Almost the same with minimal changes		Almost the same with minimal changes	
Pv				
Se				

Table 4-8: Bank 2 and Bank 3 coefficients including and excluding the variable "outstanding credit card balance"

Coefficients	Sample bank (Bank 2)				Sample bank (Bank 3)			
	Before excluding		After excluding		Before excluding		After excluding	
	Internet /Mobile	Call Centre/ Mobile	Internet/ Mobile	Call Centre/ Mobile	Internet/ Mobile	Call Centre/ Mobile	Internet/ Mobile	Call Centre/Mobile
Intercept	2.710	2.392	2.532	2.433	-2.255	-7.016	-2.259	-7.064
X1 Age	-0.025	0.001	-0.022	0.002	0.011	0.040	0.011	0.040
X2 Gender	0.378	0.249	0.561	0.380	0.364	-0.915	0.364	-0.915
X3 CA Count	0.410	-0.649	0.622	-0.540	-0.132	-0.258	-0.132	-0.263
X4 DA Count	1.521	-0.294	1.362	0.270	0.040	0.056	0.040	0.057
X5 JA Count	0.931	1.888	0.710	1.621	N/A	N/A	N/A	N/A
X6 SA Count	1.471	-0.674	1.509	-0.728	-0.248	-0.276	-0.247	-0.273
X7 Nationality	0.004	-0.007	0.004	-0.008	N/A	N/A	N/A	N/A
X8 Consumer Count	0.044	-2.542	0.019	-2.519	-0.038	0.212	-0.038	0.206
X9 Income	0.000	-0.001	0.000	-0.001	0.000	0.000	0.000	0.000
X10 Intstallment Count	-0.485	-0.898	-0.516	-0.910	-0.042	0.375	-0.041	0.389
X11 Limit	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
X12 Outstanding Balance	0.000	0.000	N/A	N/A	0.000	0.000	N/A	N/A
X13 Relationship Duration	-0.021	0.243	-0.022	0.239	0.057	0.244	0.057	0.245

Table 4-9: Bank 2 and Bank 3 p-values including and excluding the variable "outstanding credit card balance"

P-Value	Sample bank (Bank 2)				Sample bank (Bank 3)			
	Before excluding		After excluding		Before excluding		After excluding	
	Internet /Mobile	Call Centre/ Mobile	Internet/ Mobile	Call Centre/ Mobile	Internet/ Mobile	Call Centre/ Mobile	Internet/ Mobile	Call Centre/Mobile
Intercept	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

P-Value	Sample bank (Bank 2)				Sample bank (Bank 3)			
	Before excluding		After excluding		Before excluding		After excluding	
	Internet /Mobile	Call Centre/ Mobile	Internet/ Mobile	Call Centre/ Mobile	Internet/ Mobile	Call Centre/ Mobile	Internet/ Mobile	Call Centre/Mobile
X1 Age	0.000	0.854	0.000	0.643	0.000	0.000	0.000	0.000
X2 Gender	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
X3 CA Count	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
X4 DA Count	0.000	0.538	0.000	0.548	0.000	0.000	0.000	0.000
X5 JA Count	0.000	0.000	0.000	0.000	N/A	N/A	N/A	N/A
X6 SA Count	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
X7 Nationality	0.000	0.000	0.000	0.000	N/A	N/A	N/A	N/A
X8 Consumer Count	0.266	0.000	0.622	0.000	0.000	0.000	0.000	0.000
X9 Income	0.890	0.000	0.010	0.000	0.000	0.000	0.000	0.000
X10 Intstallment Count	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
X11 Limit	0.000	0.278	0.264	0.000	0.000	0.000	0.000	0.000
X12 Outstanding Balance	0.000	0.485	N/A	N/A	0.000	0.000	N/A	N/A
X13 Relationship Duration	0.002	0.000	0.001	0.000	0.000	0.000	0.000	0.000

Table 4-10: Bank 2 and Bank 3 standard errors, including and excluding the variable "outstanding credit card balance"

Standard Error	Sample bank (Bank 2)				Sample bank (Bank 3)			
	Before excluding		After excluding		Before excluding		After excluding	
	Internet /Mobile	Call Centre/ Mobile	Internet/ Mobile	Call Centre/ Mobile	Internet/ Mobile	Call Centre/ Mobile	Internet/ Mobile	Call Centre/Mobile
Intercept	0.165	0.204	0.156	0.193	0.012	0.045	0.012	0.044
X1 Age	0.003	0.004	0.003	0.004	0.000	0.001	0.000	0.001
X2 Gender	0.054	0.064	0.052	0.062	0.005	0.016	0.005	0.016
X3 CA Count	0.044	0.056	0.042	0.054	0.006	0.022	0.006	0.022
X4 DA Count	0.366	0.477	0.346	0.450	0.003	0.011	0.003	0.011

Standard Error	Sample bank (Bank 2)				Sample bank (Bank 3)			
	Before excluding		After excluding		Before excluding		After excluding	
	Internet /Mobile	Call Centre/ Mobile	Internet/ Mobile	Call Centre/ Mobile	Internet/ Mobile	Call Centre/ Mobile	Internet/ Mobile	Call Centre/Mobile
X5 JA Count	0.127	0.137	0.118	0.129	N/A	N/A	N/A	N/A
X6 SA Count	0.127	0.164	0.119	0.155	0.002	0.009	0.002	0.009
X7 Nationality	0.001	0.001	0.001	0.001	N/A	N/A	N/A	N/A
X8 Consumer Count	0.039	0.057	0.038	0.055	0.004	0.016	0.004	0.016
X9 Income	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
X10 Intstallment Count	0.027	0.037	0.026	0.036	0.003	0.007	0.003	0.007
X11 Limit	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
X12 Outstanding Balance	0.000	0.000	N/A	N/A	0.000	0.000	N/A	N/A
X13 Relationship Duration	0.007	0.008	0.006	0.008	0.001	0.003	0.001	0.003

4.3 Data training and testing

I applied four different methods in this study: Multinomial Logistic Regression (Baseline), Multivariate Linear Regression (Alt 1), Binary Decision Tree for Regression (Alt 2), and Support Vector Machine Regression (Alt 3). These were discussed in more detail in the Methodology section.

To choose the appropriate methods and validate my results, I took the out-of-sample prediction accuracy as the key performance measure (see Figure 4-1). To provide more detail about my key performance measures, I divided the data into a training sample (2/3 of the total data) and an out-of-sample (1/3 of the total data). I applied each model on the training sample. Then, I took the coefficients from the training sample to predict the output of the out-of-sample data. I calculated the deviation between the actual output and the predicted output of the

out-of-sample data in terms of probabilities. The smaller the absolute value of the deviation, the better the performance.

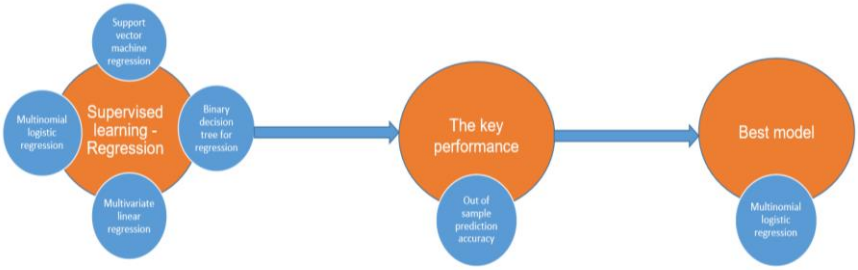


Figure 4-1: Different methods used in the research

As shown by Figure 4-2, multinomial logistic regression was the baseline, represented by the blue bars. The yellow bars represent the decision tree method, which was the least effective, so I excluded it immediately. The purple bars indicate the support vector method. This was better than the base line for automated teller machines and similar to the baseline for call centers and the internet. But it was significantly worse for mobile applications and point of sale. The results showed that multivariate linear regression, represented by the orange bars, was similar to the baseline. Overall, multinomial logistic regression was reasonably robust, though there are alternatives that could be used such as multivariate linear regression and support vector machine regression.

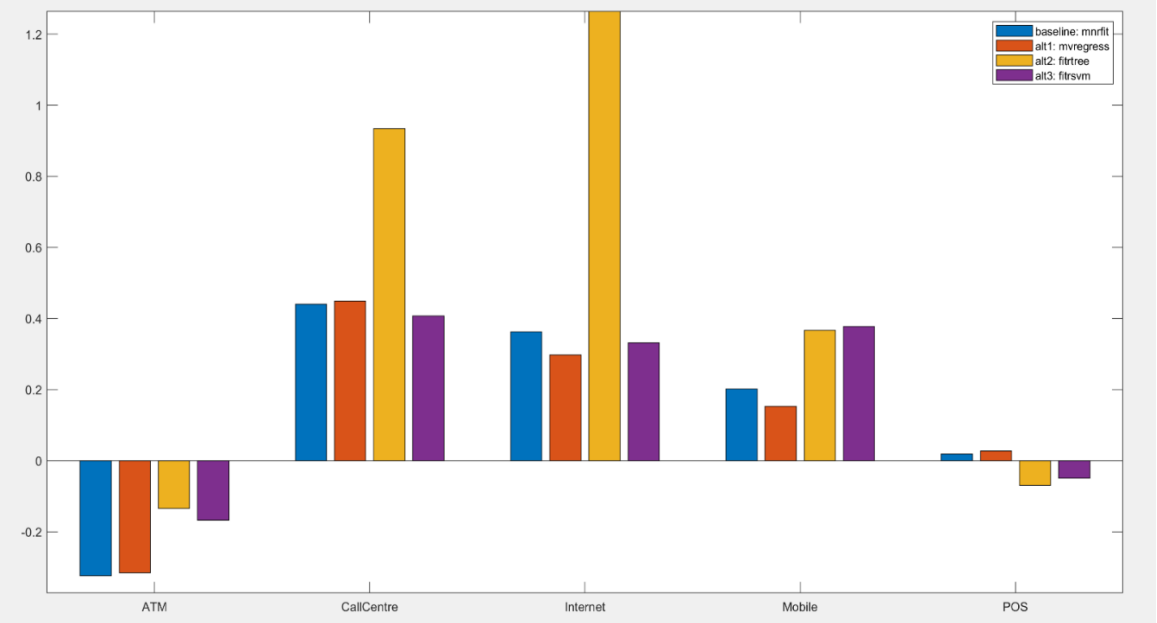


Figure 4-2: Key performance measures and validations

The validation can be seen in the graphs below. Figure 4-3 plots the graph of age vs probabilities and validates the prediction formula. If I look at how the likelihood of using different forms of banking changes with age, the results make sense. This validates the regression analysis.

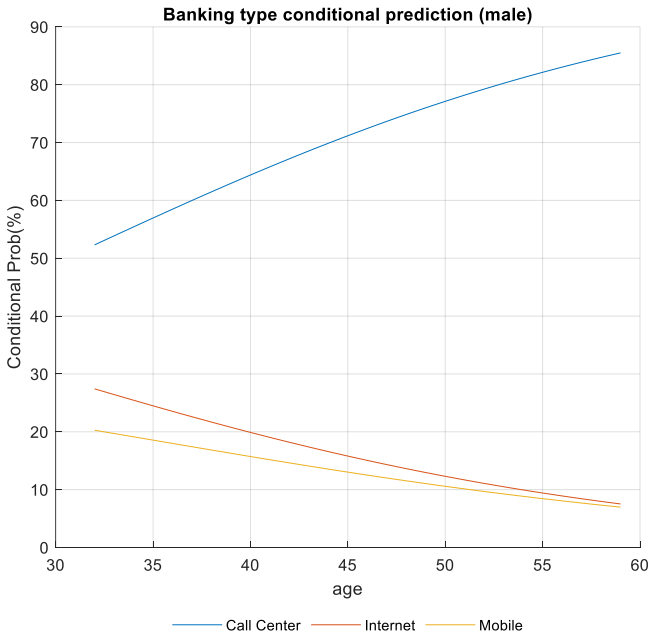


Figure 4-3: Key performance measures and validation graph

4.4 Sensitivity analysis

I also addressed the relationship between the training data and the model accuracy to see how sensitive the models were to fluctuations in the data. I used different amounts of data, beginning with 55% of the total data and ending with 80% of the total data, increasing the volume of data by 5% at a time. For each sample data set, I created the out-of-sample and calculated the accuracy of the out-of-sample prediction. The accuracy of the out-of-sample prediction was calculated based on the difference between the actual and predicted outputs in the out-of-sample data set. The smaller the gap between them, the better the out-of-sample performance. Figure 4-4 shows the relation between the sample sizes (ranging from 55% to 85%) and the accuracy of the out-of-sample measure. The results show that the multinomial logistic regression (baseline) and the multivariate linear regression (alt 1) were stable across different in sample sizes. Their performance was consistently better than the other models tested.

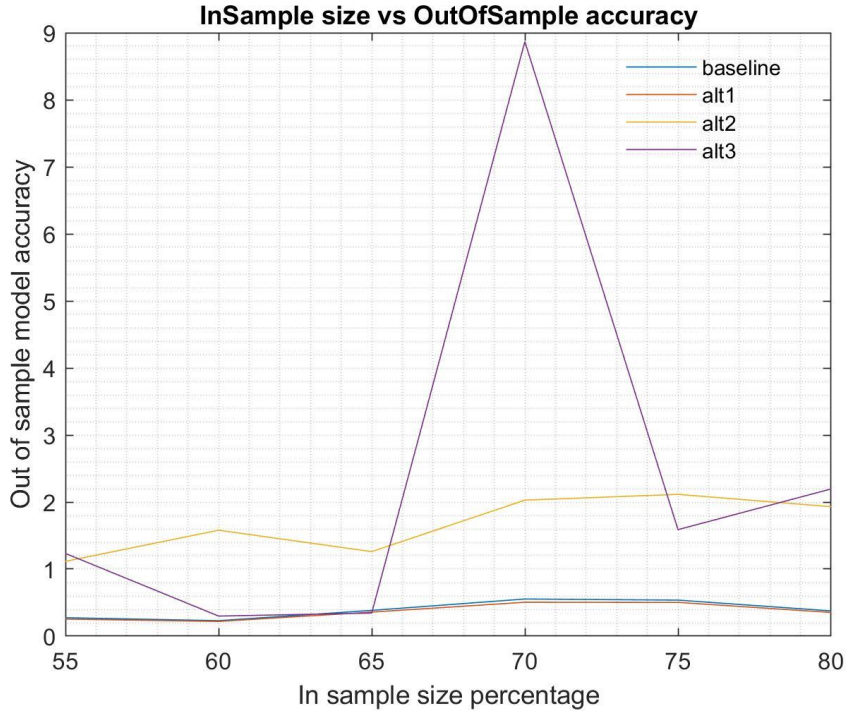


Figure 4-4: Sensitivity analysis using Bank 6 sample data (Islamic bank)

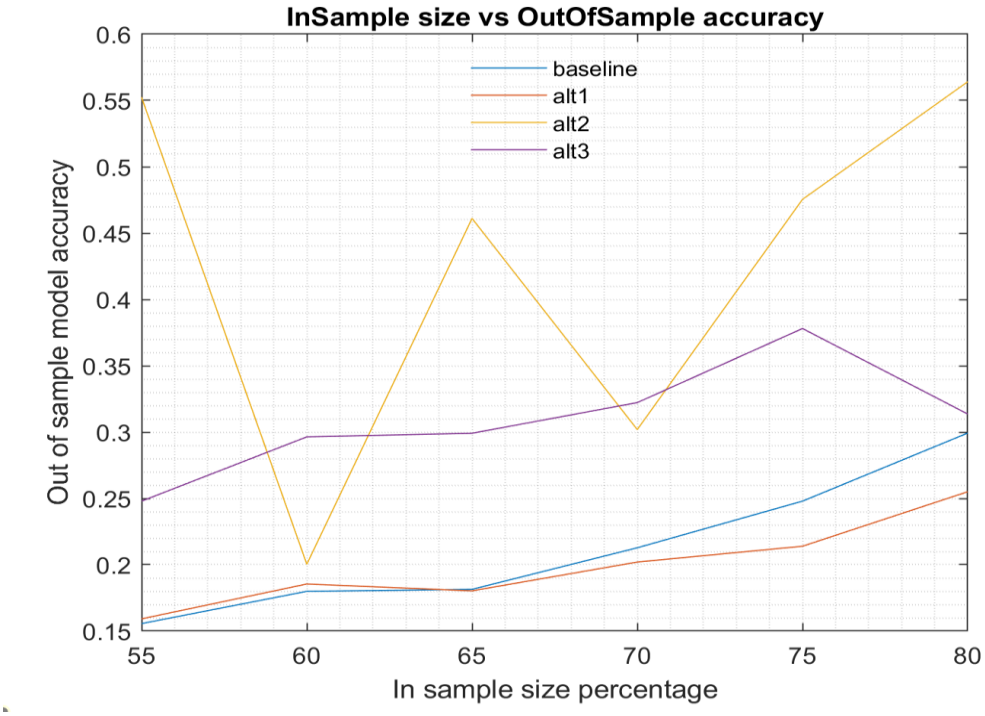


Figure 4-5: Sensitivity analysis using Bank 4 sample data (conventional bank)

To confirm my results, I repeated the sensitivity test on one of the biggest conventional banks in Kuwait. In line with the test for the Islamic bank, I excluded the binary decision tree (alt 2) for regression and the support vector machine regression (alt 3) because of the high deviation between the actual and predicted outputs in the out-of-sample sets (Figure 4-5). In the first three sample sets (55%, 60%, 65%) I found that the multinomial logistic regression was slightly better than multivariate linear regression. In the last three sets (70%, 75%, 80%) I found that the multivariate linear regression was marginally better than the multinomial logistic regression.

In summary, there are many reasons why I chose the multinomial logistic regression for this study. First, when I compared the code execution time, I found that there was an insignificant difference between the two methods. The multinomial logistic regression time was only one minute less than the multivariate linear regression. Second, when using the training model with the algorithm, multinomial logistic regression did not require as much computational power. Third, as far as analysis is concerned, multinomial logistic regression uses one of

the dependent variables as a reference category for the other dependent variables and also produces probabilities that make for better analysis.

4.5 Multinomial logistic regression

I used multinomial logistic regression to analyse the regression results (for the full results, see tables Á-1 to A-5 in appendix A). I used the Statistics and Machine Learning Toolbox in MATLAB to run a multinomial regression analysis. The aim was to predict the usage rates of electronic banking channels based on age, gender, number of current accounts, number of savings accounts, number of deposit accounts, income, number of consumer loans, number of instalment loans, credit card limit, outstanding credit card balance and the duration of the relationship with the bank. A detailed illustration of the model is provided by Equation (**4**-**1**):

$$[\boldsymbol{B}, \boldsymbol{dev}, \boldsymbol{stats}] = \boldsymbol{mnrfit}(\boldsymbol{x}, \boldsymbol{y}) \qquad (4\text{-}1)$$

Where *x* represents the parameters used for the prediction of *y*, the usage of electronic banking (Mathworks, 2019). The coefficients in the results express:

1) The effect of the independent variables on the relative risk of using one electronic channel compared to the reference electronic channel (in my example, the reference is the mobile channel).

2) The effect of the independent variables on the relative log odds of using one electronic channel compared to the reference electronic channel.

A p-value or probability value is a number that indicates the likelihood of study's data would occur by random chance. The significance as usually expressed as a p-value between 0-1 (McLeod, 2019). A small p-values indicates that the parameter (*x*) is significant, based on the relative risk of using one online channel over another one. When a variable has explanatory power, its coefficient will be highly significant, with a p-value close to zero.

The p-value for a particular result demonstrates whether there is a statistically significant or non-significant difference in the adoption of electronic banking services across the three online channels (call centre, internet and mobile). It does not give an estimation of what that difference is. This estimation is provided by the confidence interval, which tells us how effective the estimate is at predicting the population level. A narrower confidence interval indicates a more accurate estimate, whereas a wider confidence interval indicates a less accurate estimate (Tan & Tan, 2010). To calculate the confidence interval, I began by calculating the standard errors of the coefficient estimates. I then assessed the 95% confidence limits for the coefficients by computing the lower and upper limits (Equation (**4**-**2**):

LL = (beta - 1.96) (standard error) (4-2)

UL = (beta + 1.96) (standard error)

The results demonstrate that I can be 95% certain that the true regression coefficients for the independent parameters used in my analysis were between the range of the lower and upper limits.

The following sections of this chapter focus on each of the different variables in turn. A summary of the results for each of the different banks is provided, followed by an overall discussion.

4.6 Independent variable 1: age

The first variable I looked at was age. Isolating all other variables, I identified the relative risk of customers using the internet rather than a mobile application and the risk of them using a call centre rather than a mobile application. The following sections provide an overview of the results for each of the different banks.

4.6.1 Bank 1

For the first bank that I looked at, the relative risk of a customer using the internet rather than a mobile application increased 1.02 (exp 0.025) times for each additional year in age, all other things being equal. Meanwhile, the relative risk of a customer using a call centre rather than a mobile application increased 1.03 (exp 0.033) times for each unit increase in age, all other things being equal. The p-value for age in Bank 1 was close to 0, indicating that age has a significant impact on the relative risk of a customer choosing one online channel over another. The standard error was also 0 for both channels. The results of the confidence interval mean that I can be 95% sure that the lower limits of the regression coefficient for age were 0.023 for internet/mobile and 0.026 for call centre/mobile; the upper limits of the regression coefficient for age were 0.032 for internet/mobile and 0.034 for call centre/mobile.

4.6.2 Bank 2

The results of the second bank showed that the relative risk of a customer using the internet rather than a mobile application increased 1.016 (exp 0.016) times for each unit increase in age. The relative risk of a customer using a call centre rather than a mobile application increased 1.025 (exp 0.024) times for each unit increase in age. Again, the p-values that were vanishingly close to 0 and the standard errors were low: 0.003 and 0.002 for internet/mobile and call centre/mobile, respectively. The results of the confidence interval show that I can be 95% sure that the lower limits of the regression coefficient for the age I used in my test for the second bank were 0.009 and 0.022 for internet/mobile and call centre/mobile, respectively. I can be 95% sure that the upper limits of the regression coefficient for age were 0.020 and 0.029 for internet/mobile and call centre/mobile, respectively.

4.6.3 Bank 3

The results for Bank 3 showed similar increases in the risk of a customer using the internet rather than a mobile application, rising by an order of 1.010 (exp 0.010) for every unit increase in age. The relative risk of a customer using a call

centre rather than a mobile application increased by 1.040 (exp 0.039) times for each additional unit in age. The p-values were close to 0, again revealing that age had a significant impact on the relative risk of a consumer choosing one online channel over another. The standard errors for age in this bank were also zero. The results of the confidence interval mean that I can be 95% sure that the lower limits of the regression coefficient for age were 0.010 for internet/mobile and 0.037 for call centre/mobile. The upper limits of the regression coefficient for age were 0.037 and 0.041 for internet/mobile and call centre/mobile, respectively.

4.6.4 Bank 4

The results of Bank 4 showed that age had an even greater impact on the relative risk of a customer using the internet rather than a mobile application. For each unit increase in age, a customer was 1.045 (exp 0.044) times more likely to use the internet rather than a mobile application. Similarly, the relative risk of a customer using a call centre rather than a mobile application increased by 1.051 (exp 0.050) for each unit increase in age. The p-values and standard errors were 0. The 95% confidence intervals ranged from 0.044 to 0.045 for internet/mobile and from 0.049 to 0.050 for call centre/mobile.

4.6.5 Bank 5

At the fifth bank I analysed, the relative risk of a customer using the internet rather than a mobile application increased by 1.005 (exp 5.27E-03) for each additional unit of age. The relative risk of a customer using a call centre rather than their mobile increased by 1.008 (exp 0.008) for each additional unit of age. Again, the p-values for both internet/mobile and call centre/mobile were 0. The standard errors were 0 and 0.001, respectively. The confidence intervals were between 0.004 and 0.005 for internet/mobile and between 0.005 and 0.011 for call centre/mobile.

4.6.6 Bank 6

Finally, at the sixth bank I assessed, the relative risk of a customer using the internet rather than a mobile application increased 1.028 (exp 0.027) times for

each unit increase in age, and the relative risk of a customer using a call centre rather than a mobile application increases by 1.060 (exp 0.058) for each unit increase in age. P-values and standard errors were 0. The results of the confidence interval mean that I can be 95% sure that the true regression coefficient was between the range of 0.026 and 0.028 for internet/mobile and between 0.057 and 0.060 for call centre/mobile.

4.6.7 Results summary for independent variable 1: age

Table 4-11 provides an overview of the multinomial logistic regression results for the independent variable 'age' for each of the six different banks. The data for each bank is divided based on the internet/mobile (I/M) and call centre/mobile (C/M) comparisons.

Table 4-11: Multinomial logistic regression results for independent variable 1: age

	Bank 1		**Bank 2**		**Bank 3**		**Bank 4**		**Bank 5**		**Bank 6**	
Variable 1: Age	I/M*	C/M	I/M	C/M	I/M	C/M	I/M	C/M	I/M	C/M	I/M	C/M
Coefficients -relative log odds	0.025	0.033	0.016	0.024	0.010	0.039	0.044	0.050	5.27 E-03	0.008	0.027	0.058
Coefficients -relative risk exp (coefficients)	1.025	1.033	1.016	1.025	1.010	1.040	1.045	1.051	1.005	1.008	1.028	1.060
P-value	2.84 E-291	0	5.52 E-07	3.26 E-24	3.415E-312	0.00 E+00	0.00 E+00	0	2.19 E-50	1.16 E-09	0.00 E+00	0
Standard Error	0.000	0.000	0.003	0.002	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000
95% confidence limits for the coefficients - Lower Limit	0.023	0.032	0.009	0.020	0.010	0.037	0.044	0.049	0.004	0.005	0.026	0.057

	Bank 1		Bank 2		Bank 3		Bank 4		Bank 5		Bank 6	
95% confidence limits for the coefficients - Upper Limit	0.026	0.034	0.022	0.029	0.011	0.041	0.045	0.050	0.005	0.011	0.028	0.060

** I/M = Internet/Mobile; C/M = Call Centre/Mobile*

4.6.8 Overall analysis of independent variable 1: age

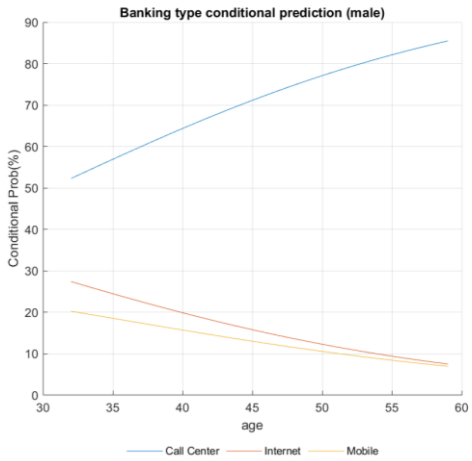
The results suggest that, overall, the model performed well in assessing whether age has a significant effect on a customer's likelihood to use the internet, a call centre, or a mobile application. The low p-values and standard error values for each of the banks show that age was a statistically significant variable, and the short distances between the confidence intervals mean that I can trust the results.

The results suggest the following:

1. As customers get older, they tend to prefer using call centres as their primary method of communication. Their second preference is to use the internet. Their least preferred option is to use a mobile phone (see Figure 4-6). This outcomes supports the findings of Gensler et al.'s (2012) study, in which younger customers tended to use mobile banking more than older ones.

2. There was no difference in customer preferences between the Islamic and conventional banks.

Sample 1: Conventional bank



Sample 2: Islamic bank

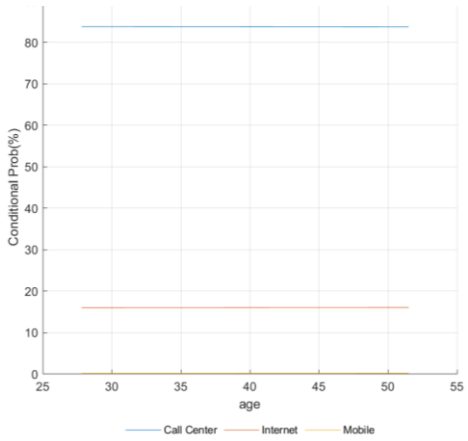


Figure 4-6: Probability of using different electronic banking channels dependent on age

4.7 Independent variable 2: gender

The second variable I examined was gender. The aim was to assess whether men or women were more likely to choose a particular channel of communication for dealing with their bank. Again, I compared the differences between customers at conventional banks and Islamic banks. The following sections give a summary of the results for each of the different banks.

4.7.1 Bank 1

The results for Bank 1 show that men were 0.85 (exp -0.15) times less likely than women to use the internet rather than a mobile application, all other factors being

equal. Men were also 0.63 (exp -0.45) times less likely than women to use a call centre rather than a mobile application. The p-values were close to 0 for internet/mobile and call centre/mobile, suggesting that gender has a significant impact on both. The standard errors were the same (0.014 for both), and the confidence interval ranges were from -0.185 to -0.127 for internet/mobile and from -0.486 to -0.428 for call centre/mobile.

4.7.2 Bank 2

The results for Bank 2 were similar, with men 0.608 (exp -0.496) times less likely than women to use the internet rather than a mobile application and 0.899 (exp -0.106) times less likely to use a call centre than a mobile application. Again, there were small p-values and standard errors, suggesting that gender was a significant variable at Bank 2. The results of the confidence interval mean that I can be 95% sure that the true regression coefficient was between the range of -0.601 and -0.390 for internet/mobile and between -0.190 and -0.022 for call centre/mobile.

4.7.3 Bank 3

The results for Bank 3, on the other hand, were slightly different. At this bank, men were 1.438 (exp 0.363) times more likely than women to use the internet rather than a mobile application. On the other hand, men showed a decrease in likelihood of 0.400 (exp -0.915) when it came to using a call centre rather than a mobile application. P-values of 0 and standard errors of 0.004 and 0.015 for internet/mobile and call centre/mobile, respectively, show that gender had a statistically significant impact at Bank 3. The confidence intervals ranged from 0.354 to 0.0373 for internet/mobile and from -0.946 to -0.884 for call centre/mobile.

4.7.4 Bank 4

The results for Bank 4 show that men are 0.729 (exp -0.315) times less likely than women to use the internet rather than a mobile application. Men are also 0.600 (exp -0.509) times less likely than women to use a call centre rather than a mobile application. The p-values for both criteria were 0, and the standard errors were

0.003 and 0.005 for internet/mobile and call centre/mobile, respectively. in case of the dependent variable internet vs mobile and call centre verses mobile, respectively. The 95% confidence intervals ranged from -0.321 to -0.309 for internet/mobile and from -0.521 to -0.498 for centre/mobile.

4.7.5 Bank 5

Male customers at Bank 5 were slightly more likely than women to use the internet rather than a mobile application (1.003 times more likely, exp 0.003). By contrast, men were 0.818 (exp -0.200) times less likely than women to use a call centre rather than a mobile application. That said, there was a high p-value of 0.6 for internet/mobile at Bank 5, suggesting that gender did not have a statistically significant impact on a person's decision to use the internet or a mobile application. There were low standard errors for both internet/mobile (0.007) and call centre/mobile (0.028). in case of the dependent variable internet vs mobile and call centre verses mobile, respectively. The results of the confidence interval mean that I can be 95% sure that the true regression coefficient was between -0.011 and 0.018 for internet/mobile and between -0.256 and -0.145 for call centre/mobile.

4.7.6 Bank 6

Examining the impact of gender on customers at Bank 6 led to some quite stark results. On the one hand, men were 1.778 (exp 0.575) times more likely than women to use the internet rather than a mobile application. On the other hand, men were 0.395 (exp -0.927) times less likely than women to use a call centre rather than a mobile application. The p-values for internet/mobile and call centre/mobile were both 0, and the standard errors for both criteria were 0.009 and 0.013, respectively. For the former, the confidence interval range was from 0.557 to 0.594. For the latter, the confidence interval range was from -0.953 to -0.902.

4.7.7 Results summary for independent variable 2: gender

Table 4-12 provides an overview of the results for the independent variable 'gender' for each of the six banks.

Table 4-12: Multinomial logistic regression results for independent variable 2: gender

	Bank 1		**Bank 2**		**Bank 3**		**Bank 4**		**Bank 5**		**Bank 6**	
Variable 2: Gender	I/M*	C/M	I/M	C/M	I/M	C/M	I/M	C/M	I/M	C/M	I/M	C/M
Coefficients -relative log odds	-0.156	-0.457	-0.496	-0.106	0.363	-0.915	-0.315	-0.509	0.003	-0.200	0.575	-0.927
Coefficients -relative risk exp (coefficients)	0.854	0.633	0.608	0.899	1.438	0.400	0.729	0.600	1.003	0.818	1.778	0.395
P-value	2.00E-26	2.64E-209	2.92E-20	1.29E-02	0	0	0.00E+00	0.00E+00	0.622	1.32E-12	0.00E+00	0.00E+00
Standard Error	0.014	0.014	0.053	0.042	0.004	0.015	0.003	0.005	0.007	0.028	0.009	0.013
95% confidence limits for the coefficients - lower limit	-0.185	-0.486	-0.601	-0.190	0.354	-0.946	-0.321	-0.521	-0.011	-0.256	0.557	-0.953
95% confidence limits for the coefficients – upper limit	-0.127	-0.428	-0.390	-0.022	0.373	-0.884	-0.309	-0.498	0.018	-0.145	0.594	-0.902

** I/M = Internet/Mobile; C/M = Call Centre/Mobile*

4.7.8 Overall analysis of independent variable 2: gender

The results suggest that gender has diverse effects on the likelihood of customers to prefer the internet, call centres, or mobile applications. There are, however, several patterns that I can identify.

There is, for example, a slight difference between the gender preferences at conventional banks compared to Islamic banks. At conventional banks, men are more likely than women to prefer using a mobile application rather than the internet. At Islamic banks, men are more likely than women to prefer using the internet rather than a mobile application or a call centre. That said, the data for one of the Islamic banks (Bank 5) showed a high p-value, suggesting that gender did not have a statistically significant effect on customer preferences.

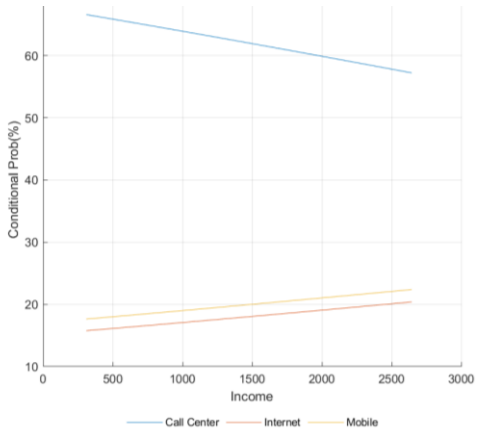
One reason for the differences between the genders may be that, as some studies have shown, women tend to be more concerned with social norms and with other people's opinions than men (Riquelme & Rios, 2010). Men, by contrast, are more likely to make decisions based on whether they perceive a particular technology to be useful, regardless of other people's opinions. This was backed up by (Haider, Changchun, Akram, & Hussain, 2018), who argued that men engaged in Islamic banking place a high priority on perceived usefulness when considering adopting a mobile technology. Women, on the other hand, are more likely to adopt accepted standards of behaviour, making mobile banking more palatable for them. These reasons may explain why men prefer to use the internet rather than mobile technology.

For Islamic banks, my results corroborate the results of previous studies. Women in these banks tend to use mobile banking more than men because men take longer to assess and adopt the mobile technology. For conventional banks, however, my findings contradict those of previous studies. I found that men were more likely than women to use mobiles and call centres rather than the internet.

Given my experience of banking, I supposed that if I considered customers' income, then this might affect the outcomes. Usually, clients with greater income, regardless of their gender, have the privilege of using more advanced facilities for free. This will result in more transactions like transfers and payments. This can shift customers' preferences regarding the adoption of certain electronic channels. This is why I studied gender in parallel with income before addressing income as an independent variable. I found that as their income increases, men

are more likely to prefer using the internet rather than their mobile (see Figure 4-7). For women, on the other hand, the results remain the same.

Sample 1: Female and Income Level



Sample 2: Male and Income Level

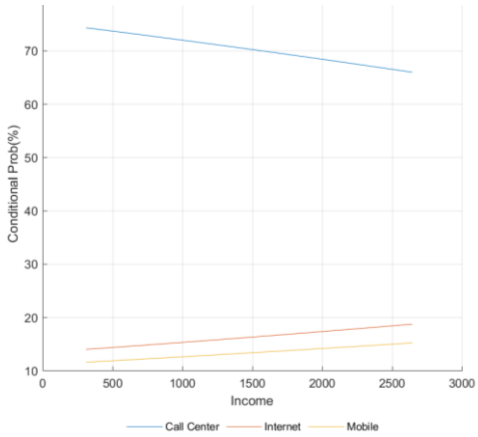


Figure 4-7: Probability of using different electronic banking channels dependent on gender and income

4.8 Independent variable 3: number of current accounts

The third variable I considered was the number of current accounts that a person had. Keeping all other variables constant, I assessed the impact of the number of current accounts on a person's likelihood to use the internet, call centres, or their mobile to communicate with their bank.

4.8.1 Bank 1

In analysing the effect of the number of current accounts on customers at Bank 1, I found that for every additional current account, a customer was 0.378 times (exp -0.971) as likely to use the internet rather than a mobile application. They

were also 0.326 (exp -1.118) times less likely to use a call centre than a mobile application. The p-values for Bank 1 were 0, and the standard errors were 0.008 for internet/mobile and 0.010 for call centre/mobile, suggesting that the number of current accounts had a statistically significant effect. The results of the confidence interval mean that I can be 95% sure that the true regression coefficient was between -0.989 and -0.953 for internet/mobile and between -1.138 and -1.098 for call centre/mobile.

4.8.2 Bank 2

The results for Bank 2 indicate that customers were 0.879 (exp -0.128) times less likely to use the internet than a mobile application for each additional current account that they had. They were also 0.346 (exp -0.061) times less likely to use a call centre than a mobile application for each additional current account. The p-values at Bank 2 were close to 0 for internet/mobile and call centre/mobile, and the standard errors were 0.043 and 0.038, respectively. The confidence intervals ranged from -0.214 to -0.042 for internet/mobile and from -1.136 to -0.985 for centre/mobile.

4.8.3 Bank 3

The results for Bank 3 were similar. For each additional current account, customers were 0.876 (exp -0.131) times less likely to use the internet than a mobile application and 0.772 (exp -0.2581) times less likely to use a call centre than a mobile application. Again, the p-values were close to 0, and the standard errors were 0.006 for internet/mobile and 0.022 for call centre/mobile. For internet/mobile, the confidence interval range was from -0.143 to -0.119. For call centre/mobile, it was from -0.301 to -0.214.

4.8.4 Bank 4

Customers at Bank 4 were 0.930 (exp -0.072) times less likely to use the internet than a mobile application for each additional current account. They were also 0.912 (exp -0.091) times less likely to use a call centre than a mobile application. Again, the p-values were close to 0, suggesting that the number of current

accounts had a statistically significant impact on customers at this bank. The standard errors for internet/mobile and call centre/mobile were 0.002 and 0.006, respectively, and the interval ranges were from -0.077 to -0.066 and from -0.103 and -0.080, respectively.

4.8.5 Bank 5

Similarly, at Bank 5, customers with more current accounts were less likely to use the internet than a mobile application (0.925 times, exp -0.077) and less likely to use a call centre than a mobile application (0.609 times, exp -0.494). The p-values were close to 0 for internet/mobile and call centre/mobile, and the standard errors were 0.009 and 0.038 respectively. in case of the dependent variable internet vs mobile and call centre verses mobile, respectively. Based on the confidence intervals, I can be 95% certain that the true regression coefficient was between the range of -0.095 and – 0.058 for internet/mobile and between -0.569 and -0.419 for call centre/mobile.

4.8.6 Bank 6

The sixth bank I investigated showed slightly different results from the other five. On the one hand, customers at Bank 6 were 1.043 (exp 0.042) times more likely to use the internet than a mobile application for each additional current account, all other factors being equal. On the other hand, they were exactly as likely to use a call centre as they were to use a mobile application, even as their number of current accounts increased. That is, they were 1.000 (exp 0.000) times more likely to use a call centre for each unit increase in the number of current accounts. The p-value for internet/mobile was close to 0, suggesting that this was statistically significant. However, the p-value for call centre/internet was 0.9, suggesting that the number of current accounts did not have a statistically significant impact on customers' decision to use a call centre or a mobile application. Both internet/mobile and call centre/mobile had low standard errors of 0.001 and 0.004, respectively. The confidence intervals ranged from 0.038 to 0.046 for internet/mobile and from -0.008 and 0.009 for call centre/mobile.

4.8.7 Results summary for independent variable 3: number of current accounts

Table 4-13 shows the impact that the number of current accounts had on the customers at each of the different banks.

Table 4-13: Multinomial logistic regression results for independent variable 3: number of current accounts

	Bank 1		**Bank 2**		**Bank 3**		**Bank 4**		**Bank 5**		**Bank 6**	
Variable 3: Current accounts	I/M*	C/M	I/M	C/M	I/M	C/M	I/M	C/M	I/M	C/M	I/M	C/M
Coefficients-relative log odds	-0.971	-1.118	-0.128	-1.061	-0.131	-0.258	-0.072	-0.091	-0.077	-0.494	0.042	0.000
Coefficients-relative risk exp (coefficients)	0.378	0.326	0.879	0.346	0.876	0.772	0.930	0.912	0.925	0.609	1.043	1.000
P-value	0	0	0.003	4.09E-167	6.27E-105	1.42E-31	4.64E-145	6.07E-52	3.07E-16	1.97E-38	1.51E-103	9.00E-01
Standard Error	0.008	0.010	0.043	0.038	0.006	0.022	0.002	0.006	0.009	0.038	0.001	0.004
95% confidence limits for the coefficients - lower limit	-0.989	-1.138	-0.214	-1.136	-0.143	-0.301	-0.077	-0.103	-0.095	-0.569	0.038	-0.008
95% confidence limits for the coefficients - upper limit	-0.953	-1.098	-0.042	-0.985	-0.119	-0.214	-0.066	-0.080	-0.058	-0.419	0.046	0.009

** I/M = Internet/Mobile; C/M = Call Centre/Mobile*

4.8.8 Overall analysis of independent variable 3: number of current accounts

The results suggest that, in general, customers with more current accounts are more likely to use their mobile to contact their bank. Their second preference is to use a call centre, and their least preferred option is to use the internet.

Only at one of the Islamic banks, Bank 6, was this variable shown not to be statistically significant. The number of current accounts did not influence the likelihood of a customer to use a call centre rather than a mobile application. Moreover, in contrast to all the other banks, customers at this bank were more likely to use the internet rather than a mobile application.

4.9 Independent variable 4: number of savings accounts

The fourth variable I considered was the total number of savings accounts that each individual customer had. As before, the following sections provide a summary of the results for this variable for each of the different banks.

4.9.1 Bank 1

The results for Bank 1 indicate that customers were 0.698 (exp -0.359) times less likely to use the internet than a mobile application for each additional savings account that they had. They were also 0.677 (exp -0.388) times less likely to use a call centre than a mobile application for each additional savings account. The p-values for internet/mobile and call centre/mobile were 0, and the standard error was 0.007 for both. The results of the confidence interval mean that I can be 95% sure that the true regression coefficient was between -0.373 and -0.345 for internet/mobile and between -0.403 and -0.373 for call centre/mobile.

4.9.2 Bank 2

Customers at Bank 2 were 0.151 (exp -1.88) times less likely to use the internet than a mobile application for each additional savings account. The same customers were 0.116 (exp -2.151) times less likely to use a call centre than a mobile application. Again, the p-values were 0, and the standard errors were

0.126 for internet/mobile and 0.110 for call centre/mobile. The 95% confidence interval for the former was between -2.138 and -1.640; for the latter, it was between -2.367 and -1.934.

4.9.3 Bank 3

There were similar results for the customers in Bank 3. For each additional savings account, these customers were 0.780 (exp -0.247) times less likely to use the internet than a mobile application and 0.758 (exp -0.276) times less likely to use a call centre. The p-value for internet/mobile was 0, and it was vanishingly close to 0 for call centre/mobile. The stand errors for both variables were 0.002 and 0.008, respectively. in case of the dependent variable internet vs mobile and call centre verses mobile, respectively. The confidence intervals were between -0.251 and – 0.243 for internet/mobile and between -0.293 and -0.259 for call centre/mobile.

4.9.4 Bank 4

The number of savings accounts had a similar effect on the customers at Bank 4. They were 0.781 (exp -0.245) times less likely to use the internet than a mobile application and 0.764 (exp -0.268) times less likely to use a call centre than a mobile application for each additional savings account. The p-values were 0, meaning that the number of savings accounts was statistically significant. The standard errors were also very low, equivalent to 2.15E-03 for internet/mobile and 3.10E-03 for call centre/mobile. The results of the confidence interval mean that I can be 95% sure that the true regression coefficient was between -0.248 and -0.243 for internet/mobile and between -0.273 and -0.262 for call centre/mobile.

4.9.5 Bank 5

Bank 5 customers were 0.906 (exp -0.098) times less likely to use the internet than a mobile application for each additional savings account. They were 0.614 (exp -0.486) times less likely to use a call centre than a mobile application for each additional savings account. Again, the p-values were very close to 0 for internet/mobile and call centre/mobile, and the standard errors were 0.003 and

0.017, respectively. For internet/mobile, the confidence interval range was from -0.105 to -0.091. For call centre/mobile, it was from 0.520 to -0.453.

4.9.6 Bank 6

As with the number of current accounts, the impact of the number of savings accounts was slightly different for the customers at Bank 6. For each additional savings account, customers at Bank 6 were 1.009 (exp 0.009) times more likely to use the internet than a mobile application and 1.002 (exp 0.002) times more likely to use a call centre than a mobile application. There was a small p-value for internet/mobile, indicating that the number of current accounts had a significant impact on customer's choice between these two options. However, a high p-value of 0.48 for call centre/mobile indicates that the number of savings accounts did not have a statistically significant impact on this choice. The standard errors for internet/mobile and call centre/mobile were 0.001 and 0.003, respectively. The results of the confidence interval mean that I can be 95% sure that the true regression coefficient was between 0.005 and 0.013 for internet/mobile and between -0.004 and 0.009 for call centre/mobile.

4.9.7 Results summary for independent variable 4: number of savings accounts

Table 4-14 provides a summary of the results for the effect of the number of savings accounts at each of the six banks.

Table 4-14: Multinomial logistic regression results for independent variable 4: number of savings accounts

	Bank 1		**Bank 2**		**Bank 3**		**Bank 4**		**Bank 5**		**Bank 6**	
Variable 4: Savings accounts	I/M*	C/M	I/M	C/M	I/M	C/M	I/M	C/M	I/M	C/M	I/M	C/M
Coefficients-relative log odds	-0.359	-0.388	-1.889	-2.151	-0.247	-0.276	-0.245	-0.268	-0.098	-0.486	0.009	0.002

	Bank 1		Bank 2		Bank 3		Bank 4		Bank 5		Bank 6	
Coefficients-relative risk exp (coefficients)	0.698	0.677	0.1511	0.116	0.780	0.758	0.781	0.764	0.906	0.614	1.009	1.002
P-value	0	0	4.42E-50	1.66E-84	0	1.86E-218	0.00E+00	0.00E+00	7.89E-153	5.07E-178	4.90E-07	0.481
Standard Error	0.007	0.007	0.126	0.110	0.002	0.008	2.15E-03	3.10E-03	0.003	0.017	0.001	0.003
95% confidence limits for the coefficients - lower limit	-0.373	-0.403	-2.138	-2.367	-0.251	-0.293	-0.248	-0.273	-0.105	-0.520	0.005	-0.004
95% confidence limits for the coefficients - upper limit	-0.345	-0.373	-1.640	-1.934	-0.243	-0.259	-0.243	-0.262	-0.091	-0.453	0.013	0.009

** I/M = Internet/Mobile; C/M = Call Centre/Mobile*

4.9.8 Overall analysis of independent variable 4: number of savings accounts

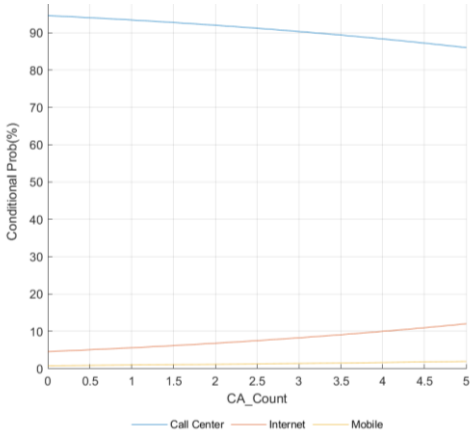
Overall, the results show that people with more savings accounts are more likely to use a mobile application to contact their bank, followed by a call centre and finally the internet.

This variable was statistically significant for all but one of the banks, Bank 6, where it was insignificant for call centre/mobile. Nevertheless, the number of savings accounts was still significant for internet/mobile at this bank.

4.9.9 Comparative analysis of current accounts and savings accounts

In comparing the impact of the number of current accounts and the number of savings accounts, neither one has a greater influence on a customer's decision to use one electronic channel. In both cases, people with more current accounts and savings accounts are more likely to use a mobile application. Their second preferred option is a call centre, and their least preferred option is the internet. From the graphs in Figure 4-8, I can see a noticeable decline in call centre usage as the number of accounts increases.

Sample 1: Current Account



Sample 2: Savings Account

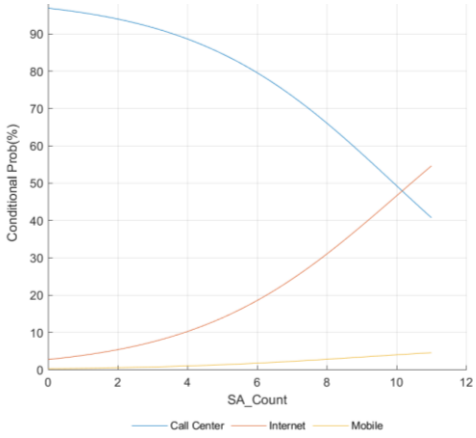
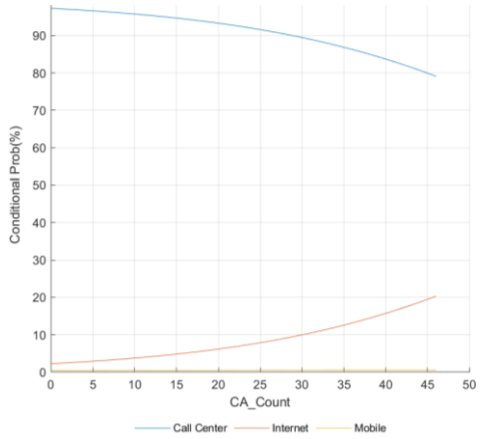


Figure 4-8: Probability of using different electronic banking channels dependent on number of current accounts and savings accounts

At one of the Islamic banks (Bank 6), the number of accounts did not have a significant effect on customers' preferences call centres over mobiles. However, the number of accounts did influence their preferences for using the internet rather than a mobile application.

Sample 1: Current Account



Sample 2: Savings Account

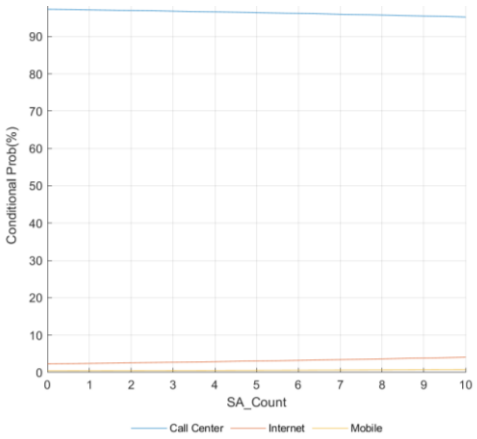


Figure 4-9: Probability of using different electronic banking channels dependent on number of current accounts and savings accounts in one Islamic bank

4.10 Independent variable 5: number of deposit accounts

The next variable I examined was a customer's number of deposit accounts. The following sections examine whether this variable had a statistically significant impact on customers' decisions at each of the different banks.

4.10.1 Bank 1

The results for Bank 1 show that the relative risk of a customer choosing the internet over a mobile application increased by 1.864 (exp 0.623) for each additional deposit account. The relative risk of a customer using a call centre rather than a mobile application increased by 1.658 (exp 0.505) for each unit increase in the number of deposit accounts. The p-values for internet/mobile and call centre/mobile were both very close to 0, and the standard errors were 0.017 and 0.019, respectively. The confidence intervals were between 0.588 and 0.658 for internet/mobile and between 0.466 and 0.544 for call centre/mobile.

4.10.2 Bank 2

The number of deposit accounts had a different effect on the customers at Bank 2. These customers were 0.4245 (exp -0.856) times less likely to use the internet than a mobile application for every additional deposit account. They were also 0.092 (exp -2.383) times less likely to use a call centre than a mobile application. Again, p-values close to 0 and standard errors of 0.366 and 0.320 for internet/mobile and call centre/mobile, respectively, indicate that this variable was statistically significant. The results of the confidence interval mean that I can be 95% sure that the true regression coefficient was between -1.574 and -0.139 for internet/mobile and between -3.010 and -1.756 for call centre/mobile.

4.10.3 Bank 3

The results for Bank 3 show that, for each unit increase in the number of deposit accounts, the relative risk of a customer choosing the internet over a mobile application increased by 1.040 (exp 0.0395), and the relative risk of a customer using a call centre rather than a mobile application increased by 1.057 (exp 0.055). The p-values were close to zero for both internet/mobile and call centre/mobile, and the standard errors were 0.003 and 0.010, respectively. The 95% confidence interval for the former shows that the true regression coefficient was in the range of 0.033 and 0.045. For the latter, it was in the range of 0.034 and 0.076.

4.10.4 Bank 4

Similarly, the results for Bank 4 show that customers are 1.074 (exp 0.072) times more likely to use the internet than a mobile application for every unit increase in the number of deposit accounts. Also, they are 1.168 (exp 0.155) times more likely to use a call centre than a mobile application for the same increase in the number of deposit accounts. P-values were close to 0 for internet/mobile and zero for call centre/mobile, and the standard errors were 0.001 and 0.002, respectively. Based on the results of the confidence interval, I can be 95% sure that the true regression coefficient for the number of deposit accounts was between 0.067 and 0.076 for internet/mobile and between 0.1492 and 0.161 for call centre/mobile.

4.10.5 Bank 5

Customers at Bank 5 were 1.003 (exp 0.003) times more likely to use the internet than a mobile application but 0.964 (exp -0.036) times less likely to use a call centre than a mobile application for every unit increase in the number of deposit accounts. However, there were high p-values for both of these results (0.46 and 0.12, respectively), suggesting that this variable was not statistically significant. The standard errors were 0.004 for internet/mobile and 0.023 for call centre/mobile, and the confidence interval ranges were from -0.005 to 0.012 and from -0.081 to 0.009, respectively.

4.10.6 Bank 6

In the sixth bank I analysed, customers were 1.026 (exp 0.026) times more likely to use the internet than a mobile application for every unit increase in the number of deposit accounts. They were also 1.014 (exp 0.014) times more likely to use a call centre than a mobile application. The p-values for internet/mobile and call centre/mobile were close to 0, and the standard errors were both 0.001 and 0.001 in case of the dependent variable internet vs mobile and call centre verses mobile, respectively. The confidence intervals ranged from 0.023 to 0.028 for internet/mobile and from 0.010 to 0.017 for call centre/mobile.

4.10.7 Results summary for independent variable 5: number of deposit accounts

Table 4-15 provides an overview of the results for variable 5: number of deposit accounts.

Table 4-15: Multinomial logistic regression results for independent variable 5: number of deposit accounts

	Bank 1		**Bank 2**		**Bank 3**		**Bank 4**		**Bank 5**		**Bank 6**	
Variable 5: Deposit accounts	I/M*	C/M	I/M	C/M	I/M	C/M	I/M	C/M	I/M	C/M	I/M	C/M
Coefficients-relative log odds	0.623	0.505	-0.856	-2.383	0.039	0.0555	0.072	0.155	0.003	-0.0036	0.026	0.014

	Bank 1		Bank 2		Bank 3		Bank 4		Bank 5		Bank 6	
Coefficients-relative risk exp (coefficients)	1.864	1.658	0.424	0.092	1.040	1.057	1.074	1.168	1.003	0.964	1.026	1.014
P-value	1.63E-265	1.21E-142	1.93E-02	9.67E-14	2.84E-35	2.82E-07	4.04E-246	0	4.61E-01	1.19E-01	7.40E-117	2.93E-15
Standard Error	0.017	0.019	0.366	0.320	0.003	0.010	0.001	0.002	0.004	0.023	0.001	0.001
95% confidence limits for the coefficients - lower limit	0.588	0.466	-1.574	-3.010	0.033	0.034	0.067	0.1492	-0.005	-0.081	0.023	0.010
95% confidence limits for the coefficients - upper limit	0.658	0.544	-0.139	-1.756	0.045	0.076	0.076	0.161	0.012	0.009	0.028	0.017

** I/M = Internet/Mobile; C/M = Call Centre/Mobile*

4.10.8 Overall analysis of independent variable 5: number of deposit accounts

Overall, the results show that customers with more than one deposit account tend to use the internet and call centres slightly more frequently than their mobile. The data also shows that as the number of deposit accounts increases, customers are less likely to use call centres and more likely to use the internet (see Figure 4-10).

Deposit account

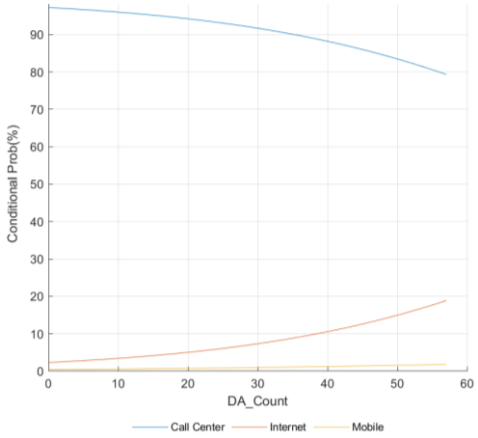


Figure 4-10: Probability of using different electronic banking channels dependent on deposit accounts

At one of the conventional banks (Bank 2), mobile usage increased as the number of deposit accounts increased (compared to both call centres and the internet). However, the increase was only very slight. On the other hand, the results for one of the Islamic banks (Bank 5) showed that the number of deposit accounts did not have a significant effect on this bank's customers' decisions to use the internet, call centres, or mobiles (as evidenced by the high p-values in these results).

4.11 Independent variable 6: income

The next variable I assessed was income. Customer income was based on their monthly salary in Kuwaiti Dinars. The results outlined in the following sections demonstrate how increases in income affected customers' decisions to communicate with their banks via the internet, a call centre, or their mobile.

4.11.1 Bank 1

The results for Bank 1 show that the relative risk of a customer using the internet rather than a mobile application increased by an order of 1.00 (exp 4.38E-05) for every additional unit of income (1 Kuwaiti Dinar). Similarly, the relative risk of a

customer using a call centre rather than a mobile application increased by 1.00 (exp3.58E-05) for every unit increase in income. The p-values for internet/mobile and call centre/mobile were very close to zero, indicating that income was statistically significant in influencing people's choice for both variables. The results also show low standard errors for income: 1.97E-06 for internet/mobile and 2.16E-06 for call centre/mobile. The results of the confidence interval show that I can be 95% certain that the true regression coefficient for income was between 4.00E-05 and 4.77E-05 for internet/mobile and between 3.15E-05 and 4.00E-05 for call centre/mobile.

4.11.2 Bank 2

In analysing Bank 2, I found that the relative risk of customers using the internet compared to a mobile application decreased by 0.999 (exp -4.15E-06) for every additional unit of income. This was similar for the relative risk of customers using their mobile rather than a call centre: for every unit increase in income, the relative risk of using a call centre rather than a mobile application decreased by a value of 0.998 (exp -1.20E-03). However, the result for internet/mobile had a high p-value (0.76), so it was not statistically significant. Call centre/mobile, though, had a p-value of zero, indicating that income had a significant effect on customers' decision to use a call centre rather than a mobile application. The standard errors were 1.35E-05 and 2.50E-05 for internet/mobile and call centre/mobile, respectively. The confidence intervals ranged from -3.06E-05 to 2.23E-05 for internet/mobile and from -0.001 to -0.001 for call centre/mobile.

4.11.3 Bank 3

The results for Bank 3 show that the relative risk of customers using the internet rather than a mobile application decreased by 0.999 (exp -1.12E-05) for every unit increase in income. Similarly, the relative risk of customers using a call centre rather than a mobile application decreased by 0.999 (exp -8.52E-05) for the same increase in income. The p-values for both internet/mobile and call centre/mobile were close to 0, and the standard errors were 1.05E-06 and 4.17E-06, respectively. The 95% confidence interval for the former shows that the true

regression coefficient was in the range of -1.33E-05 and -9.19E-06. For the latter, it was in the range of -9.33E-05 and-7.70E-05.

4.11.4 Bank 4

The relative risk of customers at Bank 4 using the internet rather than a mobile application increased by 1.000 (exp 1.37E-05) for every unit increase in income. The relative risk of customers using a call centre rather than a mobile application decreased by 0.999 (exp -0.000) for each additional unit of income. Internet/mobile had a p-value close to 0, and call centre/mobile had a p-value of 0. The standard errors for both options were 5.16E-07 and 2.65E-06, respectively, and the confidence intervals ranged from 1.27E-05 to 1.47E-05 and from -0.000 to -0.000.

4.11.5 Bank 5

The results for customers at Bank 5 were very similar. Customers were 0.999 (exp -1.62E-05) times less likely to use the internet than a mobile application and 0.999 (exp -3.48E-05) times less likely to use a call centre than a mobile application for each additional unit of income. The p-value for internet/mobile was close to 0, and the p-value for call centre/mobile was 0. The standard errors were 2.04E-06 for internet/mobile and 9.20E-06 for call centre/mobile. Based on the results of the confidence interval, I can be 95% sure that the true regression coefficient was between -2.02E-05 and -1.22E-05 for internet/mobile and between -5.29E-05 and -1.68E-05 for call centre/mobile.

4.11.6 Bank 6

For customers in Bank 6, the relative risk of choosing the internet rather than a mobile application increased by 1.000 (exp 4.92E-06) with every unit increase in income, and the relative risk of using a call centre rather than a mobile application decreased by 0.999 (exp -0.000275499) with every unit increase in income. The p-values were 0 for both variables. The standard errors were 1.35E-06 and 6.37E-06 for internet/mobile and call centre/mobile, respectively. The confidence interval

for the former ranged from 2.27E-06 and 7.58E-06 and from -0.000 to -0.000 for the latter.

4.11.7 Results summary for independent variable 6: income

Table 4-16 provides a summary of the effect of income on the customers at every bank.

Table 4-16: Multinomial logistic regression results for independent variable 6: income

	Bank 1		**Bank 2**		**Bank 3**		**Bank 4**		**Bank 5**		**Bank 6**	
Variable 6: Income	I/M*	C/M	I/M	C/M	I/M	C/M	I/M	C/M	I/M	C/M	I/M	C/M
Coefficients-relative log odds	4.38E-05	3.58E-05	-4.15E-06	-1.20E-03	-1.12E-05	-8.52E-05	1.37E-05	-0.000	-1.62E-05	-3.48E-05	4.92E-06	-0.000
Coefficients-relative risk exp (coefficients)	1.000	1.000	0.999	0.998	0.999	0.999	1.000	0.999	0.999	0.999	1.000	0.999
P-value	3.16E-110	9.32E-62	7.58E-01	0.00E+00	6.64E-27	1.17E-92	5.73E-156	0	1.79E-15	0.000	0.000	0
Standard Error	1.97E-06	2.16E-06	1.35E-05	2.50E-05	1.05E-06	4.17E-06	5.16E-07	2.65E-06	2.04E-06	9.20E-06	1.35E-06	6.37E-06
95% confidence limits for the coefficients - lower limit	4.00E-05	3.15E-05	-3.06E-05	-0.001	-1.33E-05	-9.33E-05	1.27E-05	-0.000	-2.02E-05	-5.29E-05	2.27E-06	-0.000
95% confidence limits for the coefficients - upper limit	4.77E-05	4.00E-05	2.23E-05	-0.001	-9.19E-06	-7.70E-05	1.47E-05	-0.000	-1.22E-05	-1.68E-05	7.58E-06	-0.000

** I/M = Internet/Mobile; C/M = Call Centre/Mobile*

4.11.8 Overall analysis for independent variable 6: income

Across all the banks, the results suggest that, in general, customers with higher incomes were more likely to use the internet or their mobile rather than a call centre (see Figure 4-11). However, there were some exceptions to this. One conventional bank (Bank 2) showed a weak correlation between customer income level and the choice between internet or mobile banking (as indicated by the high p-value). Nevertheless, the results did show that customers with higher incomes at this bank preferred to use their mobile rather than a call centre.

At Bank 4 and Bank 6, which considered as large banks (based on their share of the total assets of the banking sector), the customers with higher incomes preferred to use the internet rather than their mobile, but they also preferred to use their mobile rather than a call centre. Previous studies had found that high-income consumers were able to afford the latest technologies and the best internet connections, thereby increasing their uptake of new technologies and mobile applications (Chawla & Joshi, 2017).

Customers with higher incomes at two of the Islamic banks (Bank 3 and Bank 5) preferred to use their mobile rather than the internet or a call centre.

Finally, for one of the conventional banks (Bank 1), customers with a higher income were more likely to use the internet or a call centre than their mobile.

Sample 1: Conventional Bank

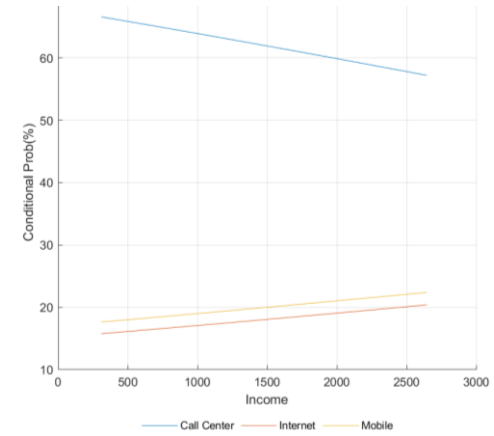


Figure 4-11: Probability of using different electronic banking channels dependent on income

4.12 Independent variable 7: number of consumer loans

I also examined the influence of the number of consumer loans on customers' digital banking preferences. A consumer loan is a medium-term personal loan given to a customer their personal finances, to pay for durable goods, educational expenses, medical expenses, or similar. Consumer loans are repayable in monthly instalments within a period of no more than 5 years (Central Bank of Kuwait, 1996).

As before, I provide a summary of the results for each of the different banks, followed by an overall consideration of the patterns across all the banks.

4.12.1 Bank 1

The results for Bank 1 show that for every additional consumer loan that a customer had taken out, the customer was 1.126 (exp 0.119) times more likely to use the internet rather than a mobile application. Those same customers were also 1.205 (exp 0.186) times more likely to use a call centre rather than a mobile application. The p-values for internet/mobile and call centre/mobile were close to 0, and the standard error for both was 0.006. The results of the confidence interval show that I can be 95% certain that the true regression coefficient was between

0.107 and 0.131 for internet/mobile and between 0.174 and 0.198 for call centre/mobile.

4.12.2 Bank 2

The results for Bank 2 were slightly different from Bank 1. At Bank 2, the relative risk of customers using the internet rather than a mobile application decreased by 0.998 (exp -0.001) for each additional consumer loan they had taken out. Also, the relative risk of them using a call centre rather than a mobile application decreased by 0.077 (exp -2.55) for every additional consumer loan. However, there was a high p-value (0.98) for internet/mobile, suggesting that the number of consumer loans did not have a statistically significant impact on the choice between the internet and a mobile application. There was a small p-value for call centre/mobile, suggesting that the number of consumer loans was statistically significant for this variable. The standard errors for internet/mobile and call centre/mobile were 0.039 and 0.045, respectively. The confidence interval ranged from -0.077 to 0.075 for the former and from -2.641 to -2.462 for the latter.

4.12.3 Bank 3

Customers at Bank 3 were 0.962 (exp -0.037) times less likely to use the internet rather than a mobile application for each additional consumer loan. By contrast, they were 1.235 (exp 0.211) times more likely to use a call centre rather than a mobile application for each additional consumer loan. The p-values for both internet/mobile and call centre/mobile were close to 0, and the standard errors were 0.004 and 0.015, respectively. Based on the results of the confidence interval, I can be 95% sure that the true regression coefficient was between -0.046 and -0.029 for internet/mobile and between 0.181 and 0.242 for call centre/mobile.

4.12.4 Bank 4

For customers at Bank 4, the relative risk of using the internet rather than a mobile application decreased by 0.991 (exp -0.008) with every additional consumer loan that they took out. The relative risk of using a call centre rather than a mobile application also decreased with each additional consumer loan, in this case by an

order of 0.945 (exp -0.056). The p-value for internet/mobile was 0, and the p-value for call centre/mobile was close to 0. The standard errors for each of them were 0.001 and 0.003, respectively, and the confidence intervals ranged from -0.012 to -0.005 for the former and from -0.063 to -0.048 for the latter.

4.12.5 Bank 5

Our analysis of Bank 5 showed that customers were 0.994 (exp -0.005) times less likely to use the internet rather than a mobile application for each additional consumer loan. By contrast, they were 1.04 (exp 0.045) times more likely to use a call centre than a mobile application with every extra consumer loan. That said, the p-value for internet/mobile was slightly high (0.07), suggesting that there was only a weak correlation between additional consumer loans and the choice between the internet or a mobile application. The p-value for call centre/mobile, by contrast, was 0, indicating statistical significance. The standard errors were 0.003 and 0.014 for internet/mobile and call centre/mobile respectively, and the results of the confidence interval mean that I can be 95% certain that the true regression coefficient was within the range of -0.011 and 0.000 for internet/mobile and within the range of 0.017 and 0.072 for call centre/mobile.

4.12.6 Bank 6

Finally, the results for Bank 6 show that the relative risk of a customer using the internet rather than a mobile application decreased by 0.910 (exp -0.093) with every additional consumer loan. The relative risk of using a call centre rather than a mobile application, however, increased by an order of 1.019 (exp 0.019). The p-values for both internet/mobile and call centre/mobile were 0, and the standard errors were 0.005 and 0.009, respectively. The 95% confidence interval for the former shows that the true regression coefficient was in the range of -0.105 and -0.082. For the latter, it was in the range of 0.001 and 0.037.

4.12.7 Results summary for independent variable 7: number of consumer loans

Table 4-17 outlines all the results of the multinomial logistic regression with the number of consumer loans as the independent variable.

Table 4-17: Multinomial logistic regression results for independent variable 7: number of consumer loans

	Bank 1		**Bank 2**		**Bank 3**		**Bank 4**		**Bank 5**		**Bank 6**	
Variable 7: Consumer loans	I/M*	C/M	I/M	C/M	I/M	C/M	I/M	C/M	I/M	C/M	I/M	C/M
Coefficients-relative log odds	4.38E-05	3.58E-05	-4.15E-06	-1.20E-03	-1.12E-05	-8.52E-05	1.37E-05	-0.000	-1.62E-05	-3.48E-05	4.92E-06	-0.000
Coefficients-relative risk exp (coefficients)	1.000	1.000	0.999	0.998	0.999	0.999	1.000	0.999	0.999	0.999	1.000	0.999
P-value	3.16E-110	9.32E-62	7.58E-01	0.00E+00	6.64E-27	1.17E-92	5.73E-156	0	1.79E-15	0.000	0.000	0
Standard Error	1.97E-06	2.16E-06	1.35E-05	2.50E-05	1.05E-06	4.17E-06	5.16E-07	2.65E-06	2.04E-06	9.20E-06	1.35E-06	6.37E-06
95% confidence limits for the coefficients - lower limit	4.00E-05	3.15E-05	-3.06E-05	-0.001	-1.33E-05	-9.33E-05	1.27E-05	-0.000	-2.02E-05	-5.29E-05	2.27E-06	-0.000
95% confidence limits for the coefficients - upper limit	4.77E-05	4.00E-05	2.23E-05	-0.001	-9.19E-06	-7.70E-05	1.47E-05	-0.000	-1.22E-05	-1.68E-05	7.58E-06	-0.000

** I/M = Internet/Mobile; C/M = Call Centre/Mobile*

4.12.8 Overall analysis for independent variable 7: number of consumer loans

All three of the Islamic banks (Bank 3, Bank 5 and Bank 6) had customers who were more likely to use a call centre rather than the internet or a mobile application if they had taken out more consumer loans.

One conventional bank (Bank 1) also demonstrated a similar pattern. Its customers were more likely to use a call centre if they had taken out more consumer loans. For these customers, their second preference was the internet, followed by using a mobile application.

The other conventional banks (Bank 2 and Bank 4) had customers who were more likely to use a mobile application if they had taken out more loans. Their second preference was to use a call centre. However, it should also be noted that the results for Bank 2 showed that the number of consumer loans did not have a statistically significant impact on whether a customer used the internet or a mobile application.

4.13 Independent variable 8: number of instalment loans

The next variable I looked at was related to the previous one: the number of instalment loans. An instalment or housing loan is a long-term personal loan that is used for non-commercial purposes, such as to restoring or purchasing a private residence. Instalment loans are repaid in monthly instalments within a period of no more than 5 years (Central Bank of Kuwait, 1996). In what follows, I provide a breakdown of the different results for each of the banks. At the end of this section, I also provide a comparison of the influence both consumer loans and instalment loans on customer preferences.

4.13.1 Bank 1

The results for the first bank I looked at show that for every additional instalment loan they had taken out, a customer was 0.822 (exp -0.195) times less likely to use the internet rather than a mobile application. On the other hand, an additional instalment loan made them 1.023 (exp 0.023) times more likely to use a call centre

rather than a mobile application. The p-values for internet/mobile and call centre/mobile were 0 or close to 0, and the standard error was the same for both: 0.003. Based on the results of the confidence intervals, I can say with 95% certainty that the true regression coefficient was between of -0.203 and -0.187 for internet/mobile and between 0.016 and 0.030 for call centre/mobile.

4.13.2 Bank 2

The customers at Bank 2 showed results that were quite different from the customers at Bank 1. The relative risk of a customer at Bank 2 using the internet rather than a mobile application increased by 1.658 (exp 0.506) with every additional instalment loan. By contrast, the relative risk of a customer using a call centre rather than a mobile application decreased by 0.689 (exp -0.371) for each additional instalment loan. The p-values for both were close to 0, and the standard errors were 0.026 and 0.030, respectively. The results of the confidence intervals mean that I can be 95% sure that the true regression coefficient was between 0.454 and 0.558 for internet/mobile and between -0.431 and -0.311 for call centre/mobile.

4.13.3 Bank 3

The results for customers at Bank 3 were different yet again. The relative risk of a Bank 3 customer using the internet rather than a mobile application decreased by 0.958 (exp -0.042) with every additional instalment loan. The relative risk of a customer using a call centre rather than a mobile application increased by 1.455 (exp 0.375) with each extra instalment loan. The p-value for internet/mobile was close to 0, and the p-value for call centre/mobile was 0. The standard errors for each of these were 0.002 and 0.007, respectively. The confidence interval ranged from -0.047 to -0.037 for the former and from 0.360 to 0.390 for the latter.

4.13.4 Bank 4

Customers at Bank 4 were 0.932 (exp -0.070) times less likely to use the internet rather than a mobile application for each unit increase in instalment loans. On the other hand, they were 1.131 (exp 0.123) times more likely to use a call centre

rather than a mobile application. The p-values for both dependent variables were 0. The standard error for internet/mobile was 1.20E-03, whereas it was 1.97E-03 for call centre/mobile. The results of the confidence interval mean that I can be 95% sure that the true regression coefficient was between -0.072 and -0.067 for internet/mobile and between 0.119 and 0.126 for call centre/mobile.

4.13.5 Bank 5

The results for Bank 5 show that the relative risk of a customer using the internet rather than a mobile application increased by 1.002 (exp 0.002) with every additional instalment loan. The relative risk of a customer using a call centre rather than a mobile application also increased, in this case by 1.037 (exp 0.036) with every additional instalment loan. However, the dependent variable of internet/mobile had a high p-value (0.11), meaning that the number of instalment loans did not have a statistically significant impact on this variable. The p-value for call centre/mobile, by contrast, was close to 0, indicating statistical significance. The standard error for internet/mobile was 0.001, and the standard error for call centre/mobile was 0.005. The confidence interval ranged from 0.000 to 0.005 for the former and from 0.024 to 0.047 for the latter.

4.13.6 Bank 6

Customers at Bank 6 were 0.976 (exp -0.023) times less likely to use the internet rather than a mobile application for every additional instalment loan. They were 1.109 (exp 0.103) times more likely to use a call centre rather than a mobile application. The p-values for both variables were close to 0, and the standard errors were 0.003 and 0.005 for internet/mobile and call centre/mobile, respectively. Based on the confidence intervals, I can be 95% certain that the true regression coefficient was within the range of -0.029 and -0.016 for internet/mobile and within the range of 0.093 and 0.114 for call centre/mobile.

4.13.7 Results summary for independent variable 8: instalment loans

Table 4-18 outlines the results of the analysis with the number of instalment loans as the independent variable.

Table 4-18: Multinomial logistic regression results for independent variable 8: number of instalment loans

	Bank 1		Bank 2		Bank 3		Bank 4		Bank 5		Bank 6	
Variable 8: Instalment loans	I/M*	C/M	I/M	C/M	I/M	C/M	I/M	C/M	I/M	C/M	I/M	C/M
Coefficients-relative log odds	-0.195	0.023	0.506	-0.371	-0.042	0.375	-0.070	0.123	0.002	0.036	-0.023	0.103
Coefficients-relative risk exp (coefficients)	0.822	1.023	1.658	0.689	0.958	1.455	0.932	1.131	1.002	1.037	0.976	1.109
P-value	0	3.34E-11	9.89E-81	7.01E-34	5.24E-55	0.00E+00	0.00E+00	0.00E+00	1.14E-01	5.33E-10	2.54E-12	1.27E-84
Standard Error	0.003	0.003	0.026	0.030	0.002	0.007	1.20E-03	1.97E-03	0.001	0.005	0.003	0.005
95% confidence limits for the coefficients - lower limit	-0.203	0.016	0.454	-0.431	-0.047	0.360	-0.072	0.119	0.000	0.024	-0.029	0.093
95% confidence limits for the coefficients - upper limit	-0.187	0.030	0.558	-0.311	-0.037	0.390	-0.067	0.126	0.005	0.047	-0.016	0.114

** I/M = Internet/Mobile; C/M = Call Centre/Mobile*

4.13.8 Overall analysis for independent variable 8: number of instalment loans

The results for this variable were relatively consistent. Customers at five of the banks were more likely to use call centres than the internet or mobiles if they had more instalment loans. Their second preference was for mobiles.

At Bank 5, instalment loans did not have a significant impact on whether customers used the internet or a mobile application.

The results for Bank 2 showed that customers were more likely to use the internet if they had more instalment loans. This was followed by mobiles and then by call centres. This could be because mobile banking was only introduced at this bank in late 2016, meaning that internet banking was more common at the time of my study.

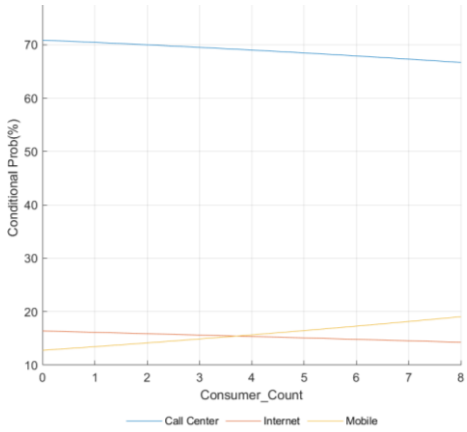
4.13.9 Comparative analysis of number of consumer loans and instalment loans

In **my** analysis, I compared two types of loans: consumer loans and instalment loans. However, the results did not suggest that one loan had more of an influence on whether a customer chose to use call centres, the internet, or their mobile. For both loans, customers tended to use call centres if they had more loans (see Figure 4-12).

This was true for both Islamic banks and conventional banks. In the Islamic banks, the more loans a customer had, the more likely they were to use call centres, followed by mobiles and then the internet.

In the conventional banks, one bank (Bank 1) showed that as the number both kinds of loans increased, customers were more likely to use call centres. Customers at the other two banks were less likely to use call centres and more likely to use mobiles as the number of loans increased.

Sample 1: Consumer Loans



Sample 2: Instalment Loans

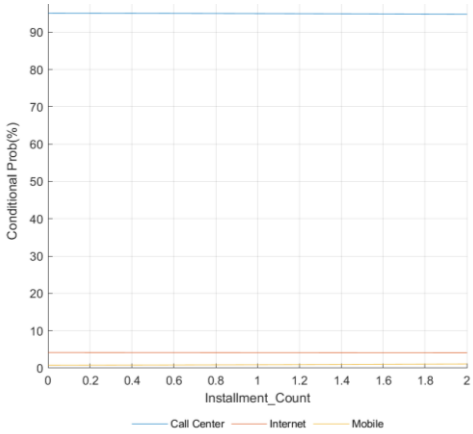


Figure 4-12: Probability of using different electronic banking channels dependent on the number of consumer loans and instalment loans

4.14 Independent variable 9: credit card limit

I also assessed whether customers' credit card limits had a significant impact on their chosen channel for communicating with their bank. The results of the analysis for each of the banks are provided in the following sections.

4.14.1 Bank 1

The results for Bank 1 show that the relative risk of a customer using the internet rather than a mobile application increased by 1.00 (exp 0.000) times for every unit increase in their credit card limit (an increase of 1 Kuwaiti Dinar). On the other hand, the relative risk of a customer using a call centre rather than a mobile application decreased by 0.99 (exp -6.84E-05) for each unit increase in their credit card limit. The p-value for internet/mobile was 0, and the p-value for call centre/mobile was close to 0, suggesting that both results were statistically significant. The standard errors were 3.59E-06 and 4.31E-06, respectively, and

the confidence intervals ranged from 0.000 to 0.000 for internet/mobile and from -7.68E-05 to -5.99E-05 for centre/mobile.

4.14.2 Bank 2

The results for customers in Bank 2 were similar, with the relative risk of a customer using their internet rather than their mobile increasing by 1.000 (exp 0.000) for each unit increase in their credit card limit. The relative risk for a customer using a call centre rather than a mobile application increased by 1.000 (exp 2.14E-04) for each unit increase in credit card limit. Again, the p-values were close to 0, indicating statistical significance, and the standard errors were 3.81E-05 for internet/mobile and 3.45E-05 for call centre/mobile. The results of the confidence interval mean that I can be 95% certain that the true regression coefficient is between 0.000 and 0.000 for internet/mobile and between 0.000 and 0.000 for centre/mobile.

4.14.3 Bank 3

Customers in Bank 3 were 0.999 (exp -2.76E-05) times less likely to use the internet rather than a mobile application for every single unit increase in their credit card limit. They were also 0.999 (exp -2.34E-05) times less likely to use a call centre rather than a mobile application. The p-values for both internet/mobile and call centre/mobile were close to 0, and the standard errors were 1.14E-06 and 4.42E-06, respectively. The confidence interval ranged from -2.98E-05 to -2.54E-05 for internet/mobile and from -3.21E-05 to -1.48E-05 for call centre/mobile.

4.14.4 Bank 4

The results for Bank 4 show that the relative risk of a customer using the internet rather than a mobile application increased by 1.00 (exp 4.35E-05) for each unit increase in the customer’s credit card limit. The relative risk of a customer using a call centre rather than a mobile decreased by 0.99984687 (exp -1.53E-04) for each unit increase in that customer’s credit card limit. The p-values for both variables were 0, and the standard errors were 6.64E-07 for internet/mobile and

2.25E-06 for call centre/mobile. The 95% confidence interval for the former shows that the true regression coefficient was in the range of 4.22E-05 and 4.48E-05. The range of the latter was from -0.000 to -0.000.

4.14.5 Bank 5

The customers for Bank 5 were 0.999 (exp -1.49E-06) times less likely to use the internet than a mobile application for each unit increase in their credit card limit. They were also 0.999 (exp -1.49E-04) times less likely to use a call centre than a mobile application. However, there was a high p-value for internet/mobile (0.62), suggesting that credit card limit did not have a statistically significant effect on this variable. The p-value for call centre/mobile was close to 0. The standard errors were 3.02E-06 and 1.39E-05for internet/mobile and call centre/mobile, respectively, and the confidence intervals ranged from -7.40E-06 to 4.43E-06 for the former and from -0.000 to -0.000 for the latter.

4.14.6 Bank 6

The results for Bank 6 show that the relative risk of a customer using the internet rather than a mobile application increased by 1.000 (exp 1.46E-05) times for each unit increase in their credit card limit. The risk of using a call centre rather than a mobile application increased by 1.000 (exp 0.000) times for each unit increase in their credit card limit. Both variables had p-values of close to 0, and the standard errors were 5.63E-06 and 8.40E-06. Based on the confidence intervals, I can be 95% certain that the true regression coefficient was within the range of 3.56E-06 to 2.56E-05 in case of dependent variable internet/mobile and 95% sure that the true regression coefficient for the variable “credit card limit” are between the range of 0.000 and 0.000 in case of call centre/mobile.

4.14.7 Results summary for independent variable 9: credit card limit

Table 4-19 provides an overview of impact of credit card limits on electronic banking channels across each of the different banks.

Table 4-19: Multinomial logistic regression results for independent variable 9: credit card limit

	Bank 1		Bank 2		Bank 3		Bank 4		Bank 5		Bank 6	
Variable 9: Credit card limit	I/M*	C/M	I/M	C/M	I/M	C/M	I/M	C/M	I/M	C/M	I/M	C/M
Coefficients-relative log odds	0.000	-6.84E-05	0.0000	2.14E-04	-2.76E-05	-2.34E-05	4.35E-05	-1.53E-04	-1.49E-06	-1.49E-04	1.46E-05	0.000
Coefficients-relative risk exp (coefficients)	1.000	0.999	1.0000	1.0000	0.9999	0.9999	1.0000	0.9999	0.9999	0.9999	1.0000	1.000
P-value	0	1.27E-56	8.52E-31	5.27E-10	3.10E-130	1.19E-07	0	0.00E+00	6.22E-01	9.49E-27	0.009	2.46E-98
Standard Error	3.59E-06	4.31E-06	3.81E-05	3.45E-05	1.14E-06	4.42E-06	6.64E-07	2.25E-06	3.02E-06	1.39E-05	5.63E-06	8.40E-06
95% confidence limits for the coefficients - lower limit	0.000	-7.68E-05	0.0000	0.0000	-2.98E-05	-3.21E-05	4.22E-05	-0.0000	-7.40E-06	-0.0000	3.56E-06	0.000
95% confidence limits for the coefficients - upper limit	0.000	-5.99E-05	0.0000	0.0000	-2.54E-05	-1.48E-05	4.48E-05	-0.0000	4.43E-06	-0.0000	2.56E-05	0.000

** I/M = Internet/Mobile; C/M = Call Centre/Mobile*

4.14.8 Overall analysis for independent variable 9: credit card limit

Customers at four out of the six banks were more likely to use the internet if they had a higher credit card limit. Their second preference was for mobiles, followed by call centres (see Figure 4-13).

Two of the Islamic banks (Bank 3 and Bank 5) had customers who were more likely to use mobiles if they had higher credit card limits. For these banks, the

likelihood of using a mobile application increased sharply in comparison with the other options.

At Bank 5, small Islamic bank, credit card limit did not have a statistically significant influence on whether a customer used the internet or a mobile application.

Consumer Loans

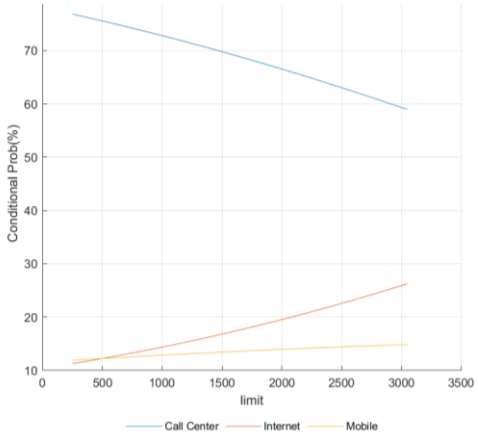


Figure 4-13: Probability of using different electronic banking channels dependent on the number of consumer loans

4.15 Independent variable 10: outstanding credit card balance

As well as looking at customers' total credit card limits, I also assessed them based on whether they had an outstanding credit card balance. The results of the analysis are detailed in each of the following sections.

4.15.1 Bank 1

Customers in Bank 1 were 0.999 (exp -0.000) times less likely to use the internet rather than a mobile application for each unit increase in their outstanding credit card balance (an increase of 1 Kuwaiti Dinar). They were 1.000 (exp 1.81E-05) times more likely to use a call centre than a mobile application for each unit increases in their outstanding credit card balance. The p-value was close to 0 for internet/mobile and 0 for centre/mobile, indicating statistical significance. The

standard errors were 4.49E-06 and 4.82E-06 respectively, and the confidence intervals ranged from -0.000 to -0.000 for the former and from 8.66E-06 to 2.76E-05 for the latter.

4.15.2 Bank 2

Similarly, customers in Bank 2 were 0.999 (exp -0.000) times less likely to use the internet rather than a mobile application for each unit increase in their outstanding credit card balance. They were also 0.999 (exp -4.45E-04) times less likely to use a call centre rather than a mobile application. The p-values were close to 0 for both internet/mobile and call centre/mobile, and the standard errors were 2.74E-05 and 2.35E-05, respectively. The results of the confidence interval mean that I can be 95% certain that the true regression coefficient was between -0.000 and -0.000 for both internet/mobile and call centre/mobile.

4.15.3 Bank 3

The results for Bank 3 show that the relative risk of a customer using the internet rather than a mobile application decreased by 0.999 (exp -1.09E-05) times for each unit increase in their outstanding credit card balance. Similarly, the risk of a customer using a call centre rather than a mobile application decreased by 0.999 (exp -1.51E-04) for each unit increase in their outstanding credit card balance. Again, the p-values for both variables were close to zero, indicating statistical significance. The standard errors were 2.59E-06 for internet/mobile and 1.03E-05 for call centre/mobile, and the confidence intervals ranged from -1.60E-05 to -5.88E-06 for the former and from -0.000 to -0.000 for the latter.

4.15.4 Bank 4

The results for Bank 4 show that the relative risk of a customer using the internet rather than a mobile application decreased by 0.999 (exp -6.69E-05) for each unit increase in their outstanding credit card balance. The relative risk of a customer using a call centre rather than a mobile application also decreased by 0.999 (exp -8.31E-05). The p-value for internet/mobile was 0, and the p-value for call centre/mobile was close to 0, indicating that outstanding credit card balance was

significant for both variables. The standard errors were 9.19E-07 and 2.58E-06, respectively. Based on the confidence intervals, I can be 95% sure that the true regression coefficient was between -6.87E-05 and -6.51E-05 for internet/mobile and between -8.82E-05 and -7.81E-05 for call centre/mobile.

4.15.5 Bank 5

Customers in Bank 5 were 1.000 (exp 1.34E-05) times more likely to use the internet than a mobile application for each unit increases in their outstanding credit card balance, but they were 0.999 (exp -0.00056736) times less likely to use a call centre than a mobile application. The p-values were close to 0 for both variables, and the standard errors were 5.42E-06 for internet/mobile and 3.04E-05 for call centre/mobile. The confidence intervals ranged from 2.72E-06 to 2.40E-05 for internet/mobile and from -0.000 to -0.000 for call centre/mobile.

4.15.6 Bank 6

The results for Bank 6 were similar, with customers being 1.000 (exp 9.97E-05) times more likely to use the internet than a mobile application for each unit increase in their outstanding credit card balance and 0.999 (exp -0.000) times less likely to use a call centre rather than a mobile application. Again, the p-values were close to 0. The standard errors were 6.39E-06 for internet/mobile and 1.01E-05 for call centre/mobile, and the confidence intervals ranged from 8.71E-05 to 0.000 for the former and from -0.000 to -0.000 for the latter.

4.15.7 Results summary for independent variable 10: outstanding credit card balance

Table 4-20 provides a summary of the results for the tenth variable I analyzed: outstanding credit card balance.

Table 4-20: Multinomial logistic regression results for independent variable 10: outstanding credit card balance

	Bank 1		Bank 2		Bank 3		Bank 4		Bank 5		Bank 6	
Variable 10: Outstanding credit card balance	I/M*	C/M	I/M	C/M	I/M	C/M	I/M	C/M	I/M	C/M	I/M	C/M
Coefficients-relative log odds	-0.000	1.81E-05	-0.000	-4.45E-04	-1.09E-05	-1.51E-04	-6.69E-05	-8.31E-05	1.34E-05	-0.000	9.97E-05	-0.000
Coefficients-relative risk exp (coefficients)	0.999	1.000	0.999	0.999	0.999	0.999	0.999	0.999	1.000	0.999	1.000	0.999
P-value	2.00E-152	0.000	3.04E-92	4.09E-80	2.31E-05	4.09E-48	0.00E+00	4.17E-228	1.38E-02	1.21E-77	8.76E-55	3.87E-259
Standard Error	4.49E-06	4.82E-06	2.74E-05	2.35E-05	2.59E-06	1.03E-05	9.19E-07	2.58E-06	5.42E-06	3.04E-05	6.39E-06	1.01E-05
95% confidence limits for the coefficients - lower limit	-0.000	8.66E-06	-0.000	-0.000	-1.60E-05	-0.000	-6.87E-05	-8.82E-05	2.72E-06	-0.000	8.71E-05	-0.000
95% confidence limits for the coefficients - upper limit	-0.000	2.76E-05	-0.000	-0.000	-5.88E-06	-0.000	-6.51E-05	-7.81E-05	2.40E-05	-0.000	0.000	-0.000

** I/M = Internet/Mobile; C/M = Call Centre/Mobile*

4.15.8 Overall analysis for independent variable 10: outstanding credit card balance

For this variable, there were different trends across each of the different banks. Two conventional banks (Bank 2 and Bank 4) and one Islamic bank (Bank 3) had customers who were more likely to use a mobile application if they had greater outstanding balances on their credit card. By contrast, one conventional bank (Bank 1) had customers who were more likely to use a call centre if they had a

larger credit card balance. Finally, the other two Islamic banks (Bank 5 and Bank 6) had customers who were more likely to use the internet if they had higher credit card balances.

4.16 Independent variable 11: duration of relationship with the bank

The final variable that I measured was the duration of a customer's relationship with their bank. This was measured in years, and the aim was to assess whether customers who had been with their bank longer tended to use different communication channels than short-term customers.

4.16.1 Bank 1

Bank 1 did not provide us with data for this variable. Due to an issue with their system, they were unable to extract the data.

4.16.2 Bank 2

Customers in the second bank that I looked at were 1.04 (exp 0.043) times more likely to use the internet than a mobile application for every additional year that they had been with their bank. They were also 1.30 (exp 0.264) times more likely to use a call centre rather than a mobile application for each additional year that they had been with the bank. The p-value for internet/mobile was 0, and the p-value for call centre/mobile was close to 0, indicating statistical significance. The standard errors were 0.006 and 0.005, respectively. Based on the confidence intervals, I can be 95% sure that the true regression coefficient was between 0.030 and 0.056 for internet/mobile and between 0.253 and 0.276 for call centre/mobile.

4.16.3 Bank 3

The results for Bank 3 were fairly similar, with customers being 1.058 (exp 0.0573) times more likely to use the internet rather than a mobile application and 1.27 (exp 0.244) times more likely to use a call centre rather than a mobile application for each year that they had been with their bank. The p-values for both variables were

0, and the standard errors were 0.000 for internet/mobile and 0.003 for call centre/mobile. The results of the confidence interval mean that I can be 95% certain that the true regression coefficient was between 0.055 and 0.059 for internet/mobile and between 0.237 and 0.250 for call centre/mobile.

4.16.4 Bank 4

The results for Bank 4 show that the relative risk of a customer using the internet rather than a mobile application decreased by 0.993 (exp -6.15E-03) for every year of their relationship with their bank. The relative risk of a customer using a call centre rather than a mobile application increased by 1.020 (exp 2.02E-02) for every year that they were with their bank. The p-values were close to 0 for internet/mobile and 0 for call centre/mobile. The standard error for both variables was 0.000, and the confidence intervals ranged from -0.006 to -0.005 for internet/mobile and from 0.019 to 0.020 for call centre/mobile.

4.16.5 Bank 5

Customers at Bank 5 were 1.007 (exp 7.23E-03) times more likely to use the internet rather than a mobile application for every additional year that they had spent with their bank. They were also 1.726 (exp 5.46E-01) times more likely to use a call centre rather than a mobile application for every year of their relationship. Again, the p-values were close to 0 for internet/mobile and 0 for call centre/mobile, and the standard errors were 0.002 and 0.008, respectively. The results of the confidence intervals show that I can be 95% certain that the true regression coefficient was within the range of 0.003 and 0.011 for internet/mobile and within the range of 0.529 and 0.562 for call centre/mobile.

4.16.6 Bank 6

The final bank that I looked at, Bank 6, had customers who were 0.996 (exp -3.61E-03) times less likely to use the internet than their mobile for every year that they had been with their bank. Those customers were also 0.988 (exp -1.15E-02) times less likely to use a call centre than a mobile application for every year they had been with their bank. The p-values for both variables were close to 0, and the

standard error for both variables was also 0.000. The confidence interval ranged from -0.004_to -0.002 for internet/mobile and from -0.013 to -0.009 for call centre/mobile.

4.16.7 Results summary for independent variable 11: duration of relationship with the bank

The following tableTable 4-21) details the results for the impact of the duration of a customer's relationship with their bank on their preferred channel of communication.

Table 4-21: Multinomial logistic regression results for independent variable 11: duration of relationship with the bank

	Bank 1		**Bank 2**		**Bank 3**		**Bank 4**		**Bank 5**		**Bank 6**	
Variable 11: Relationship duration	I/M*	C/M	I/M	C/M	I/M	C/M	I/M	C/M	I/M	C/M	I/M	C/M
Coefficients-relative log odds	N/A	N/A	0.043	0.264	0.057	0.244	-6.15E-03	2.02E-02	7.23E-03	5.46E-01	-3.61E-03	-1.15E-02
Coefficients-relative risk exp (coefficients)	N/A	N/A	1.044	1.303	1.058	1.276	0.993	1.020	1.007	1.726	0.996	0.988
P-value	N/A	N/A	1.85E-10	0	0	0.00E+00	6.02E-198	0.00E+00	4.51E-04	0.00E+00	1.14E-09	1.24E-37
Standard Error	N/A	N/A	0.006	0.005	0.000	0.003	0.000	0.000	0.002	0.008	0.000	0.000
95% confidence limits for the coefficients - lower limit	N/A	N/A	0.030	0.253	0.055	0.237	-0.006	0.019	0.003	0.529	-0.004	-0.013

	Bank 1		Bank 2		Bank 3		Bank 4		Bank 5		Bank 6	
95% confidence limits for the coefficients - upper limit	N/A	N/A	0.056	0.276	0.059	0.250	-0.005	0.020	0.011	0.562	-0.002	-0.009

* *I/M = Internet/Mobile; C/M = Call Centre/Mobile*

4.16.8 Overall analysis for independent variable 11: duration of relationship with the bank

Across almost all the banks, the results show that the longer a customer has been with a bank, the more likely they are to use a call centre as their primary source of communication. This is followed by the internet and then by mobiles. Only one Islamic bank (Bank 6) showed a different pattern. Customers who had been at this bank longer were more likely to use a mobile application to communicate.

Relationship Duration

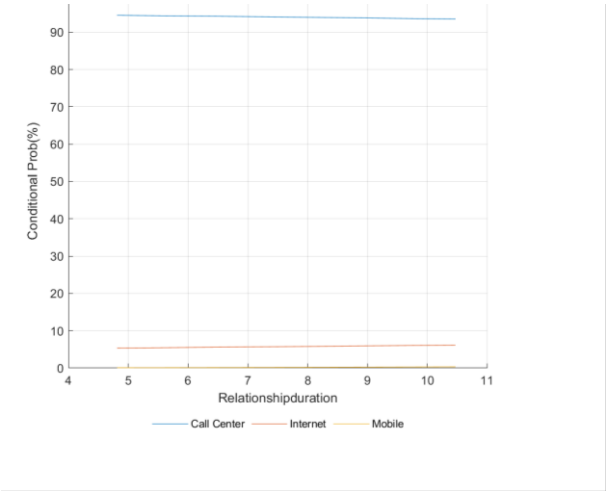


Figure 4-14: Probability of using different electronic banking channels dependent on the duration of the customer's relationship with the bank

4.17 Prediction model

After the regression analysis, I constructed a formula for predicting the difference in the likelihood that different customers would choose to use different online

channels, based on their characteristics. I used the Statistics and Machine Learning Toolbox from MATLAB to run a multinomial regression model (using equation (**4**-**1**).my model could predict the probabilities for the multinomial logistic regression model with predictors, X, and coefficient estimates, B (Equation (4-3):

[pihat,dlow,dhi] = mnrval(B,X,stats) (4-3)

[pihat,dlo,dhi] = bankingtype_prediction(age,gendernum,CA_Count,DA_Count, JA_Count,SA_Count,Continent_num,Consumer_Count, Income, Instalment_Count, limit, Outstanding_Bal)

4.18 Sample results of the prediction model

Table 4-22 provides three sample results from the model for predicting customer behaviour across each of the electronic channels, depending on the specific characteristics of the customers. In the example given here, the only variable that I adjust is the age of the individual. Everything else remains the same.

Table 4-22: Three sample results from the prediction model

Case 1	**Case 2**	**Case 3**
Customer details:	Customer details:	Customer details:
Age: 25	Age: 40	Age: 50
Gender: 1 (male)	Gender: 1 (male)	Gender: 1 (male)
Income: 2500	Income: 2500	Income: 2500
Continent: 2	Continent: 2	Continent: 2
Number of consumer loans: 1	Number of consumer loans: 1	Number of consumer loans: 1
Number of instalment loans: 0	Number of instalment loans: 0	Number of instalment loans: 0
Number of current accounts: 1	Number of current accounts: 1	Number of current accounts: 1
Number of savings accounts: 1	Number of savings accounts: 1	Number of savings accounts: 1
Number of joint accounts: 0	Number of joint accounts: 0	Number of joint accounts: 0
Number of deposit accounts: 2	Number of deposit accounts: 2	Number of deposit accounts: 2
Limit of credit card: 0	Limit of credit card: 0	Limit of credit card: 0
Outstanding balance of credit card: 0	Outstanding balance of credit card: 0	Outstanding balance of credit card: 0

Results: Probability of using Call Centre 16% Probability of using Internet 57% Probability of using Mobile 27%	Results: Probability of using Call Centre 33% Probability of using Internet 44% Probability of using Mobile 23%	Results: Probability of using Call Centre 48% Probability of using Internet 33% Probability of using Mobile 19%

4.19 Results for people from different continents and countries

Using the prediction model, I also tested whether a person's home continent or their nationality would affect their likelihood of choosing one of the three channels for communicating with their bank (call centre, internet, or mobile). As shown by Table 4-23, a person's home continent had no significant effect on their choice of communication channel.

Table 4-23: Use of electronic banking channels by people from different continents

	Africa	**Asia**	**Australia**	**Europe**
Call Centre Probability	74%	74%	75%	75%
Internet Probability	17%	16%	15%	14%
Mobile Probability	9.2%	9.9%	11%	11%

Note: In this example, I picked bank customers with the following features: male customers from a conventional bank, aged 35, with one current account, an income of 500 and no other credit facilities with the bank.

I also found that nationality had no significant effect on customers' choice of communication channel (see Table 4-24).

Table 4-24: Use of electronic banking channels by people of different nationalities

	Kuwaiti	**Indian**	**Egyptian**	**Philippines**
Prob. Call Centre	93%	93%	93%	94%
Prob. Internet	4.8%	4.7%	4.6%	5%
Prob. Mobile	1.7 %	2 %	2.4%	1.2%

Note: In this example, I picked bank customers with the following features: male customers from a conventional bank, aged 35, with one current account, an

income of 300, no other credit facilities with the bank and a three-year relationship with the bank.

4.20 Further analysis of mobile applications

Given the results of my study, it is clear that mobile banking is becoming a dominant communication channel. In my research, I discovered variables, such as customer characteristics and financial products, that affected the use of mobile banking compared to other communication channels. I found that there was increasing uptake in the use of mobile applications even among customers who were least influenced by the variables. This caught my attention and obliged us to extend my study of mobile applications (see Figure 4-15 and Figure 4-16).

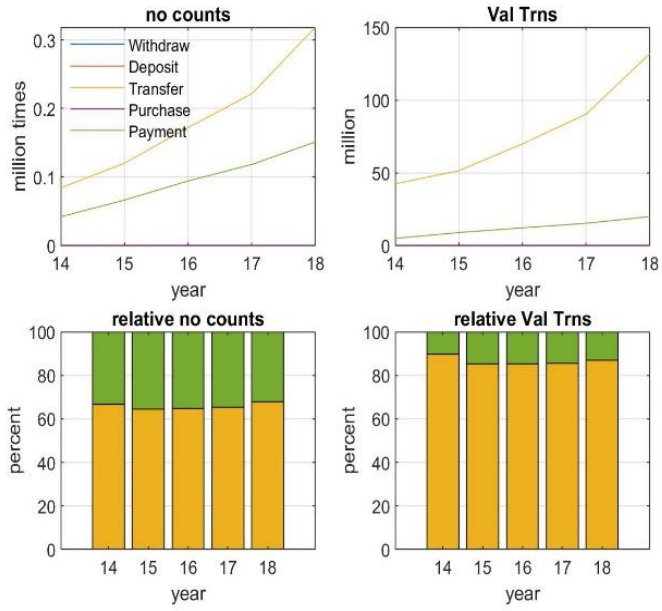


Figure 4-15: Mobile application services (conventional bank)

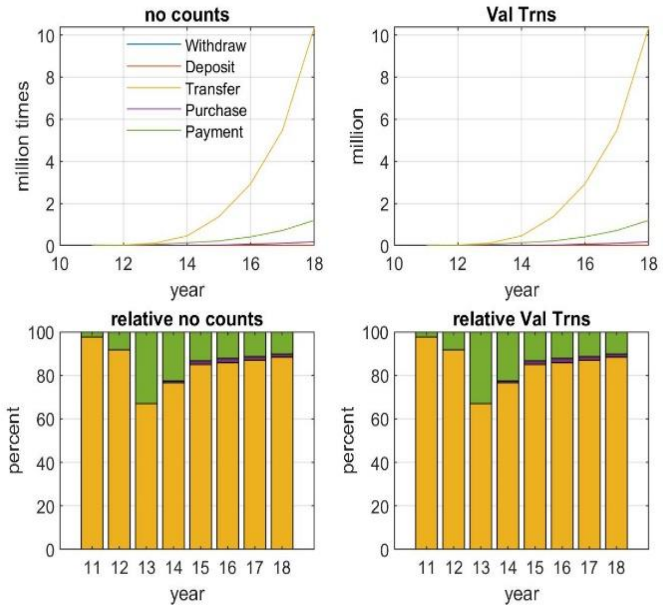


Figure 4-16: Mobile application services (Islamic bank)

Our research showed that the most common reason for using mobile banking was to make bank transfers (over 80% of the total usage). Figure 4-15 and Figure 4-16 show that payments were insignificant compared to transfers. One of the main reasons for this is that customers nowadays prefer fast and accessible methods when paying for services. This is not an advantage provided by mobile payment services.my findings have led us to propose certain solutions to help increase the number of payments conducted though mobile banking. First, banks should make it easier to access payment methods on mobile applications through shortcuts. For example, instead of logging in and navigating to the payment options that are buried in the menus, mobile banking applications should include payment options on the first page, along with a search bar. Second, banks should allow for governmental and non-governmental utility expenses and purchases to be paid for within their mobile applications. They should develop financial services that can incorporate most governmental and

private entity expenses. They should make the system visible and easy to use. Lastly, several banks in Kuwait have received the ISO 27001: 2013, International Organization for Standardization certificate. This is one of the most significant cyber security certificates in the world(Kuwait times, 2021). Banks should capitalize on this by embarking publicizing the strength of their cyber security, encouraging customers to make payments on their mobile phones. Customers will naturally carry out more payment activities on their mobile phones once they are sure that the payments are safe and secure.

4.21 Overfitting

Given my large sample size, I needed to deal with the problem of overfitting. Therefore, I ran a subsample regression to see if the results were similar. I found similar results of the subsample when compared to the full sample results. This indicated that the model used was not overfitted. The following is a sample for Bank 1.

In the full data set, the observations (number of transactions) were 108,071. I used the function with syntax “cvpartition” from the Statistics and Machine Learning Toolbox in Matlab to create a cross-validation partition for the data.

How does this function work? c=cvpartition (n, ‘holdout’, p). This function creates a random, non-stratified partition for holdout validation on n observations. Thus, this partition divides the n observations (108,071 transactions) into two sets. The parameter p must be a scalar. When $0 < p < 1$, the function randomly selects an approximate p*n for the test set. When p is an integer, the function randomly selects p observations for the test set:

1) Training set: the remaining n observations, which is 108,071 – 32,421 = 75,650

2) Test (holdout) set: here I use p = 0.3, so the test set equals 0.3*108,071 = 32,421

Thus, the full set used 108,071 observations, while my sub-sample set used 75,650 observations (Mathworks, 2019). As the subsample results were very

similar to the full sample results, the estimation in the full sample is not overfitted (see Figure 4-17).

When comparing the two prediction models, aggregate results are more meaningful. The probabilities vs age plots are aggregate results (rather than averages). I can see that they show similar trends. When used for predictions, the full sample model is more accurate as it used more data. This is how big data works: when more data are available, the model calibrated with more data is more reliable.

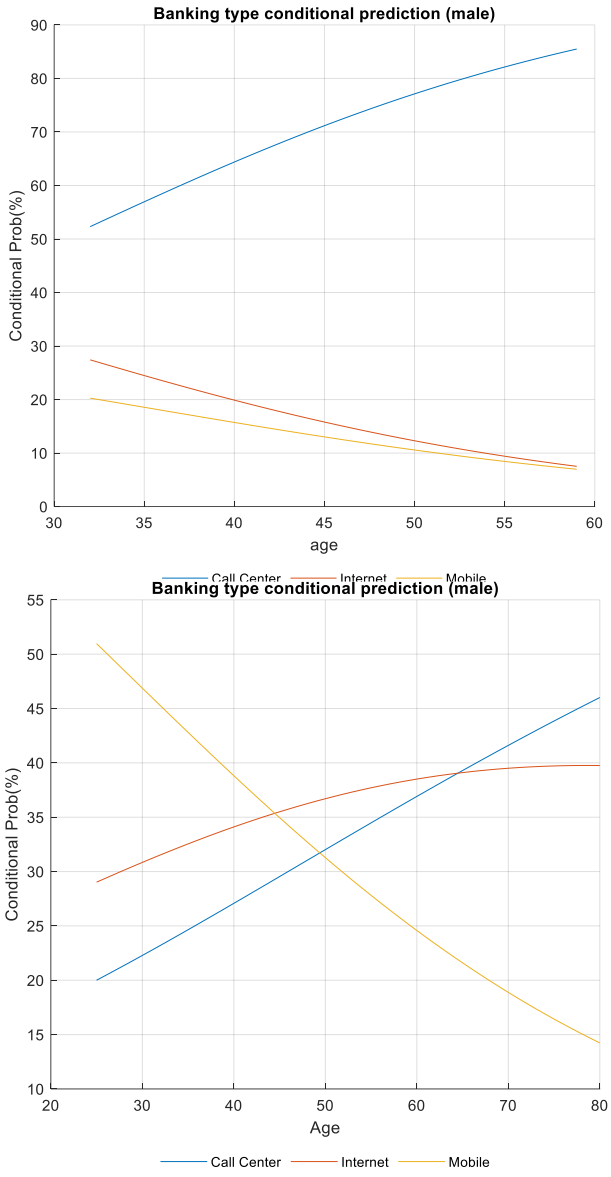


Figure 4-17: Full sample vs sub-sample

4.22 Summary

This chapter has outlined the results of the regression analysis for each of the different variables across the six banks that were the focus of this study. These variables were chosen based on my review of the literature, and several new variables were also included that had not previously been analyzed (such as the total number of loans, credit card limits, income, home continent, and nationality).

I also developed a richer data set for examining those variables that had previously been assessed in other studies. Furthermore, I differentiated between Conventional banks and Islamic banks, a distinction which has not previously been made before. Based on the results, I can identify several key findings for this study.

First, as customers get old, they tend to use call centres as their primary way of communicating with their bank. This result was expected given my experience and my review of previous studies. I can infer from this that the tools that banks use to explain their internet and mobile services are not well developed or promoted. Based on my statistical analysis, therefore, I strongly recommend that banks improve their explanatory materials, ideally by using video recordings that are available online or on a mobile platform. These video advertisements should target elderly customers.

Second, the findings from this study have provided more information about the difference between Islamic banks and conventional banks. This study focused on this difference because 50% of banks in Kuwait are Islamic, and more than 42% of banks' total assets go to Islamic banks (Central Bank of Kuwait, 2019). But Islamic banking is not just growing in Kuwait; it is becoming more widespread around the world. Nevertheless, relatively little research has been conducted into the difference between Islamic and conventional banks; this study intended to fill this research gap.my findings reveal, however, that there are very few differences between the customers of conventional banks and those of Islamic banks. For all variables except gender, customers of both banks made similar choices regarding their preferred method of communicating with their banks. Specifically, when it comes to gender, men in Islamic banks tend to use internet more frequently than women, who use the mobile. Previous studies such as that by (Riquelme & Rios, 2010) had found that women tended to be more interested in social norms, whereas men tended to be more interested in the particular advantages of a particular technology or form of communication. This suggests that men who use Islamic banks are yet to be convinced of the value of using their bank's mobile

applications. Islamic banks, therefore, should make this one of their targets as they develop their mobile applications.

Third, I was expecting that customers with more accounts would be more likely to use mobile applications. To a certain extent, this was backed up by the results, which showed that people with more current accounts and savings accounts were more likely to use their bank's mobile applications. However, one finding that surprised us was the fact that people with more deposit accounts were more likely to use the internet or call centres rather than a mobile application.

Fourth, the results also showed that people with higher incomes were more likely to use mobile applications than call centres or the internet. One possible reason for this is that people with higher incomes can afford to pay for devices and networks that are more advanced and secure. This increases their trust in mobile applications, making them more likely to use them. Another reason is that people with higher incomes tend to carry out more transactions. They therefore value ease-of-use and efficiency, making them more likely to adopt a mobile application. Finally, they are more likely to require the services offered by mobile applications, such as the ability to pay school fees, make charitable donations, pay telecom bills, and more.

Fifth, I found that, in two of the banks, customers with more loans tended to use call centres as their primary means of communication. This was true regardless of the type of loan they had. This is an important finding with implications for how banks operate and set future strategies. This finding shows that not all customers necessarily prefer using the most advanced technological solution (a mobile application) for communicating with their bank. As customers with deposit accounts and loans do not necessarily need to make as many daily transactions, they are less likely to want to use the most up-to-date technology. By contrast, customers with current accounts who need to make lots of daily transactions prefer to use mobile applications with user-friendly interfaces. Similarly, customers with higher incomes and higher credit card limits are also shifting to mobile applications.

A further finding is that customers who have been with their bank for longer are more likely to use call centres as their primary form of communication. This is probably because long-term customers tend to have a stronger relationship with their bank and place more trust in the employees. Banks need to be aware of this. Rather than simply focusing on attracting new customers, they should also pay attention to their existing customers and provide them with suitable services. For example, banks could consider enhancing their digital options by giving long-term customers access to an electronic customer service agent. Banks would need to make sure that this type of electronic agent could satisfy their loyal customers' needs.

Finally, the results showed that a person's home continent or their nationality have no significant impact on the channel of communication that they prefer. I were not anticipating this result for nationality, particularly because 65% of workers are low-income labourers with a moderate level of education (Statista Research Department, 2020). Therefore, I was expecting them to have different communication preferences compared to Kuwaiti nationals. Expanding my analysis to include customers' home continent also did not yield any significant results.

Moreover, I tested four different methods to choose the appropriate method for my research. I also used out-of-sample prediction accuracy as a key performance measure. Multinomial logistic regression analysis was used in this study as it was found to perform robustly. I also investigated if there were any correlations between the independent variables. I found one significant correlation between credit card limit and outstanding credit card balance. I examined this further, looking at time of execution and the key performance measures. I decided to keep both variables in my study.

I examined the relation between the volume of training data and the accuracy of the model to tell how sensitive the models were to variations in the size of data. The results indicated that multinomial logistic regression and multivariate linear regression performed stably across the different sample sizes. They performed better than the other models that were tested.

Given the large sample size, I ran a sub-sample regression to avoid any problems with overfitting. In comparing the results to the full sample, I concluded that the model was not overfitted.

I performed more in-depth analysis on mobile applications and found that more than 80% of the usage of mobile banking applications is directed towards transfers. I recommended some solutions on how to increase the usage of mobile banking for payments.

I also recognized that although each bank has its different customer sets, there is a similarity in the effect of each variable on the use of the bank's communication channels.

In summary, this chapter has explained the key performance parameters I used as part of the multinomial logistic regression. It has provided a summary of the outcomes for every variable across every bank. It has also given an account of the prediction model that I developed so that banks can predict their customers' preferences based on their characteristics. The chapter also considered some of the study's key findings. It discussed correlation analysis, the comparison between different models, the sensitivity testing, the risk of overfitting, and the extended analysis on mobile banking.

5 CONCLUSION

5.1 Summary of the research topic, rationale, and findings

In this study, I requested and collected substantial amounts of confidential financial data from large- and medium-sized banks in Kuwait. The data provided information about the characteristics of the banks' customers, their product portfolios, and their use of several digital banking channels. I focused on channels that provided similar services: online banking, call centres, and mobile applications. The aim was to build a prediction model for each bank to predict the preferences of customers for different electronic channels (the dependent variables) based on the customers' information (the independent variables). Given the size and type of data I collected, I decided to use machine learning techniques to process it. Bankers and financial analysts would not be able to use that kind of data without using machine learning techniques.

Electronic banking has become a very important source of revenue in the banking sector. It increases customer satisfaction, enhances efficiency, and increases profitability. Being able to predict the best online channel for each customer based on their characteristics and their banking portfolio would help to improve customer satisfaction and increase customer loyalty. This could lead to longer customer relationships and more active clients who would make use of more financial products and services. Moreover, effective prediction models would help to decrease advertising costs and other operating expenses (while also increasing revenue).

By examining the previous literature, I identified several gaps in the research. Previous studies used limited data sets, failed to offer comparisons between electronic channels, used limited variables, and did not provide prediction models. Based on these limitations, I identified my research objectives. First, I established my research question and hypotheses. Second, I determined the type and size of data needed for the study. I then began requesting and collecting the data. Third, I tested various methods for analyzing the data. Fourth, I used the chosen method,

multinomial logistic regression, to analyze the data, using the relevant parameters.

In terms of the methodology, I chose to use statistical machine learning to clean and process the data. This provided predictive analysis. I chose to use multinomial logistic regression after applying different methods such as multivariate linear regression, binary decision tree for regression, and support vector machine regression. I used out-of-sample prediction as my key performance measurement; I assessed the relative deviation between the out-of-sample predicted probability for the different banking types and their actual counterparts. The smaller the absolute value of this measure, the better the performance.my findings showed that multinomial logistic regression was relatively robust in its performance. Multinomial logistic regression also allowed us to use mobile banking (the most up-to-date technology) as my base for comparing different channels. I compared the probabilities of using mobile applications rather than another mobile channel. This made the analysis more efficient. The regression analysis was done using MATLAB, which meant that I did not have to run individual regressions and then interpret and compare the outcomes. This type of regression evaluated the parameters simultaneously, making it more efficient. I used the multinomial logistic regression model to predict how customers would use different electronic banking channels based on their age, gender, number of current accounts, number of deposit accounts, number of savings accounts, continent, nationality, number of consumer loans, income, number of instalment loans, credit card limit, outstanding credit card balance, and the length of relationship with their bank.

One of the key findings of my study was that as customers get older, they tend to prefer using call centres as their primary method of communication; their second preference is for the internet, followed by mobile applications. This finding supports the findings of previous studies, but I also found that there is no difference between customers of Islamic banks and customers of conventional banks. Another key finding related to gender. I found that male customers of conventional banks were more likely than women to prefer mobile applications to

the internet and call centres. Male customers at Islamic banks are more likely than women to prefer using the internet rather than mobile applications and call centres. These findings were discussed in more detail in the chapter on results.

Our results also show that a customer's types of account or types of loan have no influence on their decision to use a particular electronic channel. That said, a customer who has more current accounts and savings accounts is more likely to use mobile applications, followed by call centres and then the internet. The data shows that customers with more than one deposit account use the internet and call centres, followed closely by mobile applications. Furthermore, I also noticed that customers with higher incomes tended to prefer mobile applications and the internet rather than call centres. Customers with more loans tended to use call centres as their primary channel, as did customers who had a longer relationship with their bank. Neither continent nor nationality had any significant effect on customers' use of electronic channels.

5.2 Contributions and achievements

After reading the relevant literature, I recognized that electronic banking (and mobile applications in particular) was an important topic in the field. I identified the achievements of other studies and assessed the variables they had used. I made my own contribution by examining new relationships and studying new variables (such as number of loans, credit card limits, income, continent, and nationality). I also used a much larger data set. I differentiated between conventional banks and Islamic banks; a distinction that had not been made in previous studies. I made several visits to ten banks and collected a substantial amount of confidential data. I managed all the necessary contract issues and overcame all the legal obstacles. I used different types of data and a larger data set than previous studies. In the methodology, the main contribution my study made was to introduce machine learning techniques into the accounting field. I used financial data to create a prediction model for each bank to enhance their performance.

Overall, my study showed that customers' preferences for particular electronic banking channels are not based simply on what the most advanced technology

is. As customers who with deposit accounts and loans do not need to make daily transactions, for example, they are less likely to use mobile applications, even though it is the most up-to-date technology. Mobile usage increases, however, when customers need to make daily transactions. For example, customers with savings and current accounts prefer to use mobile applications because they are simple and easy to use. Furthermore, customers with higher income and higher credit card limits are also shifting to mobile applications. The results did not show any noticeable difference between Islamic and conventional banks, except when it came to gender. Men at conventional banks were more likely to prefer mobile applications over the internet and call centers. But male customers of Islamic banks were more likely than women to prefer the internet rather than mobile applications and call centres. I found that type of loan (consumer and instalment) and type of accounts (current and savings) had no effect on customers' preferences. Continent and nationality also had no significant effect on customers' preferences.

The goal of my research was to calculate the probability of a customer choosing a particular banking channel based on their information. I developed a code and used the built-in functions in MATLAB to create a prediction formula for each bank. The purpose of this was to help banks to allocate their marketing tools effectively. One of the main effects of this research is that it will provide banks with a better understanding of their customers' needs, allowing them to offer better quality products. This will increase customer engagement and trust. In everything, my prediction model will allow banks to reduce their marketing costs, capital expenditures, and employees. This will ultimately lead to increased profits. The information that I have obtained and the prediction models that I have generated will make a significant contribution to local banks in Kuwait and have a positive effect on the banking industry.

5.3 Limitations

I successfully completed the study and met my objectives. However, there were some limitations to the study, which present opportunities for future research.

- There are other very important variables that could be added to a study of this kind. I were not able to include all the variables I wanted to include. One of the problems I faced was that the banks were very reluctant to share their confidential data. Furthermore, the way that some of the bank's function made it difficult to obtain historical data. For example, some banks had updated their systems and were not able to access data on their old systems. Some banks also provided us with data for a shorter period than I had asked for. This may have stopped us from seeing how customer preferences developed over time. There were some variables that I was unable to access (like customer risk factors). I could only access the data from last year.

- Due to the limited time available, I chose to focus on the general channels used by customers. I did not examine the different services within these channels.

- I was able to perform more detailed analysis of mobile banking, but only for some of the banks. However, the other banks did not differentiate between the services they offered through their mobile applications. For example, transfers and purchases were not separated.

- In my study, I focused on three electronic channels: the internet, mobile applications, and call centres. But there are other electronic channels that could be considered, like automated teller machines (ATMs) and point-of-sale services.

- Our research was based only on the number of transactions and did not consider the value of each transaction.

5.4 Future work

There are several remaining directions for future studies.

- Future studies could include other variables like educational level, joint accounts, and customer risk factors. They could also add different parameters like security and level of trust.

- Future work could also look at the difference between two stages: before and after the evolution of electronic banking.
- Future research should compare different services within each channel. This would provide more information for the banks, helping them to improve their decisions and strategies.
- Future studies should add more dependent variables, like automated call centres and point-of-sale services.
- In my study, I only looked at the number of transactions for each channel, rather than the value of the transactions. This was because I believed that older people generate more wealth, so I thought that they would carry out higher-value transactions, skewing the analysis of the electronic channels. Future studies could include the value of transactions as well.

REFERENCES

Attaway, S. (2017). *MATLAB a practical introduction to programming and problem solving* (fourth edi). Butterworth-Heinemann.

Bataev, A. V. (2018). Comparative Analysis of Cloud Computing Application in Russian and Foreign Financial Institutions. *2018 IEEE International Conference "Quality Management, Transport and Information Security, Information Technologies" (IT&QM&IS)*, 578–582. https://doi.org/10.1109/ITMQIS.2018.8525077

Bekamiri, H., Ganji, S. F. G., Simonetti, B., & Seno, S. A. H. (2021). A new model to identify the reliability and trust of internet banking users using fuzzy theory and data-mining. *Mathematics*, *9*(9), 1–16. https://doi.org/10.3390/math9090916

Bhimani, A., & Willcocks, L. (2014). Digitisation, Big Data and the transformation of accounting information. *Accounting and Business Research*, *44*(4), 469–490. https://doi.org/10.1080/00014788.2014.910051

Burton, G. F., & Jermakowicz, E. K. (2015). International Financial Reporting Standards. In *International Financial Reporting Standards*. https://doi.org/10.4324/9780203522356

Campbell, D., & Frei, F. (2010). Cost Structure, Customer Profitability, and Retention Implications of Self-Service Distribution Channels: Evidence from Customer Behavior in an Online Banking Channel. *Management Science*, *56*(1), 4–24. https://doi.org/10.1287/mnsc.1090.1066

Carpenter, B. (2008). Lazy Sparse Stochastic Gradient Descent for Regularized Mutlinomial Logistic Regression Maximum a Posteriori Estimate. *Alias-i, Inc.*, *Tech. Rep*(5), 1–20.

Central Bank of Kuwait. (1996). *Instructions Conserning Rules and Regulations for Banks' Extention of Consumer Loans and Othr Installment Loans*. Retrieved from https://www.cbk.gov.kw/en/images/section-13-2748_v50_tcm10-2748.pdf

Central Bank of Kuwait. (2019). Financial Stability Report 2019. In *Central Bank of Kuwait*. Retrieved from http://www.cbk.gov.kw/en/statistics-and-publication/publications/financial-stability-report.jsp

Chawla, D., & Joshi, H. (2017). Role of demographics as moderator in mobile banking adoption. *AMCIS 2017 - America's Conference on Information Systems: A Tradition of Innovation*, *2017-Augus*, 1–10.

Chua, F. (2013). Big data : its power and perils. *The Association of Chartered Certified Accountants*, 40.

Council, F. R. (2016). International Standard on Auditing (UK) 200 (Revised June 2016). *Financial Reporting Council Standard*, *200*(June).

Coursera. (2019). *No Title*. Retrieved from https://www.coursera.org/learn/machine-learning/home/info

Danaher, P. J., Wilson, I. W., & Davis, R. A. (2003). A Comparison of Online and Offline Consumer Brand Loyalty. *Marketing Science*, *22*(4), 461–476. https://doi.org/10.1287/mksc.22.4.461.24907

Danyali, A. A. (2018). Factors influencing customers ' change of behaviors from online banking to mobile banking in. *Journal of Organizational Change Management*, *31*, 1226–1233. https://doi.org/10.1108/JOCM-07-2017-0269

Degeratu, A. M., Rangaswamy, A., & Wu, J. (2000). *Consumer choice behavior in online and traditional*. https://doi.org/10.1016/S0167-8116(00)00005-7

Elshawi, R., Sakr, S., Talia, D., & Trunfio, P. (2018a). Big Data Systems Meet Machine Learning Challenges : Towards Big Data Science as a Service. *Big Data Research*, *14*, 1–11. https://doi.org/10.1016/j.bdr.2018.04.004

Elshawi, R., Sakr, S., Talia, D., & Trunfio, P. (2018b). Big Data Systems Meet Machine Learning Challenges : Towards Big Data Science as a Service. *Big Data Research*, *14*, 1–11. https://doi.org/10.1016/j.bdr.2018.04.004

Emptage, M., & Dobson, A. J. (1992). An Introduction to Generalized Linear Models. In *Technometrics* (Vol. 34). https://doi.org/10.2307/1269239

Emrouznejad, A., & Anouze, A. L. (2010). Data envelopment analysis with classification and regression tree - A case of banking efficiency. *Expert Systems*, *27*(4), 231–246. https://doi.org/10.1111/j.1468-0394.2010.00516.x

Ernst & Young. (2014). Big data: Changing the way businesses compete and operate. *International Journal of Simulation: Systems, Science and Technology*, *16*(April), 28. https://doi.org/10.5013/IJSSST.a.16.5B.22

Ferreira, A. J. M. (2009). MATLAB codes for finite element analysis: Solids and structures. In *Solid Mechanics and its Applications* (Vol. 157).

Flood, J. M. (2020). *Practitioner ' s Guide to GAAP 2020*. New Jersey: WILEY.

Freedmn, D. A. (2009). *Statistical Models Theory and Practice* (Revised Ed). Cambridge University Press.

Frew, L. (2016). Impact of Technology on the Financial Sector. *Global Banking and Finance Review*. Retrieved from https://www.globalbankingandfinance.com/impact-of-technology-on-the-financial-sector/

Gamage, P. (2016). Big Data : are accounting educators ready ? *Accounting and Management Information Systems*, *15*(3), 588–604.

Gautam, V. (2012). Measuring the Impact of New Technologies through Electronic-Banking on Profitability of Banks: Evidence form Indian Banking Industry. *Romanian Journal of Marketing*, (3), 20–26. Retrieved from http://search.proquest.com/docview/1115592498?accountid=14495%5Cnhttp://diana.uca.es:4550/resserv?genre=article&issn=&title=Romanian+Journal+of+Marketing&volume=&issue=3&date=2012-07-01&atitle=Measuring+the+Impact+of+New+Technologies+through+Electronic-Ba

Gensler, S., Leeflang, P., & Skiera, B. (2012). Impact of online channel use on customer revenues and costs to serve: Considering product portfolios and self-selection. *International Journal of Research in Marketing*, *29*(2), 192–201. https://doi.org/10.1016/j.ijresmar.2011.09.004

Gepp, A., Linnenluecke, M. K., O'Neill, T. J., & Smith, T. (2018). Big data techniques in auditing research and practice: Current trends and future opportunities. *Journal of Accounting Literature*, *40*(February), 102–115. https://doi.org/10.1016/j.acclit.2017.05.003

Gibbons, J. D., & Pratt, J. W. (1975). *and Methodology * P-values : Interpretation*. *29*(1), 829–834.

Gilat, A. (2013). MATLAB An Introduction with Applications. In *Angewandte Chemie International Edition, 6(11), 951–952.* (Fifth Edit). Don Fowley.

Greene, W. H. (2012). *Econometric Analysis* (Seventh Ed). Boston: Pearson Education.

Haider, M. J., Changchun, G., Akram, T., & Hussain, S. T. (2018). Exploring Gender Effects in Intention to Islamic Mobile Banking Adoption: an empirical study. *Arab Economic and Business Journal*, *13*(1), 25–38. https://doi.org/10.1016/j.aebj.2018.01.002

Hausman, J., H.Hall, B., & Griliches, Z. (1984). Econometric models for count data with with an applicayion to patents-R&D relation-ship. *Econometrica*, *52*(4), 909–938.

Hidalgo, B., & Goodman, M. (2013). Multivariate or multivariable regression? *American Journal of Public Health*, *103*(1), 39–40. https://doi.org/10.2105/AJPH.2012.300897

Honest, N., & Patel, A. (2016). A SURVEY OF BIG DATA ANALYTICS. *International Journal of Information Sciences and Techniques (IJIST)*, *6*(1/2), 35–43.

Huerta, E., & Jensen, S. (2017). An Accounting Information Systems Perspective on Data Analytics and Big Data. *Journal of Information Systems*, *31*(3), 101–114. https://doi.org/10.2308/isys-51799

Ikhsanova, L., & Protsko, E. (2016). ELECTRONIC BANKING : TOOL OF TRANSFORMING THE INTERACTION BETWEEN BANKS AND CLIENTS

AND IMPROVING THE SERVICE QUALITY OF THE. *Journal of Economics and Economic Education Research*, *17*, 73–80.

IndustryToday. (2019). The Growth of Technology. Retrieved from https://industrytoday.com/the-growth-of-technology/

Kuwait times. (2021, March 17). Banks meet global cybersecurity standards. *Kuwait Times*. https://doi.org/https://news.kuwaittimes.net/website/banks-meet-global-cybersecurity-standards/

Loh, W. (2004). Regression Trees with. Unbiased Variable Selection. *Korean Journal of Applied Statistics*, *17*(3), 459–473. https://doi.org/10.5351/kjas.2004.17.3.459

Long, J. S., & Freese, J. (2003). Models for Ordinal Outcomes. In *Regression models for categorical dependent variables using Stata.*

Lynch, S. (2014). Dynamical systems with applications using MATLAB®, second edition. *Dynamical Systems with Applications Using MATLAB, Second Edition*, (August 2014), 1–514. https://doi.org/10.1007/978-3-319-06820-6

Mathworks. (2016a). *Applying Supervised Learning*. Natick: Mathworks. Available at: https://uk.mathworks.com/?s_tid=gn_logo (accessed 17th January 2019).

Mathworks. (2016b). *Applying Unsupervised Learning*. Natick: Mathworks. Available at: https://uk.mathworks.com/?s_tid=gn_logo (accessed 17th January 2019).

Mathworks. (2016c). *Getting Started with Machine Learning*. Natick: Mathworks. Available at: https://uk.mathworks.com/?s_tid=gn_logo (accessed 17th January 2019).

Mathworks. (2016d). *Introducing Machine Learning*. Natick: Mathworks. Available at: https://uk.mathworks.com/?s_tid=gn_logo (accessed 17th January 2019).

Mathworks. (2019). Statistic and Machine learning Toolbox: User's Guide (R2019b). Retrieved from MathWorks website: https://uk.mathworks.com/help/stats/mnrfit.html

Mathworks. (2021a). 'algorithm'_ “ecm” and 'cwls'. Retrieved from https://www.mathworks.com/help/stats/mvregress.html#namevaluepairarguments

Mathworks. (2021b). Access MATLAB Add-On Toolboxes. Retrieved from https://uk.mathworks.com/help/thingspeak/matlab-toolbox-access.html

Mathworks. (2021c). cellfun. Retrieved from https://www.mathworks.com/help/matlab/ref/cellfun.html

Mathworks. (2021d). datetime. Retrieved from https://www.mathworks.com/help/matlab/ref/datetime.html

Mathworks. (2021e). erase. Retrieved from https://www.mathworks.com/help/matlab/ref/erase.html

Mathworks. (2021f). fitrsvm. Retrieved from https://www.mathworks.com/help/stats/fitrsvm.html?s_tid=doc_ta

Mathworks. (2021g). fitrtree. Retrieved from https://www.mathworks.com/help/stats/fitrtree.html

Mathworks. (2021h). lbusdate. Retrieved from https://www.mathworks.com/help/finance/lbusdate.html

Mathworks. (2021i). mnrfit. Retrieved from https://www.mathworks.com/help/stats/mnrfit.html

Mathworks. (2021j). mvregress. Retrieved from https://www.mathworks.com/help/stats/mvregress.html

Mathworks. (2021k). outerjoin. Retrieved from https://www.mathworks.com/help/matlab/ref/outerjoin.html

Mathworks. (2021l). readtable. Retrieved from https://www.mathworks.com/help/matlab/ref/readtable.html

Mathworks. (2021m). sortrows. Retrieved from https://www.mathworks.com/help/matlab/ref/double.sortrows.html

Mathworks. (2021n). splitapply. Retrieved from https://www.mathworks.com/help/matlab/ref/splitapply.html

Mathworks. (2021o). Statistics and Machine Learning Toolbox. Retrieved from https://uk.mathworks.com/help/stats/

Mathworks. (2021p). xcorr. Retrieved from https://uk.mathworks.com/help/matlab/ref/xcorr.html

MathWorks. (2020). Multivariate Linear Regression.

MathWorks. (2021). Understanding Support Vector Machine Regression. Retrieved from MathWorks website: https://www.mathworks.com/help/stats/understanding-support-vector-machine-regression.html?searchHighlight=support vector machine regression concept&s_tid=srchtitle

McCullagh, P., & Nelder FRS, J. A. (1990). *Generalized Linear Models* (Second Edi). Chapman and Hall.

McLeod, S. (2019). What a p-value tells you about statistical significance. Retrieved from https://www.simplypsychology.org/p-value.html

Md Mujeeb, S., Praveen Sam, R., & Madhavi, K. (2021). Adaptive Exponential Bat algorithm and deep learning for big data classification. *Sadhana - Academy Proceedings in Engineering Sciences*, *46*(1). https://doi.org/10.1007/s12046-020-01521-z

Mistry, M. (2021). Digital battle: Bank customers demand more. *FintechMagazine*. Retrieved from https://fintechmagazine.com/banking/digital-battle-bank-customers-demand-more

Moga, L., Nor, K., Neculita, M., & Khani, N. (2012). Trust and Security in E-banking Adoption in Romania. *Communications of the IBIMA*, *2012*, 1–10. https://doi.org/10.5171/2012.583012

Mohammad, A.-S., & Saad, A.-W. (2011). The Impact of E- Banking on The Performance of Jordanian Banks. *Journal of Internet Banking and*

Commerce, *16*(2), 1–11.

Morgan, J. N., & Sonquist, J. A. (1963). *1 St Decision Tree*. *58*(302), 415–434.

Muller, M. (2004). Generalized Linear Models. *Research Gate*, (February 2004), 1–24. https://doi.org/10.1007/978-3-642-21551-3

Nick, I. (2018). How technology is impacting the finance and banking sector. Retrieved from Information Age website: https://www.information-age.com/technology-finance-banking-sector-123471800/

Okasha, M. K., & Shehada, M. A. (2016). CLASSIFICATION OF ANEMIC PALESTINIAN CHILDREN USING THE MULTINOMIAL LOGISTIC REGRESSION MODEL. Mahmoud. *INTERNATIONAL JOURNAL OF ADVANCED RESEARCH*, *4*(2), 560–573.

Oxford Business Group. (2019). *The Report: Kuwait 2019*. Retrieved from https://oxfordbusinessgroup.com/analysis/let's-get-digital-financial-technology-changing-local-banking-landscape

Pwc. (2017). *Redrawing the lines: FinTech's growing influence on Financial Services*. Retrieved from https://www.pwc.com/gx/en/industries/financial-services/assets/pwc-global-fintech-report-2017.pdf

Radi, B., & ELHami, A. (2018). *No TitleAdvanced Numerical Methods with Matlab® 2: Resolution of Nonlinear, Differential and Partial Differential Equations* (First Edit). ISTE Ltd and John Wiley & Sons, Inc.

Ratner, B. (2009). The correlation coefficient: Its values range between 1/1, or do they. *Journal of Targeting, Measurement and Analysis for Marketing*, *17*(2), 139–142. https://doi.org/10.1057/jt.2009.5

Rawwash, H., Masa'd, F., Enaizan, O., Eanizan, B., Adaileh, M. J., Saleh, A. M., & Almestarihi, R. (2020). Factors affecting Jordanian electronic banking services. *Management Science Letters*, *10*(4), 915–922. https://doi.org/10.5267/j.msl.2019.10.004

Richins, G., Stapleton, A., Stratopoulos, T. C., & Wong, C. (2017). Big Data

Analytics: Opportunity or Threat for the Accounting Profession? *JOURNAL OF INFORMATION SYSTEMS*, *31*(3), 63–79. https://doi.org/10.2139/ssrn.2813817

Riquelme, H. E., & Rios, R. E. (2010). The moderating effect of gender in the adoption of mobile banking. *International Journal of Bank Marketing*, *28*(5), 328–341. https://doi.org/10.1108/02652321011064872

Salman, A., & Nawaz, H. (2018). Islamic financial system and conventional banking: A comparison. *Arab Economic and Business Journal*, *13*(2), 155–167. https://doi.org/10.1016/j.aebj.2018.09.003

Samuel, A. L. (1969). Some studies in machine learning using the game of checkers. II-Recent progress. *Annual Review in Automatic Programming*, *6*(PART 1), 1–36. https://doi.org/10.1016/0066-4138(69)90004-4

Sethi, A. (2020). Support Vector Regression Tutorial for Machine Learning. Retrieved from Analytics Vidhya website: https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/

Simons, P., Masamvu, T., & Parks, P. (2013). From insight to impact - Unlocking opportunities in big data. In *CGMA*. Retrieved from http://www.cgma.org/RESOURCES/REPORTS/PAGES/INSIGHT-TO-IMPACT-BIG-DATA.ASPX?TestCookiesEnabled=redirect

Statista. (2020). Online banking penetration in Great Britain from 2007 to 2020. Retrieved July 23, 2021, from Statista website: FinTech is becoming essential to providing bank services locally. The organizations have found the digital revolution as key to productively pushing economic growth in the New Kuwait 2035 long-term development strategy. Executive director of supervision

Statista Research Department. (2020). *Population in Kuwait by citizenship 2012-2019*. Retrieved from https://www.statista.com/statistics/645790/kuwait-population-by-citizenship-status/

Steorts, R. C. (2017). *Tree Based Methods : Bagging , Boosting , and Regression Trees*. Retrieved from http://www2.stat.duke.edu/~rcs46/lectures_2017/08-trees/08-tree-advanced.pdf

Tan, S. H., & Tan, S. (2010). *The Correct Interpretation of Confidence Intervals*. *19*(3), 276–278.

The MathWorks Inc. (2020). MATLAB. *R2020a*. Retrieved from https://www.mathworks.com/

Tian, K., Zhuang, X., & Yu, B. (2021). The Incentive and Supervision Mechanism of Banks on Third-Party B2B Platforms in Online Supply Chain Finance Using Big Data. *Mobile Information Systems*, *2021*. https://doi.org/10.1155/2021/9943719

Vapnik, & N., V. (1995). The Nature of Statistical Learning. *Theory*. Retrieved from https://ci.nii.ac.jp/naid/10020951890

Vasarhelyi, M. A., Kogan, A., & Tuttle, B. M. (2015). Big data in accounting: An overview. *Accounting Horizons*, *29*(2), 381–396. https://doi.org/10.2308/acch-51071

Warren, J. D., Moffitt, K. C., & Byrnes, P. (2015). How big data will change accounting. *Accounting Horizons*, *29*(2), 397–407. https://doi.org/10.2308/acch-51069

Yang, S., Li, Z., Ma, Y., & Chen, X. (2018). Does Electronic Banking Really Improve Bank Performance? Evidence in China. *International Journal of Economics and Finance*, *10*(2), 82. https://doi.org/10.5539/ijef.v10n2p82

Yang, Y., Sun, L., & Guo, C. (2018). Aero-material consumption prediction based on linear regression model. *Procedia Computer Science*, *131*, 825–831. https://doi.org/10.1016/j.procs.2018.04.271

Yee, B., & Faziharudean, T. M. (2010). Factors Affecting Customer Loyalty of Using Internet Banking in Malaysia. *Journal of Electronic Banking Systems*, *2010*, 1–22. https://doi.org/10.5171/2010.592297

Zhang, J., Yang, X., & Appelbaum, D. (2015). Toward effective big data analysis in continuous auditing. *Accounting Horizons*, *29*(2), 469–476. https://doi.org/10.2308/acch-51070

APPENDICES

Appendix A Multinomial logistic regression results

A.1 Multinomial logistic regression results

Table A-1: Multinomial logistic regression (coefficients-relative log odds)

coefficients-relative log odds	**Bank 1**		**Bank 2**		**Bank 3**		**Bank 4**		**Bank 5**		**Bank 6**	
	Internet/Mobile	Call center/Mobile	Internet/Mobile	Call center/Mobile	Internet/Mobile	Call center/Mobile	Internet/Mobile	Call center/Mobile	Internet/Mobile	Call center/Mobile	Internet/Mobile	Call center/Mobile
Intercept	-0.154	-0.723	-3.010	-0.393	-2.255	-7.0164	-1.278	-2.854	0.227	-5.866	-4.164	-4.969
x1-age	0.025	0.033	0.016	0.024	0.010	0.039	0.044	0.050	5.27E-03	0.008	0.027	0.058
x2-gender	-0.156	-0.457	-0.496	-0.106	0.363	-0.915	-0.315	-0.509	0.003	-0.200	0.575	-0.927
x3-CA Count	-0.971	-1.118	-0.128	-1.061	-0.131	-0.258	-0.072	-0.091	-0.077	-0.494	0.042	0.000
X4-SA Count	-0.359	-0.388	-1.889	-2.151	-0.247	-0.276	-0.245	-0.268	-0.098	-0.486	0.009	0.002
X5-Da Count	0.623	0.505	-0.856	-2.383	0.039	0.055	0.072	0.155	0.003	-0.036	0.026	0.014
X6-Income	4.38E-05	3.58E-05	-4.15E-06	-1.20E-03	-1.12E-05	-8.52E-05	1.37E-05	-0.000	-1.62E-05	-3.48E-05	4.92E-06	-0.000
X7-CL Count	0.119	0.186	-0.001	-2.552	-0.037	0.211	-0.008	-0.056	-0.005	0.045	-0.093	0.019
x8-Ins.L Count	-0.195	0.023	0.506	-0.371	-0.042	0.375	-0.070	0.123	0.002	0.036	-0.023	0.103
x9-Credit card lim.	0.000	-6.84E-05	0.000	2.14E-04	-2.76E-05	-2.34E-05	4.35E-05	-1.53E-04	-1.49E-06	-1.49E-04	1.46E-05	0.000

coefficients-relative log odds	Bank 1		Bank 2		Bank 3		Bank 4		Bank 5		Bank 6	
x10-Credit card out.	-0.000	1.81E-05	-0.000	-4.45E-04	-1.09E-05	-1.51E-04	-6.69E-05	-8.31E-05	1.34E-05	-0.0000	9.97E-05	-0.0000
X11-Relationship duration in Y	N/A	N/A	0.043	0.264	0.057	0.244	-6.15E-03	2.02E-02	7.23E-03	5.46E-01	-3.61E-03	-1.15E-02

Table A-2: Multinomial logistic regression (coefficients-relative risk exp)

coefficients-relative risk exp (coefficients)	Bank 1		Bank 2		Bank 3		Bank 4		Bank 5		Bank 6	
	Internet/Mobile	Call center/Mobile	Internet/Mobile	Call center/Mobile	Internet/Mobile	Call center/Mobile	Internet/Mobile	Call center/Mobile	Internet/Mobile	Call center/Mobile	Internet/Mobile	Call center/Mobile
Intercept	0.857	0.485	0.049	0.674	0.104	0.000	0.278	0.057	1.256	0.002	0.015	0.006
x1-age	1.025	1.033	1.016	1.025	1.010	1.040	1.045	1.051	1.005	1.008	1.028	1.060
x2-gender	0.854	0.633	0.608	0.899	1.438	0.400	0.729	0.600	1.003	0.818	1.778	0.395
x3-CA Count	0.378	0.326	0.879	0.346	0.876	0.772	0.930	0.912	0.925	0.609	1.043	1.000
X4-SA Count	0.698	0.677	0.151	0.116	0.780	0.758	0.781	0.764	0.906	0.614	1.009	1.002
X5-Da Count	1.864	1.658	0.424	0.092	1.040	1.057	1.074	1.168	1.003	0.964	1.026	1.014
X6-Income	1.000	1.000	0.999	0.998	0.999	0.999	1.000	0.999	0.999	0.999	1.000	0.999
X7-CL Count	1.126	1.205	0.998	0.077	0.962	1.235	0.991	0.945	0.994	1.046	0.910	1.019
x8-Ins.L Count	0.822	1.023	1.658	0.689	0.958	1.455	0.932	1.131	1.002	1.037	0.976	1.109
x9-Credit card lim.	1.000	0.999	1.000	1.000	0.999	0.999	1.000	0.999	0.999	0.999	1.000	1.000

coefficients-relative risk exp (coefficients)	Bank 1		Bank 2		Bank 3		Bank 4		Bank 5		Bank 6	
x10-Credit card out.	0.999	1.000	0.999	0.999	0.999	0.999	0.999	0.999	1.000	0.999	1.000	0.999
X11-Relationship duration in Y	N/A	N/A	1.044	1.303	1.058	1.276	0.993	1.020	1.007	1.726	0.996	0.988

Table A-3: Multinomial logistic regression (p-value)

p-value	Bank 1		Bank 2		Bank 3		Bank 4		Bank 5		Bank 6	
	Internet/Mobile	Call center/Mobile	Internet/Mobile	Call center/Mobile	Internet/Mobile	Call center/Mobile	Internet/Mobile	Call center/Mobile	Internet/Mobile	Call center/Mobile	Internet/Mobile	Call center/Mobile
Intercept	2.27E-05	1.22E-85	2.20E-74	4.89E-03	0	0	0.00E+00	0.00E+00	1.25E-31	0	0.00E+00	0.00E+00
x1-age	2.84E-291	0	5.52E-07	3.26E-24	3.41509180109000e-312	0.00E+00	0.00E+00	0	2.19E-50	1.16E-09	0.00E+00	0
x2-gender	2.00E-26	2.64E-209	2.92E-20	1.29E-02	0	0	0.00E+00	0.00E+00	0.622	1.32E-12	0.00E+00	0.00E+00
x3-CA Count	0	0	0.003	4.09E-167	6.27E-105	1.42E-31	4.64E-145	6.07E-52	3.07E-16	1.97E-38	1.51E-103	9.00E-01
X4-SA Count	0	0	4.42E-50	1.66E-84	0	1.86E-218	0.00E+00	0.00E+00	7.89E-153	5.07E-178	4.90E-07	0.481
X5-Da Count	1.63E-265	1.21E-142	1.93E-02	9.67E-14	2.84E-35	2.82E-07	4.04E-246	0	4.61E-01	1.19E-01	7.40E-117	2.93E-15
X6-Income	3.16E-110	9.32E-62	7.58E-01	0.00E+00	6.64E-27	1.17E-92	5.73E-156	0	1.79E-15	0.000	0.000	0
X7-CL Count	1.46E-81	8.83E-197	9.80E-01	0.00E+00	3.29E-19	2.29E-42	1.84E-06	2.13E-51	0.068	0.001	3.09E-59	0.033

p-value	Bank 1		Bank 2		Bank 3		Bank 4		Bank 5		Bank 6	
x8-Ins.L Count	0	3.34E-11	9.89E-81	7.01E-34	5.24E-55	0.00E+00	0.00E+00	0.00E+00	1.14E-01	5.33E-10	2.54E-12	1.27E-84
x9-Credit card lim.	0	1.27E-56	8.52E-31	5.27E-10	3.10E-130	1.19E-07	0	0.00E+00	6.22E-01	9.49E-27	0.009	2.46E-98
x10-Credit card out.	2.00E-152	0.000	3.04E-92	4.09E-80	2.31E-05	4.09E-48	0.00E+00	4.17E-228	1.38E-02	1.21E-77	8.76E-55	3.87E-259
X11-Relationship duration in Y	N/A	N/A	1.85E-10	0	0	0.00E+00	6.02E-198	0.00E+00	4.51E-04	0.00E+00	1.14E-09	1.24E-37

Table A-4: Multinomial logistic regression (standard error)

Standard Error	Bank 1		Bank 2		Bank 3		Bank 4		Bank 5		Bank 6	
	Internet/Mobile	Call center/Mobile	Internet/Mobile	Call center/Mobile	Internet/Mobile	Call center/Mobile	Internet/Mobile	Call center/Mobile	Internet/Mobile	Call center/Mobile	Internet/Mobile	Call center/Mobile
Intercept	0.036	0.036	0.165	0.139	0.011	0.044	0.007	0.013	0.019	0.081	0.020	0.028
x1-age	0.000	0.000	0.003	0.002	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000
x2-gender	0.014	0.014	0.053	0.042	0.004	0.015	0.003	0.005	0.007	0.028	0.009	0.013
x3-CA Count	0.008	0.010	0.043	0.038	0.006	0.022	0.002	0.006	0.009	0.038	0.001	0.004
X4-SA Count	0.007	0.007	0.126	0.110	0.002	0.008	2.15E-03	3.10E-03	0.003	0.017	0.001	0.003
X5-Da Count	0.017	0.019	0.366	0.320	0.003	0.010	0.001	0.002	0.004	0.023	0.001	0.001
x7-Income	1.97E-06	2.16E-06	1.35E-05	2.50E-05	1.05E-06	4.17E-06	5.16E-07	2.65E-06	2.04E-06	9.20E-06	1.35E-06	6.37E-06
x6-CL Count	0.006	0.006	0.039	0.045	0.004	0.015	0.001	0.003	0.003	0.014	0.005	0.009

Standard Error	Bank 1		Bank 2		Bank 3		Bank 4		Bank 5		Bank 6	
x8-Ins.L Count	0.003	0.003	0.026	0.030	0.002	0.007	1.20E-03	1.97E-03	0.001	0.005	0.003	0.005
x9-Credit card lim.	3.59E-06	4.31E-06	3.81E-05	3.45E-05	1.14E-06	4.42E-06	6.64E-07	2.25E-06	3.02E-06	1.39E-05	5.63E-06	8.40E-06
x10-Credit card out.	4.49E-06	4.82E-06	2.74E-05	2.35E-05	2.59E-06	1.03E-05	9.19E-07	2.58E-06	5.42E-06	3.04E-05	6.39E-06	1.01E-05
X11-Relationship duration in Y	N/A	N/A	0.006	0.005	0.000	0.003	0.000	0.000	0.002	0.008	0.000	0.000

Table A-5: Multinomial logistic regression (95% confidence lower and upper limits)

95% confidence limits for the coefficients	Bank 1		Bank 2		Bank 3		Bank 4		Bank 5		Bank 6	
LL	Internet/Mobile	Call center/Mobile	Internet/Mobile	Call center/Mobile	Internet/Mobile	Call center/Mobile	Internet/Mobile	Call center/Mobile	Internet/Mobile	Call center/Mobile	Internet/Mobile	Call center/Mobile
Intercept	-0.225	-0.795	-3.334	-0.667	-2.278	-7.103	-1.292	-2.880	0.189	-6.025	-4.203	-5.025
x1-age	0.023	0.032	0.009	0.020	0.010	0.037	0.044	0.049	0.004	0.005	0.026	0.057
x2-gender	-0.185	-0.486	-0.601	-0.190	0.354	-0.946	-0.321	-0.521	-0.011	-0.256	0.557	-0.953
x3-CA Count	-0.989	-1.138	-0.214	-1.136	-0.143	-0.301	-0.077	-0.103	-0.095	-0.569	0.038	-0.008
X4-SA Count	-0.373	-0.403	-2.138	-2.367	-0.251	-0.293	-0.248	-0.273	-0.105	-0.520	0.005	-0.004
X5-Da Count	0.588	0.466	-1.574	-3.010	0.033	0.034	0.067	0.1492	-0.005	-0.081	0.023	0.010

95% confidence limits for the coefficients	Bank 1		Bank 2		Bank 3		Bank 4		Bank 5		Bank 6	
X6-Income	4.00E-05	3.15E-05	-3.06E-05	-0.001	-1.33E-05	-9.33E-05	1.27E-05	-0.000	-2.02E-05	-5.29E-05	2.27E-06	-0.000
X7-CL Count	0.107	0.174	-0.077	-2.641	-0.046	0.181	-0.012	-0.063	-0.011	0.017	-0.105	0.001
x8-Ins.L Count	-0.203	0.016	0.454	-0.431	-0.047	0.360	-0.072	0.119	-0.000	0.024	-0.029	0.093
x9-Credit card lim.	0.000	-7.68E-05	0.000	0.000	-2.98E-05	-3.21E-05	4.22E-05	-0.000	-7.40E-06	-0.000	3.56E-06	0.000
x10-Credit card out.	-0.000	8.66E-06	-0.000	-0.000	-1.60E-05	-0.000	-6.87E-05	-8.82E-05	2.72E-06	-0.000	8.71E-05	-0.000
X11-Relationship duration in Y	N/A	N/A	0.030	0.253	0.055	0.237	-0.006	0.019	0.003	0.529	-0.004	-0.013
UL	Internet/Mobile	Call center/Mobile	Internet/Mobile	Call center/Mobile	Internet/Mobile	Call center/Mobile	Internet/Mobile	Call center/Mobile	Internet/Mobile	Call center/Mobile	Internet/Mobile	Call center/Mobile
Intercept	-0.082	-0.650	-2.687	-0.119	-2.231	-6.929	-1.264	-2.827	0.266	-5.706	-4.125	-4.913
x1-age	0.026	0.034	0.022	0.029	0.011	0.041	0.045	0.050	0.005	0.011	0.028	0.060
x2-gender	-0.127	-0.428	-0.390	-0.022	0.373	-0.884	-0.309	-0.498	0.018	-0.145	0.594	-0.902
x3-CA Count	-0.953	-1.098	-0.042	-0.985	-0.119	-0.214	-0.066	-0.080	-0.058	-0.419	0.046	0.009
X4-SA Count	-0.345	-0.373	-1.640	-1.934	-0.243	-0.259	-0.243	-0.262	-0.091	-0.453	0.013	0.009
X5-Da Count	0.658	0.544	-0.139	-1.756	0.045	0.076	0.076	0.161	0.012	0.009	0.028	0.017

95% confidence limits for the coefficients	Bank 1		Bank 2		Bank 3		Bank 4		Bank 5		Bank 6	
X6-Income	4.77E-05	4.00E-05	2.23E-05	-0.001	-9.19E-06	-7.70E-05	1.47E-05	-0.000	-1.22E-05	-1.68E-05	7.58E-06	-0.000
X7-CL Count	0.131	0.198	0.075	-2.462	-0.029	0.242	-0.005	-0.048	0.000	0.072	-0.082	0.037
x8-Ins.L Count	-0.187	0.030	0.558	-0.311	-0.037	0.390	-0.067	0.126	0.005	0.047	-0.016	0.114
x9-Credit card lim.	0.000	-5.99E-05	0.000	0.000	-2.54E-05	-1.48E-05	4.48E-05	-0.000	4.43E-06	-0.000	2.56E-05	0.000
x10-Credit card out.	-0.000	2.76E-05	-0.000	-0.000	-5.88E-06	-0.000	-6.51E-05	-7.81E-05	2.40E-05	-0.000	0.000	-0.000
X11-Relationship duration in Y	N/A	N/A	0.056	0.276	0.059	0.250	-0.005	0.020	0.011	0.562	-0.002	-0.009