

CRANFIELD UNIVERSITY

Najlah Saleh AlBarrak

Modelling the Probability of Household Loan Default for
Conventional and Islamic Banking

School of Aerospace, Centre for Computational Engineering
Sciences

PhD

Academic Year: 2018 - 2021

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

CRANFIELD UNIVERSITY

School of Aerospace, Centre for Computational Engineering
Sciences

PhD

Academic Year 2018 – 2021

Najlah Saleh AlBarrak

Modelling the Probability of Household Loan Default for
Conventional and Islamic Banking

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

Credit rating risks have become a main indicator of the performance of banks. An effective credit assessment can make it easier to anticipate losses and manage it. Based on regulators, banks' credit exposure must have a risk weight that are assigned as a way of calculating the expected loss for each customer. In a country like Kuwait, whose economy is based on oil, advances in technology and increases in the amount of data available about banking customers have made it possible to develop a built-in credit default probability model that is more effective than the standard fixed risk model that is currently in use. Having a robust and fair model under the control of a central bank will increase supervision and make the internal risk weight framework more reliable, bringing it into line with regulations. Our aim in this study is to come up with an internal credit rating system that uses machine learning to calculate the risk weights for different households. We aim to create models that will be approved by the Central Bank of Kuwait and be acceptable to other banks. Our objectives are to develop a customized model for each of the banks under study; to account for different types of banking (conventional and Islamic); to produce different models for the different types of loans granted, to generate a general model for the Central Bank as well for each type of loan.

We compared different classification models for conventional and Islamic banks. The classification models were as follows: Logistic Regression, Fine Decision Tree, Linear Support Vector Machines, Kernel Naïve Bayes, Bagging, AdaBoostM1, and RUSBoosted. Ensemble models were used to classify the customers who took our household loans at the different banks and determine the likelihood that they would default. This led to the development of further mechanisms for assessing the models at the central bank, such as Gentelboost, Logitboost and Robustboost. The following were adopted as key performance measures: AUC curve, confusion matrix, running time, and k-fold. This ensured that the prediction models in this critical field were well structured.

Our findings met our expectations. The key performance measures for comparing the different models show that ensemble models are the most suitable. The

parameters selected for this study varied in importance depending on the prediction model. We found that several new parameters were significant and as influential as we expected. One of the key contributions that we made was that we used a substantial amount of data; this can be considered a contribution in itself. This was also the first time that a credit rating model was developed for Islamic banks. We provided an internal model for each of the six banks we studied, as well as a robust system for the central banks.

The results of our work show that banks should reduce the credit risk for each customer from 75% to 30% or less. The results should also enhance the role of the central banks given that it provides a robust system, with new and undiscovered variables. This system can be used to calculate the credit default probability. Therefore, there will be efficient supervision of the banks' internal systems, ensuring the reliability of the banks. This will also be considered in central banks through periodic stress testing that adds new impact factors for predicting default cases. A periodic run of the systems will help to prevent customers from defaulting unexpectedly. In conclusion, our models could be passed to other governmental financial bodies in Kuwait.

Keywords:

credit rating, credit risk, technology, conventional banking, Islamic banking, classification models, machine learning

ACKNOWLEDGEMENTS

This work would not have been possible without the financial support of my country, particularly the support of the Central Bank of Kuwait. I am especially indebted to his Excellency the Governor Dr. Mohammed Al-Hashel and the Deputy Governor Mr. Yousif Al-Obaid, who have supported my career goals and worked actively to give me enough time to pursue my goals.

I am grateful to all of those with whom I have had the pleasure of working during this and other related projects. The supervisors of my dissertation, Dr. Irene Moulitsas and Dr. Salvatore Filippone, have provided me with extensive personal and professional guidance. They have taught me a great deal about scientific research and life in general. I would like to thank them for their patience, motivation, enthusiasm, and knowledge. Their guidance helped me throughout my research and while writing my thesis. I could not have asked for better advisors and mentors during my PhD.

I would especially like to thank my colleague in work and studies Mrs. Hessa AlSanousi, who has been my teacher and mentor. She has taught me more than I could ever give her credit for. Through her example, she has shown me what a good scientist (and person) should be.

Nobody has been more important to me in the pursuit of this project than the members of my family. I would like to thank my parents Mr. Saleh AlBarrak and Mrs. Dalal AlZaid, whose love and guidance are with me in everything I pursue. I would also like to thank my brothers and sister, who have been excellent role models, providing endless inspiration.

Finally, I would like to thank my self for having faith in me and trusting God's plan. I thank me for following my dreams and seeing those dreams come true.

TABLE OF CONTENTS

ABSTRACT	i
ACKNOWLEDGEMENTS.....	iii
LIST OF FIGURES	vii
LIST OF TABLES	ix
LIST OF EQUATIONS.....	xiii
1 INTRODUCTION.....	15
1.1 Problem Background	17
1.2 Aim and Objectives	19
1.3 Publications and Conferences	25
2 LITERATURE REVIEW	26
2.1 Data in the Financial Sector.....	26
2.2 What is Big Data?	26
2.3 The Uses of Big Data in Banking	27
2.4 Big Data and Privacy	30
2.5 Summary	32
3 RESEARCH DEVELOPMENT	32
3.1 Mobile Money Transfers and Digital Banking Tools.....	32
3.1.1 The Importance and Uses of Mobile Money Transfers.....	32
3.1.2 Other Digital Banking Tools and Mobile Financial Services	38
3.1.3 Why Do Customers Use Mobile Payment Services?.....	40
3.2 Consumer Behaviour and Credit Risk.....	44
3.2.1 How Do Consumers Make Purchasing Decisions?	44
3.2.2 Economic and Non-Economic Factors that Affect Purchasing Decisions.....	48
3.2.3 Consumer Satisfaction after Purchasing Decisions.....	50
3.2.4 Understanding How Marketing Strategies Can Increase Customer Satisfaction.....	53
3.3 Share-of-Wallet and Consumer Spending as Credit Risk Parameters....	54
3.4 Developing an Understanding of Credit Risk	62
4 METHODOLOGY	64
4.1 Processing and Structuring the Data	64
4.1.1 Data Collection	65
4.1.2 Our literature review revealed that samples that range from 1,200 to 1,500 observations are sufficiently large, assuming the selection is random. However, smaller sample sizes can give good results if they are tested against the confidence interval (Holmes, Illowsky and Dean, 2017). , , , , and in Appendix A resent the raw data, including the different components and the size of the data set.Data Cleaning	66
4.1.3 The Limitations of Our Dataset.....	67
4.2 Analysing the Data.....	68

4.2.1 An Introduction to the Analysis of Big Data	68
4.2.2 Using Machine Learning to Analyse Big Data Sets	71
4.2.3 Which Machine Learning Methods Can Be Used to Analyse Financial Data Sets?	74
4.3 Machine Learning Methods Used in this Study	75
4.3.1 Logistic Regression	76
4.3.2 Decision Tree	78
4.3.3 Support Vector Machines	78
4.3.4 Bayesian Network	79
4.3.5 Bagging	80
4.3.6 AdaBoostM1	80
4.3.7 Random Under-Sampling (RUSBoost)	81
4.3.8 LogitBoost	82
4.3.9 GentleBoost	83
4.3.10 RobustBoost	83
4.3.11 Key performance measures	84
4.3.12 Code Used for Implementing Methods	86
5 RESULTS AND DISCUSSION	91
5.1 Correlation variables	91
5.1.1 Bank 1	92
5.1.2 Bank 2	94
5.1.3 Bank 3	96
5.1.4 Bank 4	97
5.1.5 Bank 5	99
5.1.6 Bank 6	100
5.1.7 All Banks	101
5.2 Models for Instalment Loans	103
5.2.1 Bank-by-Bank Models	103
5.2.2 Central Bank Model	121
5.3 Models for Consumer Loans	124
5.3.1 Bank-by-Bank Models	124
5.3.2 Central Bank Model	142
5.4 Results Overview	145
6 CONCLUSION	148
REFERENCES	152
APPENDICES	161

LIST OF FIGURES

Figure 1-1: Expected and unexpected losses	18
Figure 1-2: Gross loans by type of borrower	22
Figure 2-1: Investment in big data by industry.....	26
Figure 3-1: Research areas regarding the implementation of mobile financial services	38
Figure 3-2: C5.0 test results	43
Figure 3-3: Consumer behaviour and decision-making	44
Figure 3-4: The five stages of the consumer decision-making process.....	47
Figure 3-5: Demand curve.....	50
Figure 3-6: Factors that influence expenditures	57
Figure 3-7: Eigenvectors (factor loading) plots by segments.....	59
Figure 4-1: List of algorithms based on the type of data available.....	72
Figure 4-2: Machine learning process	73
Figure 4-3: Hypothetical ROC curve.....	84
Figure 5-1: Basic spectrum for interpreting the correlation coefficient (R).....	92
Figure 5-2: Bank 1 AUC results for 70% in-sample and 30% out-of-sample ..	105
Figure 5-3: Relative importance of the predictive variables for Bank 1.....	106
Figure 5-4: Bank 2 AUC results for 70% in-sample and 30% out-of-sample ..	108
Figure 5-5: The relative importance of the predictive variables for Bank 2	109
Figure 5-6: Bank 3 AUC results for 70% in-sample and 30% out-of-sample for (AdaBoostM1).....	112
Figure 5-7: Bank 3 AUC results for 70% in-sample and 30% out-of-sample for (Bag).....	112
Figure 5-8: Bank 4 AUC results for 70% in-sample and 30% out-of-sample (RUSBoosted)	115
Figure 5-9: Bank 4 AUC results for 70% in-sample and 30% out-of-sample (bagged)	115
Figure 5-10: The relative importance of the predictive variables for Bank 4...	116
Figure 5-11: Bank 5 AUC results for 70% in-sample and 30% out-of-sample	118
Figure 5-12: The relative importance of the predictive variables for Bank 5...	118

Figure 5-13: Bank 6 AUC results for 70% in-sample and 30% out-of-sample	121
Figure 5-14: The relative importance of the predictive variables for Bank 6 ...	121
Figure 5-15: AUC results for 70% in-sample and 30% out-of-sample (all banks)	123
Figure 5-16: The relative importance of the predictive variables for all banks	124
Figure 5-17: Bank 1 AUC results for 70% in-sample and 30% out-of-sample	126
Figure 5-18: The relative importance of the predictive variables for Bank 1 ...	127
Figure 5-19: Bank 2 AUC results for 70% in-sample and 30% out-of-sample	129
Figure 5-20: The relative importance of the predictive variables for Bank 2 ...	129
Figure 5-21: Bank 3 AUC results for 70% in-sample and 30% out-of-sample (AdaBoostM1).....	132
Figure 5-22: Bank 3 AUC results for 70% in-sample and 30% out-of-sample (bagging)	132
Figure 5-23: Bank 3 AUC results for 70% in-sample and 30% out-of-sample (RUSBoosted)	133
Figure 5-24: Bank 3 AUC results for 70% in-sample and 30% out-of-sample (decision tree).....	133
Figure 5-25: Bank 4 AUC results for 70% in-sample and 30% out-of-sample (AdaBoostM1).....	136
Figure 5-26: Bank 4 AUC results for 70% in-sample and 30% out-of-sample (bagging)	136
Figure 5-27: Bank 4 AUC results for 70% in-sample and 30% out-of-sample (RUSBoosted)	136
Figure 5-28: Bank 5 AUC results for 70% in-sample and 30% out-of-sample	138
Figure 5-29: The relative importance of the predictive variables for Bank 5 ...	139
Figure 5-30: Bank 6 AUC results for 70% in-sample and 30% out-of-sample	141
Figure 5-31: The relative importance of the predictive variables for Bank 6 ...	141
Figure 5-32: AUC results for 70% in-sample and 30% out-of-sample for all banks	144
Figure 5-33: The relative importance of the predictive variables for all the banks	144

LIST OF TABLES

Table 2-1: Guidelines for using big data.....	29
Table 2-2: Estimation results for model 2	31
Table 3-1: Regression results for Access-To-Finance status of the firms and the adoption of the mobile	34
Table 3-2: Mobile payment system uses in Brazil, Indonesia, and Kenya in 2011 (Percentage of uses by individuals aged 15 years and over)	37
Table 3-3: Survey results for m-commerce usage in Bangladesh	40
Table 3-4: Factors influencing customers' buying decisions	46
Table 3-5: Moderating factors for cross-country differences	53
Table 3-6: Variables used for estimation with statistical descriptions	56
Table 3-7: Estimates for parameters	58
Table 3-8: Prediction comparison between a focal bank and its competitors. ..	60
Table 3-9: Number of customers and expenditures in the 'platinum' segment of the focal bank, separated by model.....	61
Table 4-1: Confusion matrix for credit scoring	85
Table 5-1: Bank 1 correlation coefficients between selected variables for instalment loans.....	93
Table 5-2: Bank 1 correlation coefficients between selected variables for consumer loans	93
Table 5-3: Bank 2 correlation coefficients between selected variables for instalment loans.....	94
Table 5-4: Bank 2 correlation coefficients between selected variables for consumer loans	95
Table 5-5: Bank 3 correlation coefficients between selected variables for instalment loans.....	96
Table 5-6: Bank 3 correlation coefficients between selected variables for consumer loans	97
Table 5-7: Bank 4 correlation coefficients between selected variables for instalment loans.....	97
Table 5-8: Bank 4 correlation coefficients between selected variables for consumer loans	98

Table 5-9: Bank 5 correlation coefficients between selected variables for instalment loans.....	99
Table 5-10: Bank 5 correlation coefficients between selected variables for consumer loans	99
Table 5-11: Bank 6 correlation coefficients between selected variables for instalment loans.....	100
Table 5-12: Bank 6 correlation coefficients between selected variables for consumer loans	101
Table 5-13: Correlation coefficients between selected variables for instalment loans at all banks.....	102
Table 5-14: Correlation coefficients between selected variables for consumer loans at all banks.....	102
Table 5-15: Bank 1 observations and customers	103
Table 5-16: Bank 1 AUC results	104
Table 5-17: Bank 1 confusion matrix	104
Table 5-18: Bank 1 running time of models	104
Table 5-19: Bank 2 number of observations and customers	106
Table 5-20: Bank 2 AUC results	107
Table 5-21: Bank 2 Confusion matrix results.....	107
Table 5-22: Bank 2 running time of models	108
Table 5-23: Bank 3 number of observations and customers	109
Table 5-24: Area Under Curve results.....	110
Table 5-25: Bank 3 confusion matrix results.....	110
Table 5-26: Bank 4 running time of models	111
Table 5-27: Bank 4 number of observations and customers	113
Table 5-28: Bank 4 AUC results	113
Table 5-29: Bank 4 confusion matrix results.....	114
Table 5-30: Bank 4 running time of models	114
Table 5-31: Bank 5 number of observations and customers	116
Table 5-32: Bank 5 AUC results	117
Table 5-33: Bank 5 confusion matrix results.....	117
Table 5-34: Bank 5 running time of models	117

Table 5-35: Bank 6 number of observations and customers	119
Table 5-36: Bank 6 AUC results	119
Table 5-37: Confusion Matrix results	120
Table 5-38: Bank 6 running time of models	120
Table 5-39: Number of observations and customers for all banks.....	122
Table 5-40: AUC results for all banks.....	122
Table 5-41: Confusion matrix results for all banks.....	122
Table 5-42: Running time of models for all banks	123
Table 5-43: Bank 1 number of observations and customers	124
Table 5-44: Bank 1 AUC results	125
Table 5-45: Bank 2 confusion matrix results.....	125
Table 5-46: Bank 1 running time of models	126
Table 5-47: Bank 2 number of observations and customers	127
Table 5-48: Bank 2 AUC results	128
Table 5-49: Confusion Matrix results	128
Table 5-50: Bank 2 running time for models.....	128
Table 5-51: Bank 3 number of observations and customers	130
Table 5-52: Bank 3 AUC results	130
Table 5-53: Bank 3 confusion matrix results.....	130
Table 5-54: Bank 3 running time for models.....	131
Table 5-55: Bank 4 number of observations and customers	133
Table 5-56: Area Under Curve results.....	134
Table 5-57: Bank 4 confusion matrix results.....	134
Table 5-58: Bank 4 running time of models	135
Table 5-59: Bank 5 number of observations and customers	137
Table 5-60: Bank 5 AUC results	137
Table 5-61: Bank 5 confusion matrix results.....	137
Table 5-62: Bank 5 running time of models	138
Table 5-63: Bank 6 number of observations and customers	139

Table 5-64: Bank 6 AUC	140
Table 5-65: Bank 5 confusion matrix results.....	140
Table 5-66: Bank 5 running time models.....	140
Table 5-67: Number of observations and customers for all banks.....	142
Table 5-68: AUC results for all banks.....	142
Table 5-69: Confusion matrix results for all banks.....	143
Table 5-70: Running time for the models for all banks	143
Table 5-71: AUC results comparison for Instalment loans	145
Table 5-72: AUC results comparison for Consumer loans	146
Table A-1: Bank 1 raw data.....	161
Table A-2: Bank 2 raw data.....	167
Table A-3: Bank 3 raw data.....	170
Table A-4: Bank 4 raw data.....	172
Table A-5: Bank 5 raw data.....	174
Table A-6: Bank 6 raw data.....	175
Table A-7: Bank 1 transferred data from CSV to MATLAB.....	177
Table A-8: Bank 2 transferred data from CSV to MATLAB.....	180
Table A-9: Bank 2 transferred data from CSV to MATLAB.....	182
Table A-10: Bank 3 transferred data from CSV to MATLAB.....	184
Table A-11: Bank 5 transferred data from CSV to MATLAB.....	186
Table A-12: Bank 6 transferred data from CSV to MATLAB.....	188
Table A-13: Banks' code execution times	190

LIST OF EQUATIONS

(1-1).....	17
(3-1).....	49
(3-2).....	52
(3-3).....	52
(3-4).....	52
(4-1).....	76
(4-2).....	76
(4-3).....	77
(4-4).....	77
(4-5).....	77
(4-6).....	77
(4-7).....	78
(4-8).....	78
(4-9).....	78
(4-10).....	79
(4-11).....	79
(4-12).....	79
(4-13).....	79
(4-14).....	79
(4-15).....	80
(4-16).....	81
(4-17).....	81
(4-18).....	82
(4-19).....	82
(4-20).....	83
(4-21).....	85
(4-22).....	85
(4-23).....	85

(5-1).....	91
------------	----

1 INTRODUCTION

We live in an era when technology is advancing at an unprecedented rate. Banks are having to evolve to keep pace with this rapid growth in technology. This growth changes the way we live, make decisions, access information, calculate rates, and more.

Currently, success in business is achieved through smart customer engagement and data-driven decision-making (Badrinath, 2021). Technology and digitization have transformed the banking, finance, and insurance sectors by allowing them to make informed, real-time decisions that improve the consumer experience and lead to competitive advantage. They also allow banks to share potential products with customers, as well as upselling, cross-selling, planning strategically. With models based on Artificial Intelligence (AI), banks have an unprecedented opportunity to transform the experience of their customers.

AI helps banks to be more confident about issuing credit to those who pass the checks on their systems. Banks can use programs and algorithms to analyse all the available information about a potential borrower. They can study their credit history and assess how their wages have changed, which helps banks to determine the reliability of the client and the security of the loan.

In China, banks have already gone further than this, no longer limiting themselves simply to analysing data. They have begun to use technology to recognize people's micro-expressions, allowing them to find out whether customers are lying about their financial situation when asking for a loan. To do this, they have developed AI systems that use smartphone cameras to detect minute changes in facial expressions that are invisible to the naked eye. This helps banks to identify potential fraudsters, and they have already reduced their losses from unpaid loans by 60% (Badrinath, 2021).

In the UK, the Bank of England (BoE) and the Financial Conduct Authority (FCA) have shown a keen interest in investing how machine learning can be used by financial institutions (Bank of England, 2019). That is why they conducted a joint survey in 2019 to better understand the use of machine learning in UK financial

services. The survey was sent to almost 300 firms, including banks, credit brokers, e-money institutions, financial market infrastructure firms, investment managers, insurers, non-bank lenders, and principal trading firms. A total of 106 responses were received. The key findings from the survey were:

- Machine learning is increasingly being used in UK financial services. Two thirds of respondents reported that they already used machine learning in some form. The median firm uses live machine learning applications in two business areas, and this is expected to more than double within the next three years.
- In many cases, machine learning has passed the initial stage of development and is entering the more advanced stage of deployment. One third of machine learning applications are used for a considerable share of the activities in each business area. Deployment is most advanced in the banking and insurance sectors.
- From front office to back office, machine learning is now used across a range of business areas. machine learning is most used in anti-money laundering and fraud detection schemes, as well as in customer-facing applications (such as customer services and marketing). Some firms also use machine learning in areas like credit risk management, trade pricing and execution, general insurance pricing, and underwriting.
- Regulation is not seen as a barrier, but some firms stress the need for additional guidance on how to interpret the current regulations. The biggest reported constraints are internal to firms. They include legacy IT systems and data limitations.
- Firms did not think that machine learning was creating new risks, but they noted that it could amplify existing ones. For example, they stated that machine learning applications might not work as intended if the frameworks for validating models and managing machine learning processes could not keep pace with new, technological developments.

The banking sector in general, and in Kuwait in particular, faces several risks (credit, market, and operational). The major threat is credit risk, which is our focus

here. This is the risk of not being paid for loans that have been granted. Banks manage this risk by setting aside part of their profits as a regulatory capital to cover any risks that might occur. In calculating the portion that should be deducted from the profits to manage the credit risk, banks must multiply the risk weight of each credit transaction by its exposure.

1.1 Problem Background

The Bank for International Settlements in Basel, Switzerland, developed the Basel III regulation in 2010. This rule stipulates that banks should have enough regulatory capital to cover 8% of their Risk Weighted Assets (RWAs). Of all the RWAs at Kuwaiti banks, 60.7% of them are loans (Central Bank of Kuwait, 2020). Those loans are granted for corporations, small- and medium-sized businesses, and households. Each of these categories has its own risk rating weight. There are two types of risk rating: the standard model, in which risk weights are fixed, conservative, and very high; and the internal rating model that banks use to calculate the appropriate risk weight per customer. In Kuwait, such risk weights are calculated using the standard model. Loans are considered unrated by the top international rating agency, so their risk weight is the highest. Our aim in this study is to come up with an internal credit rating system that uses machine learning to calculate the risk weights for households. We aim to create models that will be approved by the Central Bank of Kuwait and be acceptable to the banks that use them.

To calculate the final risk weight, banks can use the following equation:

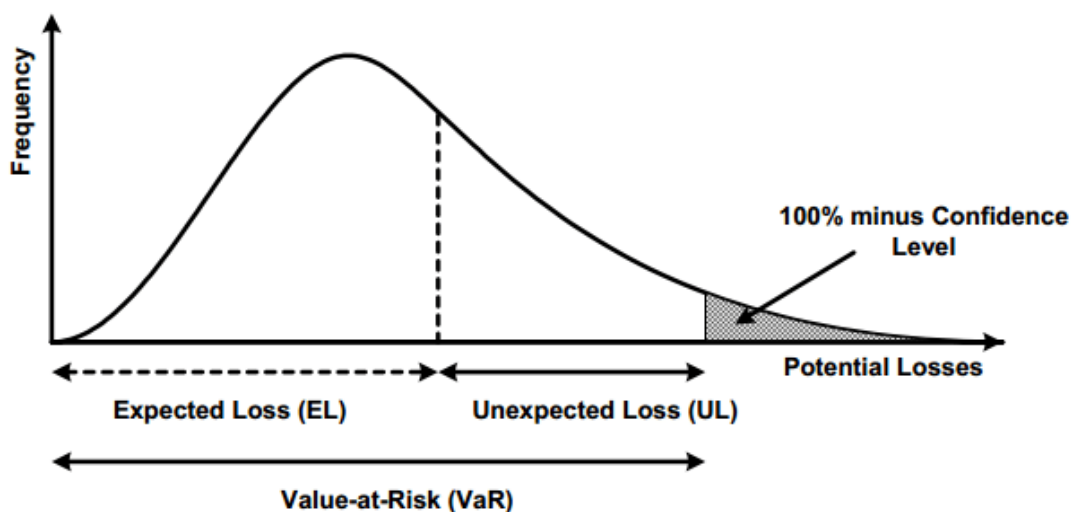
$$\begin{aligned} \text{Expected loss} \\ &= \text{Probability of Default} * \text{Exposure At Default} \quad \textbf{(1-1)} \\ &\quad * \text{Loss Given Default} \end{aligned}$$

From the probability of default, there will be categories of ratings that can be assigned specific risk weights. The main difficulty with applying Equation 1-1 is

knowing what the probability of defaulting is. Banks can manipulate this to obtain low risk weights when estimating any unexpected losses.

From Figure 1-1, we can calculate the percentage of unexpected loss, which is equal to the standard deviation of the expected loss. Expected losses are covered by the provision charges, whereas unexpected losses require capital charges to be anticipated (Basel Committee on Banking Supervision, 2005).

Figure 1-1: Expected and unexpected losses



Source: (Basel Committee on Banking Supervision, 2005: 3).

A critical problem for banks is how they should calculate the appropriate weight of the risk for each customer. As mentioned, this can be done either by using a standard approach, fixed rates, or by using an internal rating system, calculating the probability that each customer will default. Our research responds to the shortcomings of the standard approach by building a functioning internal model and adding several new parameters to previous models. These new parameters are:

- The behaviour of consumers (net cash in and out), using sizes and types of data that have never been used for this purpose before.
- The number of transactions that customers make.
- The customers' credit card exposure.
- The length of the customers' relationship with the bank.

In any case, not all the work done in this context will be suitable for implementation in Kuwait unless it is properly customized for Kuwaiti residents. This is another major gap in the literature. Our research extends the current practice by collecting a significant amount of data that spans more than ten years and covers every local bank in Kuwait.

At this point, we should emphasise that 50% of the Kuwaiti banking sector is made up of Islamic banks, and 50% of it is made up of conventional banks. The credit rating models of Islamic banks have never been studied before. Conventional banks borrow money from depositors at a low interest rate and lend them to borrowers at a high interest rate. However, interest is forbidden in Islam, so Islamic banks enter into profit-sharing arrangements with both depositors and borrowers (Qian and Velayutham, 2017). Given the nature of the credit facility granted by Islamic banks, such financing is considered sales to be paid. The intention is that they will be fully repaid in the future. Hence, the existing models are best suited to deal with 'ordinary loans', rather than with Islamic products such as 'Murabaha' and 'Tawarruq'. Our model has been customized for these Islamic products. Finally, there has been a delay in research into Islamic banks as fewer people have adapted to this kind of banking. This is the first time that data for Islamic banking has been collected on such a broad scale, which will help to facilitate future studies.

1.2 Aim and Objectives

Our aim is to develop a framework that calculates the probability that a customer will default on a loan. It is hoped that the system will meet the approval of the Central Bank of Kuwait and be convenient for other banks to use with their existing data sets.

Our objective for the work involved the production of customized prediction models for each of the banks that we studied, as well as an aggregated prediction model for the Central Bank. The research was structured as follows:

- Identify the research question

- Identify the hypotheses to be tested based on a literature review and the author's experience of the banking sector
- Determine the type and size of data to be collected
- Gather and clean the data. Almost half of the overall time was spent:
 - Importing data from CSV and Excel files to MATLAB
 - Amending and changing the codes based on popup issues that appeared while cleaning the data
 - Building structures and tables
- Apply the relevant methods and choose the most suitable method from the key performance indicators
- Produce seven models in total: six for each of the banks studied, due to data limitations, and one for the overall banking system
- Analyse the results by using the key performance parameters

The research was separated into two parts: first, we calculated the probability of defaulting for customers at each of the banks; second, we combined and aggregated all the banks' data to develop a model for the central bank. Given that we were predicting categorical binary classes (default or non-default), we chose to use classification models for our research. For each bank, we used the following models: logistic regression, decision fine tree, support vector machine, Bayesian inference, ensemble bagged trees, ensemble RUSBoost, and ensemble AdaBoostM1. The key performance indicator for the study was the area under curve (AUC) for the receiver operating characteristic (ROC) curve. AUC measures the entire two-dimensional area underneath an ROC curve from (0,0) to (1,1). From the results of our analysis and the literature review, we decided that the most successful methods were decision fine tree analysis and the three ensemble methods. For cross-validation, we used k-fold cross, which folds the data set and takes various portions of it at random to train the model and test it using the remaining instances. This helps to check whether the model works for different training samples and test samples. We used 70% of the data as our training set and 30% as our test set.

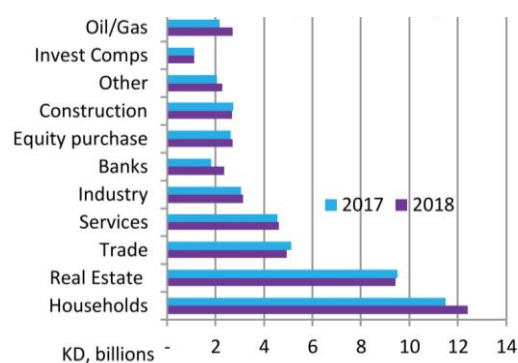
For the aggregated- and system-wise data, we performed the same comparison described above. The results from the AUC suggested that the ensemble bagged tree method was the most appropriate, followed by decision fine tree. Due to aggregation, the large number of observations, and the skewed data sets, we ran the models again and focused on almost all the ensemble methods (bagged trees, AdaBoostM1, RUSBoost, LogitBoost, GentelBoost, and Robust Boost). This was because the later methods are a combination of several approved methods. They were more effective, particularly when we compared them with the AUC results from our data. They were based on several random forest trees that had been shown to be efficient for our data set.

Given the nature of the ensemble methods (which subcategorized the data into several random groups when running the random forest tree analysis, providing the most accurate solution for each subgroup), it was necessary to test for overfitting. The overfitting of boosting techniques is a topic that is not yet theoretically understood. However, empirical results show that boosting seems to be very robust against overfitting (Lim *et al.*, 2018; Friedman *et al.*, 2004). As a contribution to the used toolbox, we automated the analysis and set the comparison of the AUC as the key performance measure. The code revealed what the best model was, based on the 70% training sample and the 30% holdout sample.

Financial crises are not a new phenomenon. They have been part of the financial world since its early stages, and there are many signs for identifying a potential financial crisis. The main obstacle is how much it costs to develop the financial market and whose responsibility it is to cover any sudden losses. Banks in Kuwait have gone through several crises, including the 2008 crisis and an impactful drop in oil prices. Kuwait stood strong, without the need for any support from regulators to back up the financial sector. This is due to the stringent regulations in place and the timely adoption of international regulations with more buffers than the internationally accepted benchmarks. Although Kuwait has overcome its financial crises, there is still room for improvement to ensure a more stable financial climate in the future. Due to the globalized nature of the economy, it is impossible

to avoid financial crises altogether, but countries can introduce safeguards to ensure they are prepared for them. After the 2008 crisis, the Bank of International Settlements emphasized the importance of having a minimum amount of capital available to cover the risk of a bank's assets and investments. The banking industry primarily operates by using raised capital and borrowed funds to lend money, profiting from the different interest rates. In other words, the main activity of a bank is lending money. Lending accounts for almost 60% of the total assets of banks in Kuwait (Central Bank of Kuwait, 2018). Moreover, the proportion of loans given to households has grown in each of the last four years. It recorded growth of 7.8% in 2018, making it the sector that benefited the most from an increase in credit lending. This is shown in Figure 1-2.

Figure 1-2: Gross loans by type of borrower



Source: Central Bank of Kuwait (2018: 5).

Household lending has grown in importance because it allows banks to distribute the risk of loan repayments to a wider range of customers. This is in contrast to big loans given to a small number of corporations. However, excessive lending in one sector might increase the concentration risk. This suggests that more work is needed to analyse the riskiness of lending in the household sector. Therefore, it is necessary to develop a better credit scoring system for rating the probability that a customer will default on a loan. This is necessary for all kinds of customers (households, small- and medium-sized enterprises, and corporations). This study uses machine learning to categorize customers with the potential to default on their household loans.

A large amount of data was solicited from ten local banks in Kuwait. Five of these banks were conventional, and five were Islamic. We requested and analysed data from 2008 to 2018. Our review of the literature revealed that previous studies had tested a very small number of parameters, including age, gender, nationality, income level, education, and number of loans. Previous studies had only looked at conventional banks. In this study, we included the following parameters alongside those that had already been studied: the length of a customer's relationship with the bank, customer credit card exposure (their outstanding balance), the average net cash flow of a customer each month, and the average number of transactions of a customer each month. We included Islamic banks alongside conventional banks.

We examined two types of loans in our study: consumer loans and instalment loans. Consumer loans are loans for personal needs and durable goods. They are capped at 25,000 KWD, or 15 times a person's salary (whichever is less). Instalment loans are loans for the maintenance or purchase of private property. They are capped at 70,000 KWD. To accommodate the differences between loans from conventional banks and contracts with Islamic banks, we combined all forms of lending as either 'consumer loans' or 'instalment loans' based on the reason why the funds were granted. This provided consistency and allowed for comparisons using definitions provided by the Central Bank of Kuwait.

Though we studied both types of loans, we chose only to share the results for instalment loans because they are far more common and are typically used in the real estate sector. Default cases are cases in which the customer fails to meet their monthly obligation to meet their loan and interest repayments for three consecutive months. This means that the customer is defaulting (that is, they are a bad customer). A customer who meets their ongoing monthly payments is non-defaulting (a good customer).

To ensure that our model could accommodate a large amount of data. Our review of the literature showed that samples with between 1,200 and 1,500 observations were considered large enough and strong enough, assuming the data selection was random. We also found that smaller sample sizes could also provide good

results when tested against the confidence intervals (Holmes, Illowsky and Dean, 2017). We calculated the confidence intervals for our models.

From our analysis of the data for Bank 2, we realised that a customer's age and the length of their relationship with the bank were the most important variables when it came to predicting the likelihood that they would default. This was followed by their number of loans and their level of education. One unanticipated finding was that neither the average net cash flow per month nor the number of transactions in the account were significant predictors for defaulting. There was only a small amount of data for Bank 3, which made it difficult to make predictions. This shows that a large data set is required. Furthermore, we found that the results from Bank 6 were different from those from the other Islamic banks. For Bank 6, a customer's number of loans was the least influential variable, and income level was the most influential variable. The new variables that we introduced (credit card exposure, length of relationship, number of monthly transactions, and net monthly cash flow) were all influential. They all had the same weight in our prediction model.

In the final step of the analysis, we combined the data from seven banks (70% of all the banks in Kuwait), including both conventional and Islamic banks. We realised discovered that when the data was consolidated, the length of a customer's relationship with a bank was the most significant variable when predicting their likelihood of defaulting on a loan. This confirmed our initial hypothesis. The second most significant variable was the number of loans. This was mainly driven by data from the Islamic banks.

The rest of this study is divided into the following chapters Chapter 2 provides a literature review of other relevant studies. Chapter 3 demonstrates how this research develops the literature. Chapter 4 discusses the different methods that were used to analyse the data. Chapter 5 summarizes the results and discusses them in relation to the literature review. Finally, Chapter 6 offers a conclusion and makes several suggestions for future research and policy decisions.

1.3 Publications and Conferences

The following research articles were published as part of this PhD work:

1. Albarrak N., Alsanousi H., Moulitsas I. & Filippone S. (2020) Using big data to compare classification models for household credit ratings in Kuwait, *International Journal of Soft Computing and Artificial Intelligence*, 8(2) 1-7.
2. Albarrak N, Alsanousi H, Moulitsas I & Filippone S (2021) Using Big Data to compare classification models for household credit rating in Kuwait. In: 6th International Congress on Information and Communication Technology, London, 25-26 February 2021.
3. Albarrak N., Alsanousi H., Moulitsas I., & Filippone S. Credit rating models for Central Banks. *International Journal of Business and Social Science*. (Accepted)

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

1. The 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. The Sixth International Congress on Information and Communication Technology, London, United Kingdom.

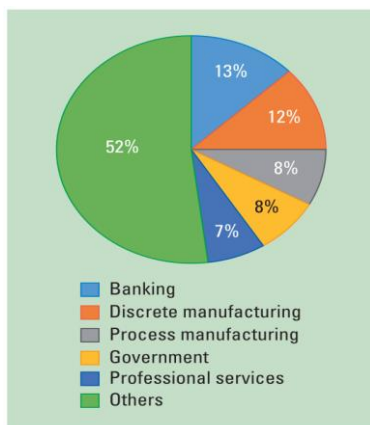
2 LITERATURE REVIEW

This chapter provides an overview of previous research into how data is used in the financial sector, focusing on the rise of big data and its effect on business. It examines the data sources, variables, and methodologies employed in other studies of prediction models in the banking sector. This lays the foundation for the account of how this study further developed the research in the field, which is outlined in the following chapter.

2.1 Data in the Financial Sector

The most important development in terms of data in the financial sector is the rise of big data. Trelewicz (2017) has demonstrated that the banking sector has invested more in big data than any other industry (Figure 2-1).

Figure 2-1: Investment in big data by industry



Source: Trelewicz (2017: 9).

2.2 What is Big Data?

According to Trelewicz (2017), big data has three characteristics: volume, velocity, and variety. The volume of big data refers to its size: in the financial sector, big data contains a large amount of information about trade and the market. The velocity of big data refers to its rapid speed: in the stock market, for example, new data is issued every second. The variety of big data refers to the different types of data that are contained in big data sets.

2.3 The Uses of Big Data in Banking

The reason for this investment is likely because big data is deemed to have several uses in banking. Kshetri (2016) has examined the role of big data in improving access to financial products for economically active, low-income families, as well as small- and medium-sized enterprises in China. An objective of this study was to investigate how formal and informal institutions can facilitate the use of big data by expanding access to financial services in China. Kshetri's methodology built on multiple case studies; it sampled organizations on a substantive rather than a statistical basis. Building on the work of previous researchers (Golder, 2000), Kshetri analysed the evidence and its sources using the standards developed by Gottschalk (1945). The author examined the case study of Alibaba. This was a very useful example because it is an example of a funding institution that is not a bank. There are several examples of financing entities that are not banks. Alibaba is the most relevant and can be used as a reference point. Alibaba has been able to use big data to gain a significant market share, leading to the development of Alibaba as a financial institution.

Kshetri's (2016) research also focused on examples of online-only banks. These online banks are e-commerce businesses that use platforms to generate big data sets. The paper outlined five aspects of big data (adding to the three characteristics outlined above): volume, velocity, variety, variability, and complexity. Furthermore, the paper showed how big data companies and financial institutions overlap in their control of the data they use. The study provided evidence that big data reduces opacity and the cost of transactions.

O'Leary (2009) has also illustrated the importance of big data in the financial sector by providing a computerised text analysis of previous publications in the field. He used the findings to identify common themes and guidelines. Using open-source software (KH Coder), O'Leary took a data-driven approach, letting 'the data speak', as Mayer-Schönberger and Cukier (2013) put it. O'Leary (2009) outlined patterns related to big data. A map of the results was created using co-occurrence maps in the KH Coder. The map revealed three themes for businesses relating to the use of big data:

- Decision-making. The studies showed that data enhances decision-making by providing managers with actionable insights. Organizations that rely on big data can gain competitive advantage. Nevertheless, to make the most of big data, companies must overcome certain challenges, such as a lack of a data-driven organizational culture, a lack of an analytics strategy, and a lack of leadership.
- Innovation. Big data leads to data-driven innovation and faster innovation cycles. It is also helpful because it is crowd-sourced and cost-efficient. The case studies revealed that big data sets could often be used for another purpose than the one they were originally collected for.
- Business value. Big data adds value to a business. Using big data can save on costs and energy, transform business performance, increase revenue, improve customer satisfaction, and make operations more transparent.

The study also revealed two themes relating to information and communications:

- Data management. Big data is difficult to manage and shows up problems that need to be fixed. Many of the cases studied had inconsistent or poor data-handling techniques. There were also occasional privacy issues with privacy and data security. Furthermore, new types of data typically require special analytical tools for examining sources like social media posts or text messages.
- New data sources. New, free data sources are emerging rapidly, but they often require a lot of processing.

Table 2-1 summarizes the above thematic areas and provides examples of research for each.

Table 2-1: Guidelines for using big data

Theme	Guidelines	Examples
Decision-making	• Embed analytics into decision-making processes. • Be prepared for organizational side effects.	Bekmamedova and Shanks (2014); Cai <i>et al.</i> (2014); Dutta and Bose (2015); Phillips-Wren and Hoskisson (2015)
Innovation	• Trust the data. • Search for new methods.	Ciulla <i>et al.</i> (2012); Amatriain (2013); Jetzek <i>et al.</i> (2014); Martinez and Walton (2014)
Business value	• Look for value in various directions; experiment with the data. • Enable business transformation with the data. • Consider secondary usage of the data.	Amatriain (2013); O'Leary (2013a); Prescott (2014); Bettencourt-Silva <i>et al.</i> (2015); Dutta and Bose (2015)
Data management	• Expect to face technical and data-related challenges. • Plan for security.	Shen and Varvel (2013); Halamka (2014); Krumeich <i>et al.</i> (2014); Dutta and Bose (2015); Prinsloo <i>et al.</i> (2015)
New data sources	• Experiment with new data types. • Consider potential privacy issues.	He <i>et al.</i> (2013); Yu <i>et al.</i> (2014); Marine-Roig and Clavé (2015)

Source: O'Leary (2009: 29).

The following Table 2-2 synthesize how banks use big data

Table 2-2: Big data utilization in banking

Theme	Uses	Examples
Analysis of clients' incomes and expenditures	• Banks make informed decisions about the possibility of credit extensions, assess the risks. • Consider whether the client is interested in benefits or investments.	Tekaya <i>et al.</i> (2020), Soltani Delgosha, Hajiheydari and Fahimi (2021),
Risk assessment and fraud prevention	• Data Science in banking can be used to assess risks when trading stocks or when checking the creditworthiness of a loan applicant. • Big Data analysis also helps banks cope with processes that require compliance	Tekaya <i>et al.</i> (2020), Dagilienė and Klovienė (2019), Shabani, Munir and Mohanty (2022)

	verification, auditing, and reporting.	
Feedback management to increase customer loyalty	Better understand customer needs and make accurate decisions. Hence they can be more efficient and prompt in responding to market demands	Chau and Xu (2012), Chen, Chiang and Storey (2012), Almoqren and Altayar (2016)

2.4 Big Data and Privacy

Given the large amount of data collected from customers by financial institutions, privacy and security are important concerns. Lim *et. al.* (2018) discussed this in a paper on the perspectives of consumers in South Korea regarding data security. The paper used a quantitative survey to examine how people valued their posts on personal social networking services (SNSs) in monetary terms, comparing this with personal information. The test was conducted using two models:

1. The model assumed a homogenous preference structure for personal information leakage situations, using a binary logit model.
2. The authors estimated that consumer preferences consisted of two classes by using a latent class model (LCM).

The statistical outcomes showed that the second model was more appropriate for estimating consumer preferences. Therefore, two hypothetical scenarios were created relating to the leakage of personal data. In line with the two scenarios, an increase of compensation was detected and gradually increased to examine respondents' willingness to accept the scenario (WTA in Table 2-2). The results show that consumers were more sensitive to the leakage of data types. This is shown in Table 2-3.

Table 2-3: Estimation results for model 2

Class	Variables	Coefficients	Standard error	RI	WTA
Class 1	A. Basic personal information	−2.59197***	0.300504	16.76%	KRW 666,586
	B. Purchase list and payment information	−2.39647***	0.236968	15.50%	KRW 616,310
	C. Personal medical information	−2.41402***	0.298495	15.61%	KRW 620,824
	D. Personal Internet search history	−1.3958***	0.196882	9.03%	KRW 358,964
	E. Personal SNS posts	−1.46718***	0.121413	9.49%	KRW 377,321
	F. Location information	−1.70052***	0.210901	11.00%	KRW 437,329
	G. Compensation for personal information leakage	0.038884***	0.004205	22.63%	—
Class 2	A. Basic personal information	−0.67246***	0.074399	12.60%	KRW 184,541
	B. Purchase list and payment information	−0.37825***	0.07748	7.09%	KRW 103,802
	C. Personal medical information	−0.12706*	0.076443	2.38%	KRW 34,868
	D. Personal Internet search history	−0.14541*	0.076517	2.72%	KRW 39,903
	E. Personal SNS posts	−0.25522***	0.06958	4.78%	KRW 70,040
	F. Location information	−0.47952***	0.074104	8.98%	KRW 131,593
	G. Compensation for personal information leakage	0.036439***	0.001348	61.44%	—

Source: Lim *et al.* (2018: 66)

The results of both models show that consumers:

- Placed a high value on basic personal information, purchase lists, and payment information.
- Placed a medium value on locational information.
- Placed a low value on internet search results and SNS posts.

One can infer from this that customers are concerned about their personal data being leaked. The paper made the following suggestions for the Korean government:

1. The compensation provided by the courts should be increased.
2. Laws should be introduced to protect people's data being leaked or abused, particularly in the healthcare sector.
3. The differences in the services people use and the way they use them should be considered when developing data protection regulations.

In our research, we took privacy into consideration. All the data gathered about the customers were anonymised. The data was encrypted in an external hard drive. The passwords for the encrypted data was sent differently through our work emails. We worked of the network during the programming part to insure no data leakage.

2.5 Summary

This chapter has provided a brief overview of the understanding and uses of big data in the banking sector, demonstrating its importance in a rapidly evolving business world. It has also discussed the importance of data privacy. By looking at other studies that have used big data sets, the following chapter develops the research further by looking at digital banking tools and customer purchasing decisions. It develops a new theoretical framework for understanding purchasing decisions that relates to the study's wider aim of analysing credit risk in banks that use digital platforms.

3 RESEARCH DEVELOPMENT

This chapter focuses specifically on previous research into mobile money transfers, digital banking tools, and consumer purchasing decisions. Mobile money transfers and digital banking are two of the main reasons why big data has become so important in the banking sector. Data from mobile money transfers and digital banking can be used to understand how and why consumers use digital banking tools and engage in certain purchasing decisions. This chapter develops previous research in the field to create a theoretical framework for the study's methodology and its approach to analysing data, both of which are outlined in Chapter 4.

3.1 Mobile Money Transfers and Digital Banking Tools

3.1.1 The Importance and Uses of Mobile Money Transfers

One of the areas in which big data has been useful has been in tracking the data from mobile money transfers. Mobile money transfers have gained momentum in recent years because they allow for fast, cheap transactions (Gosavi, 2018). Businesses use mobile transfers to pay suppliers and employees, and customers use them to pay for purchases.

Recent studies have shown how mobile money transfers have changed the business models in Kenya, Tanzania, Uganda, and Zambia, as well as in other

places around the world (Gosavi, 2018). Banks have sensed the potential for profit, but they have not been able to adapt quickly due to heavy regulations and geographical limitations. In response, therefore, banks have teamed up with mobile money companies to benefit from how the latter have expanded their capital and charged interest on their customers' accounts. Gosavi (2018) tested whether the adoption of mobile money solutions would help make finance more accessible. He measured the status of customers' relationship with their bank, categorizing people based on whether they had a bank account and a credit line, a bank account only, or neither. He used the World Bank's Enterprise Surveys as his source of information. He examined whether the firms who provided access to financial services were more productive.

Gosavi's (2018) survey showed that mobile money transfers were used because they were much faster than other methods, they satisfied customers' needs, and they reduced costs. The survey showed that small- and medium-sized businesses did not tend to have lines of credit and relied on their own capital for purchases and expansions. Alternatively, these small- and medium-sized businesses would use trade credit, which is short term and can be used to finance long-term investment, creating liquidity risks. Mobile money is popular because it details firms' historical transactions, substituting for the audits required by banks when granting loans or lines of credit. The results of the regression show a positive correlation between firms with mobile money accounts and the likelihood of having a loan or a line of credit. The results were more significant among suppliers than customers. Table 3-1 displays the results of the regression analysis.

Table 3-1: Regression results for Access-To-Finance status of the firms and the adoption of the mobile

Measurement of access-to-finance status of the firms	Usage of mobile money by all firms Column 1	Usage of mobile money by suppliers Column 2	Usage of mobile money by customers Column 3
Independent Variables	Regression Coefficients	Regression Coefficients	Regression Coefficients
Use of mobile money for financial transactions	0.236 (4.25)**	0.317 (4.62)**	0.238 (3.92)**
Firm-level-control variables			
Firm age (log age)	0.077 (2.28)*	0.073 (2.17)*	0.079 (2.33)*
Firm size [dummy, the omitted size is small firms (5 to 19 workers)]			
Firm size: medium (20 to 99 workers)	0.279 (4.70)**	0.280 (4.71)**	0.280 (4.71)**
Firm size: large (100 or more workers)	0.519 (5.44)**	0.527 (5.51)**	0.529 (5.52)**
Sector type [dummy, the omitted sector is services sector]			
Sector type: manufacturing	−0.005 (0.08)	−0.011 (0.16)	−0.012 (0.18)
Sector type: retail	−0.014 (0.19)	−0.016 (0.22)	−0.019 (0.27)
Export [dummy]	0.073 (1.20)	0.070 (1.15)	0.073 (1.21)
Audited accounts [dummy]	0.453 (7.86)**	0.448 (7.78)**	0.447 (7.77)**
Firm's legal status [dummy, the omitted type is limited liability]			
Sole proprietorship firm	−0.053 (0.86)	−0.064 (1.04)	−0.056 (0.90)
Partnership firms	−0.271 (3.31)**	−0.281 (3.43)**	−0.274 (3.35)**
Country dummies	Yes	Yes	Yes
Pseudo R-squared	0.0889	0.0897	0.0883
Number of observations	2,451	2,451	2,451

Source: Gosavi (2018: 355)

The regression results for the second model indicate that firms that had mobile money accounts had a more productive labour force. This might be because easy access to funds affects the productivity of the labour force. From those results of this study, we can infer that easy access of funds affect the productivity behaviour of labour. That means that they will care to work harder because access of funds is easier. Which means spending is expected to be more volatile which may effect meeting credit obligations.

Relatedly, in a study of three mobile payment projects in Brazil, Kenya, and Indonesia, Iman (2018) questioned whether it was still necessary to develop mobile payment systems, given that so many of them have failed. The author proposed several solutions to the current problems with the complexity of mobile payment systems:

1. Ensure that there is forwards and backwards integration in the supply chain and encourage large players to increase their domination of the

market. This promotes independence and means that businesses are not reliant on the supply chain.

2. Rely on national operators like telecommunications firms to provide the mobile payment services.
3. Note that openness may be positive or negative. If the number of entities in the supply chain declines, this might lead to a monopoly and high prices. However, on the positive side, this might lead to greater liberty.

The paper also noted that the following factors can increase the growth of mobile payment systems:

1. Socioeconomic conditions. There is no alternative to cash as an accepted medium. People in emerging markets do not have access to bank accounts, credit cards, or debit cards. Cash is the preferred mechanism to avoid theft.
2. Cost efficiency. Mobile systems make expansion easier because branchless banks can avoid fixed expenses. Thus, the fees for money transfer services in South Africa, for example, can be reduced from \$30-\$50 per transaction to \$0.05 per transaction.
3. The diffusion of mobile phones. The availability of mobile phones and mobile services is an important factor.
4. Convenience. In developed countries, it is more convenient to carry out transactions online than to travel to a branch.
5. New initiatives. Organisations must support the new services.

The factors that limit the growth of mobile payments are as follows:

1. Heavy regulations and restrictions: banks still want to play the role for all payments which limits the spread of such technology given the high existent regulations on banks.
2. Limited collaboration: it's not easy to transfer mobile payments to cash as it took 5 years for "M-PESA" to convince commercial banks to withdraw money from ATMs

3. An underdeveloped ecosystem: currently the status of mobile payment infrastructure is not ready for expansion, minimum mobile features, absence of standards
4. Security problems: developing countries are prone to cybercriminals due to poor frameworks and enforcement as well as customers lack knowledge and pay less attention on security measures.

To implement a mobile payment system successfully, the following points must be taken into consideration:

1. Usability and simplicity. The system must be user-friendly, capable of being personalized, and easy to use.
2. Universality. The system must be able to perform all types of transactions and be suitable for all users.
3. Interoperability. The system should be flexible and smooth for all uses on all forms of hardware.
4. Cost and speed. The system should reduce the cost and increase the speed of transactions.
5. Integration. The system should be connected to existing banking systems.
6. Security, trust, and privacy. The system should be secure so that users feel confident entering their bank details on the platform.
7. Local market understanding. The system should understand and meet the needs of the local market.
8. Cross-border payments. To make the system appeal to users, ensure that cross-border transactions are smooth and seamless.

The study analysed quantitative data from formal databases and qualitative data from interviews, company reports, and articles. Uptake of mobile payment systems had been strong in Kenya, moderate in Indonesia, and low in Brazil respectively. Iman (2018) illustrated the usage of mobile payment systems in the three countries. The thing that will explain the impact of easy access to funds on volatility in spending thus burdens in meeting credit obligations as described in Table 3-2.

**Table 3-2: Mobile payment system uses in Brazil, Indonesia, and Kenya in 2011
(Percentage of uses by individuals aged 15 years and over)**

Usage Categories	Brazil	Indonesia	Kenya
Receive money	0.83	0.62	66.65
Receive money, male	1.21	0.80	71.59
Receive money, female	0.47	0.45	61.92
Receive money, rural	0.00	0.10	65.13
Receive money, young adults (15–24)	0.79	0.00	62.07
Receive money, young adults (25 +)	0.84	0.83	69.29
Receive money, income, poorest 2/5ths	0.31	0.00	53.24
Receive money, income, richest 3/5ths	1.18	1.04	75.65
Receive money, primary education or less	0.35	0.18	49.60
Receive money, secondary education or more	1.27	1.10	81.90
Pay bills	1.25	0.23	13.43
Pay bills, male	1.44	0.10	16.67
Pay bills, female	1.07	0.35	10.33
Pay bills, rural	0.00	0.19	12.83
Pay bills, young adults (15–24)	0.79	0.00	11.37
Pay bills, older adults (25 +)	1.40	0.30	14.62
Pay bills, income, poorest 2/5ths	0.72	0.18	4.41
Pay bills, income, richest 3/5ths	1.62	0.26	19.49
Pay bills, primary education or less	0.76	0.00	5.11
Pay bills, secondary education or more	1.70	0.47	20.88
Send money	0.00	0.57	60.48
Send money, male	0.00	0.69	69.16
Send money, female	0.00	0.45	52.16
Send money, rural	0.00	0.26	58.69
Send money, young adults (15–24)	0.00	0.00	57.25
Send money, older adults (25 +)	0.00	0.76	62.33
Send money, income, poorest 2/5ths	0.00	0.00	43.27
Send money, income, richest 3/5ths	0.00	0.95	72.03
Send money, primary education or less	0.00	0.00	41.32
Send money, secondary education or more	0.00	1.19	77.60

Source: Iman (2018: 79)

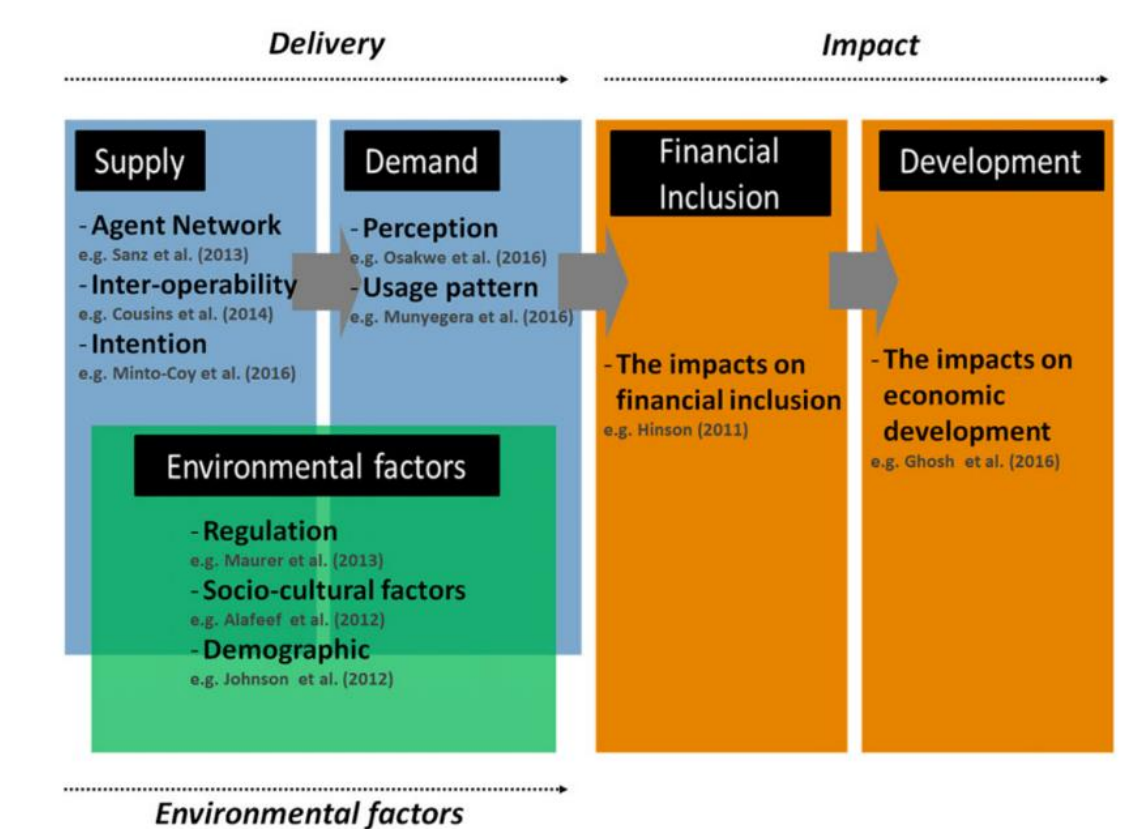
There are several reasons why Brazil, Indonesia, and Kenya developed mobile payment systems. Brazil aimed to provide money solutions for unbanked citizens. Indonesia wanted to move from a cash-based to a cashless system. Kenya wanted to increase financial inclusion. The case studies show that there are challenges when it comes to adopting mobile payment systems, such as: commercial challenges (managing costs and profits), technical challenges (finding technical solutions), legal challenges (laws and regulations), and market challenges (reaching a high percentage of the market). Gosavi (2018) stated that mobile payment systems should be operated by mobile companies rather than banks. Mobile payment systems cannot develop in the highly regulated environment that banks operate in.

3.1.2 Other Digital Banking Tools and Mobile Financial Services

Other tools that could help with mobile banking are Radio Frequency Identification (RFI) and Near Field Communication (NFC) (Rahman and Sloan, 2015). These could reduce the use of handwriting and relay information through barcodes. The paperwork for a loan could be assigned a barcode and then shared with different departments. This would make the overall process more efficient.

Kim *et al.* (2018) provided a review of the literature relating to mobile financial services, identifying areas for future research. The review showed that there were three general areas of interest in the field: delivery, environmental factors, and the impact of mobile financial services. The study showed that mobility (connecting people cross countries) and flexibility (capable of being used with multiple operating systems) were features that enhanced the development of mobile financial systems. Mobile systems help to provide financial inclusion for more people, regardless of their income or location. This can lead to economic development and attract investment opportunities from inside or outside the country. Figure 3-1 shows how the research areas were categorized.

Figure 3-1: Research areas regarding the implementation of mobile financial services



Source: Kim *et al.* (2018: 7)

By applying Heeks' (2014) model, this study showed that research into mobile financial services has focused primarily on countries' readiness to adopt financial services. Other topics, such as availability, uptake, and impact, have been discussed to a lesser extent. More research is needed in areas such as business models, demand, and the benefits and risks of mobile financial services. The paper also noted that the supply of mobile financial services was typically analysed qualitatively, whereas the demand was typically analysed quantitatively. This suggests the need for future qualitative and quantitative studies in each of these respective areas. In our research, we have included both type of data, quantitative and qualitative, for predicting customers' risk weight. The quantitative data are the monthly cash movements in the accounts and monthly number of transactions made. The qualitative data is the business relationship duration between the customer and the bank.

3.1.3 Why Do Customers Use Mobile Payment Services?

To understand the uptake of mobile money services and digital banking, it is important to recognize people's understanding of these services. Rahman and Sloan (2015) examined people's perception of the advantages and disadvantages of mobile commerce (m-commerce) in Bangladesh. They used surveys, face-to-face meetings, accounts of the infrastructure of m-commerce, and secondary resources. The results of their survey are detailed in Table 3-3.

Table 3-3: Survey results for m-commerce usage in Bangladesh

	Female		Male		Total	
	Count	%	Count	%	Count	%
I used mobile phone for m-Railway ticketing	29	18%	94	22.6%	123	21.4%
I used mobile phone for m-Billing	38	23.7%	126	30.3%	164	28.5%
I used mobile phone for local Remittance	11	7%	48	11.5%	59	10.2%
I used mobile phone for foreign	6	3.7%	29	7%	35	6%

Remittance						
I used mobile phone for m-banking	12	7.5%	48	11.5%	60	10.4%
I used mobile phone for balance transfer	62	38.7%	171	41.2%	233	40.5%
I used mobile phone for listening radio	132	82.5%	329	79.2%	461	80%
I used mobile phone for updated news	68	42.5%	188	45.3%	256	44.5%
I used mobile phone for internet modem	43	26.8%	126	30.3%	169	29.3%
I used mobile phone for my mobile phone	101	63%	246	59.3%	347	60.3%
I used mobile phone for facebook	85	53%	188	45.3%	273	47.4%
I used mobile phone for email	62	38.7%	146	35.2%	208	36%
I used mobile phone for reading newspaper	44	27.5%	110	26.5%	154	26.8%

I used mobile phone for ringtones download	66	41.2	206	49.6	272	47.3
I used mobile phone for video games	22	13.7	106	25.5	128	22.2
I used mobile phone for videos	24	15	104	25	128	22.2

Source: Rahman and Sloan (2015: 9-10).

As part of the survey, the authors had the opportunity to ask face-to-face questions about m-commerce, such as why a person did or did not intend to use m-commerce. The authors grouped the responses into three categories:

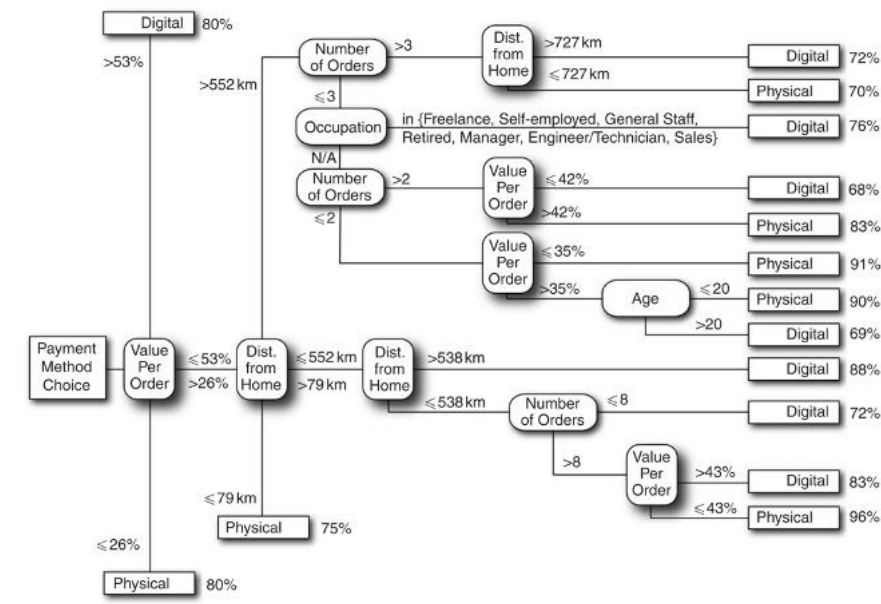
- Lack of trust. This factor was the major barrier for m-commerce adoption. There was an inverse relationship between trust and risk; the higher the risk, the lower the trust.

- Lack of awareness. Though m-commerce is well advertised across different media, adoption remains low.
- Lack of literacy. Illiterate people may struggle with e-commerce. Though they may favour using e-commerce rather than paper transactions. The paper proposed a solution for illiterate people by using a customer care point (CCP). Similarly, an interactive voice recorder (IVR) might help illiterate people in mobile banking. However, it is important to take into consideration that a machine will respond, and it might be difficult to distinguish between customer accents.

Addressing these obstacles for adopting m-commerce, will promote further purchases, nevertheless, with easy access of funds. Thus, may influence the credit obligation repayments instead. It is important to understand the line between substituting physical purchases to online/m-commerce. Adding more temptation for purchasing above the need through online availability may affect credit obligations.

Focusing on China, rather than Bangladesh, Lloyd *et al.* (2013) described the responses to the digital world in China by using the term 'digital divide.' Customer data shows that there are demographics who are capable of paying electronically but who are reluctant to do so. The paper tries to explain why this is the case by using data mining but fails to do so. It fails to provide policy suggestions for how to fix the problem. The author used C5.0, which is thought to be the most important data mining algorithm (Wu *et al.*, 2008). It has been used for supply chain management (Emerson, Zhou and Piramuthu, 2009), economic modelling (Chen, Chi and Wang, 2015) detecting fraud (Song *et al.*, 2014), and pricing in financial markets (Basti, Kuzey and Delen, 2015). C5.0 can be used to interpret the how much of the behaviour is explained by the attributes. The operation is repeated until the sample cannot be divided any further. The test results of the C5.0 analysis are shown in Figure 3-2.

Figure 3-2: C5.0 test results



Source: (Lloyd, Antonioletti and Sloan, 2013: 266)

From the analysis, four general rules were established:

1. When a shop is within 79 kilometres from a consumer's home address, the consumer will usually pay using a physical payment.
2. When a shop is greater than 79 kilometres but less than or equal to 552 kilometres from a consumer's home address, the consumer will tend to favour digital payments unless the number of items is high and the value per order is low.
3. When a shop is greater than 552 kilometres from a consumer's home address and the number of orders is greater than three, the consumer will tend to prefer a digital form of payment.
4. When a shop is greater than 552 kilometres from a consumer's home address and the number of orders is less than three, the consumer will tend to prefer a digital form of payment. This is also true for young customers, who prefer physical payments for high-value orders.

It was possible to make the following assumptions from the data:

1. Customers generally prefer to use digital payments for high-value orders.

2. Customers generally prefer to use physical payments for low-value orders.
3. Customers who live close to the shop they are visiting tend to prefer physical payments.

Therefore, it is important to distinguish between customers who are willing but unable to use digital payments and customers who are not willing to use digital payments. This will help governments to design effective legislation. One of the reasons for the delay in adopting digital payments is poor legislation. So, people might want to adopt this measure, but they are unable to. However, laws might also be too advanced, meaning that people are unwilling to adopt the measure. It is therefore necessary to explore why governments delay adopting digital payment methods. It is also important to understand consumer behaviour and its impact on the probability that consumers will default on their loans.

3.2 Consumer Behaviour and Credit Risk

To understand the relationship between customer behaviour and credit risk, it is first necessary to understand how customers make purchasing decisions.

3.2.1 How Do Consumers Make Purchasing Decisions?

Stejskal and Stavkova (2011) have examined the factors that affect the behaviour of consumers when making a purchasing decision. Figure 3-3 summarises consumer behaviour during the decision-making process.

Figure 3-3: Consumer behaviour and decision-making



Source: (Stejskal and Stavkova, 2011: 45)

The authors conducted a survey of 2,000 residents in the Czech Republic, leading to a total of 1,418 accurate responses. The factors chosen to identify consumers were age, social group, occupation, and level of education. The factors that were found to influence purchasing decisions were:

1. Consumer buying habits
2. Need
3. Previous experience with similar products
4. Recommendations and word-of-mouth advice from relatives
5. Recommendations from experts
6. Characteristics of previously bought products
7. Quality
8. Trademark
9. Price
10. Discount and promotions
11. Trying new things
12. Fashion trends
13. Product design

A detailed breakdown of the influencing factors for each of the variables is provided in Table **3-4**.

Table 3-4: Factors influencing customers' buying decisions

Commo- dities / Factors	neces.of need	prev. exp.	recom. of relatives	recom. of experts	char., paramet.	quality	trademark	price	discount action	design
Groceries and non- alcoholic beverages	9.02	7.57	6.33	5.25	7.07	8.06	5.60	7.32	6.75	4.78
Alcoholic beverages, tobacco	4.40	6.48	5.43	4.10	6.13	7.28	6.55	6.65	6.11	3.83
Clothing, footwear	7.87	7.00	5.09	4.53	8.14	8.55	6.65	7.63	7.07	8.05
Healthcare	8.39	7.68	6.49	7.33	7.87	8.30	5.42	5.80	5.03	3.09
Transport	8.08	7.04	5.26	5.04	7.19	7.33	5.69	7.58	5.87	5.38
Mailing and telecom.	7.29	6.59	5.21	4.66	6.38	6.76	4.91	6.96	5.54	4.14
Recreation, culture, sports	5.48	7.39	7.05	5.19	7.37	7.92	5.18	7.70	6.74	4.74
Education	7.31	6.58	6.45	6.11	7.27	8.19	5.50	6.48	5.25	3.76
Alimenta- tion, acco- modation	7.71	7.65	6.67	5.02	7.49	8.02	5.27	7.29	5.94	4.99
Other goods and services	6.66	6.72	6.14	5.31	7.01	7.50	5.44	7.27	6.41	5.70
Furnishing	7.30	6.67	5.66	5.36	8.23	8.28	6.02	7.80	6.80	7.78

Source: (Stejskal and Stavkova, 2011: 48)

Gray shaded cells under each column represent the factors that had the most influence on purchasing the goods in each row. In general, need, previous experience, characteristics of previous products, quality, and price were the main factors that influenced purchasing decisions. However, there are some outliers that influenced the purchase of products.

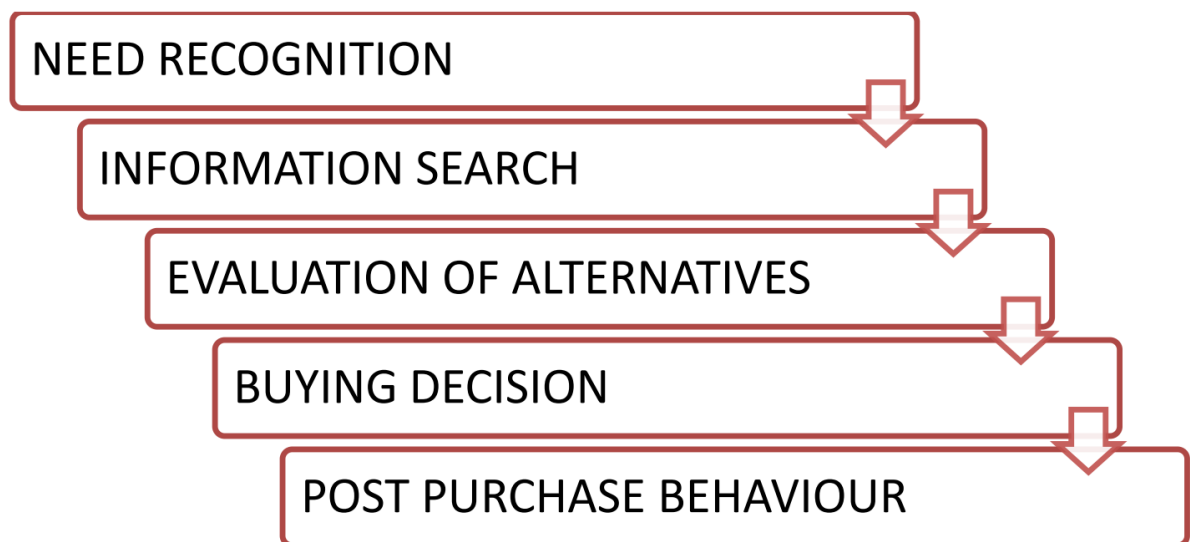
Singh *et al.* (2014) studied the consumer purchasing process and the factors that influence consumer decision-making. They provided a review of the relevant literature. Kotler and Keller (2011) explained that the purchasing process and the exchange of goods and services is intended to satisfy the needs and wants of customers. The paper listed four major factors that influence consumer buying behaviour:

- Cultural factors. The culture, social class, and inherited values of a consumer will influence their decision-making.

- Social factors. Family members, friends, and others will influence people's decisions directly or indirectly.
- Personal factors. Demographics and lifestyle will impact decisions. Such factors (which include age, interests, and opinions) will change over time.
- Psychological factors: external and internal factors based on Maslow's theory of needs (Mcleod, 2007) and including motivations, beliefs, and attitudes.

After listing the factors that influence the purchasing behaviour of consumers, the paper listed five stages that every consumer goes through in their decision-making process (see Figure 3-4).

Figure 3-4: The five stages of the consumer decision-making process



Source: (Singh, Dhayal and Shamim, 2014: 18)

Wu and Chen (2000) found that the time it takes to make a purchase can be extended, making it easy to estimate. Their paper showed that it is important to consider the period before and after the transaction process. Particularly when it comes to high value customers, it is important for companies to consider how their customers will learn after making a purchase. This is known as ensuring easy 'entry' and 'exit' to the market. Juster (1966) claimed that surveys of consumers' intention to buy cannot be used as predictors of purchase rates

because they fail to account for the probability of a purchase. They only provide information about intentions. Similarly, Lawrence (1966) stated that each individual has his or her own pattern for making decisions in a market that contains multiple brands. Shepherd *et al.* (2005) studied how organic foods influence consumer decision-making. They found that the health benefits of food had more influence on people than the environmental benefits. Additional factor that its impact may affect credit obligations. They concluded that purchasing behaviour has a greater influence on a person's purchasing habits than their specific belief system. Broderick *et al.* (2007) proposed consumer involvement as an input to account for behavioural homogeneity in market segments. They reached the conclusion that consumer involvement influences purchasing behaviour. Vitell *et al.* (2006) stated that religiosity and financial ethics have a significant impact on consumers. Kohly *et al.* (2004) examined online consumer behaviour and found that saving money and time increased consumer satisfactions. Carrington *et al.* (2010) found that ethically-minded consumers rarely have ethical products on their shopping list. Therefore, credit obligation repayments may be affected although you are ethically-minded consumer. Pilati, *et al.* (1995) stated that in situations where there was limited information, price could be used to predict the nature of people's purchasing behaviour, as well as the overall cashflow. These papers show the importance of studying consumer behaviour in the purchasing process. They demonstrate why it is difficult to have an exact understanding of why people make purchasing decisions, particularly since behaviours change continuously and affect people's decisions.

3.2.2 Economic and Non-Economic Factors that Affect Purchasing Decisions

Another reason for the difficulty of understanding purchasing decisions is the combination of economic and non-economic factors that goes into any purchasing decision. Chitnis and Hunt (2011) studied the factors that affect the expenditure of UK consumers across 12 categories of goods and services. The study examined the economic factors (price and income) and non-economic factors (technical progress, consumer taste, preferences, socio-demographic, geographic, lifestyle, and values) in every quarter from 1964 to 2006, using a

structural time-series model. The study showed that policymakers should consider economic and non-economic factors to help move towards sustainable consumption.

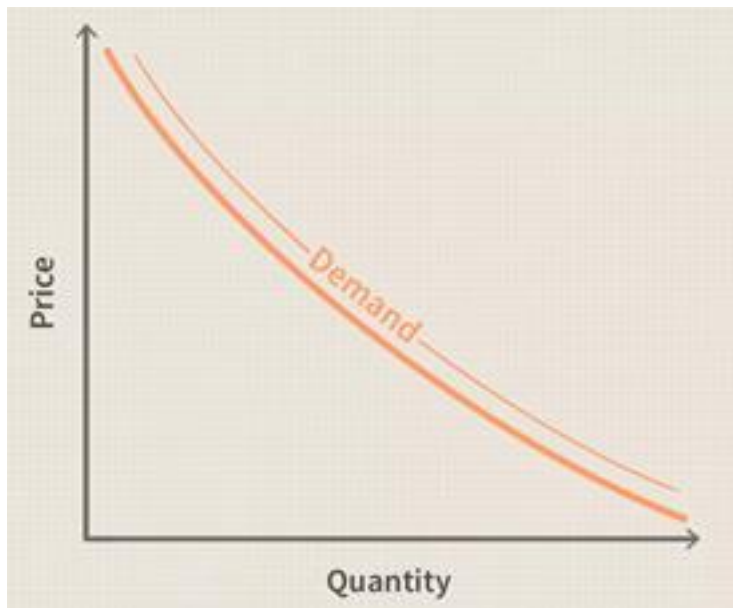
There have been several studies that have examined economic factors but few that have examined non-economic factors. For example, Selvanathan and Selvanathan (2004) examined the demand for 11 of the 12 product categories in South Africa without including any non-economic factors. They used the Almost Ideal Demand System. Similarly, Attfield (2005) used the same model to study 11 of the 12 product categories in the UK. They examined demographic and income factors, but non-economic factors were not included. Lula and Antille (2007) studied Swiss expenditures for 12 products, but there was little focus on non-economic factors. The authors used a regression model with the following equation (3-1):

$$\begin{aligned} exp_t - exp_{t-4} = & D'(L)(\hat{\mu}_t - \hat{\mu}_{t-4}) + E'(L)(\hat{\lambda}_t - \hat{\lambda}_{t-4}) + \hat{B}'(L)(p_t - p_{t-4}) \quad (3-1) \\ & + \hat{C}'(L)(y_t - y_{t-4}) + F'(\hat{L})(\hat{v}_t - \hat{v}_{t-4}) \end{aligned}$$

With this equation, Chitnis and Hunt (2011) calculated the impact of economic variables (income and price) and exogenous non-economic variables for defining the UK's annual household expenditure. They found that there is inelasticity for price and income in either the short term or the long term, except for when it comes to recreation and culture, for which income is elastic in the long term. They found that the expenditure trends were nonlinear, so there was no trend determining where and when price and income might stabilize as the demand curve shifts to the right in

Figure 3-5.

Figure 3-5: Demand curve



Source: (Kenton, 2021)

In this curve, less is being spent on alcoholic drinks, tobacco, and narcotics. There is a stochastic trend for expenditures on clothing, footwear, and house repairs. There were no seasonal changes except for communication services and education. In general, the results showed that exogenous non-economic factors have a significant impact on expenditure. Policymakers could manage expenditure by increasing or decreasing the prices or taxes for certain products, while also accounting for the exogenous non-economic factors that have an influence on certain other products. In this case, we can manage the access of need purchases and ensure funds are available for credit obligation repayments.

3.2.3 Consumer Satisfaction after Purchasing Decisions

Yeung *et al.* (2013) examined consumer satisfaction after purchases in nine European countries from 1999 to 2011. The paper showed the positive impact of consumer satisfaction on consumer expenditure, as well as the challenges posed by economic structures and cultural, political, and socio-economic factors. However, the impact of consumer satisfaction on countries' economies could not be quantified. There has been recent research into the macro effects of consumer

satisfaction on national economies, though there is still insufficient data for exploring the topic. Fornell *et al.* (2010) tested how changes in the American Customer Satisfaction Index (ACSI) affected consumer expenditure. Ramasamy and Yeung (2010) tested the influence of the ACSI and the Consumer Confidence Index (CCI) on national consumption. They aimed to answer two questions: first, is there a consumer satisfaction index (CSI) for the aggregate consumption of nine European countries? Second, are there differences in how consumer satisfaction affects consumer expenditure in those European countries? The findings show that CSI is a significant predictor for expenditure on personal consumption and that CSI is a more valuable policy option than CCI.

Yeung *et al.* (2013) have demonstrated several points about consumer satisfaction. First, customer satisfaction measures the quality of economic output, so consumption has a positive effect on the quality of goods and services. Second, consumption is a relatively significant proportion of GDP (in 2010, it was 70.1% in the US, 61.8% in the UK, and 55.9% in Germany (Yeung *et al.*, 2013)). The greater the satisfaction, the higher the level of consumption. This also highlights the importance of customer satisfaction in ensuring the competitiveness and welfare of citizens. Therefore, the study proposed the following hypotheses:

1. The aggregate customer satisfaction has a significant impact on future consumer expenditure in the group of European countries selected.
2. In countries where the services sector is larger than the products sector, customer satisfaction has a larger impact on consumer expenditure.
3. Consumer satisfaction has a more significant impact on consumer spending in self-expressive societies (with high levels of wellbeing, good health, and a large amount of interpersonal trust) than it does on survival societies (with economic and physical security, low levels of wellbeing, poor health, and little interpersonal trust).
4. Customer satisfaction has less impact on consumer expenditure when the consumer's level of education increases.
5. Customer satisfaction has more impact on consumer expenditure when the consumer's income increases.

6. Customer satisfaction has more impact on consumer expenditures in countries with more open competition and greater economic freedom.

To examine the first hypothesis, a regression model was developed that used Equation(3-2). This equation was first developed in (Spanos, 1989):

$$\Delta PCE_{it} = (\nu_t + \varepsilon_{it}) + \beta_2 \Delta PDI_{it-1} + \beta_3 ECT_{it-1} + \delta \Delta CSI_{it-1} + \theta CCI_{it-1} + \beta_4 \Delta CPI_{it-1} + \beta_5 DEBT_{it-1} \quad (3-2)$$

Where PCE is personal consumer expenditure, PDI is personal disposable income, and ECT is the error correction term, derived from Equation (3-3

$$PCE_{it} = (\alpha_i + u_{it}) + \beta_0 PDI_{it} \quad (3-3)$$

In other words, PCE-Beta PDI. CSI is the customer satisfaction index, CCI is the consumer confidence index, CPI is the inflation, and DEBT is the consumer debt. However, to examine hypotheses H2-H6, the following equation (3-4) was adopted:

$$\Delta PCE_{it} = (\alpha_t + u_{it}) + \beta_1 \Delta PDI_{it-1} + \beta_2 ECT_{it-1} + \delta \Delta CSI_{it-1} + \theta \lambda CCI_{it-1} + \beta_3 \Delta CPI_{it-1} + \beta_4 DEBT_{it-1} + \beta_5 MF_{it-1} + \beta_6 (MF_{it-1} \times \Delta CSI_{it-1}) \quad (3-4)$$

Where MF is the moderating factor and MF*change in CSI is the interacting variable. There were 5 five moderating variables to test each of the hypotheses, as outlined in

Table 3-5.

Table 3-5: Moderating factors for cross-country differences

Symbol	Moderating factor	Description	Source
SERV	Proportion of the services sector in GDP	Proxy for the differences in terms of economic structure.	World Bank's World Development Indicators
PDI	Income per capita	Proxy for socio-economic differences. Disposable income measured as gross income minus social security contributions and income tax, in constant US dollars (in per capita values).	World Bank's World Development Indicators
EDU	Education levels	Proxy for socio-economic differences. Measured by the proportion of the country's labor force with only primary education.	World Bank's World Development Indicators
SURV	Survival vs. self-expression values	Proxy for cultural differences among countries. Measured based on a battery of questions as per Inglehart & Welzel, 2005.	World Values Survey
FREE	Economic freedom	Proxy for political-economy differences. Similar to Morgeson et al. (2011), we use trade and business freedom to represent economic freedom.	Heritage Foundation

Source: Yeung *et al.* (2013: 413).

A Monte Carlo test showed that the model performed better in the current form with panel data estimations for time invariant regressors. The results showed that hypotheses 2, 4, and 6 were significant, whereas hypotheses 3 and 5 were not significant. For hypothesis 3, there was a negative relation between satisfaction and consumption in highly self-expressive societies. The results also showed that CSI was a good predictor of PCE, so the first hypothesis was accepted. A summary of the findings is as follows:

1. Customer satisfaction has a more significant impact on consumer expenditure in countries where the services sector dominates.
2. Consumer satisfaction is more significant in countries with lower levels of education.
3. Consumer satisfaction is more significant in countries with higher levels of competition.
4. Consumer satisfaction has a negative impact on consumption in highly self-expressive countries.

3.2.4 Understanding How Marketing Strategies Can Increase Customer Satisfaction

In general, marketing strategies have an impact on national economic growth. Marketers need to know which strategies to use to increase customer satisfaction and sales, thereby directly affecting growth.

Buneeva *et al.* (2016) examined the effect of consumer lifestyle on consumption in Russia. Their study assessed the reasons behind household consumption and the influence of consumption on people's lives. They used statistical analysis to assess how the number of children in each Russian household affected consumption. The paper also focused on how other variables like social status, education, and age affect expenditure, as well as the impact of place of residency.

The study showed that personal 'need' and economic 'quality' of life contradict each other. The first comes from what a person thinks about their quality of life. The second refers to a consumer's actual quality of life. The paper discussed the objectives of the consumption process which are the ideas formed by a person about the quality of life they desire and that help to define the kinds of purchases they make.

The quality of life was calculated as resulting from three conditions in the process of consumption:

1. The extent to which the needs of consumers are met
2. The extent to which consumers are satisfied
3. How a person's consumption relates to the wider social standards that are reflected by consumption norms.

3.3 Share-of-Wallet and Consumer Spending as Credit Risk Parameters

A study by Jang *et al.* (2016) defined share of wallet as the percentage of a consumer's total expenditure (or size-of-wallet). This is important information for marketing purposes and for understanding consumer expenditure. Share-of-wallet is used as a measurement of loyalty (Bowman and Narayandas, 2004) and a segmentation standard (Reinartz and Kumar, 2003). It also has an impact on

profits (Reinartz, Thomas and Kumar, 2005). As Du *et al.* (2007) note, calculating share of wallet requires companies to have information about competitors. They raise three questions about share of wallet:

- Is it possible to improve benchmark share of wallet models by studying the interrelationships between purchases by firms in specific categories?
- How many consumer segments are there and how can they be categorised based on different consumption patterns?
- What strategies could be inferred from the results to help with targeting consumers?

The research covers the following areas:

1. The cover of multiple firms' purchases, rather than that for one single firm. Cross category usage, rather than cross-selling.
2. The interactions between irregular, complementary, and substitute products, rather than symmetric interactions or interactions between substitutes.
3. How to estimate competing firms' expenditures and calculate the share of wallet.
4. Making financial decisions based on credit card usage.

The study used Bayesian calculations based on simultaneous equations, using a Tobit model. The two advantages of the Bayesian approach were that it made it easier to estimate models with many endogenous variables and to estimate all the different types of simultaneous equations. Therefore, the study proposed a methodology using simultaneous equations with a Tobit model and latent classes. It focused on three issues:

- Heterogeneity in spending patterns.
- Simultaneity created by the interrelationship of expenditures across firms and product categories.
- Data censoring that happens when consumers have no expenditures in each category.

The study also estimated credit card data using Bayesian inference. To estimate the benchmark for expenditures, the study used multiple discreteness (KAR) and

‘Model C’ proposed by Du *et al.* (2007), known as the DKM model. The study examined Korean financial services, focusing on two kinds of credit card expenditures: purchases and cash advances. The research focused on one bank, its competitors, and credit card companies. Credit card transactions were defined as ones in which the customer paid the outstanding balance with interest rather than a revolving option, which was less popular. Revolving credit is a type of credit that can be used repeatedly up to a certain limit as long as the account is open and payments are made on time. With revolving credit, the amount of available credit, the balance, and the minimum payment can go up or down depending on the purchases and payments made to the account. In this model, interest was paid in instalments on purchases and cash advances. Table 3-6 details the variables used in the study.

Table 3-6: Variables used for estimation with statistical descriptions

(A) Endogenous and Exogenous Variables					
Type	Variables*	Mean	SD	Min	Max
Endogenous	Purchases at competitors (\$1 K)	0.903	1.00	0	7.44
	Cash advances at competitors (\$1 K)	1.393	2.61	0	15.83
	Purchases at the focal bank (\$1 K)	0.370	0.48	0	3.20
	Cash advances at the focal bank (\$1 K)	0.335	0.65	0	3.52
Exogenous	Age	40.42	9.15	22.0	79.0
	Male (dummy)	0.62	0.48	0	1
	Annual income (\$1 K)	25.57	17.26	0.004	239.2
	Salaried person (dummy)	0.647	0.48	0	1
	Past delinquent amount at competitors (\$1 K)	0.022	0.11	0	1.22
	Past delinquent amount at the focal bank (\$1 K)	0.005	0.03	0	0.35
	Loan interest rate	0.125	0.03	0.05	0.27
	Past purchases at competitors (\$1 K)	0.846	0.94	0	6.80
	Past cash advances at competitors (\$1 K)	1.364	2.55	0	15.53
	Past purchases at the focal bank (\$1 K)	0.347	0.46	0	3.04
	Past cash advances at the focal bank (\$1 K)	0.354	0.67	0	3.70
*: Transaction amounts are the average of amounts for three months					
(B) Usage Rate and Distribution of Shares					
Category	Usage	Share = 0	0 < Share < 1	Share = 1	
Purchases	Users	11,258 (99.3%)	12.78%	80.57%	6.64%
	Non-users	79 (0.7%)			
Cash advances	Users	7,032 (62.03%)	31.26%	57.05%	11.69%
	Non-users	4,305 (37.97%)			
(C) Censoring Percentage of Expenditures Variables					
Expenditure Types	Recent Period		Past Period		
Purchases at competitors	7.29%		7.06%		
Cash advance at competitors	45.22%		45.75%		
Purchases at the focal firm	13.39%		14.55%		
Cash advance at the focal firm	57.36%		57.17%		

Source: (Jang, Prasad and Ratchford, 2016).

A study by Jang *et al.* (2016) also outlined the factors that should be expected to influence customers’ expenditures (see Figure 3-6).

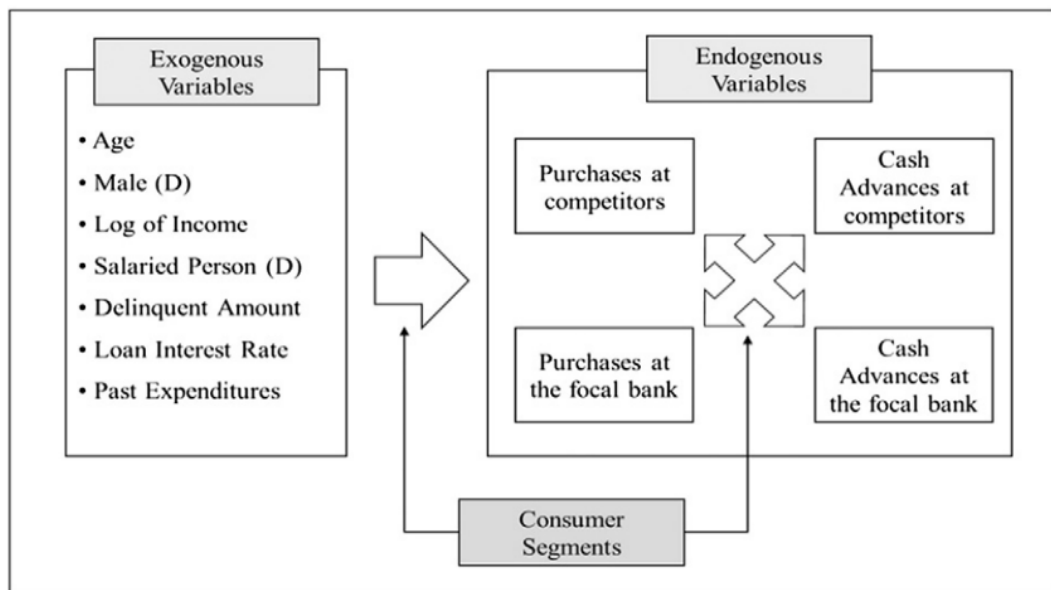


Figure 3-6: Factors that influence expenditures

Source: (Jang, Prasad and Ratchford, 2016).

The results indicate that there were two segments: large and small. The first relied on past expenditures, whereas the second relied on factors to do with interrelationships, such as income. Table 3-7 shows the results of the interrelationships between the expenditures, listing the coefficients of the endogenous variables in terms of their derivatives (labelled 'C' in the table).

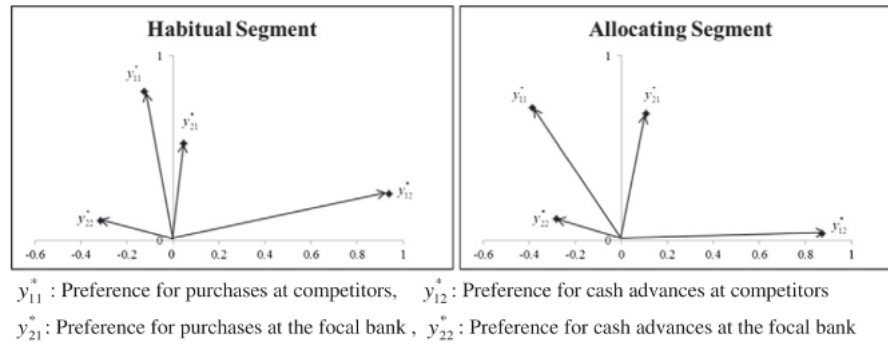
Table 3-7: Estimates for parameters

LHS	RHS	Habitual Segment			Allocating Segment			
		Estimates	2.5%	97.5%	Estimates	2.5%	97.5%	
(A) Endogenous variables: Expenditures at competitors								
Purchases at competitors	Exogenous	Intercept	−0.034	−0.21	0.13	−0.732	−1.60	0.09
		Age	−0.016	−0.03	0.00	−0.025	−0.08	0.03
		Male (dummy)	−0.009	−0.03	0.01	−0.033	−0.14	0.07
		Log of income	0.021	0.00	0.04	0.170	0.09	0.26
		Salaried person (dummy)	−0.003	−0.03	0.02	−0.136	−0.24	−0.03
		Delinquent amount at competitors (\$1 K)	0.006	−0.10	0.11	−0.353	−0.75	0.03
		Delinquent amount at focal bank (\$1 K)	−0.479	−0.80	−0.17	−0.893	−2.93	1.20
		Loan interest rate	0.005	−0.29	0.30	−0.526	−1.94	0.91
		Past purchases at competitors	0.817	0.80	0.84	0.531	0.49	0.57
	Other endogenous	Purchases at competitors						
		Cash advances at competitors	0.017	0.01	0.02	0.086	0.06	0.11
		Purchases at focal bank	−0.039	−0.08	0.00	0.007	−0.07	0.10
		Cash advances at focal bank	−0.012	−0.04	0.00	−0.263	−0.37	−0.18
Cash advances at competitors	Exogenous	Intercept	−0.511	−1.03	0.01	−0.358	−1.97	1.23
		Age	−0.051	−0.09	−0.01	−0.044	−0.15	0.06
		Male (dummy)	0.023	−0.04	0.09	−0.063	−0.26	0.12
		Log of income	−0.037	−0.09	0.01	−0.072	−0.23	0.09
		Salaried person (dummy)	0.103	0.03	0.18	0.063	−0.13	0.26
		Delinquent amount at competitors (\$1 K)	0.071	−0.20	0.35	−0.416	−1.12	0.26
		Delinquent amount at focal bank (\$1 K)	0.589	−0.32	1.48	−0.767	−4.44	2.82
		Loan interest rate	2.610	1.73	3.49	2.785	0.12	5.39
		Past cash advances at competitors	1.079	1.06	1.10	0.854	0.80	0.90
	Other endogenous	Purchases at competitors	0.188	0.13	0.24	0.393	0.29	0.55
		Cash advances at competitors						
		Purchases at focal bank	−0.362	−0.40	−0.34	−0.494	−0.77	−0.30
		Cash advances at focal bank	0.306	0.26	0.35	0.478	0.34	0.63
Bold estimates are significant at the 0.05 level.								
(B) Endogenous variables: Expenditures at the focal bank								
Purchases at competitors	Exogenous	Intercept	0.002	−0.10	0.10	−0.127	−0.59	0.32
		Age	−0.014	−0.02	−0.01	−0.016	−0.05	0.02
		Male (dummy)	−0.004	−0.02	0.01	0.015	−0.04	0.07
		Log of income	0.009	0.00	0.02	0.061	0.02	0.11
		Salaried person (dummy)	−0.001	−0.01	0.01	−0.107	−0.16	−0.05
		Delinquent amount at competitors (\$1 K)	−0.007	−0.06	0.05	−0.085	−0.29	0.12
		Delinquent amount at focal bank (\$1 K)	−0.137	−0.32	0.04	−1.896	−3.10	−0.70
		Loan interest rate	0.068	−0.10	0.24	0.405	−0.37	1.17
		Past purchases at focal bank	0.776	0.75	0.80	0.503	0.46	0.55
	Other endogenous	Purchases at competitors	−0.012	−0.02	0.00	−0.041	−0.08	0.00
		Cash advances at competitors	−0.009	−0.01	−0.01	−0.040	−0.06	−0.02
		Purchases at focal bank						
		Cash advances at focal bank	0.000	−0.01	0.01	0.082	0.03	0.15
Cash advances at competitors	Exogenous	Intercept	−0.200	−0.42	0.02	0.093	−0.46	−0.20
		Age	−0.017	−0.03	0.00	−0.026	−0.06	−0.02
		Male (dummy)	0.012	−0.02	0.04	−0.007	−0.07	0.01
		Log of income	−0.017	−0.04	0.00	−0.030	−0.09	−0.02
		Salaried person (dummy)	−0.009	−0.04	0.02	−0.044	−0.11	−0.01
		Delinquent amount at competitors (\$1 K)	−0.003	−0.12	0.11	0.104	−0.13	0.00
		Delinquent amount at focal bank (\$1 K)	0.008	−0.38	0.40	−0.390	−1.71	0.01
		Loan interest rate	0.657	0.28	1.03	0.436	−0.50	0.66
		Past cash advances at focal bank	1.096	1.07	1.12	0.748	0.70	1.10
	Other endogenous	Purchases at competitors	−0.068	−0.10	−0.05	−0.104	−0.15	−0.07
		Cash advances at competitors	0.035	0.03	0.04	0.066	0.05	0.04
		Purchases at focal bank	0.083	0.06	0.14	0.100	0.03	0.08
		Cash advances at focal bank						
Bold estimates are significant at the 0.05 level.								
(C) Derivatives Implied by Coefficients of Endogenous Variables								
Types	Purchases			Cash Advances				
	Derivatives	Habitual Segment	Allocating Segment	Derivatives	Habitual Segment	Allocating Segment		
(1) Within a bank–Across categories	$\partial PUR_C / \partial CASH_C$	0.017	0.086	$\partial CASH_C / \partial PUR_C$	0.188	0.393		
	$\partial PUR_F / \partial CASH_F$	0.0002	0.082	$\partial CASH_F / \partial PUR_F$	0.083	0.100		
(2) Within a category–Across banks	$\partial PUR_C / \partial PUR_F$	−0.039	0.007	$\partial CASH_C / \partial CASH_F$	0.306	0.478		
	$\partial PUR_F / \partial PUR_C$	−0.012	−0.041	$\partial CASH_F / \partial CASH_C$	0.035	0.066		
(3) Across banks and categories	$\partial PUR_C / \partial CASH_F$	−0.012	−0.263	$\partial CASH_C / \partial PUR_F$	−0.362	−0.494		
	$\partial PUR_F / \partial CASH_C$	−0.009	−0.040	$\partial CASH_F / \partial PUR_C$	−0.068	−0.104		

Source: (Jang, Prasad and Ratchford, 2016: 132).

The second explanation is that the interrelationship stems from the assumption that consumer expenditure equals the utility of expenditure if it is higher than zero, and there is no spending in categories where the utility is negative. This is shown in Figure 3-7.

Figure 3-7: Eigenvectors (factor loading) plots by segments



Source: (Jang, Prasad and Ratchford, 2016: 133).

The same study calculated the size-of-wallet of the expenditure at the focal bank and the expected expenditure at the bank's competitors. To calculate the expected expenditure of the bank's competitors, a Monte Carlo integration was used, as well as KAR and DKM. The paper quantified the share of wallet by segmenting the sample into three groups:

- Group 1: consumers who spend at competitors but not at the focal bank.
- Group 2: consumers who may or may not spend at competitors but who spend at the focal bank.
- Group 3: consumers who spend at neither the focal bank nor its competitors.

To examine accuracy of the model, a comparison of the median absolute error (MAE) and median absolute percentage error (MAPE) was used to limit the influence of any outliers. Table 3-8 shows the results, based on purchases or cash advances measured in \$1,000 increments.

Table 3-8: Prediction comparison between a focal bank and its competitors.

Prediction Error	Group	Purchases				Cash Advances			
		N	Proposed	KAR Model	DKM Model	N	Proposed Model	KAR Model	DKM Model
(A) Prediction Errors in the Size-of-Wallet									
MAE	1	439	0.270	0.653	0.404	700	0.318	0.537	0.707
	2	3201	0.221	0.441	0.500	1618	0.465	0.794	0.822
	3	21	0.212	0.129	–	1343	0.124	0.530	–
MAPE	1	439	0.359	0.822	0.456	700	0.585	0.822	0.793
	2	3201	0.421	0.541	0.611	1618	0.405	0.435	0.518
	3	21	–	–	–	1343	–	–	–
(B) Prediction Errors in the Share-of-Wallet									
MAE	1	439	0	0	0	700	0	0	0
	2	3201	0.066	0.247	0.115	1618	0.086	0.073	0.218
	3	21	–	–	–	1343	–	–	–
MAPE	1	439	–	–	–	700	–	–	–
	2	3201	0.280	1.080	0.433	1618	0.338	0.303	0.731
	3	21	–	–	–	1343	–	–	–
(C) Medians of Individual's Differences between Our Model and Benchmark Models									
Prediction Error	Group	Size				Share			
		KAR		DKM		KAR		DKM	
		Purchases	Cash Advances	Purchases	Cash Advances	Purchases	Cash Advances	Purchases	Cash Advances
MAE	1	–0.355	–0.176	–0.088	–0.358	–	–	–	–
	2	–0.192	–0.221	–0.186	–0.179	–0.165	–0.001	–0.038	–0.076
	3	0.082	–0.397	–	–	–	–	–	–
MAPE	1	–0.457	–0.335	–0.104	–0.327	–	–	–	–
	2	–0.151	0.060	–0.213	–0.124	–0.796	–0.009	–0.109	–0.211
	3	–	–	–	–	–	–	–	–

Source: (Jang, Prasad and Ratchford, 2016: 136).

The model showed that studies of the interrelationships between expenditures have better results with lower prediction errors compared to other studies. Thus, the model could be used to improve the efficiency of marketing to customers with a large size-of-wallet and those with a low share of wallet at the focal bank, there is a ‘platinum’ segment, whose size-of-wallet ranked in the top 20%, but their share of wallet is below average. Table 3-9 shows the real numbers of customers and expenditures from the real target and the proposed model. It shows that it is more accurate than the KAR, DKM, and benchmark models.

Table 3-9: Number of customers and expenditures in the ‘platinum’ segment of the focal bank, separated by model

(A) Purchases Category				
Types	Real Target	Proposed Model	KAR*	DKM
Number	539 (14.8%)**	409 (11.2%)	158 (4.3%)	298 (8.2%)
Percentage of customers in the real target		48.2%	16.1%	37.9%
Expenditures (\$1 K)	1,692.37 (36.7%)	1,039.88 (22.5%)	369.33 (8.0%)	838.77 (18.2%)
*: Results of KAR model in the purchases category is based on the different criterion for the platinum segment (we used size-of-wallet > 50% instead of 20%). See footnote 7.				
**: Percentages in the parentheses are the ratios of the number or expenditures of platinum segment to those of the total customers in the validation sample.				
(B) Cash Advances Category				
Types	Real Target	Proposed Model	KAR	DKM
Number	386 (16.7%)	341 (14.7%)	215 (9.3%)	223 (9.6%)
Percentage of customers in the real target		71.0%	50.4%	46.6%
Expenditures (\$1 K)	3,210.99 (50.8%)	2,633.31 (41.6%)	1,705.55 (27.0%)	1,878.05 (29.7%)

Source: (Jang, Prasad and Ratchford, 2016: 137).

The researchers mentioned three managerial benefits in the study:

- The study showed that increased purchases were correlated with increased cash advances, suggesting that managers could engage in cross-selling.
- The results showed that there was a positive relation in expenditures (purchases and cash advances) in the focal bank and its competitors. Cash advances probably increased because clients needed higher cash limits. So, managers could increase cash limits or offer other products such as loans. If clients required increases to pay off other debts, then the bank could reduce cash limits and exposure to reduce the risk of the client defaulting.
- By using the method proposed by the authors of the study, managers could create a targeting strategy suitable for generating higher revenues from customers in the ‘platinum’ category. For example, the study shows that there are two segments: habitual and allocating segments. On average, the latter spends twice as much as the former, so managers could target them.

The limitations of the study were:

- The data used to explain current expenditures were restricted to a single period.
- Variables that might have enhanced the study (such as credit limits, interest rates, and minimum required payments) were not included.
- If there was panel data, there could have been a dynamic calculation of the size and share of wallet.
- Daily data for different credit cards would have made it possible to examine if there were any patterns of expenditures across different categories or banks.

3.4 Developing an Understanding of Credit Risk

Recent events have showcased the importance of credit risk, which is an important factor when it comes to managing credit portfolios efficiently. There are several machine learning processes that make it possible to examine credit risk (Wang *et al.*, 2012). A recent paper by the International Monetary Fund discussed the advantages of financial technology (Bazarbash, 2019). It considered how machine learning solutions could reduce the cost of credit and provide clearer solutions for non-technical audiences than the current 'black box' templates.

Recent studies of both corporate and household customers have recognized the importance of programming in the process of granting loans. Advances in technology have led to collaboration between computer science and finance, helping the latter to benefit from technological efficiency. There are a wide range of options for using classification models to calculate the probability of a customer defaulting on their loan. Developing this into a credit rating model would help regulators and banks make informed decisions (Petropoulos *et al.*, 2018).

Baesens *et al.* (2003) performed a comparison of 17 different mechanisms for analysing credit classes. They found that classification methods are a suitable way of analysing credit rating. However, they did not use heavily unbalanced data as we do in this study. The data in our study is unbalanced insofar as it is not well distributed across the two classes. In our data set, almost 85% of the data is from

customers who are considered 'good' and the remaining 15% is from customers who are considered 'bad' (likely to default).

Given the importance of the topic in banking sector, commercially and regulatory. And due to the lack of available risk weight models in the region. The significance of the risk weight model imposed itself through the advancement in programming and machine learning as solution tools.

4 METHODOLOGY

In this study, it was difficult to select, collect, clean, test, and analyse the data. We visited ten banks several times, travelling back and forth between Kuwait and the UK. We collected a substantial amount of confidential data, so we had to ensure that we had followed all the requisite legal procedures. This involved signing non-disclosure agreements with several independent banks, the Central Bank of Kuwait, and Cranfield University. We collected approximately 1TB of financial data from ten large and medium-sized Kuwaiti banks. The data included information about customer characteristics and credit portfolios. Our aim was to build a model to predict the probability that customers would default (this was the basis for our dependent variables) depending on their information (this was the basis for our independent variables). Given the style of data we collected and the kinds of big data sets being used in banking today (as detailed in Chapter 2), we decided to use machine learning to analyse this data. This chapter describes our processes for collecting and cleaning the data. Then, it provides some background about the use of machine learning for analysing big data. Finally, it describes the specific machine learning techniques that we adopted for analysing our data.

4.1 Processing and Structuring the Data

Structuring the data for this study was difficult. Before this study, there were no machine learning solutions for predicting the likelihood of people in Kuwait to default. No work of this kind had been done on any other platform. Gathering the required data took a lot of time and effort. Based on our review of the literature and our banking experience, we labelled the data categories that we required (our independent variables) to make it easy for the banks to provide us with the right information.

4.1.1 Data Collection

After listing the information, we met with our supervisors several times, we contacted the Central Bank of Kuwait to help us liaise with the local banks through formal letters. The Central Bank approved the project under the supervision of Cranfield University. We obtained the necessary legal approval to obtain, store, and analyse the data and to publish the results. The contracts were signed by all the relevant parties, and we began to gather the data.

We travelled to Kuwait to collect the data and ensure that the banks understood our intentions for gathering, storing, and analysing the data. The data set was very detailed, covering a period of ten years. We asked the banks to provide us with the data within a month. However, after our first collection and scan of the data, we had to go back and forth with the banks to correct the data, liaising between Kuwait and Cranfield.

For security purposes, we used four two-terabyte hard drives to save the data in an encrypted format. We were sent the password to unencrypt the data separately by email. We were not allowed to connect to the network while working with the data to guarantee its security.

Substantial data was obtained from all ten banks in Kuwait. However, as discussed below, four banks were eventually excluded. Five of the banks were conventional and five were Islamic. We requested panel data from 2008 to 2018. As noted in the literature review, previous studies had only looked at conventional banks and had only tested the following variables: age, gender, nationality, income, education, and number of loans. In our study, we looked at these parameters but based on research by (Albarrak *et al.*, 2020), we also added the following: client-bank relationship duration, credit card exposure ('outstanding balance'), monthly average net cashflow, and monthly average number of transactions. We also included Islamic banks.

There are two types of loan in Kuwait: consumer loans and instalment loans. Consumer loans are loans for personal needs and durable goods. They are limited to 25,000 KWD or 15 times a person's salary (whichever is less).

Instalment loans are for maintaining or purchasing private residences. They are limited to 70,000 KWD. Our work looked at both types of loan. However, we chose to only share the results relating to instalment loans because they are granted more often and are exposed to the real estate sector.

Defaulting is when a customer fails to meet their obligation to pay back their loan and interest for three consecutive months. A customer that defaults is a bad customer. A customer who makes their payments and does not default is a good customer.

Our literature review revealed that samples that range from 1,200 to 1,500 observations are sufficiently large, assuming the selection is random. However, smaller sample sizes can give good results if they are tested against the confidence interval (Holmes, Illowsky and Dean, 2017). Table A-6-1, Table A-6-2, Table A-6-3, Table A-6-4, Table A-6-5, Table A-6-6 in Appendix A resent the raw data, including the different components and the size of the data set.

4.1.2 Data Cleaning

To clean the data, we first had to educate ourselves about MATLAB and learn C++ through one-on-one sessions, classes, books, online courses, and learning via the MATLAB website. We then had to work with data we had obtained and tailor the solutions to suit our needs. After several days of hard work, we had to deal with errors during the process. Variations in the types of data we had collected prevented some of the functions from running effectively. We had to adjust the code and deal with the fact that there was not enough memory. Even after finalizing the work for one of the banks, it was not straightforward to handle the data for the other banks. We had to make alterations for all of the banks' data sets depending on the specific errors and the structure of the data. During the cleaning process, we were also forced to return to the banks and request amendments and corrections. This was time consuming given their schedules, the amount of time required to obtain new data, the staff shortages caused by Covid-19, and the general situation with pandemic globally. All the data we received from the banks were in the form of Comma Separated Value (CSV) files.

We changed them to MATLAB files to analyse the data in Appendix A show the data format, the number of observations, and the size of the data sets.

The cleaning process was conducted several times. Though there were initially errors, the correct processes were calculated in time to allow the work to continue. Table A-6-13 details code execution times for all the six banks that were studied.

4.1.3 The Limitations of Our Dataset

Initially, we gathered the data for all ten banks in Kuwait with the aim of developing a comprehensive system. We examined the data and tweaked some of the items, working both in Kuwait and the United Kingdom. However, given the situation with Covid-19, we realised that the process was going to be too time consuming. We therefore eliminated four banks for the following reasons:

- Bank 7 did not provide any examples of customers who had defaulted. They only provided information about people who had not defaulted. Therefore, it was not possible for us to develop a default prediction model for the bank. However, their data was included for the model for the Central Bank.
- Bank 8's data did not tally between the tables, particularly when it came to CBK5: Loans. The principles and the interest on the loans did not tally with the outstanding balance for the loans. This may have been because the calculations for customers who defaulted were incorrect. The bank stated that this was the only data they had available, so we decided to remove them.
- After cleaning the data for Bank 9, we realised that they had not provided historical outstanding balances for CBK5: Loans. Bank 9's system updates itself to the current outstanding balance with every principle or interest payment, thereby eliminating the historical data for defaulting and non-defaulting customers.
- Due to lack of time, Bank 10 could not be included. Given the type of data, the data gathering, and data cleaning, as well as the time required to

program the models for each bank, it was sometimes not possible to include all of the data.

4.2 Analysing the Data

4.2.1 An Introduction to the Analysis of Big Data

Due to the characteristics of big data described in Chapter 2, different methods are required to analyse big data sets from those that are used with normal data sets. Honest and Patel (2016) have noted that there are three reasons why most current methods cannot be used to analyse big data: first, the methods are not designed to handle complex data sets; second, the methods do not offer dynamic analysis that matches the data inputs; third, most of the methods use the same kind of input.

4.2.1.1 Tools for Handling Big Data

Honest and Patel (2016) listed three tools for handling big data: NoSQL, Massively Parallel Processing (MPP), and MapReduce and Storage. NoSQL makes cloud computing more efficient by lowering costs and increasing the speed of computation. MPP is a multiple processor that has its own storage and operating systems. It uses processors to communicate through messaging. MapReduce breaks data into small parts, processing each part separately. Then, the results are scanned and aggregated into nodes and clusters. Although both methods (MPP and MapReduce) work, similarly, making multiple calculations at the same time, they differ in the following ways:

- MPP requires expensive CPU storage, whereas MapReduce uses commodity servers and discs.
- MPP uses structured query language, whereas MapReduce uses java code.
- MPP loads data slower than MapReduce.
- Querying is easier in MPP, whereas forming maps and reducing them is a complex procedure in MapReduce.
- MPP is good for structured data, whereas MapReduce is good for semi-structured data.

- MPP can handle a large amount of structured data and uses iterative queries, whereas MapReduce can handle unstructured or semi-structured data in its native format.

Another study by Elshawi *et al.* (2018) examined how software such as Hadoop and machine learning can improve data analysis. The paper provided the following:

1. An evaluation of the supporting software used to analyse big data on cloud storage systems.
2. A summary of the main features of big data.
3. An assessment of the software used to analyse big data.
4. An overview of the latest developments and challenges in this field.

4.2.1.2 Storing Big Data

Nowadays, data can be stored in the cloud using platforms provided by big companies like Google and Amazon. These companies provide solutions such as Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS).

- IaaS is a virtualised resource available for developers to implement their big data analysis systems from scratch.
- PaaS offers a support platform for developers to build new big data applications or extend current ones.
- SaaS provides fully structured data analysis processes ready to use as an online service.
- There is also another way of storing and analysing data, which is known as Everything as a Service (EaaS). This handles everything from storing the data to configuring the applications, analysing the data, and providing the final results.

Thus, cloud computing helps provide big data storage and computational solutions. However, it is subject to the following requirements in all cases:

- Standardized access. The infrastructure should use a standard technology to make it easier to add services and applications.

- Heterogeneous/distributed data support. The infrastructure should be able to handle large data sets with several dimensions that may be stored in different formats and locations.
- Scalability. The infrastructure should be able to grow (as the size of the data set increases or more sophisticated algorithms are used). It should be able to allocate and redistribute resources as and when required.
- Efficiency. The infrastructure should reduce the number of resources used when performing a task.
- Security. The infrastructure should have a proper security system to protect the data and ensure privacy.

The requirements for the architectural level are:

- Service-orientation. The infrastructure should be suitable for use with different services. It should be capable of being reused.
- Openness and extensibility. Big data science should be open to new tools and services. Current services should be open for extensions but closed for modifications. Given the size of data and the time required for processing, it is not supposed to be set out for modifications. Big data science architecture should be open to the integration of new tools and services. Moreover, existing services should be open for extension but closed for modification, according to the open-closed principle.

4.2.1.3 Operating Big Data Systems

The main task when it comes to operating big data applications is the management of data sources, analytical tools, and outcomes. To manage these, the following requirements must be met:

- Data management. The system must be able to handle data inputs and outputs that are in different formats.
- Tool management. The system must be able to manage the different kinds of analytical tools.

Big data systems must be able to perform analytical tasks and control how analyses are implemented. This requires:

- Design management. Data mining processes should be effectively designed so that users can follow them.
- Execution management. The system should include a parallel execution environment to process a large amount of data.

According to Elshawi *et al.* (2018), the main challenges when it comes to operating big data systems are:

- Data availability and sharing.
- Operational challenges, such as making sure that the data format and the analytical tools are suitable for conducting operations.
- Managing the size of the data, which might require additional hardware storage or cloud storage.
- Handling the data as it changes continuously, which might require the development of special mechanisms or algorithms.
- Keeping track of the model management process by saving the metadata that allows users to follow the process from inputs to outputs.
- Dealing with programming issues if the algorithms are not suitable for use with big data.
- Ensuring usability since machine learning is still an emerging subject area for data scientists and not easily understood by the public.

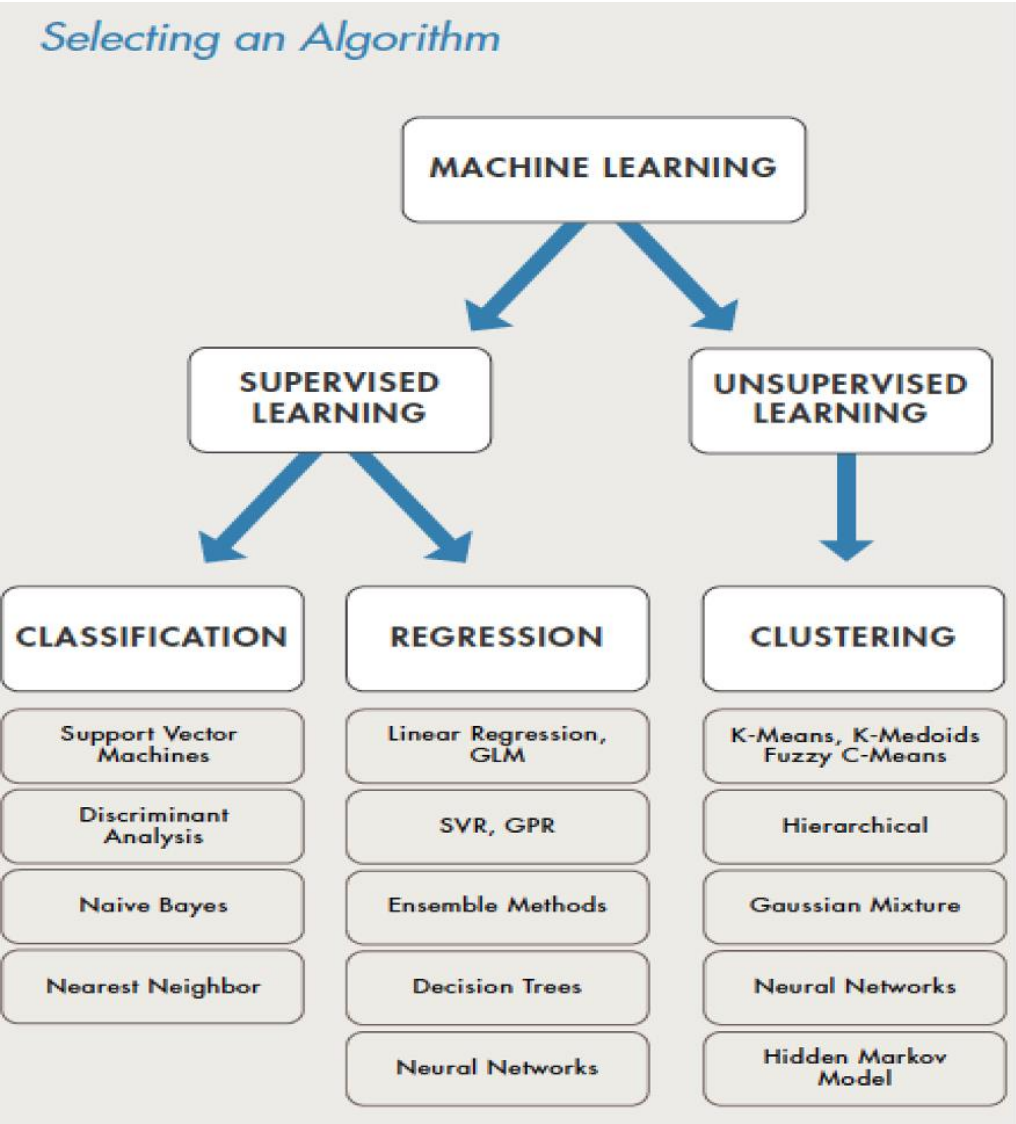
4.2.2 Using Machine Learning to Analyse Big Data Sets

4.2.2.1 What is Machine Learning?

To understand how machine learning can be used to analyse big data sets, it is important to understand what machine learning is. According to Phitchayanon (2020), machine learning involves using calculations to analyse data and generate a future outlook based on the current horizon of experience provided by the data.

Machine learning can be supervised or unsupervised. Supervised machine learning is used for data that has outputs that can be used to create future outputs in the form of regression or classification. Unsupervised machine learning searches for patterns in a data set in the form of clustering. There are different types of algorithms for each, but not even an experienced data scientist will always be able to predict which model is the most efficient. The best way to determine the correct model is to use trial and error. Machine learning is used for large data sets that are continuously increasing and for which there is no equation with which to make future predictions. Figure 4-1 is a list of algorithms used in machine learning, depending on the type of data available.

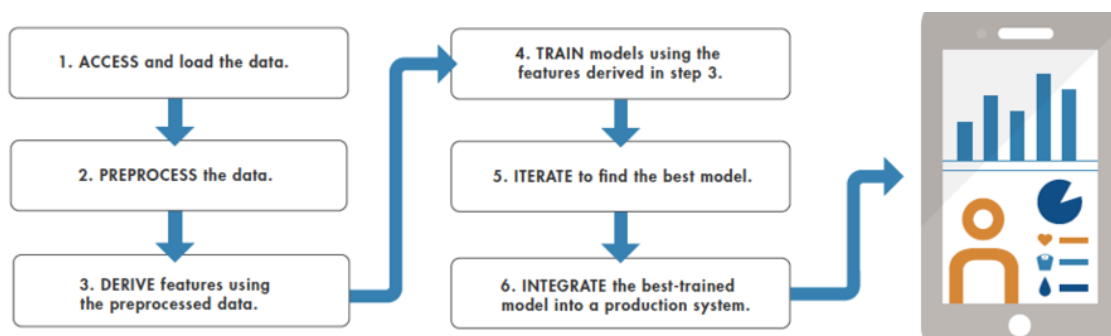
Figure 4-1: List of algorithms based on the type of data available



Source: (Phitchayanon, 2020: 9).

Phitchayanon (2020) stated that the challenges when it comes to machine learning are handling different types of data and different sizes of data sets; using specialist techniques for dealing with different types of data; and investing a lot of time to determine the most efficient model. The machine learning process involves several steps. These are outlined in Figure 4-2.

Figure 4-2: Machine learning process



Source: (Phitchayanon, 2020): 19).

According to Phitchayanon (2020), unsupervised learning tools are suitable for testing data when it is not clear what kind of results they might produce. The technique of clustering is used in unsupervised learning, which involves grouping similar pieces of data together. There are two forms of clustering:

- Hard clustering, in which each data point belongs to one cluster group
- Soft clustering, in which data points can belong to several cluster groups

To deal with the complexity of big data sets, machine learning models must downsize features or reduce dimensionality. It is important to note that unsupervised techniques could either be used for data without an end goal or be used as a step towards supervised learning, helping to reduce the number of features.

Phitchayanon (2020) also notes that supervised learning is used when there are known inputs and outputs. It is used to predict the future of a data set. There are two forms of supervised learning:

- Classification, which is used to predict separate responses and classify data into categories.
- Regression, which is used to predict continuous responses.

If the model's outcomes are weak or the model needs improvement (if, for example, it cannot differentiate between noise and data points), then feature selection, feature transformation, or hyperparameter tuning can be used to improve the outcomes:

- Feature selection involves adding or removing variables that do not improve the model.
- Feature transformation involves transforming features through technical solutions.
- Hyperparameter tuning involves identifying the parameters that provide the best model.

4.2.3 Which Machine Learning Methods Can Be Used to Analyse Financial Data Sets?

A comparison of machine learning methods used to classify customers who might default on their loans showed that RUSBoost was the most effective method to use (Munkhdalai *et al.*, 2019). Seiffert *et al.* (2010) showed that RUSBoost is significantly better at handling imbalanced data.

The current study aims to overcome the problem that previous investigations have relied on small research samples (Wang *et al.*, 2012). As recommended in the literature, ensemble learning and gradient boosting decision trees can be used to make up for the disadvantages of decision trees, especially for large data sets that cover a long period, as is true in our case (Bazarbash, 2019). Similar studies have had a research period of three years (Petropoulos *et al.*, 2018), whereas our aim has been to expand it to between five and eleven years.

Another problem with previous studies is that they have relied on similar data sets gathered from customers, disregarding the data available from banks. Our study, by contrast, will rely on current data available from banks (Nyathi *et al.*, 2014). The paper will also provide observations across multiple periods because it will

examine customers with loans for which the term is 15 years. The parameters assessed in this study were derived from the literature (Samreen, Zaidi and Sarwar, 2013). These were supplemented by new variables that had not been used before (such as number of monthly transactions and monthly average cash flow). For aggregated- and system-wide data, we redid the same comparisons and used the results from the AUC. We found that ensemble bag and decision fine tree were the most appropriate methods.

Due to aggregation and the large number of observations, as well as the skewed data sets, we reran the models and focused on the ensemble methods (Bag, AdaBoostM1, RUSBoost, LogitBoost, GentelBoost, and Robust Boost). The reason for this was that those later methods were a combination of several approved methods. They were enhanced when we compared the AUC results with our data. They were also based on several random forest trees that had been shown to be efficient for our data set.

The ensemble methods relied on subcategorizing the data into several random groups to run the random forest tree analysis and obtain the most accurate solution for each subgroup. Therefore, it was necessary to test for overfitting.

The overfitting of boosted techniques is a topic that is not yet fully understood in theoretical terms, but empirical results have shown that boosting seems to be very robust against overfitting (Breiman, 2001; Friedman *et al.*, 2004). As a contribution to the used toolbox, we automated the analysis and set the comparison of the AUC as the key performance measure. The code revealed the best model is, using a 70% training sample and a 30% holdout sample. Overall, this makes our work suitable for financial analysts and economists.

4.3 Machine Learning Methods Used in this Study

The aim of this study was to develop the most appropriate credit rating model to predict when customers were likely to default on household loans. To achieve this, we compared different classification models for each bank. The models were logistic regression, decision tree, support vector machine, Bayesian network,

Bag, AdaBoostM1, and RUSBoosted. Based on the performance of the ensemble models for each bank and our initial analysis of the Central Bank, we revised our analysis by using further ensemble models: Logitboost, Gentelboost, and Robustboost. It is important to highlight that such work for modelling default risk weights have been done through advancements in machine learning tools. The thing that has never been done for country Kuwait, Islamic banks, customised model per bank and a comprehensive model for central bank for regulatory purposes.

These methods were selected because they were easy to implement, efficient, fast, flexible, and allowed for interpretability. Previous studies supported applying the selected methods for credit default classifications. Furthermore, the data type used in our study (skewed data) dictated that we should use ensemble methods.

4.3.1 Logistic Regression

Logistic regression is a method of classification. Instead of fitting a straight line or hyperplane, logistic regression uses the logistic function to keep the output of a linear equation between 0 and 1. The logistic function is defined by Equation (4-1)

$$\text{logistic}(\eta) = \frac{1}{1 + \exp(-\eta)} \quad (4-1)$$

The move from linear regression to logistic regression is direct. Using the linear regression model, we demonstrated the relationship between the outcome and the features using a linear equation (4-2)..

$$\hat{y}^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_p x_p^{(i)} \quad (4-2)$$

For classification, we preferred the odds to be between 0 and 1, so we changed the right side of the equation into a logistic function, forcing the output to remain between 0 and 1. This is shown in Equation (4-3)

$$P(y^{(i)} = 1) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_p x_p^{(i)}))} \quad (4-3)$$

The interpretation of the weights in logistic regression differs from the interpretation of the weights in linear regression because the outcome in logistic regression is a probability between 0 and 1. In this study, we differentiated between good and bad creditors (Brown and Mues, 2012). The weights no longer affected the probability linearly. The weighted sum was changed into a probability by using the logistic function. The equation was rewritten so that only the linear term was on the right side of the formula (4-4).

$$\log\left(\frac{P(y = 1)}{1 - P(y = 1)}\right) = \log\left(\frac{P(y = 1)}{P(y = 0)}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p \quad (4-4)$$

We call the term in the log function the ‘log odds’, which is the probability of an event divided by the probability of no event.

Equation 4-4 shows that the logistic regression model is a linear model for the log odds. However, we can work out how the prediction changes when one of the features x_j is changed by one unit. To do that, we apply the $\exp()$ function to both sides of the equation (4-5).

$$\frac{P(y = 1)}{1 - P(y = 1)} = odds = \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p) \quad (4-5)$$

Then, we compare the chances when we increase one of the x ’s by one unit. Except, instead of looking at the difference, we look at the ratio between the two predictions (4).

$$\frac{odds_{x_{j+1}}}{odds} = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_j (x_j + 1) + \beta_p x_p)}{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_j x_j + \dots + \beta_p x_p)} \quad (4-6)$$

We apply the following rule (4-7):

$$\frac{\exp(a)}{\exp(b)} = \exp(a - b) \quad (4-7)$$

And we remove several terms ((4-8).

$$\frac{\text{odds}_{x_j+1}}{\text{odds}} = \exp(\beta_j(x_j + 1) - \beta_j x_j) = \exp(\beta_j) \quad (4-8)$$

In the end, we arrive at $\exp(\)$ as a feature weight. A change in a feature by one unit alters the odds ratio (multiplicative) by a factor of $\exp(\beta_j)$. So, a change in x_j by one unit increases the log odds ratio by the value of the corresponding weight.

4.3.2 Decision Tree

In classification decision trees, a single node is the preliminary point, and this is followed by binary differences (1, 0). This provides the most evidence about a given class (Speybroeck, 2012). The procedure is repeated with the resultant new node until we arrive at a place to stop. Usually, the tree is too big, so it is back-tested using cross-validation. The dependent variable Y is categorical in type, so we can use information theory to calculating how much we know about it given the value of another separate variable A , as shown in Equation (4-9)

$$I[Y; A] = \sum_a \Pr(A = a) I[Y; A = a] \quad (4-9)$$

Where $I[Y; A]$ is the value of the uncertainty about Y , which decreases from knowing that $A = a$, given that we go from the full population to a sample where $A = a$.

4.3.3 Support Vector Machines

By assuming a training set of $N \{(X_i, Y_i)\}$ $N_i = 1$ with input data $X_i \in R^n$ and consistent binary class labels $Y_i \in \{-1, +1\}$, the support vector machines classifier in Vapnik's (Vapnik, 2000) theory satisfies Equation (4-10)

$$y_i[w^T \varphi(x_i) + b] \geq 1, \quad i = 1, \dots, N \quad (4-10)$$

The non-linear function $\varphi(\cdot)$ designates the input space for a high-dimensional feature space (Baesens et al., 2003). In this space that has been created, the variations are used to construct a hyperplane $w^T \varphi(x) + b = 0$, distinguishing between two classes. In the original weighted space, the following equation is used for the classifier (4-11):

$$y(x) = \text{sign}[w^T \varphi(x) + b] \quad (4-11)$$

However, it is never assessed in this form where the curved optimization problem could be defined as follows (4-12):

$$\min_{w, b, \xi} j(w, b, \xi) = \frac{1}{2} w^T w + C \sum_{i=1}^N \xi_i \quad (4-12)$$

This is subject to (4-13):

$$\begin{cases} y_i[w^T \varphi(x_i) + b] \geq 1 - \xi_i, & i = 1, \dots, N \\ \xi_i \geq 0, & i = 1, \dots, N \end{cases} \quad (4-13)$$

The variables used in ξ_i are loose variables that are necessary to allow for misclassifications to occur in the set of inequalities due to the overlying distribution. The first segment of the objective function is set to maximize the margin between the two classes in the feature space. The second part is set to minimize errors from misclassification.

4.3.4 Bayesian Network

Bayesian Network is a simple, high-performance classifier (Baesens et al., 2003). The model works by learning the class condition probability $p(x_i|y)$ from each input variable x_i , where $i = 1 \dots n$, given the class label y . A new observation is classified using Bayes' rule to calculate the probability of each class of y , given the vector of observed feature values (4-14).

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \quad (4-14)$$

To make the process easier, there is an assumption about the naïve Bayes classifier that the features are independent given the class label, which gives Equation 4-15.

$$p(x|y) = \prod_{i=1}^n p(x_i|y) \quad (4-15)$$

The probabilities $p(x_i|y)$ are then estimated using frequency counts for the discrete features and a normal method for the continuous features.

4.3.5 Bagging

Bootstrap aggregating, often known as bagging, involves giving each model in the ensemble an equal weight in the random forest tree. To encourage model variance, bagging trains each model in the ensemble using a random subset of the training set, leading to very high classification accuracy (Breiman, 1996).

In bagging, the samples are created in such a way that they differ from each other, however, replacement is allowed. This can happen in multiple samples at multiple times. Alternatively, it may not show up in some samples at all. These samples are then given to several learners, and the results from each learner are combined in the form of voting.

4.3.6 AdaBoostM1

The outputs of the other learning algorithms (known as ‘weak learners’) are combined into a weighted sum that represents the final output of the boosted classifier (Friedman, Tibshirani and Hastie, 2000). Originally, AdaBoost was designed so that at every step the sample distribution was adapted to place more weight on misclassified samples and less weight on correctly classified samples. The final prediction is then weighted as an average of all the weak learners, with more weight placed on stronger learners (4-16).

$$\sum_{n=1}^N w_n \exp(-y_n f(X_n)) \quad (4-16)$$

Where

- $y_n \in \{-1, +1\}$ is the true class label.
- w_n are observation weights normalized to add up to 1.
- $f(x_n) \in (-\infty, +\infty)$ is the predicted classification score.

4.3.7 Random Under-Sampling (RUSBoost)

Gradient boosting, which is an ensemble algorithm developed by (Friedman, 2002), is used to calculate the probability of a person defaulting on their loan. It relies on incremental minimization of the error term, which makes the prediction function more precise (Brown and Mues, 2012). Once the learner base is set, every tree calculated is fit to the ‘pseudo residuals’, which are deviations from the median rather than the expectation. This generally lowers the level of error. This is shown by Equation (4-17).

$$F(x) = G_0 + \beta_1 T_1(x) + \beta_2 T_2(x) + \dots + \beta_n T_n(x) \quad (4-17)$$

G_0 is the initial value for the set. $T_1 \dots T_n$ are the trees and $\beta_1 \dots \beta_n$ are the coefficients for particular tree nodes calculated by the algorithm. For the gradient boosting classifier to work, a maximum branch size needs to be set. For this study, 30 learning cycles and a 0.1 learn rate were set.

Random under-sampling boosting (RUSBoost) is particularly effective at classifying imbalanced data. It works by setting K classes. Then, for each weak learner in the ensemble, RUSBoost takes a subset of the data with N observations from each of the K classes. The boosting process follows that in AdaBoostM1 for reweighting and constructing the ensemble (Seiffert *et al.*, 2008).

Due to aggregation, the large number of observations, and the skewed data sets, we ran the models and focused almost entirely on the ensemble methods (Bag, AdaBoostM1, RUSBoost, LogitBoost, GentelBoost and Robust Boost). The

reason for this was that these methods combined several of the approved methods. They created some enhancements, particularly when we compared the AUC results from our data. Furthermore, they were based on several Random Forest Trees that had been shown to be efficient for use with our data (Albarrak *et al.*, 2020).

4.3.8 LogitBoost

LogitBoost works similarly to AdaBoostM1, except that it minimizes binomial deviance (Friedman, Tibshirani and Hastie, 2000). Binomial deviance assigns less weight to badly misclassified material. On average, LogitBoost can provide more accuracy than AdaBoostM1 for data with classes that are difficult to separate (4-18).

$$\sum_{n=1}^N w_n \log(1 + \exp(-2y_n f(x_n))) \quad (4-18)$$

Where

- $y_n \in \{-1, +1\}$ is the true class label.
- w_n are observation weights normalized to add up to 1
- $f(x_n) \in (-\infty, +\infty)$ is the predicted classification score

Binomial deviance allocates less weight to seriously misclassified observations (observations with large negative values for $y_n f(x_n)$). LogitBoost can give better average accuracy than AdaBoostM1 for data with poorly distinguishable classes. Learner t in a LogitBoost ensemble fits a regression model to respond to specific values, as shown by Equation (4-19).

$$\tilde{y}_n = \frac{y_n^* - p_t(x_n)}{p_t(x_n)(1 - p_t(x_n))} \quad (4-19)$$

Where

- $y_n^* \in \{0, c + 1\}$ are re-labelled classes (0 instead of -1)

- $p_t(x_n)$ is the current ensemble estimate of the probability for observation x_n to be of class 1

4.3.9 GentleBoost

GentleBoost combines features of AdaBoostM1 and LogitBoost. Like AdaBoostM1, GentleBoost minimizes exponential loss. But its numeric optimization is set up differently (Friedman, Tibshirani and Hastie, 2000). Like LogitBoost, every weak learner fits a regression model to response values, as in Equation (4-20)

$$\sum_{n=1}^N d_n^{(t)} (\tilde{y}_n - h_t(X_n))^2 \quad (4-20)$$

Where

- $d_n^{(t)}$ are observation weights at step t (the weights add up to 1).
- $h_t(x_n)$ are predictions of the regression model h_t , fitted to the response values y_n .

4.3.10 RobustBoost

Boosting algorithms such as AdaBoostM1 and LogitBoost increase the weights for misclassified observations at every phase of the boosting process. These weights can become very large. Thus, boosting algorithms occasionally focus on a few misclassified observations and disregard the bulk of the training data. Accordingly, the average classification accuracy suffers. In this case, the robust boosting comes into play. This algorithm does not allocate the full data weight to badly misclassified observations, thereby leading to better average classification accuracy.

Unlike AdaBoostM1 and LogitBoost, RobustBoost does not minimize a loss function. Instead, it maximizes the number of observations with the classification margin above a certain threshold (Freund, 2009).

RobustBoost develops as it evolves over time. The algorithm starts at $t = 0$. At every step, RobustBoost resolves an optimization problem to find a positive step in time Δt and a consistent positive change in the average margin for training data Δm . RobustBoost stops training and exits if at least one of these three conditions is true: time t reaches 1; it can find no solution to the optimization problem with positive updates Δt ; or Δm grows to as many learners as requested.

Results are usable for any exit condition. Cross-validation is advised to estimate the classification accuracy.

4.3.11 Key performance measures

The selection of the appropriate model will be through an evaluation of the AUC for the Receiver Operating Characteristic (ROC) curve, as shown in Figure 4-3 (Yang and Berdine, 2017). To interpret these results more effectively, we could define the ROC as a curve connecting more predictive data points starting at a point less than the lowest value observed (0,0) and ending at a point greater than the highest value observed (1,1).

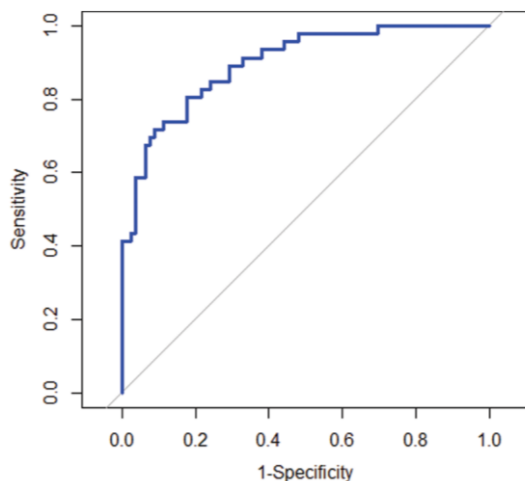


Figure 4-3: Hypothetical ROC curve

Source: (Yang and Berdine, 2017: 34).

The AUC is the entire area underneath the ROC curve. The diagonal line between points (0,0) and (1,1) indicates that the values on this line are not indicative of a better estimate than a random guess (AUC = 0.5). The further along the point is

in the ROC space above the diagonal line, the better the predictive value of the test.

Next, an overlook for the true positive rates and false negative rates from the confusion matrix is conducted to enhance the analysis. Through the information developed in the confusion matrix (Table 4-1), we can calculate the accuracy rate, type I error, and type II error.

Table 4-1: Confusion matrix for credit scoring

		Actual condition	
		Positive (non-risk)	Negative (risk)
Test result	Positive (non-risk)	True positive (TP)	False positive (FP)
	Negative (risk)	False negative (FN)	True negative (TN)

The three measures are defined by Equations (4-21), (4-22) and (4-23)

$$\text{Average accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \quad (4-21)$$

$$\text{Type I Error} = \frac{FN}{TP + FN} \quad (4-22)$$

$$\text{Type II Error} = \frac{FP}{TN + FP} \quad (4-23)$$

To cross-validate the results, we used a k-fold cross, which folds the dataset and takes various portions of it to train the model at random. It then tests the remaining instances. This is an effective way of checking whether the model works well for the training samples and the test samples. We picked a training set that was 70% of the dataset and a test set that was 30% of the dataset. A range of greater than 0.7 and less than or equal to 0.8 of the AUC is considered excellent discrimination in the medical sector (Yang & Berdine, 2017). The medical field is worth considering because it involves high levels of risk.

4.3.12 Code Used for Implementing Methods

4.3.12.1 Default Prediction Modelling (Individual Banks)

1. Default cases collection: the main function for this step is 'loandefault_calculation.m'. This function loads 'CBK5_raw' and 'CBK5' in the data folder, then generates the default indicators (with 1 indicating default, and 0 indicating non-default) for 'Consumer' loans and 'Instalment' loans separately. More concretely, this function takes outstanding loans ('CBK5.outstanding'), loan principles ('CBK5.principle'), and interest (CBK5.interest) from CBK5, then takes 'RiskFactorNumber' from 'CBK5_raw' to generate the default indicators by checking if the 'RiskFactorNumber' is 1 or not (if it is not 1 then it is default, whereas anything else means non-default). It combines all these variables into a new table ('tablewithdef') and saves the table to the data folder. This table has seven columns with the following headers: 'Customer ID', 'Dates', 'Outstanding', 'Principle', 'Interest', 'RiskFactorNumber', and 'defindicators'.

2. Explanatory variables collection: two functions are used in this step: 'add_creditcardexp.m' and 'add_othervars.m'. In 'add_creditcardexp.m', 'CBK6' is loaded from the data folder to calculate credit exposure (for which the variable name is 'exposure'). This equals 'CBK6.Outstanding_Bal' minus 'CBK6.limit'. This 'exposure' variable is added to the 'tablewithdef' from the function 'loandefault_calculation.m'. The new table is renamed 'tablewithdef_creditexp' and saved in the data folder for 'Consumer' and 'Instalment' separately. In 'add_othervars.m', 'CBK3' is loaded to collect data on monthly incomes. 'CBK2' is loaded to collect data regarding age, gender, and length of relationship with

the bank (calculated as the number of calendar days between 'CBK2.DateOfRelationship' and 2019-12-31, divided by 365). 'CBK5' is loaded to collect data about the number of accounts. 'CBK1' is loaded to collect cash flow data. These variables are added to the table 'tablewithdef_creditexp' from the function 'add_creditcardexp.m'. The new table is renamed 'fulltable' and duplicate are removed. The table is saved to the data folder for 'Consumer' and 'Instalment' separately.

3. Explanatory variables correlation matrix: the main function in this step is 'vars_correlation_credit_default.m'. This function takes the input 'credittype', which can be either 'Instalment' or 'Consumer'. Then, it loads the corresponding 'fulltable' generated in the function 'add_othervars.m'. MATLAB's built-in 'corr.m' function is used to compute the correlation matrix for all the explanatory variables: 'exposure', 'income', 'age', 'gender', 'relationshipduration', 'mthly_num_tx', 'mthly_ave_cf', and 'count'. This correlation matrix is saved as a table named 'corr_tbl' in the data folder (MathWorks, 2021c).

4. Default prediction model calibration: the main function in this step is 'credit_default_prob_withmodelinput.m'. This function takes the inputs 'credittype' (which can be 'Instalment' or 'Consumer') and 'modelstr' (which can be 'Bag', 'RUSBoost', or 'AdaBoostM1'). The function loads the corresponding 'fulltable' generated with the function 'add_othervars.m', specifying the predictors as 'exposure', 'income', 'age', 'gender', 'rRelationshipduration', 'mthly_num_tx', and 'mthly_ave_cf', 'count'. The response variable is 'defindicators'. When training the model, 30% of the data is reserved for cross-validation. Once the model is trained, it is saved as 'trainedClassifier' in the data folder. The outputs of this function include: in-sample and out-of-sample ROC curves with AUC values printed in the titles; a plot showing the relative importance of all the predictor variables; a plot of the decision tree based on the calibrated machine learning model (Course, 2020).

The machine learning models used in this function are based on ensemble algorithms from MATLAB (MathWorks, 2021b; MathWorks, 2021e)

5. Default prediction for practical use: the main function in this step is 'credit_default_prob_prediction_withmodelinput.m'. This function uses the following inputs: 'modelstr', 'exposure', 'income', 'age', 'gender', 'relationshipduration', 'mthly_ave_cf', 'mthly_num_tx', 'count'. The first input, 'modelstr', can be 'Bag' or 'tree'. The rest of the inputs are predictors that the user needs to collect from the individuals whose credit risks are being assessed. This function loads the calibrated machine learning models ('trainedClassifier') for 'Instalment' and 'Consumer' from the data folder. These models use the following inputs: 'exposure', 'income', 'age', 'gender', 'relationshipduration', 'mthly_ave_cf', 'mthly_num_tx', and 'count'. They calculate the probability of credit defaulting for both 'Instalment' and 'Consumer' loan types. The values of the predictors and the calculated likelihood of defaulting are printed in the MATLAB command window.

4.3.12.2 Default Prediction Modelling (All Banks)

1. All data collection: the main function for this step is 'get_all_data.m'. This function takes the input 'creditttype' (which can be 'Instalment' or 'Consumer'), loops through the data folders for the seven banks, and loads the corresponding 'fulltable'. The function then finds the common variables in 'fulltable' across all seven banks. These common variables are collected in a big table named 'alldata' and saved in the data folder for 'Instalment' or 'Consumer' depending on the original 'creditttype' input.

2. Find the best model: the main function in this step is 'credit_default_prob_best_model.m'. This function takes input 'creditttype' (which can be 'Instalment' or 'Consumer') and loads 'alldata' saved in 'get_all_data.m'. Six machine learning models are THEN used as candidates to find the best model. These six models are 'RobustBoost', 'GentleBoost', 'LogitBoost', 'Bag', 'RUSBoost', and 'AdaBoostM1' (MathWorks, 2021f)(MathWorks, 2021d)(MathWorks, 2021a). As with 'credit_default_prob_withmodelinput.m', when training each machine learning model, 30% of the data are held out for cross-validation. Out-of-sample AUC ('AUC_validation') is recorded for each model in the table 'all_AUC_validation'. The best model is the one with the highest 'AUC_validation' and its information is collected in 'model_best'. It is

saved in the data folder with the 'trainedClassifier'. The outputs of this function include in-sample and out-of-sample ROC curves with the best model's name and the AUC value printed in the titles; a plot showing the relative importance of all the predictor variables; and a plot of the decision tree based on the best machine learning model if it is 'tree', 'RUSBoost', or 'AdaBoostM1'.

3. Default prediction based on the best model for practical use: the main function in this step is 'credit_default_prob_prediction_bestmodel.m'. This function takes the following inputs: 'exposure', 'income', 'age', 'gender', 'relationshipduration', 'mthly_ave_cf', 'mthly_num_tx', 'count'. The inputs are predictors the user needs to collect from the individuals whose credit risks are being assessed. This function loads the best machine learning models ('trainedClassifier') for 'Instalment' and 'Consumer' from the data folder. These models use the inputs 'exposure', 'income', 'age', 'gender', 'relationshipduration', 'mthly_ave_cf', 'mthly_num_tx', and 'count' to calculate the probability of defaulting for both 'Instalment' and 'Consumer' types. The values of the predictors and the calculated probabilities of defaulting are printed in the MATLAB command window.

5 RESULTS AND DISCUSSION

The aim of this study is to explore how computationally intelligent data could be handled by algorithms in the banking sector. We wanted to know how likely customers were to default based on their characteristics, income, and product portfolio. This would help banks to select the most suitable risk weight for customers, thereby boosting the performance and profitability of the bank.

We began by identifying the independent variables: age, gender, nationality, income, education, and number of loans, client-bank relationship duration, credit card exposure ('outstanding balance'), monthly average net cashflow, and monthly average number of transactions. Several of these variables had been analysed in previous studies, as described in the literature review. Others, however, had not previously been analysed and were included here to provide additional data.

The study included six banks, three of which were 'Islamic' and three of which were 'Conventional'. We examined the AUC results, confusion matrix, running time, training samples, and testing samples. This chapter gives an account of the results for each of the six banks. It also provides an overview of the key findings for the central bank.

5.1 Correlation variables

Several studies have used correlation analysis to discover the level of connection between variables. Linear correlation analysis is a tool for demonstrating the closeness between two associated variables. The linear correlation coefficient (r or R) is a measure of how close two variables are to each other. (Hauke and Kossowski, 2011) support using the Pearson's product-moment correlation coefficient (R or r) as a measure of the linear connection between variables. This coefficient measures the degree of linear connection between variables and demonstrates whether it falls into a normal distribution. Pearson's formula for quantifying the degree of relationship (R) between variables is given by (5-1)

$$R = \frac{n(\sum XY) - (\sum X) \cdot (\sum Y)}{\sqrt{(n(\sum X^2) - (\sum X)^2)(n(\sum Y^2) - (\sum Y)^2)}} \quad (5-1)$$

Where:

n = Number of observations

X = Measures of variable 1

Y = Measures of variable 2

$\sum XY$ = Sum of product of respective variable measures

$\sum X$ = Sum of the measures of variable 1

$\sum Y$ = Sum of the measures of variable 2

$\sum X^2$ = Sum of squared values of the measures of variable 1

$\sum Y^2$ = Sum of squared values of the measures of variable 2

Based on the outcome of the calculation, the degree of correlation can be positive, zero, or negative. According to (Gogtay and Thatte, 2017), R can be understood based on its rate, as shown by Figure 5-1.

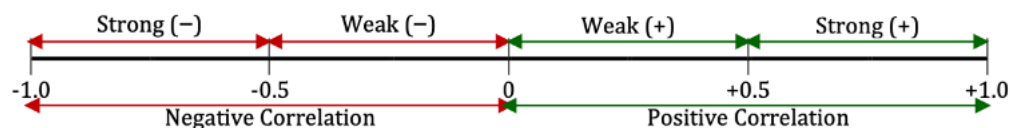


Figure 5-1: Basic spectrum for interpreting the correlation coefficient (R).

Source: Gogtay and Thatte (2017: 79)

We analysed the correlation coefficients for each of the variables for every bank. We also performed the same analysis with the consolidated data set to test the closeness of the variables for the Central Bank model. The details of the analysis are provided in the following sections.

5.1.1 Bank 1

Table 5-1 gives the correlation analysis for instalment loans at Bank 1. It shows that there is a positive linear relation between age and the length of a customer's relationship with the bank (0.44). There is a slight positive relation between income and the monthly number of transactions (0.22).

Table 5-1: Bank 1 correlation coefficients between selected variables for instalment loans

	'exposure'	'Income'	'age'	'gender'	'education'	'Nationality'	'Relationship duration'	'monthly num tx'	'monthly ave cf'	'count'
'exposure'	1.00	-0.16	0.06	-0.06	0.01	0.00	-0.07	-0.11	-0.02	0.01
'Income'	-0.16	1.00	0.03	0.09	-0.04	0.02	0.11	0.22	0.00	0.11
'age'	-0.06	0.03	1.00	0.08	0.11	0.03	0.44	-0.13	0.02	-0.10
'gender'	-0.06	0.09	0.08	1.00	0.19	0.03	0.03	0.11	0.00	0.02
'education'	0.01	-0.04	0.11	0.19	1.00	0.11	0.08	-0.04	0.00	-0.01
'Nationality'	0.00	0.02	0.03	0.03	0.11	1.00	-0.06	0.03	-0.01	0.08
'Relationship duration'	-0.07	0.11	0.44	0.03	0.08	-0.06	1.00	-0.01	0.03	0.05
'monthly num tx'	-0.11	0.22	0.13	0.11	-0.04	0.03	-0.01	1.00	-0.03	0.12
'monthly ave cf'	-0.02	0.00	0.02	0.00	0.00	-0.01	0.03	-0.03	1.00	-0.01
'count'	0.01	0.11	0.10	0.02	-0.01	0.08	0.05	0.12	-0.01	1.00

The same is true for consumer loans (see Table 5-2), for which there is a positive correlation between age and length of relationship with the bank (0.56) and a positive correlation between monthly number of transactions and income (0.34). There is a negative correlation between credit card exposure and income (-0.22).

Table 5-2: Bank 1 correlation coefficients between selected variables for consumer loans

	'exposure'	'Income'	'age'	'gender'	'education'	'Nationality'	'Relationship duration'	'monthly num tx'	'monthly ave cf'	'count'
'exposure'	1.00	-0.22	0.05	-0.05	0.01	-0.01	-0.07	-0.13	-0.01	0.05

	'exposure'	'Income'	'age'	'gender'	'education'	'Nationality'	'Relationship duration'	'monthly num tx'	'monthly average'	'count'
'Income'	-0.22	1.00	0.03	0.08	-0.04	0.06	0.15	0.34	0.00	-0.06
'age'	-0.05	0.03	1.00	0.17	0.14	0.00	0.56	-0.08	0.01	0.29
'gender'	-0.05	0.08	0.17	1.00	0.17	-0.02	0.09	0.13	-0.02	0.02
'education'	0.01	-0.04	0.14	0.17	1.00	0.06	0.12	-0.04	-0.01	0.02
'Nationality'	-0.01	0.06	0.00	-0.02	0.06	1.00	0.03	0.08	-0.01	0.00
'Relationship duration'	-0.07	0.15	0.56	0.09	0.12	0.03	1.00	0.04	0.01	0.24
'monthly num tx'	-0.13	0.34	-0.08	0.13	-0.04	0.08	0.04	1.00	-0.05	-0.06
'monthly average'	-0.01	0.00	0.01	-0.02	-0.01	-0.01	0.01	-0.05	1.00	0.00
'count'	0.05	-0.06	0.29	0.02	0.02	0.00	0.24	-0.06	0.00	1.00

The correlation results for both types of loans at Bank 1 suggest that there is no strong relation between variables. The results can be used for the study because they are far from 1 and -1.

5.1.2 Bank 2

Table 5-3 gives the correlation analysis for instalment loans at Bank 2. It shows that there is a positive linear relation between age and length of relationship with the bank (0.64). There is a negative linear relationship between income and credit card exposure (-0.45).

Table 5-3: Bank 2 correlation coefficients between selected variables for instalment loans

	'exposure'	'Income'	'age'	'gender'	'education'	'Relationship duration'	'monthly num tx'	'monthly average'
'exposure'	1.00	-0.45	0.26	-0.13	0.11	-0.18	-0.10	-0.01
'Income'	-0.45	1.00	0.26	0.21	-0.12	0.20	0.20	0.01

	'exposure'	'Income'	'age'	'gender'	'education'	'Relationship duration'	'mthly num tx'	'mthly ave cf'
'age'	-0.26	0.26	1.00	0.04	-0.01	0.64	-0.12	0.00
'gender'	-0.13	0.21	0.04	1.00	0.14	0.05	0.01	0.00
'education'	0.11	-0.12	-	0.14	1.00	0.04	-0.09	0.00
'Relationship duration'	-0.18	0.20	0.64	0.05	0.04	1.00	-0.01	0.00
'mthly num tx'	-0.10	0.20	-	0.01	-0.09	-0.01	1.00	0.00
'mthly ave cf'	-0.01	0.01	0.00	0.00	0.00	0.00	0.00	1.00
'count'	0.08	0.03	-	0.03	0.00	0.03	0.01	0.00

Table 5-4 details consumer loans for Bank 2. It shows similar trends. There is a positive linear correlation between age and length of relationship (0.55) and a negative linear correlation between income and credit card exposure (-0.45).

Table 5-4: Bank 2 correlation coefficients between selected variables for consumer loans

	'exposure'	'Income'	'age'	'gender'	'education'	'Relationship duration'	'mthly num tx'	'mthly ave cf'
'exposure'	1.00	-0.45	-	-0.11	0.08	-0.23	-0.13	-0.01
'Income'	-0.45	1.00	0.23	0.15	-0.07	0.26	0.26	0.01
'age'	-0.25	0.23	1.00	0.10	-0.05	0.55	-0.14	0.00
'gender'	-0.11	0.15	0.10	1.00	0.07	-0.01	-0.03	0.00
'education'	0.08	-0.07	-	0.07	1.00	0.05	-0.04	0.00
'Relationship duration'	-0.23	0.26	0.55	-0.01	0.05	1.00	0.06	0.00
'mthly num tx'	-0.13	0.26	-	-0.03	-0.04	0.06	1.00	0.00

	'exposure'	'Income'	'age'	'gender'	'education'	'Relationship duration'	'mthly num tx'	'mthly ave cf'
'mthly ave cf'	-0.01	0.01	0.00	0.00	0.00	0.00	0.00	1.00
'count'	0.06	-0.02	0.05	0.04	-0.05	-0.04	-0.03	0.00

The correlation analysis allows for the use of the selected variables because the maximum coefficients are close to the mean for the accepted range.

5.1.3 Bank 3

The results for Bank 3 were similar to those for Banks 1 and 2, but the hierarchy of relationships was slightly different. The positive linear correlation between credit card exposure and income (0.33) surpassed that between age and length of relationship (0.28). The results are shown in Table 5-5.

Table 5-5: Bank 3 correlation coefficients between selected variables for instalment loans

	'exposure'	'Income'	'age'	'gender'	'Relationship duration'	'mthly num tx'	'mthly ave cf'
'exposure'	1.00	0.33	0.10	0.10	0.01	0.09	0.03
'Income'	0.33	1.00	0.15	0.08	-0.04	0.26	0.04
'age'	0.10	0.15	1.00	0.00	0.28	-0.19	0.01
'gender'	0.10	0.08	0.00	1.00	-0.05	-0.03	0.00
'Relationship duration'	0.01	-0.04	0.28	-0.05	1.00	-0.10	0.00
'mthly num tx'	0.09	0.26	-0.19	-0.03	-0.10	1.00	0.02
'mthly ave cf'	0.03	0.04	0.01	0.00	0.00	0.02	1.00
'count'	-0.02	0.09	-0.11	0.00	-0.04	0.28	0.00

Table 4-6 shows that the same is true for consumer loans. There is a positive linear relationship between credit card exposure and income (0.51) and a positive linear relationship between age and length of relationship (0.35). These figures are larger than those for instalment loans.

Table 5-6: Bank 3 correlation coefficients between selected variables for consumer loans

	'exposure'	'Income'	'age'	'gender'	'Relationship duration'	'mthly num tx'	'mthly ave cf'
'exposure'	1.00	0.51	0.10	0.07	0.20	0.22	0.01
'Income'	0.51	1.00	0.17	0.09	0.30	0.40	0.02
'age'	0.10	0.17	1.00	0.14	0.35	-0.05	0.01
'gender'	0.07	0.09	0.14	1.00	0.10	-0.01	0.01
'Relationship duration'	0.20	0.30	0.35	0.10	1.00	0.16	0.00
'mthly num tx'	0.22	0.40	-0.05	-0.01	0.16	1.00	0.00
'mthly ave cf'	0.01	0.02	0.01	0.01	0.00	0.00	1.00
'count'	0.04	0.09	0.00	0.02	0.11	0.15	0.00

The correlation coefficient results for Bank 3 mean that we can include all the available variables since they are not highly correlated either positively or negatively.

5.1.4 Bank 4

The correlation coefficient results for instalment loans at Bank 4 are presented in Table 5-7. They show that there is a slight positive linear relation between age and length of relationship (0.29) and between age and income (0.25). There is a slight negative linear relation between income and credit card exposure (-0.25). If someone's income level increases, then their credit card usage decreases. This result is different from the other banks.

Table 5-7: Bank 4 correlation coefficients between selected variables for instalment loans

	'exposure'	'Income'	'age'	'gender'	'Relationship duration'	'mthly num tx'	'mthly ave cf'	'count'
'exposure'	1.00	-0.25	-0.19	-0.10	-0.12	-0.05	-0.01	-0.01
'Income'	-0.25	1.00	0.25	0.18	0.12	0.06	0.02	0.01

	'exposure'	'Income'	'age'	'gender'	'Relationship duration'	'mthly num tx'	'mthly ave cf'	'count'
'age'	-0.19	0.25	1.00	-0.01	0.29	-0.05	0.03	-0.02
'gender'	-0.10	0.18	-0.01	1.00	0.14	0.02	0.01	0.09
'Relationship duration'	-0.12	0.12	0.29	0.14	1.00	-0.04	0.01	0.11
'mthly num tx'	-0.05	0.06	-0.05	0.02	-0.04	1.00	0.01	0.02
'mthly ave cf'	-0.01	0.02	0.03	0.01	0.01	0.01	1.00	-0.01
'count'	-0.01	0.01	-0.02	0.09	0.11	0.02	-0.01	1.00

Table 5-8 shows that the results for consumer loans are similar. There is a slightly positive linear correlation between age and length of relationship (0.33) and between age and income (0.24). There is a slightly negative linear correlation between credit card exposure and income (-0.25).

Table 5-8: Bank 4 correlation coefficients between selected variables for consumer loans

	'exposure'	'Income'	'age'	'gender'	'Relationship duration'	'mthly num tx'	'mthly ave cf'	'count'
'exposure'	1.00	-0.25	-0.22	-0.16	-0.14	-0.04	-0.03	-0.04
'Income'	-0.25	1.00	0.24	0.20	0.13	0.05	0.03	0.07
'age'	-0.22	0.24	1.00	0.10	0.33	-0.06	0.05	0.12
'gender'	-0.16	0.20	0.10	1.00	0.20	0.02	0.02	0.11
'Relationship duration'	-0.14	0.13	0.33	0.20	1.00	-0.05	0.03	0.11
'mthly num tx'	-0.04	0.05	-0.06	0.02	-0.05	1.00	0.03	0.00
'mthly ave cf'	-0.03	0.03	0.05	0.02	0.03	0.03	1.00	0.01
'count'	-0.04	0.07	0.12	0.11	0.11	0.00	0.01	1.00

The correlation analysis for Bank 4 does not prevent the use of any of the selected variables.

5.1.5 Bank 5

Table 5-9 shows that the most strongly correlated variables at Bank 5 are length of relationship and number of loans. There was a positive linear relation between these variables (0.29). This may be because the bank is new, so there is not the same correlation between age and the length of a customer's relationship with the bank. This is also why there is a high correlation between the length of people's relationship with the bank the number of loans granted. There is also a slight negative linear correlation between credit card exposure and income (-0.23) as well as between credit card exposure and length of relationship (-0.24).

Table 5-9: Bank 5 correlation coefficients between selected variables for instalment loans

	'exposure'	'Income'	'age'	'gender'	'Relationship duration'	'mthly num tx'	'mthly ave cf'	'count'
'exposure'	1.00	-0.23	-0.15	-0.09	-0.24	-0.12	0.00	-0.06
'Income'	-0.23	1.00	0.16	0.12	0.11	0.18	0.00	0.04
'age'	-0.15	0.16	1.00	0.04	0.03	-0.11	0.00	-0.09
'gender'	-0.09	0.12	0.04	1.00	-0.04	-0.01	0.00	0.05
'Relationship duration'	-0.24	0.11	0.03	-0.04	1.00	0.06	0.00	0.29
'mthly num tx'	-0.12	0.18	-0.11	-0.01	0.06	1.00	0.00	0.09
'mthly ave cf'	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
'count'	-0.06	0.04	-0.09	0.05	0.29	0.09	0.00	1.00

The strongest positive linear relationship for consumer loans at Bank 5 was between age and income (0.18). The negative correlation between credit card exposure and income was stronger here than at any other bank (-0.23). There was also negative correlation between credit card exposure and age (-0.24).

Table 5-10: Bank 5 correlation coefficients between selected variables for consumer loans

	'exposure'	'Income'	'age'	'gender'	'Relationship duration'	'mthly num tx'	'mthly ave cf'	'count'
'exposure'	1.00	-0.23	-0.24	-0.12	-0.19	-0.10	0.00	0.02
'Income'	-0.23	1.00	0.18	0.15	0.13	0.15	0.00	0.03
'age'	-0.24	0.18	1.00	0.09	0.11	-0.13	0.01	0.03
'gender'	-0.12	0.15	0.09	1.00	0.03	0.02	0.01	0.00
'Relationship duration'	-0.19	0.13	0.11	0.03	1.00	0.08	-0.01	-0.20
'mthly num tx'	-0.10	0.15	-0.13	0.02	0.08	1.00	-0.01	0.05
'mthly ave cf'	0.00	0.00	0.01	0.01	-0.01	-0.01	1.00	0.00
'count'	0.02	0.03	0.03	0.00	-0.20	0.05	0.00	1.00

The correlation analysis shows that all the selected variables can be included because of the low correlation between them.

5.1.6 Bank 6

Table 5-11 shows that length of relationship and age have the strongest positive linear relationship for consumer loans at Bank 6 (0.52). There is also a positive linear relationship between income and monthly number of transactions (0.31) and between length of relationship and income (0.26).

Table 5-11: Bank 6 correlation coefficients between selected variables for instalment loans

	'exposure'	'Income'	'age'	'gender'	'Relationship duration'	'mthly num tx'	'mthly ave cf'	'count'
'exposure'	1.00	-0.16	-0.09	-0.02	-0.12	-0.09	-0.02	0.03
'Income'	-0.16	1.00	0.16	0.20	0.26	0.31	0.03	0.12
'age'	-0.09	0.16	1.00	-0.02	0.52	-0.02	0.01	0.00
'gender'	-0.02	0.20	-0.02	1.00	0.00	0.11	0.02	0.13
'Relationship duration'	-0.12	0.26	0.52	0.00	1.00	0.17	0.01	0.08
'mthly num tx'	-0.09	0.31	-0.02	0.11	0.17	1.00	0.01	0.11

'mthly ave cf'	-0.02	0.03	0.01	0.02	0.01	0.01	1.00	0.00
'count'	0.03	0.12	0.00	0.13	0.08	0.11	0.00	1.00

The correlation between variables was similar for consumer loans, as shown in Table 5-12. There is a positive linear relationship between length of relationship and age (0.52) and between monthly number of transactions and income (0.31).

Table 5-12: Bank 6 correlation coefficients between selected variables for consumer loans

	'exposure'	'Income'	'age'	'gender'	'Relationship duration'	'mthly num tx'	'mthly ave cf'	'count'
'exposure'	1.00	-0.23	-0.12	-0.03	-0.13	-0.07	-0.02	-0.03
'Income'	-0.23	1.00	0.18	0.25	0.22	0.31	0.02	0.13
'age'	-0.12	0.18	1.00	0.05	0.52	-0.05	0.00	0.10
'gender'	-0.03	0.25	0.05	1.00	0.00	0.10	0.02	0.16
'Relationship duration'	-0.13	0.22	0.52	0.00	1.00	0.15	0.01	0.09
'mthly num tx'	-0.07	0.31	-0.05	0.10	0.15	1.00	0.01	0.08
'mthly ave cf'	-0.02	0.02	0.00	0.02	0.01	0.01	1.00	0.00
'count'	-0.03	0.13	0.10	0.16	0.09	0.08	0.00	1.00

From both loans' correlation analysis, it has been indicated that the variables with the highest correlation are almost 50% away from the maximum range of 1. This supports the use of all variables as independent parameters for the research into this bank.

5.1.7 All Banks

The correlation analysis for the data for instalment loans from all the banks is provided in Table 5-13. As with the individual banks, the strongest linear correlation is between age and the length of relationship (0.49). There is also a slightly negative correlation between credit card exposure and income (-0.30). The data shows that there is no correlation between average monthly net cashflow and the other variables.

Table 5-13: Correlation coefficients between selected variables for instalment loans at all banks

	'exposure'	'Income'	'age'	'gender'	'Relationship duration'	'mthly num tx'	'mthly ave cf'	'count'
'exposure'	1.00	-0.30	-0.09	-0.09	-0.02	0.00	-0.01	0.03
'Income'	-0.30	1.00	0.13	0.15	0.10	0.15	0.01	0.04
'age'	-0.09	0.13	1.00	0.04	0.49	-0.09	0.00	-0.04
'gender'	-0.09	0.15	0.04	1.00	0.00	0.01	0.00	0.06
'Relationship duration'	-0.02	0.10	0.49	0.00	1.00	0.15	0.00	-0.10
'mthly num tx'	0.00	0.15	-0.09	0.01	0.15	1.00	0.00	0.03
'mthly ave cf'	-0.01	0.01	0.00	0.00	0.00	0.00	1.00	0.00
'count'	0.03	0.04	-0.04	0.06	-0.10	0.03	0.00	1.00

Table 4-14 provides the data for consumer loans at all banks. Again, the strongest positive linear correlation is between length of relationship and age (0.50). There is also a slightly negative linear relationship between credit card exposure and income (-0.36). The correlation data supports the addition of the new research variable, average monthly net cashflow, which has no correlation with the other variables.

Table 5-14: Correlation coefficients between selected variables for consumer loans at all banks

	'exposure'	'Income'	'age'	'gender'	'Relationship duration'	'mthly num tx'	'mthly ave cf'	'count'
'exposure'	1.00	-0.36	-0.10	-0.09	-0.09	-0.06	-0.01	0.06
'Income'	-0.36	1.00	0.09	0.13	0.15	0.23	0.00	-0.05
'age'	-0.10	0.09	1.00	0.11	0.50	-0.15	0.00	0.19
'gender'	-0.09	0.13	0.11	1.00	0.00	0.00	0.00	0.04
'Relationship duration'	-0.09	0.15	0.50	0.00	1.00	0.14	0.00	-0.01

'mthly num tx'	-0.06	0.23	-0.15	0.00	0.14	1.00	0.00	-0.09
'mthly ave cf'	-0.01	0.00	0.00	0.00	0.00	0.00	1.00	0.00
'count'	0.06	-0.05	0.19	0.04	-0.01	-0.09	0.00	1.00

Having conducted all the correlation analysis, the results show that there is only weak correlation between the variables. Therefore, all the listed variables can be included as parameters predicting the probability of a customer defaulting on their household loans (and this is the case for both consumer loans and instalment loans).

5.2 Models for Instalment Loans

5.2.1 Bank-by-Bank Models

5.2.1.1 Bank 1

Bank 1 is a medium-sized, conventional bank with 8,186 customers with instalment loans. 70 of these are at risk of defaulting, giving a default rate of 0.85%. This is shown in Table 5-15. There were 347,977 observations.

Table 5-15: Bank 1 observations and customers

Bank 1	Instalment Loans
Observations	347,977
Number of default cases per year	99
Non-Default cases	347,878
Rate of default cases	0.028%
Number of customers	8,186
Number of default customers	70
Number of non- default customers	8,116
Rate of default customers	0.855%

After analysing the AUC for the different methods (see Table 5-16) and the accuracy rate, type I error, and type II error (see

Table 5-17), it was decided that RUSBoost was the most effective model for this bank's data.

Table 5-16: Bank 1 AUC results

Bank 1	AUC Instalment loans
Logistic Regression	0.7
Decision Tree	0.72
Linear Support Vector Machines	0.59
Bayesian Network	0.66
Bagged	0.63
RUSBoosted	0.83
AdaBoostM1	0.81

Table 5-17: Bank 1 confusion matrix

Bank 1	True Positive	True Negative	False Positive	False Negative	Average Accuracy	Type I Error	Type II Error
Logistic Regression	99%	N/A	N/A	1%	N/A	1%	N/A
Decision Tree	99%	5%	95%	1%	52%	1%	49%
Linear Support Vector Machines	99%	N/A	N/A	1%	N/A	1%	N/A
Bayesian Network	99%	N/A	100%	1%	N/A	1%	50%
Bagged	99%	N/A	100%	1%	N/A	1%	50%
RUSBoosted	99%	1%	99%	1%	50%	1%	50%
AdaBoostM1	99%	N/A	N/A	1%	N/A	1%	N/A

Although the accuracy rate of decision tree was marginally higher than RUSBoosted, RUSBoosted surpassed decision tree in AUC results. From the running time in Table 5-18, it performed as well as the other methods, taking less than one minute to train. RUSBoosted could performed more observations per second, meaning that it would be more efficient with more observations.

Table 5-18: Bank 1 running time of models

Bank 1	Instalment Loans	
	Prediction Speed	Training Time
Logistic Reg.	~370000 obs/sec	684.77 sec
Tree	~1200000 obs/sec	20.712 sec
SVM	~230000 obs/sec	41.375 sec
Bayesian	~ 850000 obs/sec	9.2544 sec
RUSBoosted	~1300000 obs/sec	41.875 sec
Bagged	~150000 obs/sec	153.09 sec
AdaBoostM1	~140000 obs/sec	210.7 sec

To ensure our results were accurate, we performed a k-fold measurement for the RUSBoosted model, using 70% of the data set as an in-sample training set. This gave an AUC of 0.98. The remaining 30% of the data set was used as an out-of-sample test, producing an AUC of 0.92. These results are illustrated in Figure 5-2.

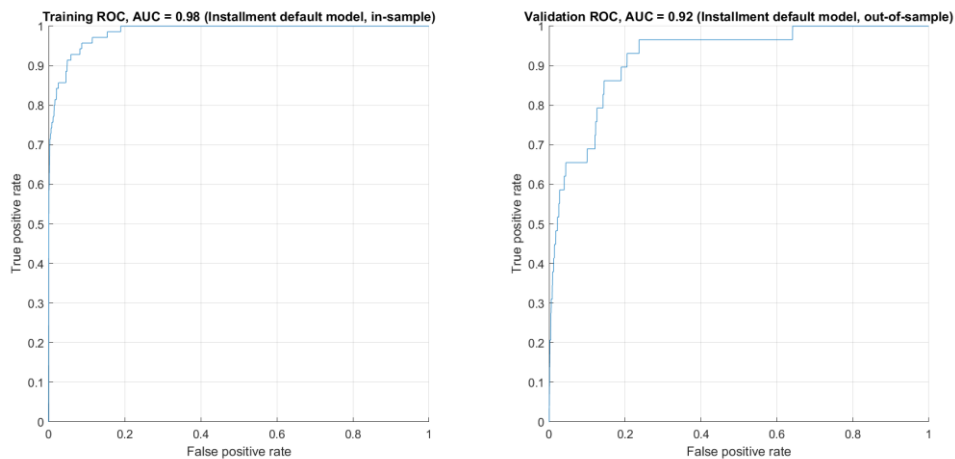


Figure 5-2: Bank 1 AUC results for 70% in-sample and 30% out-of-sample

After training the classifier with the model, we calculated the relative importance of the parameters to acknowledge the importance of each variable and its impact on the prediction model in the future. Figure 5-3 shows the predictive importance for each of the input variables.

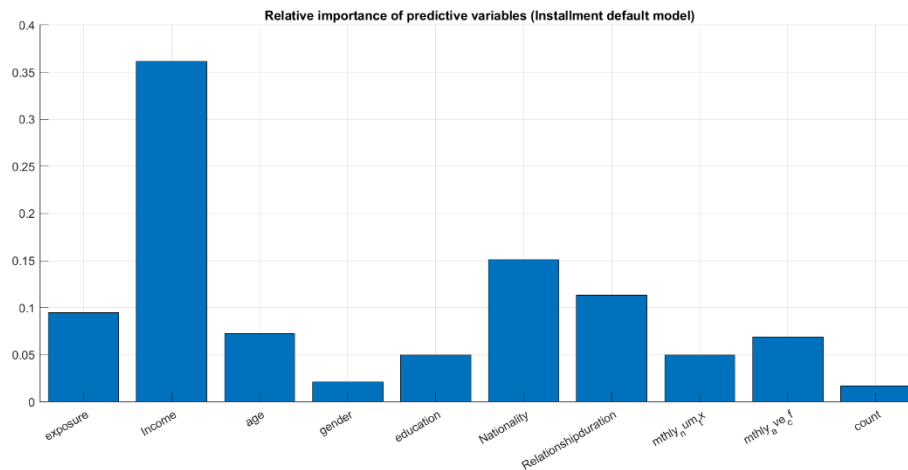


Figure 5-3: Relative importance of the predictive variables for Bank 1

As expected, customers' monthly income level was the most influential variable when it came to classifying the customers at this bank. Nationality was the second most influential variable, which could be because people from Kuwait tend to have higher incomes than non-Kuwaitis. The length of a customer's relationship with the bank was the third most influential variable, which is understandable because it is a measure of their loyalty to the bank. The fourth most influential variable was credit card exposure, which is an indicator of total monthly obligations. Customers' monthly average cash flow from their account had a medium level of importance. This was not expected because we thought that purchasing behaviour might influence people's attempts to meet their monthly obligations.

5.2.1.2 Bank 2

Bank 2 is a large conventional bank. Of the 100,000 instalment loan customers we sampled, we examined 12,048 and 7,136 were default cases, a default rate of 59.23%. There were 664,707 observations. These results are shown in Table 5-19.

Table 5-19: Bank 2 number of observations and customers

Bank 2	Instalment Loans
Observations	664,707
Default cases	17,515
Non-Default cases	647,192
Rate of default cases	2.635%

Number of customers	12,048
Number of default customers	7,136
Number of non- default customers	4,912
Rate of default customers	59.230%

After analysing the AUC for the different methods, as well as the accuracy rate, type I error, and type II, we concluded that bagging was the most effective model for this bank's data (see Table 5-20

Table 5-21).

Table 5-20: Bank 2 AUC results

Bank 2	AUC Instalment loans
Logistic Regression	0.66
Decision Tree	0.75
Linear Support Vector Machines	0.49
Bayesian Network	0.68
Bagged	0.8
RUSBoosted	0.76
AdaBoostM1	0.75

Table 5-21: Bank 2 Confusion matrix results

Bank 2	True Positive	True Negative	False Positive	False Negative	Average Accuracy	Type I Error	Type II Error
Logistic Regression	97%	N/A	100%	3%	N/A	3%	51%
Decision Tree	97%	89%	11%	3%	93%	3%	10%
Linear Support Vector Machines	97%	N/A	N/A	3%	N/A	3%	N/A
Bayesian Network	97%	6%	94%	3%	52%	3%	49%
Bagged	98%	97%	3%	2%	98%	2%	3%
RUSBoosted	99%	6%	94%	1%	53%	1%	49%
AdaBoostM1	97%	N/A	N/A	3%	N/A	3%	N/A

After calculating the running time of each model, it became clear that bagging took longer than the second most suitable model, decision tree. This is shown in Table 5-22. Bagging is derived from the random forest model, which is an enhancement of decision tree. It should be noted here that this is a one-time, training classifier run that then takes a couple of seconds to calculate the probability of defaulting equally across all models.

Table 5-22: Bank 2 running time of models

Bank 2	Instalment Loans	
	Prediction Speed	Training Time
Logistic Reg.	~890000 obs/sec	74.31 sec
Tree	~1500000 obs/sec	30.939 sec
SVM	~5400 obs/sec	3193.5 sec
Bayesian	~870000 obs/sec	14.949 sec
RUSBoosted	~67000 obs/sec	220.7 sec
Bagged	~24000 obs/sec	1524.6 sec
AdaBoostM1	~210000 obs/sec	284.32 sec

The selective criteria performed well using bagging, so we ran a k-fold analysis to assess the performance of the model. 70% of the total data set was used as our in-sample training set. It produced an AUC of 0.75. The remaining 30% of the data set was used as an out-of-sample test and produced an AUC of 0.76. It is important to highlight that in the medical centre, an AUC range between 0.70 and 0.80 is considered as excellent discrimination (Yang & Berdine, 2017). The results are illustrated in Figure 5-4.

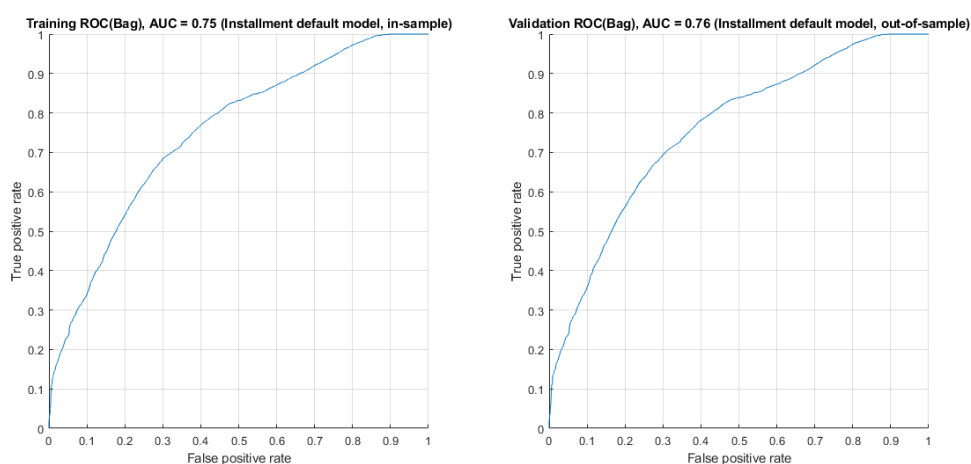


Figure 5-4: Bank 2 AUC results for 70% in-sample and 30% out-of-sample

The results for the relative importance of the predicted variables were slightly different than those for Bank 2, as shown in Figure 5-5.

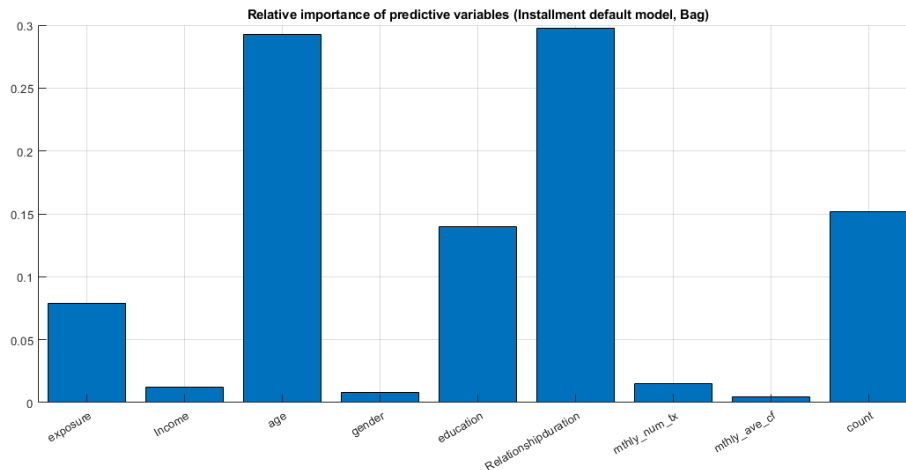


Figure 5-5: The relative importance of the predictive variables for Bank 2

Age and length of relationship were the most influential predictors. This was as expected. These were followed by number of loans and education level. The former makes sense because multiple obligations would make it more difficult not to default on a loan and those with higher levels of education would be more aware of the importance of not defaulting. We did not expect that monthly average cash transactions would not help to predict whether a customer was likely to default.

5.2.1.3 Bank 3

Bank 3 is a medium-sized, conventional bank. There were 1,245 customers with instalment loans in the sample of 100,000 that we tested. 10 were default cases, giving a default rate of 0.80%. The total number of observations was 21,211, as shown in as Table 5-23.

Table 5-23: Bank 3 number of observations and customers

Bank 3	Instalment Loans
Observations	21,211
Default cases	140
Non-Default cases	21,071
Rate of default cases	0.660%
Number of customers	1,245

Number of default customers	10
Number of non- default customers	1,235
Rate of default customers	0.803%

Analysing the AUC for the different methods showed that several models were suitable, as shown in Table 5-24.

Table 5-24: Area Under Curve results

Bank 3	AUC Instalment loans
Logistic Regression	0.82
Decision Tree	1
Linear Support Vector Machines	0.63
Bayesian Network*	N/A
Bagged	1
RUSBoosted	1
AdaBoostM1	1

* Error: A normal distribution did not fit the combination of class 1 and predictor count. The data had zero variance.

After analysing the accuracy rate, type I error, and type II error, we found that we could choose between bagging and AdaBoostM1 (see

Table 5-25). The Bayesian Network method could not be performed because the data had zero variance.

Table 5-25: Bank 3 confusion matrix results

Bank 3	True Positive	True Negative	False Positive	False Negative	Average Accuracy	Type I Error	Type II Error
Logistic Regression	99%	N/A	100%	1%	N/A	1%	50%

Bank 3	True Positive	True Negative	False Positive	False Negative	Average Accuracy	Type I Error	Type II Error
Decision Tree	99%	77%	23%	1%	88%	1%	19%
Linear Support Vector Machines	99%	N/A	N/A	1%	N/A	1%	N/A
Bayesian Network	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Bagged	99%	87%	13%	1%	93%	1%	12%
RUSBoosted	99%	52%	48%	1%	76%	1%	33%
AdaBoostM1	99%	87%	13%	1%	93%	1%	12%

* Error: A normal distribution did not fit the combination of class 1 and predictor count. The data had zero variance.

We compared the training times to help us make our selection. As can be seen from Table 5-26, all the models performed in a time quicker than 10 seconds.

Table 5-26: Bank 4 running time of models

Bank 3	Instalment Loans	
	Prediction Speed	Training Time
Logistic Reg.	~410000 obs/sec	5.8875 sec
Tree	~200000 obs/sec	3.8103 sec
SVM	~270000 obs/sec	3.7087 sec
Bayesian*	FAILED	FAILED
RUSBoosted	~130000 obs/sec	3.2497 sec
Bagged	~110000 obs/sec	5.2065 sec
AdaBoostM1	~120000 obs/sec	7.183 sec

* Error: A normal distribution did not fit the combination of class 1 and predictor count. The data had zero variance.

We used a k-fold test in which 70% of the data set was used as our in-sample training set. This produced an AUC of 1. The remaining 30% of the data set was used as our out-of-sample test and produced an AUC of 1. This is shown in Figure 4-6.

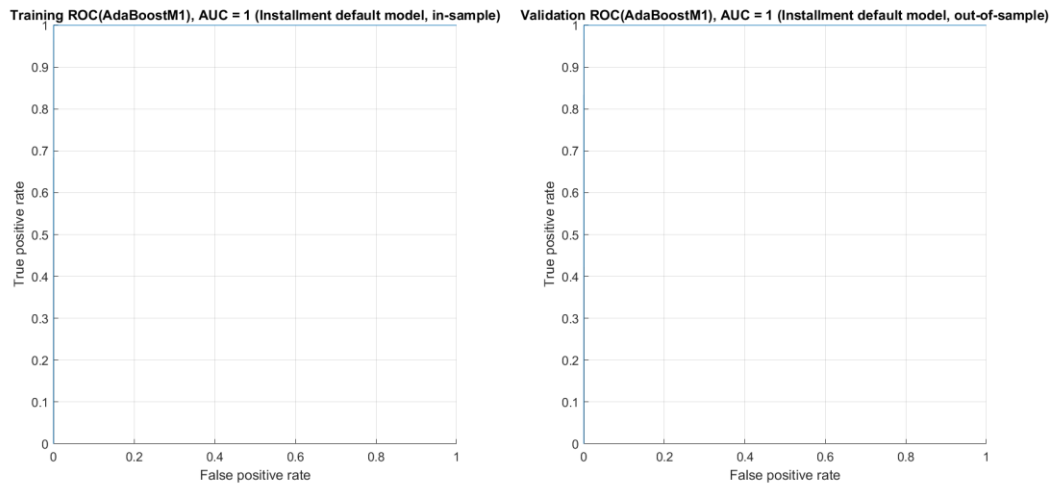


Figure 5-6: Bank 3 AUC results for 70% in-sample and 30% out-of-sample for (AdaBoostM1)

From our study, we realised that to have a functional classification model for predicting loan default cases, a minimum default rate is required. We calculated this by rerunning the models with a decreasing trend of default cases for each of the other banks under study until the comparison of models were insufficient. Error! Not a valid bookmark self-reference. is an illustration from our in-sample and out-of-sample accuracy results in both models, which shows that the tested models were suitable with 100% accuracy for this bank.

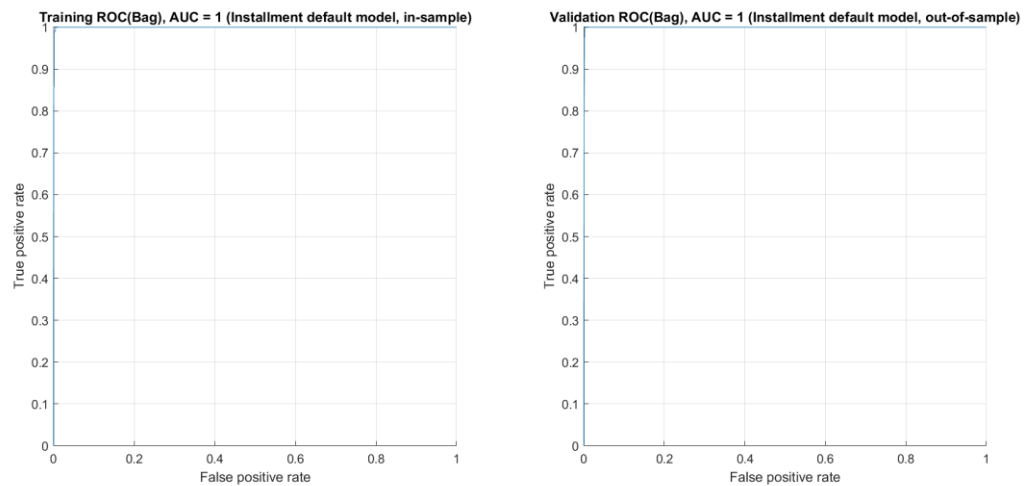


Figure 5-7: Bank 3 AUC results for 70% in-sample and 30% out-of-sample for (Bag)

5.2.1.4 Bank 4

Bank 4 is a medium-sized Islamic bank. Of our sample of 100,000 customers, 15,108 had instalment loans. 165 were default cases, giving a default rate of 1.09%. There were 249,567 observations, as shown in Table 5-27.

Table 5-27: Bank 4 number of observations and customers

Bank 4	Instalment Loans
Observations	249,567
Default cases	214
Non-Default cases	249,353
Rate of default cases	0.086%
Number of customers	15,108
Number of default customers	165
Number of non- default customers	14,943
Rate of default customers	1.092%

After analysing the AUC for each of the different methods, we decided that RUSBoost was the most effective model for this bank's data set. The results of the analysis are shown in Table 5-28.

Table 5-28: Bank 4 AUC results

Bank 4	AUC Instalment loans
Logistic Regression	0.73
Decision Tree	0.71
Linear Support Vector Machines	0.6
Bayesian Network	0.71
Bagged	0.71
RUSBoosted	0.81
AdaBoostM1	0.77

Analysing the accuracy rate, type I error, and type II error showed that bagging was slightly more effective than RUSBoosting, in contrast to the AUC outcomes. The accuracy rate for bagging was 12% better. The type I error results were equal, and the type II error results were 7% better for bagging, as shown by

Table 5-29.

Table 5-29: Bank 4 confusion matrix results

Bank 4	True Positive	True Negative	False Positive	False Negative	Average Accuracy	Type I Error	Type II Error
Logistic Regression	99%	N/A	100%	1%	N/A	1%	50%
Decision Tree	99%	10%	90%	1%	55%	1%	48%
Linear Support Vector Machines	99%	N/A	N/A	1%	N/A	1%	N/A
Bayesian Network	99%	1%	99%	1%	50%	1%	50%
Bagged	99%	25%	75%	1%	62%	1%	43%
RUSBoosted	99%	1%	99%	1%	50%	1%	50%
AdaBoostM1	99%	N/A	100%	1%	N/A	1%	50%

We used training time to compare the two best models. RUSBoosting outperformed bagging. The training time for the former was 44 seconds, whereas it was 3 minutes and 10 seconds for the latter (see Table 5-30).

Table 5-30: Bank 4 running time of models

Bank 4	Instalment Loans	
	Prediction Speed	Training Time
Logistic Reg.	~1200000 obs/sec	33.625 sec
Tree	~1500000 obs/sec	16.525 sec
SVM	~280000 obs/sec	28.171 sec
Bayesian	~640000 obs/sec	11.439 sec
RUSBoosted	~110000 obs/sec	43.694 sec
Bagged	~80000 obs/sec	186.15 sec
AdaBoostM1	~160000 obs/sec	191.81 sec

As with the other banks, we performed a k-fold test with 70% the data set as our in-sample training set. This produced an AUC of 0.91 for the RUSBoosted data and 0.84 for the bagged data. The remaining 30% was our out-of-sample test set, which produced an AUC of 0.91 for the RUSBoosted set and 0.86 for the bagged set. These results are shown in Figure 5-8 and Figure 5-9. This result favoured RUSBoosting over bagging, but we created a prediction code to allow the bank to use either model to calculate the default rate.

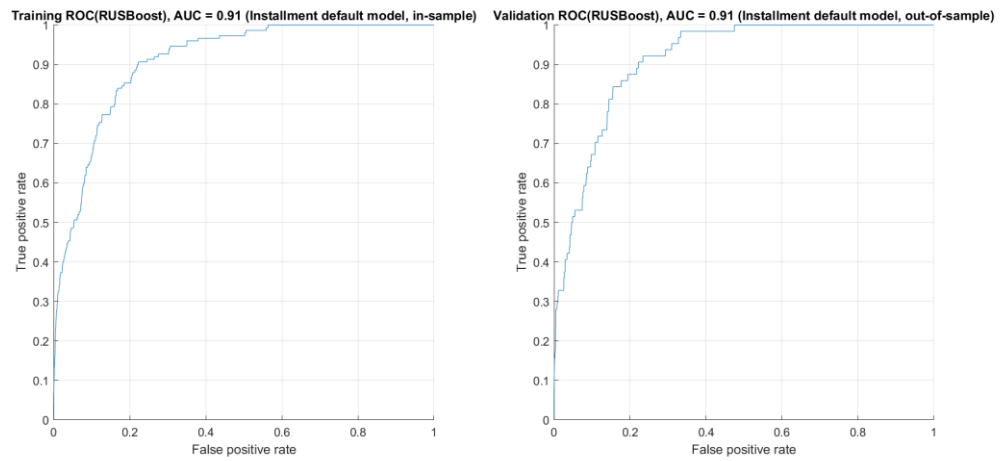


Figure 5-8: Bank 4 AUC results for 70% in-sample and 30% out-of-sample (RUSBoosted)

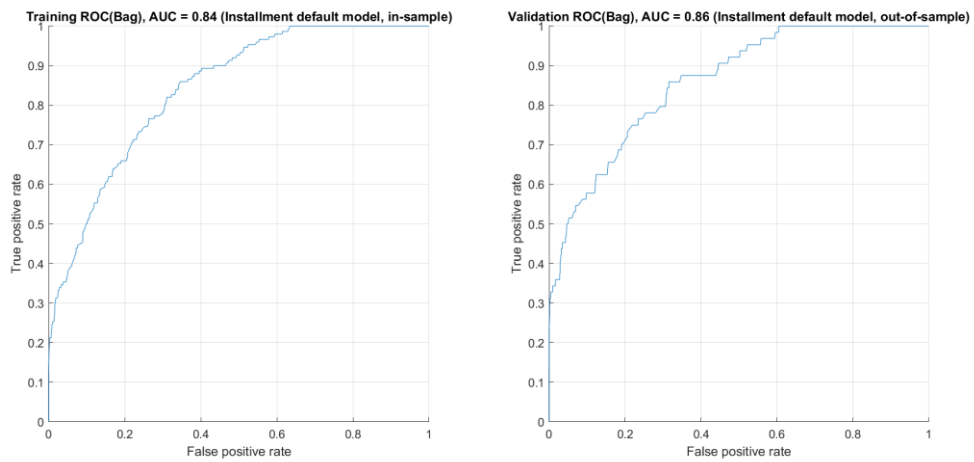


Figure 5-9: Bank 4 AUC results for 70% in-sample and 30% out-of-sample (bagged)

After training the classifier for RUSBoosted, we assessed the relative importance of the variables. The results are shown in Figure 5-10.

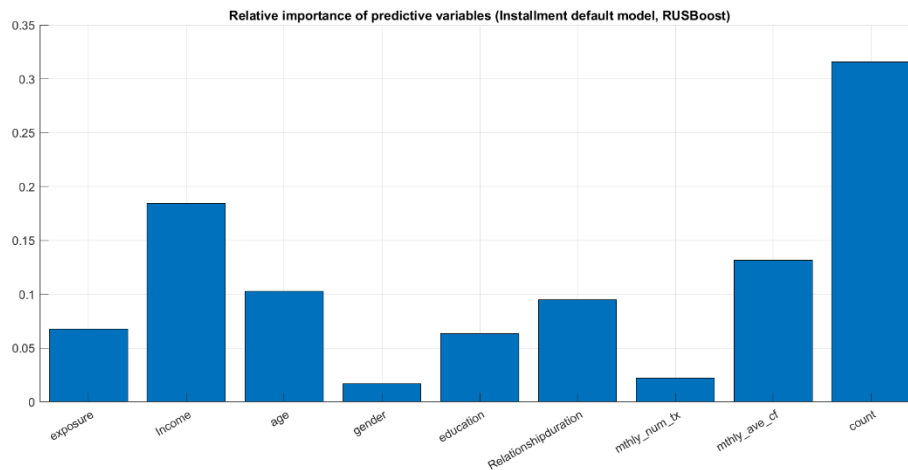


Figure 5-10: The relative importance of the predictive variables for Bank 4

As expected from our hypothesis, the number of loans had the highest impact on the prediction model, followed by the level of monthly income. The selected variables were expected to be able to strongly predict whether customers would default. This result can be clarified by the fact that there are more instalment loans than consumer loans (70,000 compared to 25,000). Thus, customers with multiple instalment loans had more to deal with. Furthermore, of the newly introduced variables, monthly average cashflow and length of relationship with the bank were relatively important. This was also in line with our expectations.

5.2.1.5 Bank 5

Bank 5 is a small Islamic bank. Out of our sample of 100,000 customers, 4,857 had instalment loans. 1,002 were default customers with a default rate of 20.63%. The number of observations was 13,879, as shown by Table 5-31.

Table 5-31: Bank 5 number of observations and customers

Bank 5	Instalment Loans
Observations	13,879
Default cases	8,491
Non-Default cases	5,388
Rate of default cases	61.179%
Number of customers	4,857
Number of default customers	1,002
Number of non- default customers	3,855

Rate of default customers	20.630%
---------------------------	---------

After analysing the AUC, the accuracy rate, type I error, type II error, and running time, we concluded that bagging was the most effective model for this bank's data set. The AUC was 0.98 (see Table 5-32), and the accuracy rate was 89% (see Table 5-33).

Table 5-32: Bank 5 AUC results

Bank 5	AUC Instalment loans
Logistic Regression	0.9
Decision Tree	0.93
Linear Support Vector Machines	0.89
Bayesian Network	0.81
Bagged	0.98
RUSBoosted	0.93
AdaBoostM1	0.93

Table 5-33: Bank 5 confusion matrix results

Bank 5	True Positive	True Negative	False Positive	False Negative	Average Accuracy	Type I Error	Type II Error
Logistic Regression	95%	56%	44%	5%	76%	5%	32%
Decision Tree	97%	62%	38%	3%	80%	3%	28%
Linear Support Vector Machines	93%	N/A	N/A	7%	N/A	7%	N/A
Bayesian Network	97%	15%	85%	3%	56%	3%	47%
Bagged	98%	80%	20%	2%	89%	2%	17%
RUSBoosted	99%	48%	52%	1%	74%	1%	34%
AdaBoostM1	97%	66%	34%	3%	82%	3%	26%

Bagging took longer than most of the other models (see Table 5-34), but the time of 162 seconds was still acceptable.

Table 5-34: Bank 5 running time of models

Bank 5	Instalment Loans	
	Prediction Speed	Training Time
Logistic Reg.	~1600000 obs/sec	5.5067 sec
Tree	~1000000 obs/sec	4.9541 sec
SVM	~19000 obs/sec	364 sec
Bayesian*	~690000 obs/sec	3.4149 sec
RUSBoosted	~85000 obs/sec	38.475 sec

Bagged	~39000 obs/sec	161.83 sec
AdaBoostM1	~69000 obs/sec	141.36 sec

We use 70% of the total data as our in-sample set, giving an AUC of 0.93. The remaining 30% was used as an out-of-sample test, giving an AUC of 0.93, as shown by Figure 5-11.

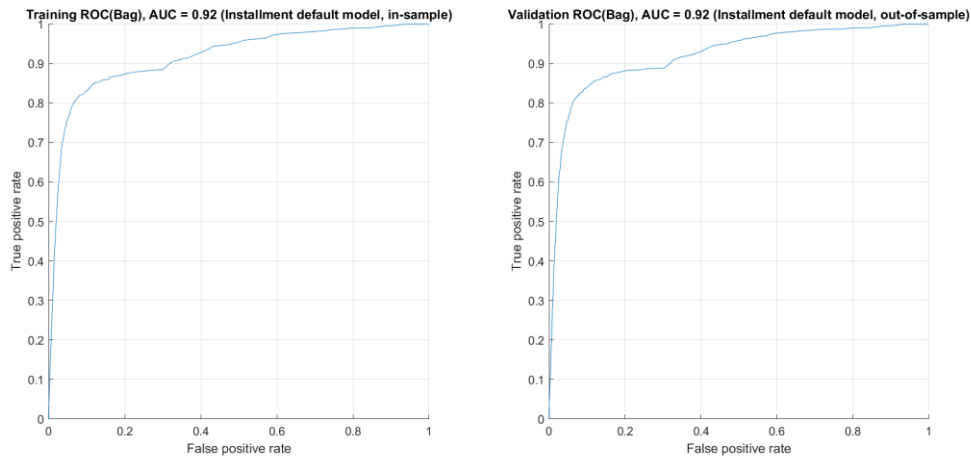


Figure 5-11: Bank 5 AUC results for 70% in-sample and 30% out-of-sample

Figure 5-12 shows the importance of the input variables for Bank 5, using the bagging model.

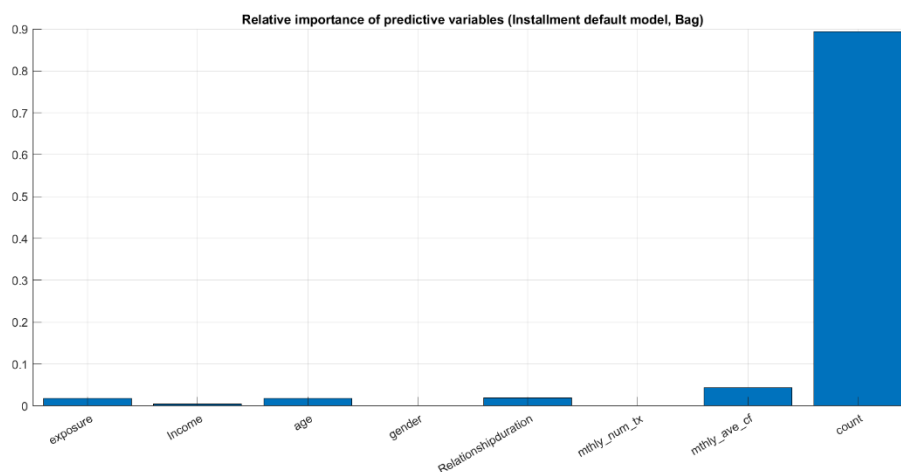


Figure 5-12: The relative importance of the predictive variables for Bank 5

The most influential variable for this bank was the number of loans. The second most influential variable was the length of a customer's relationship with the bank. The rest were all of equal importance. As this bank is new, it is understandable that the length of relationship would not have much influence, but the bank should

not disregard the other variables because they were influential for other Islamic banks.

5.2.1.6 Bank 6

Bank 6 is the large Islamic bank. 6,180 customers of the 100,000 we sampled had instalment loans. 355 were default customers, giving a default rate of 5.74%. There were 62,340 observations, as shown by Table 5-35.

Table 5-35: Bank 6 number of observations and customers

Bank 6	Instalment Loans
Observations	62,340
Default cases	2,071
Non-Default cases	60,269
Rate of default cases	3.322%
Number of customers	6,180
Number of default customers	355
Number of non- default customers	5,825
Rate of default customers	5.744%

After analysing the AUC, accuracy rate, type I error, and type II error, we decided that bagging was the most successful model, with an AUC of 0.95, closely followed by AdaBoostM1, with an AUC of 0.91, as shown in Table 5-36.

Table 5-36: Bank 6 AUC results

Bank 6	AUC Instalment loans
Logistic Regression	0.69
Decision Tree	0.74
Linear Support Vector Machines	0.51
Bayesian Network	0.65
Bagged	0.95
RUSBoosted	0.85
AdaBoostM1	0.91

The confusion matrix (see

Table 5-37) shows that AdaBoostM1 and bagging models had the highest accuracy rates (76% and 68%, respectively).

Table 5-37: Confusion Matrix results

Bank 6	True Positive	True Negative	False Positive	False Negative	Average Accuracy	Type I Error	Type II Error
Logistic Regression	97%	N/A	N/A	3%	N/A	3%	N/A
Decision Tree	97%	39%	61%	3%	68%	3%	39%
Linear Support Vector Machines	99%	N/A	N/A	3%	N/A	3%	N/A
Bayesian Network	97%	7%	93%	3%	52%	3%	49%
Bagged	98%	37%	63%	2%	68%	2%	39%
RUSBoosted	99%	13%	87%	1%	56%	1%	47%
AdaBoostM1	97%	55%	45%	3%	76%	3%	32%

The training times for each model were both acceptable at less than 40 seconds, as shown in Table 5-38. However, the prediction speed of AdaBoostM1 (~140000 observations/second) was better than bagging (~78000 observations/second). Thus, AdaBoostM1 may be more suitable when there are more observations, or when dealing with larger loan portfolios, as is sometimes the case with corporate lending.

Table 5-38: Bank 6 running time of models

Bank 6	Instalment Loans	
	Prediction Speed	Training Time
Logistic Reg.	~1400000 obs/sec	2.2461 sec
Tree	~1300000 obs/sec	2.1761 sec
SVM	~75000 obs/sec	17.67 sec
Bayesian	~9200000 obs/sec	1.5538 sec
RUSBoosted	~130000 obs/sec	10.066 sec
Bagged	~78000 obs/sec	35.925 sec
AdaBoostM1	~140000 obs/sec	30.397 sec

We performed a k-fold with 70% of the data set as our in-sample training set, producing an AUC of 0.93. We used the remaining 30% as our out-of-sample test set, giving an AUC of 0.92 (see Figure 5-13).

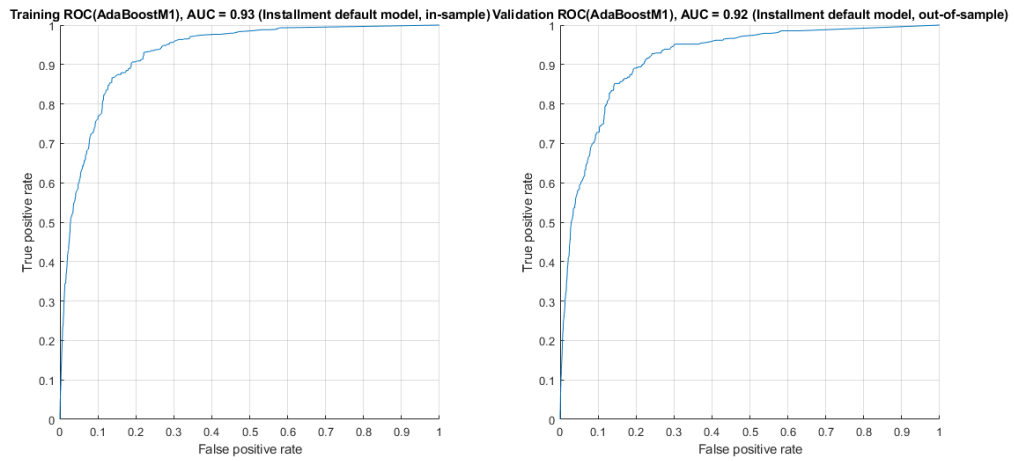


Figure 5-13: Bank 6 AUC results for 70% in-sample and 30% out-of-sample

Figure 5-14 shows the importance of the input variables for Bank 6.

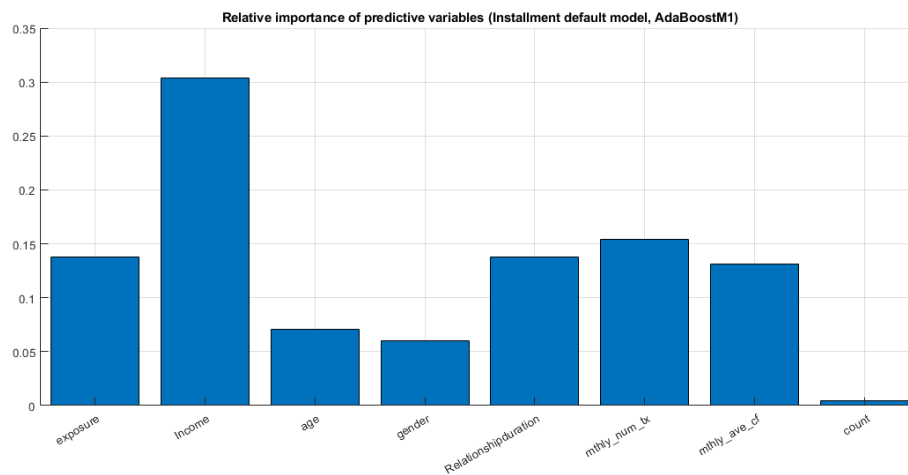


Figure 5-14: The relative importance of the predictive variables for Bank 6

Unlike the other Islamic banks, the number of loans had the least impact of all the variables for Bank 6. Income levels had the most impact. Credit card exposure, length of relationship, and the value and number of monthly cash transactions all had the same level of impact. The reason for this is probably because the default cases occurred with customers with very few loans. On the other hand, customers with multiple loans had low rates of defaulting.

5.2.2 Central Bank Model

After combining the data from all seven banks, including all the conventional and Islamic banks, we had a total of 48,341 customers who had instalment loans. 8,738 customers were defaulting, giving a default rate of 18.07%. There were 1,496,121 observations in total, as shown in Table 5-39.

Table 5-39: Number of observations and customers for all banks

All Banks	Instalment Loans
Observations	1,496,121
Default cases	28,530
Non-Default cases	1,467,591
Rate of default cases	1.907%
Number of customers	48,341
Number of default customers	8,738
Number of non- default customers	39,603
Rate of default customers	18.076%

After analysing the AUC alongside the models for the individual banks, we found that the ensemble models, such as bagging, RUSBoost, and AdaBoostM1, were the most effective models, as shown in Table 5-40.

Table 5-40: AUC results for all banks

All Banks	AUC Instalment loans
Logistic Regression	0.62
Decision Tree	0.74
Linear Support Vector Machines	0.54
Bayesian Network	0.67
Bagged	0.87
RUSBoosted	0.79
AdaBoostM1	0.79

We used the results from the confusion matrix (see Table 5-41) and the running time calculations (see Table 5-42) to compare the performance of the ensemble models. We included a combination of different models, using LogitBoost and GentelBoost along with the existing models. A robust boost was also selected to add an additional angle to check that we had not overlooked its effectiveness.

Table 5-41: Confusion matrix results for all banks

All Banks	True Positive	True Negative	False Positive	False Negative	Average Accuracy	Type I Error	Type II Error
Logistic Regression	98%	N/A	100%	2%	N/A	2%	51%
Decision Tree	98%	83%	17%	2%	91%	2%	15%

Linear Support Vector Machines	98%	N/A	N/A	2%	N/A	2%	N/A
Bayesian Network	98%	6%	94%	2%	52%	2%	49%
Bagged	99%	87%	13%	1%	93%	1%	12%
RUSBoosted	99%	4%	96%	1%	52%	1%	49%
AdaBoostM1	98%	N/A	N/A	2%	N/A	2%	N/A

Table 5-42: Running time of models for all banks

All Banks	Instalment Loans	
	Prediction Speed	Training Time
Logistic Reg.	~1800000 obs/sec	72.58 sec
Tree	~1500000 obs/sec	72.159 sec
SVM	~5800 obs/sec	9112.6 sec
Bayesian	~940000 obs/sec	34.087 sec
RUSBoosted	~110000 obs/sec	269.86 sec
Bagged	~55000 obs/sec	2663.9 sec
AdaBoostM1	~120000 obs/sec	1102.2 sec

We concluded that a deep comparison of all the ensemble models was the most suitable for the Central Bank model. By comparing the AUC of the test sets, we found that GentleBoost was the most effective model for the data set that contained all the banks. We used 70% of the data set as our in-sample training set, giving an AUC of 0.89. We used the remaining 30% as our out-of-sample test set, giving an AUC of 0.89. The results are shown in Figure 5-15.

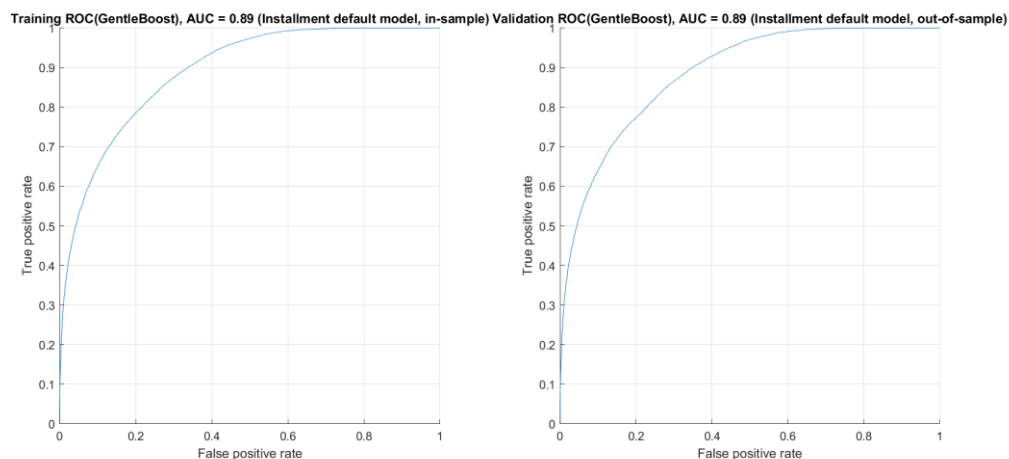


Figure 5-15: AUC results for 70% in-sample and 30% out-of-sample (all banks)

Figure 5-16 shows the importance of the input variables for our prediction model for all banks.

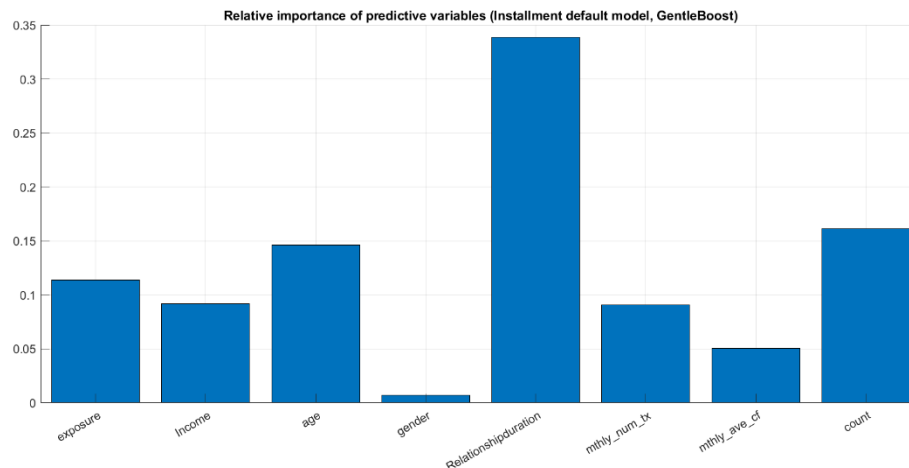


Figure 5-16: The relative importance of the predictive variables for all banks

As expected, the length of a customer’s relationship with the bank was the most influential variable for predicting a person’s likelihood to default. This was followed by the number of loans, which was more of a factor in the Islamic banks. Although the monthly average number of transactions was significant for the individual banks, it was not as significant when we combined all of the banks’ data. However, we would encourage individual banks to continue to use this variable as part of their prediction model.

5.3 Models for Consumer Loans

5.3.1 Bank-by-Bank Models

5.3.1.1 Bank 1

Bank 1 has 6,166 customers with consumer loans. 36 are default customers, with a default rate of 0.58%. The number of observations was 244,253 (see Table 5-43).

Table 5-43: Bank 1 number of observations and customers

Bank 1	Consumer Loans
Observations	244,253
Default cases	104
Non-Default cases	244,149
Rate of default cases	0.043%

Number of customers	6,166
Number of default customers	36
Number of non- default customers	6,130
Rate of default customers	0.584%

Our analysis of the AUC, the accuracy rate, type I error, and type II error, we decided that bagging was the most effective model. Although its AUC (0.86) was lower than RUSBoosted (0.96) and AdaBoostM1 (0.94), its accuracy rate was 92%. The results are shown in Table 5-44 and Table 5-45.

Table 5-44: Bank 1 AUC results

Bank 1	AUC Consumer loans
Logistic Regression	0.8
Decision Tree	0.82
Linear Support Vector Machines	0.57
Bayesian Network	0.79
Bagged	0.86
RUSBoosted	0.96
AdaBoostM1	0.94

Table 5-45: Bank 2 confusion matrix results

Bank 1	True Positive	True Negative	False Positive	False Negative	Average Accuracy	Type I Error	Type II Error
Logistic Regression	99%	N/A	N/A	1%	N/A	1%	N/A
Decision Tree	99%	30%	70%	1%	65%	1%	41%
Linear Support Vector Machines	99%	N/A	N/A	1%	N/A	1%	N/A
Bayesian Network	99%	1%	99%	1%	50%	1%	50%
Bagged	99%	85%	15%	1%	92%	1%	13%
RUSBoosted	99%	1%	99%	1%	50%	1%	50%
AdaBoostM1	99%	80%	20%	1%	90%	1%	17%

The training time for bagging was the second longest at just under two minutes, but this was still acceptable (see

Table 5-46).

Table 5-46: Bank 1 running time of models

Bank 1	Consumer Loans	
	Prediction Speed	Training Time
Logistic Reg.	~370000 obs/sec	465.14 sec
Tree	~110000 obs/sec	12.828 sec
SVM	~230000 obs/sec	32.807 sec
Bayesian	~700000 obs/sec	10.178 sec
RUSBoosted	~140000 obs/sec	31.37 sec
Bagged	~140000 obs/sec	107.51 sec
AdaBoostM1	~170000 obs/sec	123.72 sec

As before, we used 70% of the data as an in-sample training set, giving an AUC of 1. The remaining 30%, used as an out-of-sample test set, gave an AUC of 0.95, as shown by Figure 5-17.

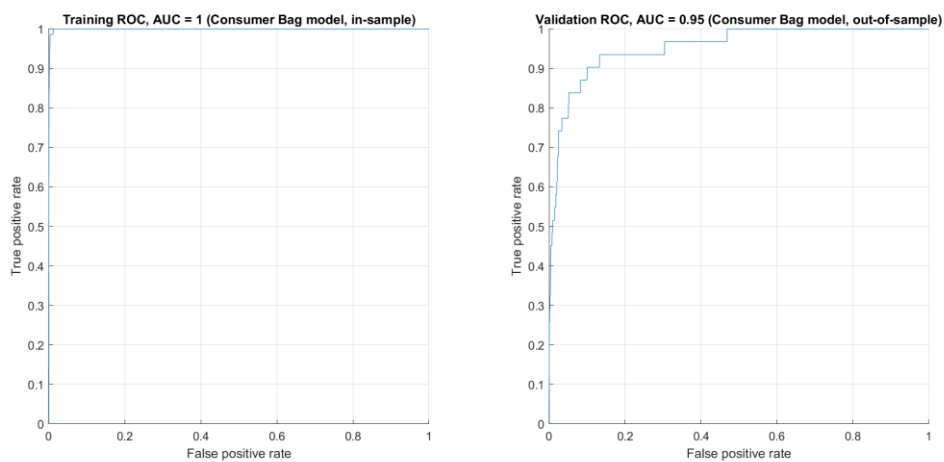


Figure 5-17: Bank 1 AUC results for 70% in-sample and 30% out-of-sample

Figure 4-18 shows the relative importance of the predictive variables for Bank 1.

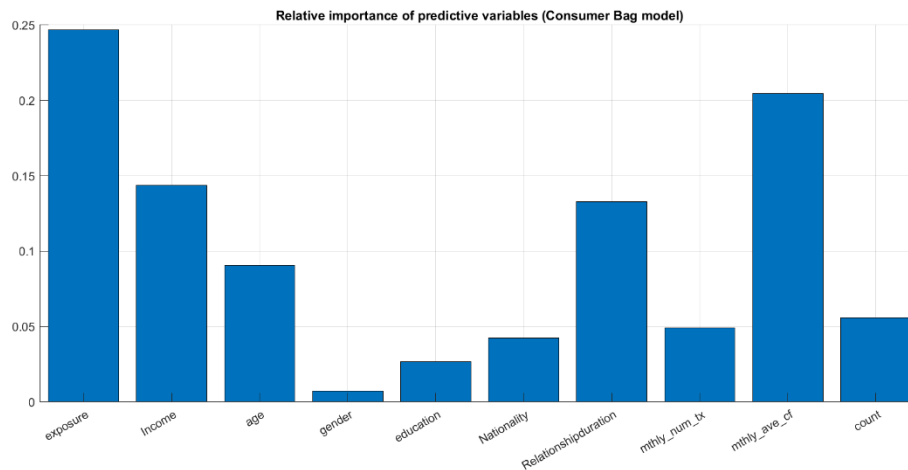


Figure 5-18: The relative importance of the predictive variables for Bank 1

As expected, customers' credit card exposure was the most influential variable, followed by the average monthly net cash flow. The purpose of these loans was for leisure, so this suggests that purchasing behaviour can affect people's ability to meet their monthly obligations. Income level was the third most influential variable, followed closely by the length of a customer's relationship and their age. This was in line with our expectations.

5.3.1.2 Bank 2

Bank 2 sample were 11,072 customers, of which 5,421 were default customers (a default rate of 48.96%). The data obtained produced 575,268 observations, as shown in Table 5-47.

Table 5-47: Bank 2 number of observations and customers

Bank 2	Consumer Loans
Observations	575,268
Default cases	10,136
Non-Default cases	565,132
Rate of default cases	1.762%
Number of customers	11,072
Number of default customers	5,421
Number of non- default customers	5,651
Rate of default customers	48.961%

The analysis of the AUC for each of the different methods revealed that AdaBoostM1 was the most effective, with an AUC of 0.77 (see Table 5-48).

Table 5-48: Bank 2 AUC results

Bank 2	AUC Consumer loans
Logistic Regression	0.67
Decision Tree	0.74
Linear Support Vector Machines	0.54
Bayesian Network	0.66
Bagged	0.73
RUSBoosted	0.76
AdaBoostM1	0.77

However, examining the accuracy rate, type I error, and type II error indicated that bagging was the most effective model, with an accuracy rate of 95% (see Table 5-49).

Table 5-49: Confusion Matrix results

Bank 2	True Positive	True Negative	False Positive	False Negative	Average Accuracy	Type I Error	Type II Error
Logistic Regression	98%	N/A	100%	2%	N/A	2%	51%
Decision Tree	98%	76%	24%	2%	87%	2%	20%
Linear Support Vector Machines	98%	N/A	N/A	2%	N/A	2%	N/A
Bayesian Network	98%	3%	97%	2%	51%	2%	50%
Bagged	99%	91%	9%	1%	95%	1%	8%
RUSBoosted	99%	3%	97%	1%	51%	1%	49%
AdaBoostM1	98%	78%	22%	2%	88%	2%	18%

When comparing the training time of the AdaBoostM1 and bagging models, we found that AdaBoostM1 was more efficient, requiring 6.7 minutes compared to 11 minutes for bagging.

Table 5-50: Bank 2 running time for models

Bank 2	Consumer Loans	
	Prediction Speed	Training Time
Logistic Reg.	~1300000 obs/sec	51.403 sec
Tree	~19000000 obs/sec	23.454 sec

SVM	~9000 obs/sec	1302.4 sec
Bayesian	~920000 obs /sec	12.853 sec
RUSBoosted	~140000 obs/sec	89.358 sec
Bagged	~50000 obs/sec	656.42 sec
AdaBoostM1	~130000 obs/sec	406.5 sec

70% of the data was used as an in-sample training set, giving an AUC of 0.78, while the remaining 30% was used as a test set, giving an AUC of 0.78 (see Figure 4.19)

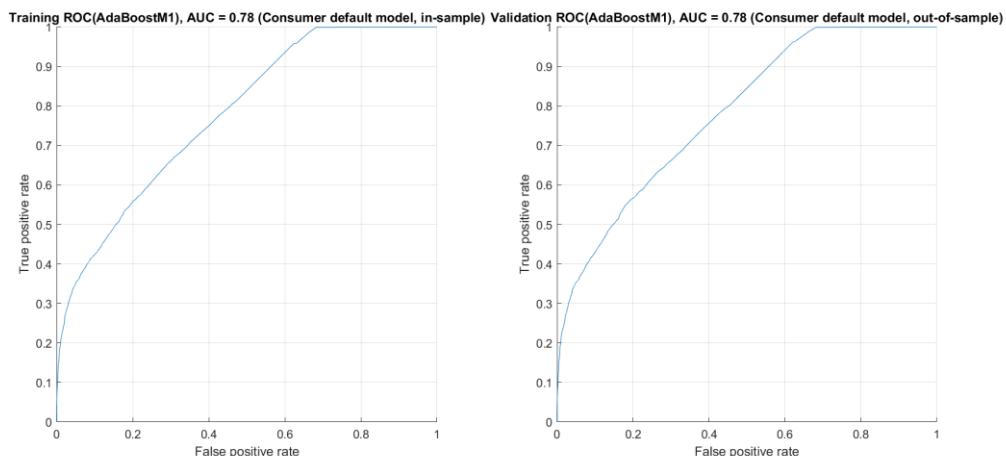


Figure 5-19: Bank 2 AUC results for 70% in-sample and 30% out-of-sample

The relative importance of the input variables differed from Bank 1, as can be seen from Figure 5-20.

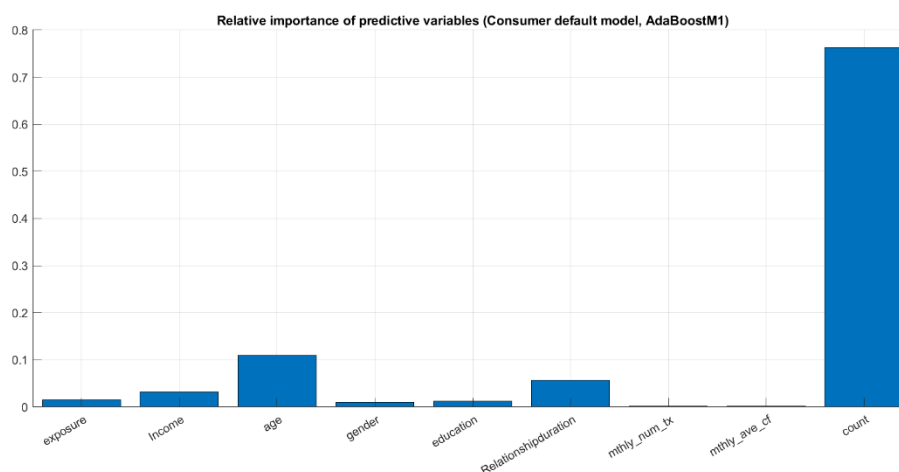


Figure 5-20: The relative importance of the predictive variables for Bank 2

Number of loans was the most influential variable for this bank. This was as we expected because multiple obligations make people more likely to default. In response, banks could consider limiting the loans they offer to customers who

already have multiple loans, even if they are within the mandated cap of 40% of their salary. These results could suggest mismanagement, given the high default rate and the large number of loans that have been approved by the bank.

5.3.1.3 Bank 3

Bank 3 is a medium-sized conventional bank. In our sample, 2,996 people had consumer loans, and 8 were default customers, giving a default rate of 0.26%. The number of observations was 46,843, as shown in Table 5-51.

Table 5-51: Bank 3 number of observations and customers

Bank 3	Consumer Loans
Observations	46,843
Default cases	101
Non-Default cases	46,742
Rate of default cases	0.216%
Number of customers	2,996
Number of default customers	8
Number of non- default customers	2,988
Rate of default customers	0.267%

After analysing the AUC, accuracy rate, type I error, type II error, and running time, we decided that several models were suitable as they all had high AUCs (see Table 5-52), high accuracy rates (see Table 5-53), and short running times (see Table 5-54). As was the case for instalment loans, the Bayesian Network model was not functional due to the data input.

Table 5-52: Bank 3 AUC results

Bank 3	AUC Consumer loans
Logistic Regression	0.92
Decision Tree	0.99
Linear Support Vector Machines	0.63
Bayesian Network*	N/A
Bagged	0.98
RUSBoosted	1
AdaBoostM1	1

* Error: A normal distribution did not fit the combination of class 1 and predictor count. The data had zero variance.

Table 5-53: Bank 3 confusion matrix results

Bank 3	True Positive	True Negative	False Positive	False Negative	Average Accuracy	Type I Error	Type II Error
Logistic Regression	99%	N/A	100%	1%	N/A	1%	50%
Decision Tree	99%	78%	22%	100%	59%	50%	18%
Linear Support Vector Machines	99%	N/A	N/A	1%	N/A	1%	N/A
Bayesian Network*	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Bagged	99%	87%	13%	1%	93%	1%	12%
RUSBoosted	100%	25%	75%	N/A	N/A	N/A	43%
AdaBoostM1	99%	88%	12%	1%	94%	1%	11%

* Error: A normal distribution did not fit the combination of class 1 and predictor count. The data had zero variance.

Table 5-54: Bank 3 running time for models

Bank 3	Consumer Loans	
	Prediction Speed	Training Time
Logistic Reg.	~900000 obs/sec	5.9285 sec
Tree	~1600000 obs/sec	1.8861 sec
SVM	~750000 obs/sec	3.2262 sec
Bayesian*	FAILED	FAILED
RUSBoosted	~190000 obs/sec	4.2644 sec
Bagged	~160000 obs/sec	10.228 sec
AdaBoostM1	~160000 obs/sec	15.695 sec

* Error: A normal distribution did not fit the combination of class 1 and predictor count. The data had zero variance.

To reach a better conclusion, we performed a k-fold prediction test using 70% of the data as our in-sample training set and 30% as our out-of-sample test set. Both gave an AUC of 1 for AdaBoostM1, bagging, RUSBoost, and decision tree, which made it difficult to choose between them. The results are presented in Figure 5-21, Figure 5-22, Figure 5-23, and Figure 5-24. It is important to note that the data sample for this bank was very low, making it difficult to use machine learning solutions. Therefore, we concluded that a minimum data set is required for the model to work efficiently. This is necessary when using a classification model to predict default loan cases.

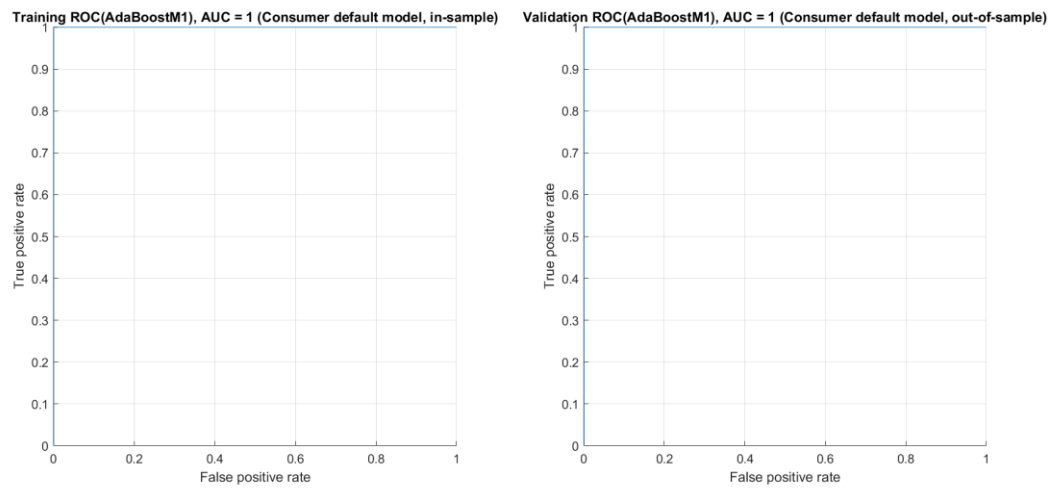


Figure 5-21: Bank 3 AUC results for 70% in-sample and 30% out-of-sample (AdaBoostM1)

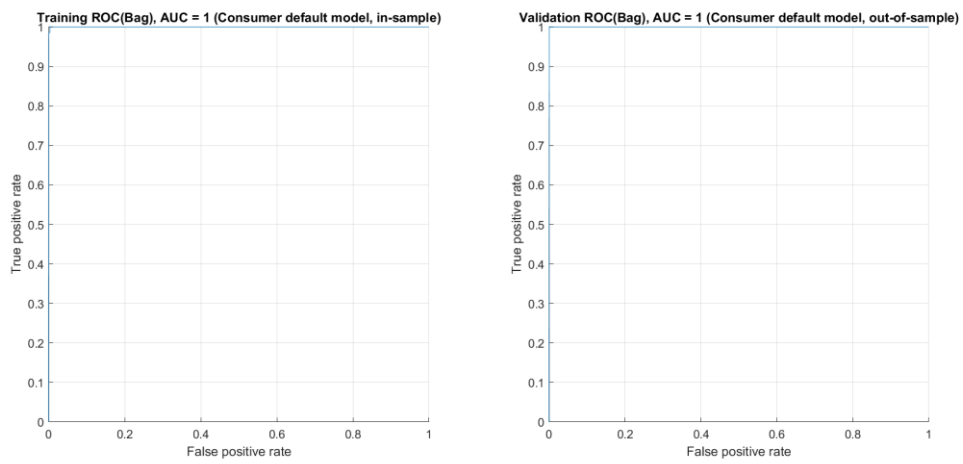


Figure 5-22: Bank 3 AUC results for 70% in-sample and 30% out-of-sample (bagging)

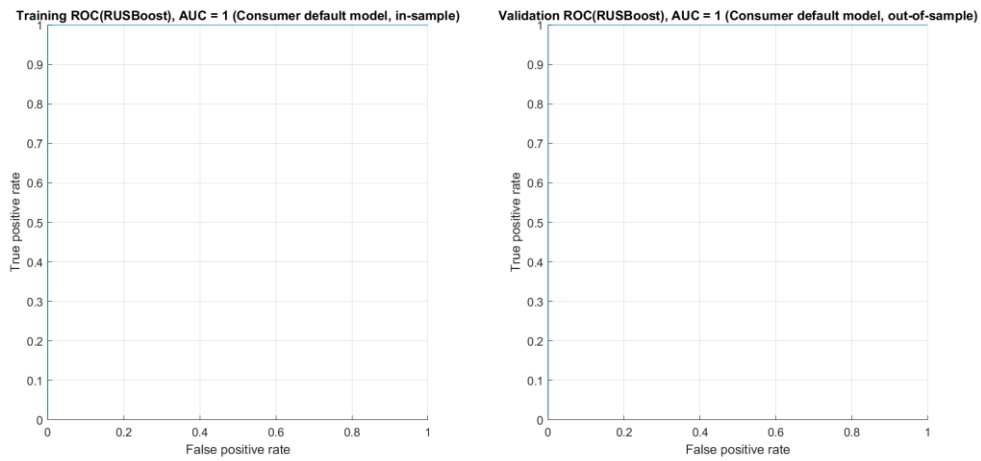


Figure 5-23: Bank 3 AUC results for 70% in-sample and 30% out-of-sample (RUSBoosted)

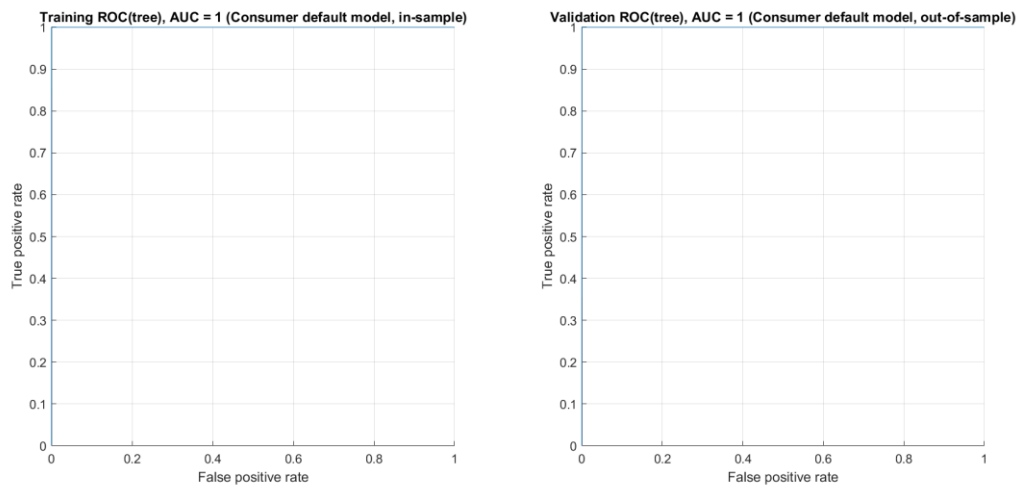


Figure 5-24: Bank 3 AUC results for 70% in-sample and 30% out-of-sample (decision tree)

As there are several good models that would provide information about different variables, more data would be required for us to determine the successful model.

5.3.1.4 Bank 4

Bank 4 is a medium-sized Islamic bank with 6,541 customers with consumer loans. Of those, 22 are at risk of defaulting, giving a default rate of 0.34%. The total number of observations was 92,377, as shown in Table 5-55.

Table 5-55: Bank 4 number of observations and customers

Bank 4	Consumer Loans
Observations	92,377

Default cases	26
Non-Default cases	92,351
Rate of default cases	0.028%
Number of customers	6,451
Number of default customers	22
Number of non- default customers	6,429
Rate of default customers	0.341%

The AUCs for almost all the different methods were between 0.55 and 0.60, making them difficult to compare. However, RUSBoosted gave an AUC of 0.80 (see Table 5-56). The Bayesian Network model did not work with this data set.

Table 5-56: Area Under Curve results

Bank 4	AUC Consumer loans
Logistic Regression	0.56
Decision Tree	0.52
Linear Support Vector Machines	0.52
Bayesian Network*	N/A
Bagged	0.57
RUSBoosted	0.8
AdaBoostM1	0.61

* Error: A normal distribution did not fit the combination of class 1 and predictor count. The data had zero variance.

The accuracy rate, type I error, and type II error results showed that several of the models were not suitable, with no average accuracy rate that could be used as a comparison. The 50% accuracy of RUSBoosted meant that it did perform, but its accuracy rate was 50%, which is not very reassuring (see Table 5-57).

Table 5-57: Bank 4 confusion matrix results

Bank 4	True Positive	True Negative	False Positive	False Negative	Average Accuracy	Type I Error	Type II Error
Logistic Regression	99%	N/A	100%	1%	N/A	1%	50%
Decision Tree	99%	N/A	100%	1%	N/A	1%	50%
Linear Support Vector Machines	99%	N/A	N/A	1%	N/A	1%	N/A

Bayesian Network*	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Bagged	99%	N/A	100%	1%	N/A	1%	50%
RUSBoosted	99%	1%	99%	1%	50%	1%	50%
AdaBoostM1	99%	N/A	100%	1%	N/A	1%	50%

* Error: A normal distribution did not fit the combination of class 1 and predictor count. The data had zero variance.

The training time was less than 41 seconds for all the models, as seen in Table 5-58.

Table 5-58: Bank 4 running time of models

Bank 4	Consumer Loans	
	Prediction Speed	Training Time
Logistic Reg.	~660000 obs/sec	26.914 sec
Tree	~580000 obs/sec	7.8839 sec
SVM	~600000 obs/sec	6.4949 sec
Bayesian*	FAILED	FAILED
RUSBoosted	~180000 obs/sec	9.4567 sec
Bagged	~94000 obs/sec	40.859 sec
AdaBoostM1	~130000 obs/sec	65.668 sec

* Error: A normal distribution did not fit the combination of class 1 and predictor count. The data had zero variance.

As with Bank 3, there were not enough cases to distinguish between the relative models. The AUC results showed that several models would be suitable. We performed a k-fold analysis using 70% of the data set as our in-sample training set, which produced an AUC of 1. We used the remaining 30% as our out-of-sample to test, which gave an AUC of 1. The results are shown in Figure 5-25, Figure 5-26, and Figure 5-27.

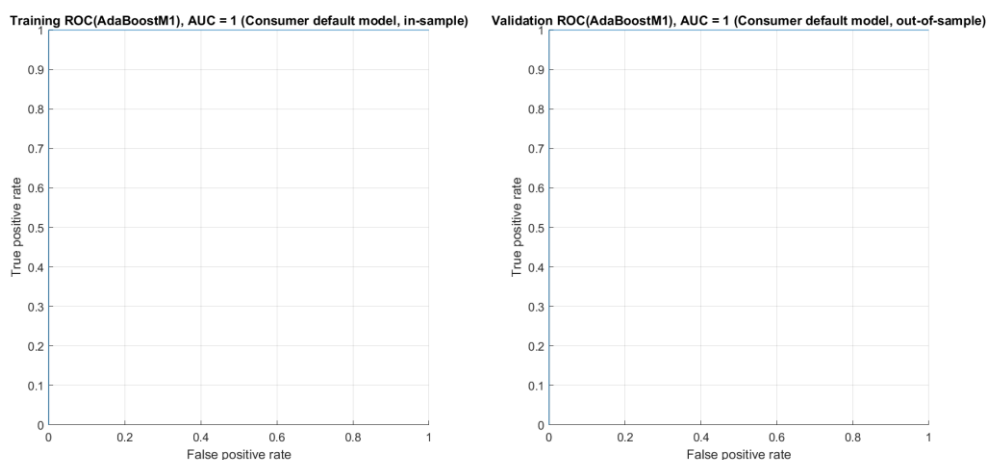


Figure 5-25: Bank 4 AUC results for 70% in-sample and 30% out-of-sample (AdaBoostM1)

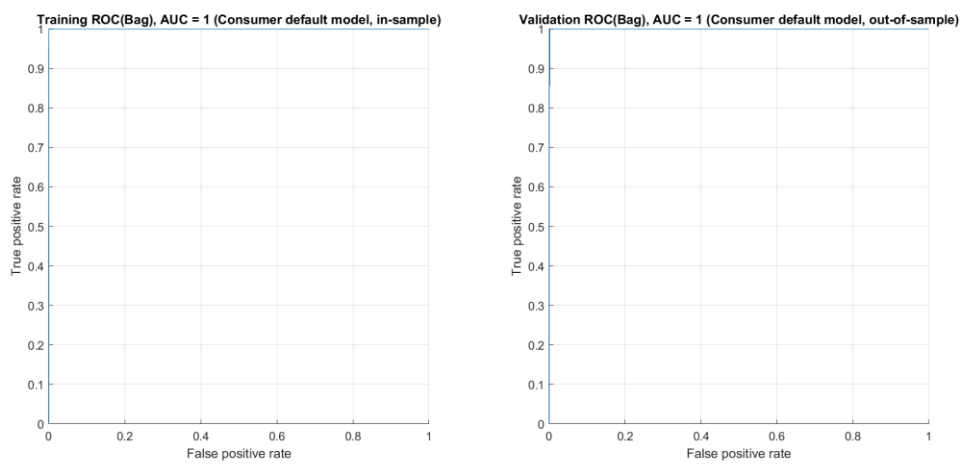


Figure 5-26: Bank 4 AUC results for 70% in-sample and 30% out-of-sample (bagging)

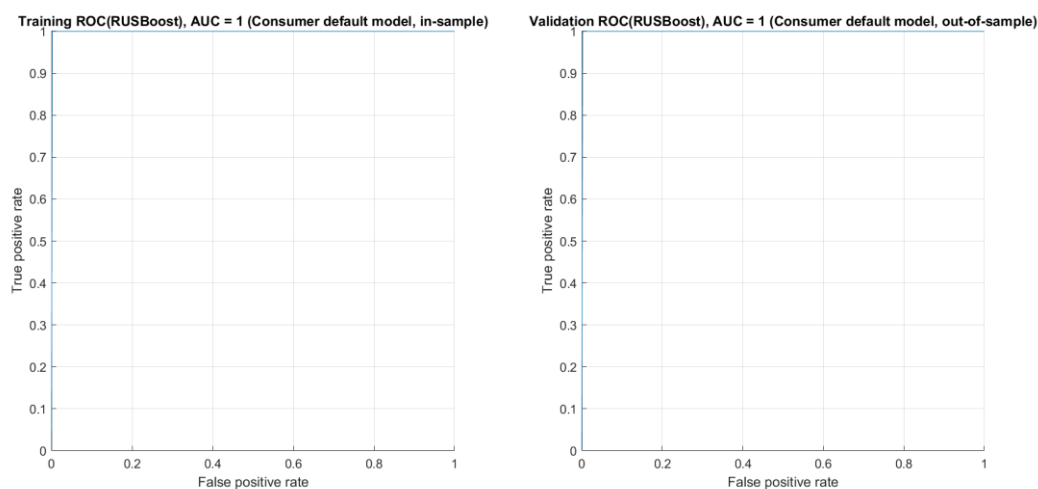


Figure 5-27: Bank 4 AUC results for 70% in-sample and 30% out-of-sample (RUSBoosted)

As with Bank 3, there are several good models that would provide information about different variables, so more data would be required for us to determine the successful model.

5.3.1.5 Bank 5

Bank 5 is a small Islamic bank. It has 2,154 customers with consumer loans, 714 are default customers. That gives a default rate of 33.14%. We made 45,022 observations, as shown in Table 5-59.

Table 5-59: Bank 5 number of observations and customers

Bank 5	Consumer Loans
Observations	45,022
Default cases	10,957
Non-Default cases	34,065
Rate of default cases	24.337%
Number of customers	2,154
Number of default customers	714
Number of non- default customers	1,440
Rate of default customers	33.148%

All the AUC results for the different methods were greater than 0.84. The best result was for bagging (0.98), as shown in Table 5-60.

Table 5-60: Bank 5 AUC results

Bank 5	AUC Consumer loans
Logistic Regression	0.86
Decision Tree	0.9
Linear Support Vector Machines	0.86
Bayesian Network	0.84
Bagged	0.98
RUSBoosted	0.89
AdaBoostM1	0.9

The results for accuracy rate, type I error, type II error, and running time also show that bagging was is the most effective model for this bank's data set. The maximum accuracy rate was 93%, as shown in Table 5-61.

Table 5-61: Bank 5 confusion matrix results

Bank 5	True Positive	True Negative	False Positive	False Negative	Average Accuracy	Type I Error	Type II Error
Logistic Regression	85%	65%	35%	15%	75%	15%	29%
Decision Tree	90%	69%	31%	10%	80%	10%	26%
Linear Support Vector Machines	86%	64%	36%	14%	75%	14%	30%

Bayesian Network	83%	60%	40%	17%	72%	17%	33%
Bagged	96%	90%	10%	4%	93%	4%	9%
RUSBoosted	93%	59%	41%	7%	76%	7%	31%
AdaBoostM1	89%	71%	29%	11%	80%	11%	25%

The training time for bagging was 36 seconds, showing that it was efficient (see Table 5-62).

Table 5-62: Bank 5 running time of models

Bank 5	Consumer Loans	
	Prediction Speed	Training Time
Logistic Reg.	~660000 obs/sec	5.1727 sec
Tree	~550000 obs/sec	2.6579 sec
SVM	~16000 obs/sec	645.29 sec
Bayesian	~250000 obs/sec	4.4691 sec
RUSBoosted	~97000 obs/sec	23.301 sec
Bagged	~42000 obs/sec	35.776 sec
AdaBoostM1	~94000 obs/sec	30.475 sec

Using 70% of the data as an in-sample training set gave an AUC of 0.9. The remaining 30%, used as an out-of-sample test, gave an AUC of 0.89, as shown by Figure 5-28.

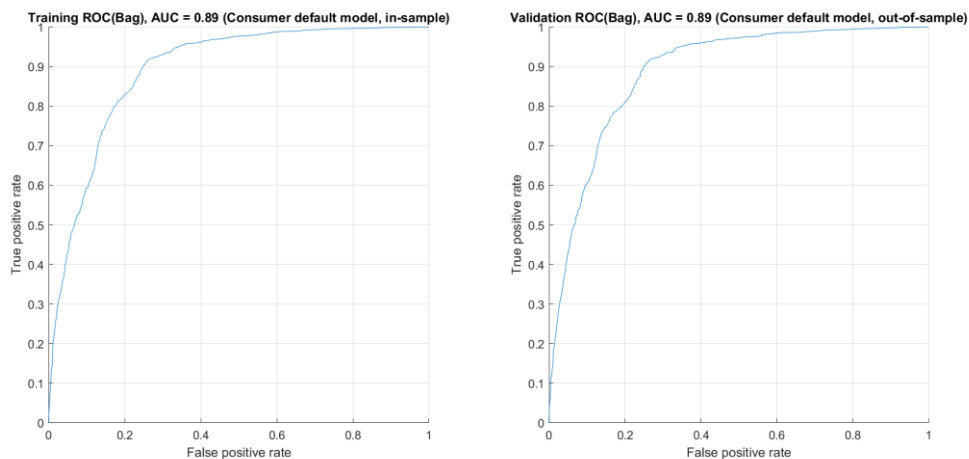


Figure 5-28: Bank 5 AUC results for 70% in-sample and 30% out-of-sample

Figure 5-29 shows the importance of the input variables for the prediction model for Bank 5.

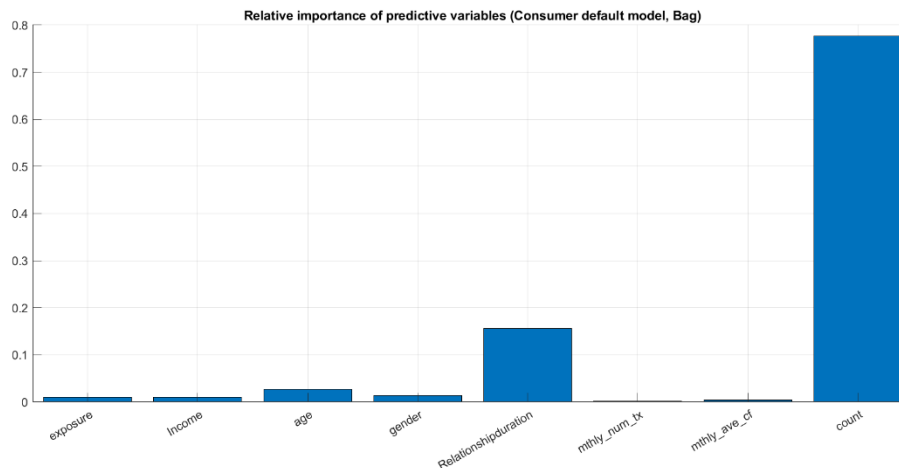


Figure 5-29: The relative importance of the predictive variables for Bank 5

The most significant variable for Bank 5 was the number of loans. This was followed by the length of a customer's relationship with the bank. The rest of the variables were of almost equal importance. For this bank, the length of relationship was more important for consumer loans than it was for instalment loans. This makes sense given that consumer loans last for 5 years, whereas instalment loans last for 15 years. This is the reason why the relationship variable is used as evidence for consumer loans rather than instalment loans because consumer loans faster will be repaid faster than instalment loans. This justifies our inclusion of different types of loans in the model because they can be affected by different variables.

5.3.1.6 Bank 6

Bank 6 is the large bank. The number of customers with consumer loans was 4,859, 142 of which were default cases (a default rate of 2.92%). There were 44,754 observations, as shown by Table 5-63.

Table 5-63: Bank 6 number of observations and customers

Bank 6	Consumer Loans
Observations	44,754
Default cases	432
Non-Default cases	44,322
Rate of default cases	0.965%
Number of customers	4,859

Number of default customers	142
Number of non- default customers	4,717
Rate of default customers	2.922%

Analysing the AUC for the different methods showed that bagging, RUSBoosted, and AdaBoostM1 were the most effective, all of which had AUCs of 0.92 (see Table 5-64).

Table 5-64: Bank 6 AUC

Bank 6	AUC Consumer loans
Logistic Regression	0.8
Decision Tree	0.85
Linear Support Vector Machines	0.46
Bayesian Network	0.71
Bagged	0.92
RUSBoosted	0.92
AdaBoostM1	0.92

The accuracy rate, type I error, type II error, and running time showed that AdaBoostM1 was the most effective model with a maximum accuracy rate of 78% (see Table 5-65) and a training time of 19 seconds (see Table 5-66).

Table 5-65: Bank 5 confusion matrix results

Bank 6	True Positive	True Negative	False Positive	False Negative	Average Accuracy	Type I Error	Type II Error
Logistic Regression	99%	N/A	N/A	1%	N/A	1%	N/A
Decision Tree	99%	40%	60%	1%	70%	1%	38%
Linear Support Vector Machines	99%	N/A	N/A	1%	N/A	1%	N/A
Bayesian Network	99%	3%	97%	1%	51%	1%	49%
Bagged	99%	46%	54%	1%	73%	1%	35%
RUSBoosted	99%	6%	94%	1%	53%	1%	49%
AdaBoostM1	99%	56%	44%	1%	78%	1%	31%

Table 5-66: Bank 5 running time models

Bank 6	Consumer Loans	
	Prediction Speed	Training Time

Logistic Reg.	~1100000 obs/sec	1.7813 sec
Tree	~1200000 obs/sec	1.6639 sec
SVM	~310000 obs/sec	4.8462 sec
Bayesian	~810000 obs/sec	1.1588 sec
RUSBoosted	~170000 obs/sec	5.6821 sec
Bagged	~88000 obs/sec	18.443 sec
AdaBoostM1	~150000 obs/sec	18.749 sec

Using 70% of the data set as an in-sample training set gave an AUC of 0.97 and using the remaining 30% as an out-of-sample test gave an AUC of 0.96. This is shown in Figure 5-30.

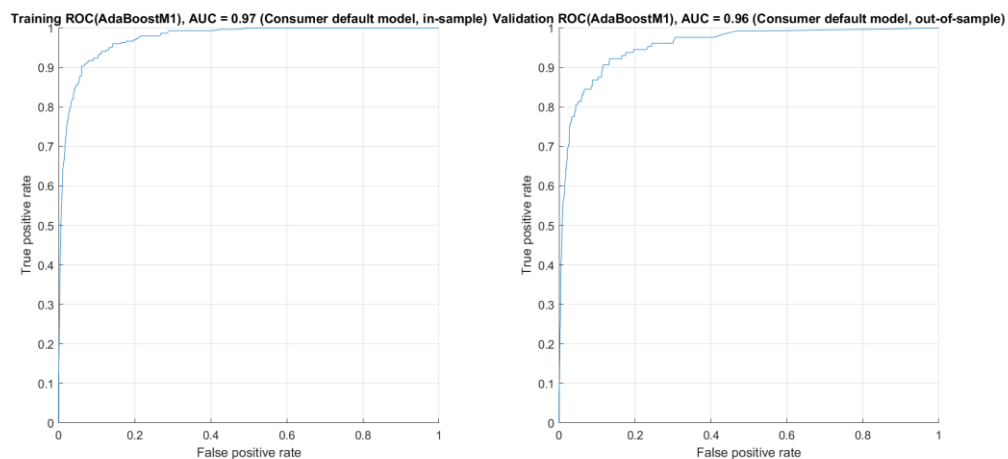


Figure 5-30: Bank 6 AUC results for 70% in-sample and 30% out-of-sample

The following Figure 5-31 displays the relative importance of the variables for Bank 6.

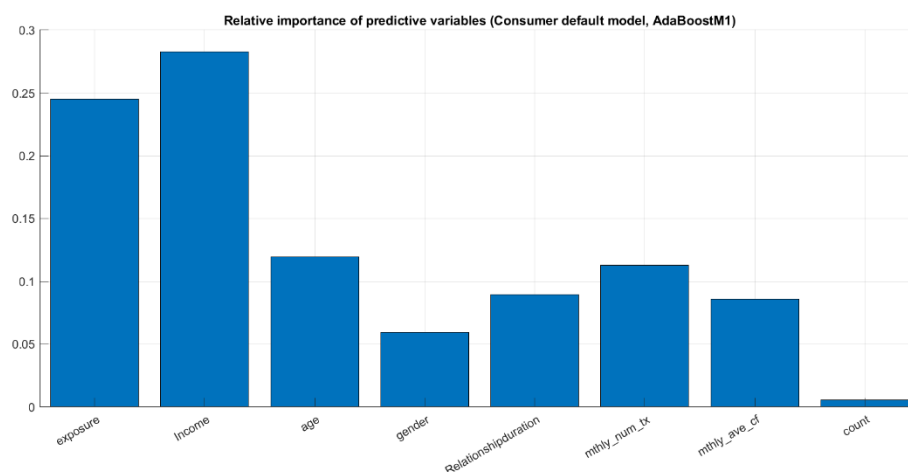


Figure 5-31: The relative importance of the predictive variables for Bank 6

For Bank 6, income was the most influential variable, followed by credit card exposure. Length of relationship with the bank, monthly number of transactions, and value of transactions were fairly equally weighted. Unlike the other Islamic banks, the number of loans was the least significant variable. This is probably because default cases at this bank occurred among customers with relatively few loans. Those with more loans had lower default rates.

5.3.2 Central Bank Model

For the Central Bank model, we combined the data from seven banks, including all the conventional and Islamic banks. In total, there were 33,699 customers with instalment loans, 6,343 of whom were default customers. This was equivalent to a default rate of 18.82%. The total number of observations was 1,048,793, as shown in Table 5-67.

Table 5-67: Number of observations and customers for all banks

All Banks	Consumer Loans
Observations	1,048,793
Default cases	21,756
Non-Default cases	1,027,037
Rate of default cases	2.074%
Number of customers	33,699
Number of default customers	6,343
Number of non- default customers	27,356
Rate of default customers	18.823%

Analysing the AUC showed that the ensemble models, bagging, RUSBoosted, and AdaBoostM1 all performed well (see Table 5-68).

Table 5-68: AUC results for all banks

All Banks	AUC Consumer loans
Logistic Regression	0.73
Decision Tree	0.82
Linear Support Vector Machines	0.5
Bayesian Network	0.71
Bagged	0.88
RUSBoosted	0.85

AdaBoostM1	0.85
------------	------

Examining the accuracy rates, type I errors, and type II errors showed that bagging had the highest accuracy rate (96%), as shown by Table 5-69.

Table 5-69: Confusion matrix results for all banks

All Banks	True Positive	True Negative	False Positive	False Negative	Average Accuracy	Type I Error	Type II Error
Logistic Regression	98%	9%	91%	2%	54%	2%	48%
Decision Tree	98%	70%	30%	2%	84%	2%	23%
Linear Support Vector Machines	98%	N/A	N/A	2%	N/A	2%	N/A
Bayesian Network	98%	5%	95%	2%	52%	2%	49%
Bagged	99%	92%	8%	1%	96%	1%	7%
RUSBoosted	99%	9%	91%	1%	54%	1%	48%
AdaBoostM1	98%	17%	83%	2%	58%	2%	46%

Bagging took longer than 25 minutes to complete its operations (see Table 5-70).

Table 5-70: Running time for the models for all banks

All Banks	Consumer Loans	
	Prediction Speed	Training Time
Logistic Reg.	~2300000 obs/sec	42.308 sec
Tree	~1900000 obs/sec	45.771 sec
SVM	~6700 obs/sec	5632.5 sec
Bayesian	~1000000 obs/sec	17.954 sec
RUSBoosted	~110000 obs/sec	184.68 sec
Bagged	~53000 obs/sec	1514 sec
AdaBoostM1	~120000 obs/sec	764.34 sec

Given that we required a robust model for the Central Bank, and based on the individual results, we thought it would be advisable to rerun the prediction model with a focus on the new ensemble models (LogitBoost and GentelBoost) that combined the results of the other models. Following our analysis, we concluded that a deep comparison of almost all the ensemble models would be suitable for the Central Bank model.

We also performed a k-fold test by the AUC of the test sets. We found that GentleBoost was the most effective model for that data for all the banks data. We

took 70% of the data as our in-sample training set, and it produced an AUC of 0.91. We used the remaining 30% as our out-of-sample test set, and it produced an AUC of 0.92. This is illustrated in Figure 4-32.

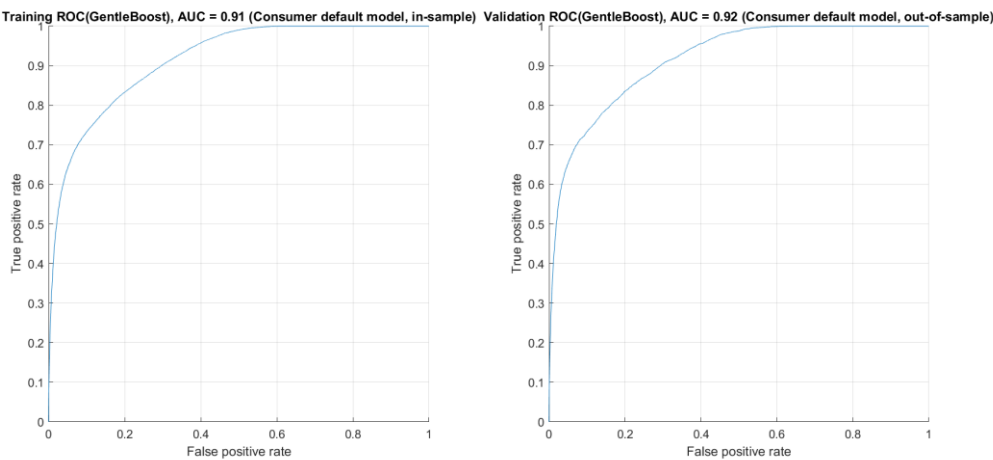


Figure 5-32: AUC results for 70% in-sample and 30% out-of-sample for all banks

Figure 5-33 shows the importance of the input variables for our prediction model for the Central Bank.

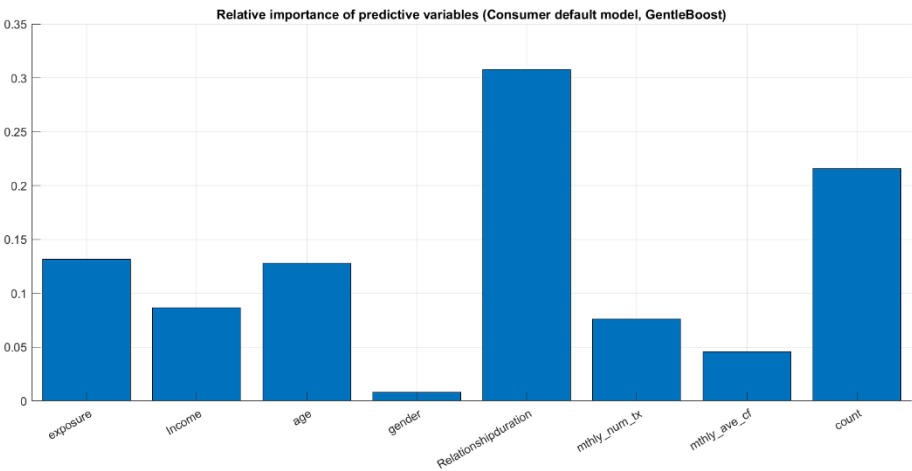


Figure 5-33: The relative importance of the predictive variables for all the banks

As expected, the length of a customer’s relationship with a bank was the most influential variable for predicting the likelihood that they would default. This was followed by the number of loans, which was driven by both bank types (Islamic and conventional). Although the average number of transactions per month was significant for some of the individual banks, it was not as influential when all the data was combined. However, we would still recommend that banks include it in

their prediction model because it can be an indicator of behaviour, spending, and money management.

5.4 Results Overview

As seen, each bank has its own data sets with individual default rates as inputs for the prediction model. Different marketing strategies can result in different acceptable exposure risks, which will lead to different acceptable rates of defaulting. Due to the different lengths of time that the banks have been established for, there were different relative importance rates for the parameters for the banks (this was the case for Bank 5). Furthermore, the different types of loans resulted in different results for the classification models used at the same bank. Nevertheless, the regulations related to each type of loans create a special circumstance for each bank. There were different performances for different customers with the same characteristics. This implies that one model is not suitable for all. This demonstrates the need for customized models for each bank. Bank 3 indicates that further data is required since almost all the models were suitable. Table 5 71 and Table 5 72 illustrate the AUC model results for each bank and the central bank for the two kinds of loans.

Table 5-71: AUC results comparison for Instalment loans

Model	Bank 1	Bank 2	Bank 3	Bank 4	Bank 5	Bank 6
Logistic Regression	0.7	0.66	0.82	0.73	0.9	0.69
Decision Tree	0.72	0.75	1	0.71	0.93	0.74
Linear Support Vector Machines	0.59	0.49	0.63	0.6	0.89	0.51
Bayesian Network	0.66	0.68	N/A	0.71	0.81	0.65
Bagged	0.63	0.8	1	0.71	0.98	0.95
RUSBoosted	0.83	0.76	1	0.81	0.93	0.85
AdaBoostM1	0.81	0.75	1	0.77	0.93	0.91

Table 5-72: AUC results comparison for Consumer loans

Model	Bank 1	Bank 2	Bank 3	Bank 4	Bank 5	Bank 6
Logistic Regression	0.8	0.67	0.92	0.56	0.86	0.8
Decision Tree	0.82	0.74	0.99	0.52	0.9	0.85
Linear Support Vector Machines	0.57	0.54	0.63	0.52	0.86	0.46
Bayesian Network	0.79	0.66	N/A	N/A	0.84	0.71
Bagged	0.86	0.73	0.98	0.57	0.98	0.92
RUSBoosted	0.96	0.76	1	0.8	0.89	0.92
AdaBoostM1	0.94	0.77	1	0.61	0.9	0.92

The results of our work have shown that the credit risk weight for each customer should be reduced for banks from the standardised approach that is currently adopted in Kuwait for unrated customers. Rather than being 75% of the total loan amount, it should be reduced to approximately 30% or less. This could help the Central Bank to adjust its regulations on banks when it comes to calculating their capital adequacy.

The results can also be used to assess the amount of liquidity needed within the banking system. The Central Bank could lower its requirements regarding capital adequacy, allowing banks to use the excess funds to manage their liquidity. This would reduce their reliance on the Central Bank for liquidity support. Interbank activities could be reduced leading to more efficient and reliable pricing for those activities. For example, banks could use interbank loans to help each other meet their liquidity risks. Lowering the connectedness between different banks would help to lower the levels of systemic risk. This would lead to more reliable interest rates.

The role of the Central Bank could be enhanced by using the robust system created in this study. That system relies on previously unused variables to calculate the likelihood that a person will default on their loan. By using a thorough

system, like the model presented, the Central Bank will be able to assess the risk assessment of clients when they apply for credit, adding more detail to the system that is already in place.

In evaluating the relative importance of the independent variables selected for our model, the Central Bank can use periodic stress testing to assess them and add new impact factors as necessary.

Nevertheless, it will be more efficient to inspect the internal models used by banks once the inspection team have a clear view of the selected variables and their relative significance. Also, a reasonable range of credit risk for specific clients with structured criteria could be approved and selected as a baseline for local banks to follow.

Some general guidelines from the results are as follows:

- The regulations for granting loans need to be revisited because multiple loans are being granted even though they are capped at 40% of a clients' total monthly income.
- From the models for the Central Bank, it is clear that the number of previous loans has a greater impact on consumer loans than instalment loans. This is likely due to the time frame of the two loans: consumer loans last for up to 5 years, whereas instalment loans last for up to 15 years.
- The modelling performed in this study could be implemented for individual banks or the Central Bank to provide early warnings about customers who might be likely to default on their loans.
- These models could be used as a preliminary stage in the process of developing credit rating scores for new loan applicants.

6 CONCLUSION

Machine learning is a robust mechanism for solving problems in many fields, not just banking. This study created satisfactory models for several banks in Kuwait, and an aggregated model to be used by the Central Bank of Kuwait was also developed. The models built for this study could be used by Islamic banks for risk assessments when issuing credit.

The results of the study show that the credit risk weight for each customer should be reduced for banks in Kuwait. Whereas the standard approach in Kuwait is to provide unrated customers with 75% of the total loan, this should be reduced to approximately 30% or less. This could help the Central Bank to adjust its requirements for other banks when calculating their capital adequacy.

This model can also be used to assess the level of liquidity needed in the banking system. The Central Bank can have lower requirements for capital and allow other banks to use the excess funds to manage their liquidity. This would reduce their need to rely on the Central Bank. Also, interbank activities could be reduced leading to more efficient and reliable pricing for those activities.

The role of the Central Bank will be enhanced by the robust system provided here, incorporating new variables to calculate the likelihood of a customer defaulting. The model can be used as a risk assessment for clients when they apply for credit, thereby making the system more detailed.

In evaluating the relative importance of the independent variables selected for our model, the Central Bank can use periodic stress testing to assess them and add new impact factors as necessary. But it will be more efficient to inspect the internal models used by banks once the inspection team have a clear view of the selected variables and their relative significance. There should be a reasonable range of credit risk weightings for specific clients with structured approval criteria that the banks can follow.

The overall takeaways from the study are as follows:

- The regulations for granting loans need to be revisited because multiple loans are being granted even though they are capped at 40% of a clients' total monthly income.
- From the models for the Central Bank, the number of previous loans has a greater impact on consumer loans than instalment loans. This is likely due to the time frame of the two loans: consumer loans last for up to 5 years, whereas instalment loans last for up to 15 years.
- The modelling performed in this study could be implemented for individual banks or the Central Bank. This would provide early warnings about customers who might be likely to default on their loans.
- These models could be used as a preliminary stage in the process of developing credit rating scores for new loan applicants.
- The variables for the individual models and the aggregated models should be maintained as they all have varying levels of significance.

In our study, we have defended the idea of using a customized prediction model, as stated in the international requirements for internal risk weighting models, which state that:

- Each bank should have its own data sets with individual default rates as inputs for the prediction model. This means that the model makes different predictions.

- Different marketing strategies can result in different levels of acceptable exposure, leading to different levels of acceptable default rates.
- Due to the different times that banks are open for, there are different levels of importance for the parameters at each of the banks.
- The different type of loans resulted in different results for the classification models under the same bank. Nevertheless, the regulations related to each type of loan create a special circumstance for each bank. This shows that there are different credit performances for different customers with the same characteristics.

This implies that one model does not fit all. This means that there should be customized models for each bank. Further data is required for Bank 3, since almost all the models were suitable.

One of this study's major contributions was its expansion of the size of the data set. Previous studies had used smaller data ranges, whereas this study used a larger data set, thereby providing more robust solutions and opening up opportunities for future models to perform better. This was also the first time in which an attempt had been made to develop a credit rating model for Islamic banks. We provided an internal model for each of the six banks that we looked at, as well as a robust system for the Central Bank that used the data from seven banks. We combined several methods by using ensemble models and testing them for new credit ratings. We automated the analysis of the suitable models through code. Rather than running the application, comparing the AUC results, and selecting the most efficient model, we automated the process so that financial analysts and bankers would be able to run the code easily. This was the first time that this kind of work had been done for Kuwaiti banks.

One of the limitations of this study was that we could not develop a model for Bank 3. This was due to the low number of default cases. It is recommended that future studies set a minimum benchmark for credit cases as observations for the prediction models. This would help with the fact that the performance of some of the classification models in the k-fold analysis were fairly similar, particularly for banks with relatively few customers.

Moreover, our work was limited to focusing on household customers. However, this work could be recommended and tested for other types of clients, such as

corporations in the Middle East that are not graded by rating agencies like Moody's or S&Ps. The internal rating system proposed here would help to solve this problem of a lack of rating. The work could be tested with small- and medium-sized companies.

Regarding Islamic products, one limitation was that we did not include all products for Islamic households. Therefore, the models should be adjusted to work with other forms of investment financing products that are used in Islamic banking.

This work could also be implemented by the specialist Kuwaiti bank that is backed by the government for financing Kuwaiti business projects, as well as the governmental institute for financing small- and medium-sized enterprises. Implementing our model would enhance the acceptance criteria used when issuing credit, helping to reward better businesses and control the default rating. However, it should be noted that one of the limitations of this study is that it was not possible to use the data from all of the banks studied. Future studies could expand the research by looking at a wider range of banks.

REFERENCES

- Albarrak, N., Alsanousi, H., Moulitsas, I. and Filippone, S. (2020) 'USING BIG DATA TO COMPARE CLASSIFICATION MODELS FOR HOUSEHOLD CREDIT RATING IN KUWAIT', *International Journal of Soft Computing and Artificial Intelligence*, 8(2), pp. 1–7.
- Almoqren, N. and Altayar, M. (2016) 'The motivations for big data mining technologies adoption in saudi banks', *2016 4th Saudi International Conference on Information Technology (Big Data Analysis), KACSTIT 2016*. IEEE. doi: 10.1109/KACSTIT.2016.7756075.
- Attfield, C. L. F. (2005) 'A Time Series Aggregate Demand Model with Demographic and Income Distribution Indices', (December).
- Badrinath, S. (2021) 'How Emerging Technologies Are Enabling The Banking Industry', *inc42*.
- Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J. and Vanthienen, J. (2003) 'Benchmarking state-of-the-art classification algorithms for credit scoring', *Journal of the Operational Research Society*, 54(6), pp. 627–635. doi: 10.1057/palgrave.jors.2601545.
- Bank of England (2019) 'Machine learning in UK financial services', *Bank of England*.
- Basel Committee on Banking Supervision (2005) *An Explanatory Note on the Basel II IRB Risk Weight Functions*, Bank for International Settlements. Available at: www.bis.org/bcisirbriskweight.pdf.
- Basti, E., Kuzey, C. and Delen, D. (2015) 'Analyzing initial public offerings' short-term performance using decision trees and SVMs', *Decision Support Systems*, 73, pp. 15–27. doi: 10.1016/j.dss.2015.02.011.
- Bazarbash, M. (2019) 'FinTech in Financial Inclusion: Machine Learning Applications in Assessing Credit Risk', *IMF Working Papers*, 19(109), p. 1. doi: 10.5089/9781498314428.001.

Bowman, D. and Narayandas, D. A. S. (2004) 'Linking Customer Management Effort to Customer Profitability in Business Markets', *XLI(November)*, pp. 433–447.

Breiman, L. (1996) 'Bagging predictors', *Machine Learning*, 24(2), pp. 123–140. doi: 10.1007/BF00058655.

Breiman, L. (2001) *Random Forests*. Berkeley, California. Available at: <https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf>.

Broderick, A. J., Greenley, G. E. and Mueller, R. D. (2007) 'The behavioural homogeneity evaluation framework: Multi-level evaluations of consumer involvement in international segmentation', *Journal of International Business Studies*, 38(5), pp. 746–763. doi: 10.1057/palgrave.jibs.8400296.

Brown, I. and Mues, C. (2012) 'An experimental comparison of classification algorithms for imbalanced credit scoring data sets', *Expert Systems with Applications*. Elsevier Ltd, 39(3), pp. 3446–3453. doi: 10.1016/j.eswa.2011.09.033.

Buneeva, R. I., Buneeva, M. V., Toropov, V. V. and Khryuchkina, E. A. (2016) 'Analysis of Researching Dynamics and Structure of Consumer Expenditure of the Russian Population', *International Journal of Economics and Financial Issues*, 6(S2), pp. 31–36.

Carrington, M. J., Neville, B. A. and Whitwell, G. J. (2010) 'Why ethical consumers don't walk their talk: Towards a framework for understanding the gap between the ethical purchase intentions and actual buying behaviour of ethically minded consumers', *Journal of Business Ethics*, 97(1), pp. 139–158. doi: 10.1007/s10551-010-0501-6.

Central Bank of Kuwait (2018) 'Financial Stability Report 2018', *Central Bank of Kuwait*, pp. 1–69. doi: 10.1007/s13398-014-0173-7.2.

Central Bank of Kuwait (2020) *Financial Stability Report*.

Chau, M. and Xu, J. (2012) 'Business Intelligence and Analytics', *MIS*

Quarterly, 36(4), pp. 1165–1188.

Chen, F. H., Chi, D. J. and Wang, Y. C. (2015) 'Detecting biotechnology industry's earnings management using Bayesian network, principal component analysis, back propagation neural network, and decision tree', *Economic Modelling*. Elsevier B.V., 46, pp. 1–10. doi: 10.1016/j.econmod.2014.12.035.

Chen, H., Chiang, R. H. L. and Storey, V. C. (2012) 'Business Intelligence and Analytics: From Big Data to Big Impact', *MIS Q.*, 36, pp. 1165–1188.

Chitnis, M. and Hunt, L. C. (2011) 'Modelling UK household expenditure: Economic versus noneconomic drivers', *Applied Economics Letters*, 18(8), pp. 753–767. doi: 10.1080/13504851.2010.496721.

Course, M. L. C. (2020) *Classification: ROC Curve and AUC*. Available at: <https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc>.

Dagillienė, L. and Klovienė, L. (2019) 'Motivation to use big data and big data analytics in external auditing', *Managerial Auditing Journal*, 34(7), pp. 750–782. doi: 10.1108/MAJ-01-2018-1773.

Du, R. Y., Kamakura, W. A. and Mela, C. F. (2007) 'Size and Share of Customer Wallet', 2429.

Elshaw, R., Sakr, S., Talia, D. and Trunfio, P. (2018) 'Big Data Systems Meet Machine Learning Challenges : Towards Big Data Science as a Service', *Big Data Research*. Elsevier Inc., 14, pp. 1–11. doi: 10.1016/j.bdr.2018.04.004.

Emerson, D., Zhou, W. and Piramuthu, S. (2009) 'Goodwill, inventory penalty, and adaptive supply chain management', *European Journal of Operational Research*. Elsevier B.V., 199(1), pp. 130–138. doi: 10.1016/j.ejor.2008.11.007.

Fornell, C., Rust, R. T. and Dekimpe, M. G. (2010) 'The Effect of Customer Satisfaction on Consumer Spending Growth', *Journal of Marketing Research*, 47(1), pp. 28–35. doi: 10.1509/jmkr.47.1.28.

Freund, Y. (2009) 'A more robust boosting algorithm'. Available at:

<http://arxiv.org/abs/0905.2138>.

Friedman, J. H. (2002) 'Stochastic gradient boosting', *Computational Statistics and Data Analysis*, 38(4), pp. 367–378. doi: 10.1016/S0167-9473(01)00065-2.

Friedman, J., Hastie, T., Rosset, S., Tibshirani, R. and Zhu, J. (2004) 'Discussion of boosting papers', *Ann Stat*, 28(2), pp. 337–407. Available at: <http://www.yaroslavvb.com/papers/friedman-discussion.pdf%5Cnpapers2://publication/uuid/C33BE497-39B4-4315-A405-3BBE029833BF>.

Friedman, J., Tibshirani, R. and Hastie, T. (2000) 'Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors)', *The Annals of Statistics*, 28(2), pp. 337–407. doi: 10.1214/aos/1016120463.

Gogtay, N. J. and Thatte, U. M. (2017) 'Principles of correlation analysis', *Journal of Association of Physicians of India*, 65(MARCH), pp. 78–81.

Golder, P. N. (2000) 'Historical Method in Marketing Research with New Evidence on Long-Term Market Share Stability', *Journal of Marketing Research*, 37(2), pp. 156–172. doi: 10.1509/jmkr.37.2.156.18732.

Gosavi, A. (2018) 'Can Mobile Money Help Firms Mitigate the Problem of Access to Finance in Eastern sub-Saharan Africa?', *Journal of African Business*. Routledge, 19(3), pp. 343–360. doi: 10.1080/15228916.2017.1396791.

Gottschalk, L. R. (1945) 'Understanding history: A primer of historical method', *Social science research council*, 53.

Hauke, J. and Kossowski, T. (2011) 'Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data', *Quaestiones Geographicae*, 30(2), pp. 87–93. doi: 10.2478/v10117-011-0021-1.

Holmes, A., Illowsky, B. and Dean, S. (2017) *Introductory Business Statistics*, openstax.

Honest, N. and Patel, A. (2016) 'A Survey of Big Data Analytics', *International Journal of Information Sciences and Techniques (IJIST)*, 6(1), pp. 35–43.

Iman, N. (2018) 'Is mobile payment still relevant in the fintech era?', *Electronic Commerce Research and Applications*, 30(May), pp. 72–82. doi: 10.1016/j.elerap.2018.05.009.

Jang, S., Prasad, A. and Ratchford, B. T. (2016) 'Consumer spending patterns across firms and categories: Application to the size- and share-of-wallet', *International Journal of Research in Marketing*. Elsevier B.V., 33(1), pp. 123–139. doi: 10.1016/j.ijresmar.2015.06.008.

Juster, F. T. (1966) 'Consumer Buying Intentions and Purchase Probability: An Experiment in Survey', *Journal of the American Statistical Association*, 61(315), pp. 658–696. doi: 10.2307/2282779.

Kenton, W. (2021) *Demand Curve*, Investopedia. Available at: <https://www.investopedia.com/terms/d/demand-curve.asp>.

Kim, M., Zoo, H., Lee, H. and Kang, J. (2018) 'Mobile financial services, financial inclusion, and development: A systematic review of academic literature', *Electronic Journal of Information Systems in Developing Countries*, (January), pp. 1–17. doi: 10.1002/isd.12044.

Kohly, R., Devaraj, S. and Mahood, M. A. (2004) 'Understanding Determinants of Online Consumer Satisfaction: A Decision Process Perspective', *Journal of Management Information Systems*, 21(1), pp. 115–136. doi: 10.1080/07421222.2004.11045796.

Kotler, P. T. and Keller, K. L. (2011) *Marketing Management*. Pearson Education.

Lawrence, R. J. (1966) 'Models of Consumer Purchasing Behaviour', *Journal of the Royal Statistical Society. Series C: Applied Statistics* 1, 15(3).

Lim, S., Woo, J., Lee, J. and Huh, S.-Y. (2018) 'Consumer valuation of personal information in the age of big data', *Journal of the Association for Information Science and Technology*, 69(1), pp. 60–71. doi: 10.1002/asi.23915.

Lloyd, A. D., Antonioletti, M. and Sloan, T. M. (2013) 'Able but not willing? Exploring divides in digital versus physical payment use in China', *Information*

Technology & People, 29(2016), pp. 250–279. doi: 10.1108/ITP-10-2014-0243.

Lula, J. and Antille, G. (2007) 'Estimation of private consumption functions for Switzerland Jonela Lula'.

MathWorks (2021a) *Adaptive Logistic Regression*. Available at: <https://uk.mathworks.com/help/stats/ensemble-algorithms.html#bsw8au3>.

MathWorks (2021b) *Bootstrap Aggregation (Bagging) and Random Forest*. Available at: <https://uk.mathworks.com/help/stats/ensemble-algorithms.html#bsw8at7>.

MathWorks (2021c) *corr*.

MathWorks (2021d) *Gentle Adaptive Boosting*. Available at: <https://uk.mathworks.com/help/stats/ensemble-algorithms.html#bsw8avb>.

MathWorks (2021e) *Random Undersampling Boosting*. Available at: <https://uk.mathworks.com/help/stats/ensemble-algorithms.html#btfwpd3>.

MathWorks (2021f) *Robust Boosting*. Available at: <https://uk.mathworks.com/help/stats/ensemble-algorithms.html#bsw8avy>.

Mayer-Schönberger, V. and Cukier, K. (2013) *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Mcleod-, S. (2007) 'HCC Certificate in Counselling Skills Maslow's Hierarchy of Needs', (December), p. 3. doi: 10.1016/B978-0-88415-752-6.50250-2.

Munkhdalai, L., Munkhdalai, T., Namsrai, O. E., Lee, J. Y. and Ryu, K. H. (2019) 'An empirical comparison of machine-learning methods on bank client credit assessments', *Sustainability (Switzerland)*, 11(3), pp. 1–23. doi: 10.3390/su11030699.

Nyathi, K., Ndlovu, S., Moyo, S. and Nyathi, T. (2014) 'Optimisation of the Linear Probability Model for Credit Risk Management', *International Journal of Computer and Information Technology*, 03(06), pp. 1340–1345.

Petropoulos, A., Siakoulis, V., Stavroulakis, E. and Klamargias, A. (2018) 'A

robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting', *The use of big data analytics and artificial intelligence in central banking*, 50(August), pp. 30–31. Available at: https://www.bis.org/ifc/publ/ifcb49_49.pdf.

Phitchayanon, P. (2020) 'Machine Learning with MATLAB', *MathWorks*, pp. 0–65. Available at: <https://uk.mathworks.com/solutions/machine-learning.html>.

Pilati, L. and Ricci, G. (1995) 'Consumer behaviour under conditions of incomplete information on quality', *Rivista Internazionale di Scienze Sociali*, 103(2), pp. 317–332.

Qian, D. J. and Velayutham, S. (2017) 'Conventional Banking and Islamic Banking: Do the Different Philosophies Lead to Different Financial Outcomes?', *Journal of Wealth Management & Financial Planning*, 4(June), pp. 3–14. Available at: <https://www.islamicfinance.com/2014/12/>.

Rahman, M. M. and Sloan, T. (2015) 'Opportunities and challenges of M-commerce adoption in Bangladesh: An empirical study', *Journal of Internet Banking and Commerce*, 20(3). doi: 10.4172/1204-5357.1000124.

Ramasamy, B. and Yeung, M. (2010) 'Customer Satisfaction and the Consumption Function', *Economics Research International*, 2010, pp. 1–5. doi: 10.1155/2010/202014.

Reinartz, W. J. and Kumar, V. (2003) 'The Impact of Customer Relationship Characteristics on', 67(January), pp. 77–99.

Reinartz, W., Thomas, J. S. and Kumar, V. (2005) 'Balancing Acquisition and Retention Resources to Maximize Customer', 69(January), pp. 63–79.

Samreen, A., Zaidi, F. B. and Sarwar, A. (2013) 'Design and development of credit scoring model for the commercial banks of Pakistan: forecasting creditworthiness of individual borrowers', *International Journal of Business and Social Science*, 2(5), pp. 1–26.

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J. and Napolitano, A. (2008)

'RUSBoost: Improving classification performance when training data is skewed', *Proceedings - International Conference on Pattern Recognition*, (December). doi: 10.1109/icpr.2008.4761297.

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J. and Napolitano, A. (2010) 'RUSBoost: A hybrid approach to alleviating class imbalance', *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans*, 40(1), pp. 185–197. doi: 10.1109/TSMCA.2009.2029559.

Selvanathan, S. and Selvanathan, E. A. (2004) 'Empirical regularities in South African consumption patterns', *Applied Economics*, 36(20), pp. 2327–2333. doi: 10.1080/0003684042000280328.

Shabani, N., Munir, A. and Mohanty, S. P. (2022) 'A Study of Big Data Analytics in Internal Auditing', *Lecture Notes in Networks and Systems*, 295(September), pp. 362–374. doi: 10.1007/978-3-030-82196-8_27.

Shepherd, R., Mahnusson, M. and Sjöden, P.-O. (2005) 'Determinants of Consumer Behaviour Related to Organic Foods', *Ambio*, 34(4/5), pp. 352–359.

Singh, A., Dhayal, N. and Shamim, A. (2014) 'Consumer Buying Behaviour', *International Research Journal of Management Sociology & Humanity*, 5(12), pp. 17–21.

Soltani Delgosha, M., Hajiheydari, N. and Fahimi, S. M. (2021) 'Elucidation of big data analytics in banking: a four-stage Delphi study', *Journal of Enterprise Information Management*, 34(6), pp. 1577–1596. doi: 10.1108/JEIM-03-2019-0097.

Song, X. P., Hu, Z. H., Du, J. G. and Sheng, Z. H. (2014) 'Application of machine learning methods to risk assessment of financial statement fraud: Evidence from China', *Journal of Forecasting*, 33(8), pp. 611–626. doi: 10.1002/for.2294.

Spanos, A. (1989) 'Early Empirical on Findings the Consumption Facts or Stylized Fiction : a Retrospective', *Oxford Economic Papers*, 41(1), pp. 150–169.

Speybroeck, N. (2012) 'Classification and regression trees', *International Journal*

of Public Health, 57(1), pp. 243–246. doi: 10.1007/s00038-011-0315-z.

Stejskal, L. and Stavkova, J. (2011) 'Structure and Determinants of Consumer Expenditures', *E & M Ekonomie a Management*, 14(2), pp. 45–50.

Tekaya, B., Feki, S. El, Tekaya, T. and Masri, H. (2020) 'Recent applications of big data in finance', *PervasiveHealth: Pervasive Computing Technologies for Healthcare*, (May). doi: 10.1145/3423603.3424056.

Trelewicz, J. Q. (2017) 'Big Data and Big Money: The Role of Data in the Financial Sector', *IT Professional*, 19(3), pp. 8–10. doi: 10.1109/MITP.2017.45.

Vapnik, N. V. (2000) *The Nature of Statistical Learning Theory*. Second. Springer.

Vitell, S. J., Paolillo, J. G. P. and Singh, J. J. (2006) 'The role of money and religiosity in determining consumers' ethical beliefs', *Journal of Business Ethics*, 64(2), pp. 117–124. doi: 10.1007/s10551-005-1901-x.

Wang, G., Ma, J., Huang, L. and Xu, K. (2012) 'Two credit scoring models based on dual strategy ensemble trees', *Knowledge-Based Systems*. Elsevier B.V., 26, pp. 61–68. doi: 10.1016/j.knosys.2011.06.020.

Wu, C. and Chen, H. L. (2000) 'A consumer purchasing model with learning and departure behaviour', *Journal of the Operational Research Society*, 51(5), pp. 583–591. doi: 10.1057/palgrave.jors.2600903.

Wu, X., Kumar, V., Ross, Q. J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z. H., Steinbach, M., Hand, D. J. and Steinberg, D. (2008) *Top 10 algorithms in data mining*, *Knowledge and Information Systems*. doi: 10.1007/s10115-007-0114-2.

Yang, S. and Berdine, G. (2017) 'The receiver operating characteristic (ROC) curve', *The Southwest Respiratory and Critical Care Chronicles*, 5(19), p. 34. doi: 10.12746/swrccc.v5i19.391.

Yeung, M. C. H., Ramasamy, B., Chen, J. and Paliwoda, S. (2013) 'Customer satisfaction and consumer expenditure in selected European countries', *International Journal of Research in Marketing*. Elsevier B.V., 30(4), pp. 406–416.

doi: 10.1016/j.ijresmar.2013.06.001.

APPENDICES

Appendix A Data

A.1 Raw data component and size

Table A-6-1: Bank 1 raw data

Name	Consists of (column heading)	Size (MB)
CBK1: Branch Daily Transactions	Customer ID Transaction Date Value of Transaction Currency of Transactions Card Type Transaction Type Year	245.56
CBK1: Cards, debit and credit, Daily Transactions (2009-2018)	Customer ID Transaction Date Value of Transaction Currency of Transaction Card Type Transaction Type Year	1,993.36
CBK1_Cards_2008	Customer ID Transaction Date Value of Transaction Currency of Transaction Card Type Transaction Type Year	138.03
CBK2: Customer Data, Customer General Information	Customer ID Age Gender Address Relationship date	12.99
	Customer ID	24.25

Name	Consists of (column heading)	Size (MB)
CBK2: Customer Data, Customer General Information_Nationality	AGE GENDER ADDRESS Relationship Date Nationality Country description	
CBK2_Education_Level	Customer ID Age Gender Address Education level	11.08
CBK3: Income	Customer ID The income from January until December (each month has its column) year	62.61
CBK4_Current Account Balance	Customer ID The balance from January until December (each month has its column) year	20.71
CBK4_Number of Current Account	Customer ID The Current Account Number from January until December (each month has its column) year	8.76
CBK4_Deposit Account Balance	Customer ID The balance from January until December (each month has its column) year	6.11
CBK4_Number of Deposit Account	Customer ID	2.72

Name	Consists of (column heading)	Size (MB)
	The Deposit Account Number from January until December (each month has its column) year	
CBK4_Joint Account Balance	Customer ID The balance from January until December (each month has its column) year	0.90
CBK4_Number of Joint Account	Customer ID The Joint Account Number from January until December (each month has its column) year	0.38
CBK4_Saving Account Balance	Customer ID The balance from January until December (each month has its column) Year	84.14
CBK4_Number of Saving Account	Customer ID The Saving Account Number from January until December (each month has its column) year	37.34
CBK5_Consumer_Loan_Count	Customer ID The count from January until December (each month has its column) Year	4.54
CBK5_Consumer_Loan_Interest	Customer ID The Interest from January until December (each month has its column) Year	5.56

Name	Consists of (column heading)	Size (MB)
CBK5_Consumer_Loan_Outstanding Balance	Customer ID The Outstanding Balance from January until December (each month has its column) Year	10.68
CBK5_Consumer_Loan_Principle_Payment	Customer ID The Principle Payment from January until December (each month has its column) Year	5.88
CBK5_Installment_Loan_Count	Customer ID The count from January until December (each month has its column) Year	6.36
CBK5_Instalment_Loan Interest	Customer ID The Interest from January until December (each month has its column) Year	22.72
CBK5_Instalment_Loan Outstanding Balance	Customer ID The Outstanding Balance from January until December (each month has its column) Year	19.35
CBK5_Installment_Loan_Principle_Payment	Customer ID The Principle Payment from January until December (each month has its column) Year	23.75
CBK6: Credit Cards limit	Customer ID The Credit Card Limit from January until December (each month has its column) Year	12.95
CBK6: Credit Cards Outstanding Balance	Customer ID	12.36

Name	Consists of (column heading)	Size (MB)
	The Credit Card Outstanding Balance from January until December (each month has its column) Year	
CBK6_Risk_Factor	Customer ID Customer risk score (1-100) Customer risk rating (low-medium-high)	5.64
CBK7_ATM	Customer 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)	43.85
CBK7_Call_Center	Customer ID Year and month Monthly transfers (counts and values) Monthly purchases (count and values) Monthly payments (counts and values)	3.86
CBK7_Internet	Customer ID Year and month Monthly transfers (counts and values) Monthly purchases (count and values) Monthly payments (counts and values)	7.86
CBK7_Mobile	Customer ID Year and month Monthly transfers (counts and values) Monthly purchases (count and values)	5.09

Name	Consists of (column heading)	Size (MB)
	Monthly payments (counts and values)	
CBK7_POS	Customer ID Year and month Monthly transfers (counts and values) Monthly purchases (count and values) Monthly payments (counts and values)	44.68
Total raw data size for bank 1		2,884.22

Table A-6-2: Bank 2 raw data

Name	Consists of (column heading)	Size (MB)
CBK1: Daily Account Transactions	Customer ID Transaction Date Value of Transactions Billing Currency Type of Card Transaction Type Year	32,101
CBK2: Customer Data, Customer General Information - All Active Customers	Customer ID Age Gender Address Relationship date	30.41
CBK2_Education_Level Customer Education 2014-2018 - All Customers	Customer ID Monthly Education level	534.11
CBK3: Income Customer Income 2014-2018 - All Active Customers	Customer ID The income from January until December for each month in the 5 years	286.28
	Customer ID	2.12

Name	Consists of (column heading)	Size (MB)
CBK4_Current Account Balance Current Account - End Balance 2014-2018 - 100k Customers	The balance from January until December for each month in the 5 years	
CBK4_Number of Current Account Current Account - No. of Account 2014-2018 - 100k Customers	Customer ID The Current Account Counts from January until December for each month in 5 years	0.77
CBK4_Deposit Account Balance Term Deposits - End Balance 2014-2018 - 100k Customers	Customer ID The balance from January until December for each month in the 5 years	0.43
CBK4_Number of Deposit Account Term Deposits - No. of Account 2014-2018- 100k Customers	Customer ID The Current Account Counts from January until December for each month in 5 years	0.201
CBK4_Saving Account Balance Savings Account - End Balance -2014-2018 - 100k Customers	Customer ID The balance from January until December for each month in the 5 years	36.99
CBK4_Number of Saving Account Savings Account - No. of Account 2014-2018 - 100k Customers	Customer ID The Current Account Counts from January until December for each month in 5 years	11.96
CBK5_Consumer_Loan_Count Consumer - No. of Account 2014-2018 - 100k Customers	Customer ID The count from January until December for each month in the 5 Years	1.57
CBK5_Consumer_Loan_Interest Loans - Consumer - Interest Paid 2014-2018 - 100k Customers	Customer ID The Interest from January until December for each month in the 5 Years	7.53
CBK5_Consumer_Loan_Outstanding Balance Loans - Consumer - Outstanding Balance 2014- 2018- 100k Customers	Customer ID The Outstanding Balance from January until December for each month in the 5 Years	5.78
	Customer ID	10.11

Name	Consists of (column heading)	Size (MB)
CBK5_Consumer_Loan_Principle_Payment Loans - Consumer - Principal Paid 2014-2018 - 100k Customers	The Principle Payment from January until December for each month in the 5 Years	
CBK5_Installment_Loan_Count Loans - Instalment - No. of Account 2014-2018 - 100k Customers	Customer ID The count from January until December for each month in the 5 Year	1.92
CBK5_Instalment_Loan Interest Loans - Instalment - Interest Paid 2014-2018 - 100k Customers	Customer ID The Interest from January until December for each month in the 5 Years	15.72
CBK5_ Instalment _Loan Outstanding Balance Loans - Instalment - Outstanding Balance 2014-2018 - 100k Customers	Customer ID The Outstanding Balance from January until December for each month in the 5 Years	8.19
CBK5_Installment_Loan_Principle_Payment Loans - Instalment - Principal Paid 2014-2018 - 100k Customers	Customer ID The Principle Payment from January until December for each month in the 5 Years	14.62
CBK6: Credit Cards limit Credit Card - Total Limit 2014-2018 - 100k Customers	Customer ID The Credit Card Limit from January until December for each month in the 5 Years	6.28
CBK6: Credit Cards Outstanding Balance Credit Card - Balance 2014-2018 - 100k Customers	Customer ID The Credit Card Outstanding Balance from January until December for each month in the 5 Years	10.31
CBK6: Credit Card - Credit Score -2016-2018 - 100k Customers	Customer ID Credit card credit scoring	0.43
	Customer ID	0.41

Name	Consists of (column heading)	Size (MB)
CBK5 Loans - Credit Score -2016-2018 - 100k Customers	Loan credit score	
CBK7_ATM ATM Transactions 2014-2018 - All Active Customers	Customer ID Year and month Monthly Withdrawals (counts & values) Monthly deposits (counts and values)	533.72
CBK7_Call_Center Automated Call Centre Transactions 2014-2018 - All Active Customers	Customer ID Year and month Monthly transfers (counts and values) Monthly payments (counts and values)	32.43
CBK7_Internet Online Banking Transactions 2014-2018 - All Active Customers	Customer ID Year and month Monthly transfers (counts and values) Monthly payments (counts and values)	81.75
CBK7_Mobile Mobile Banking Transactions 2014-2018 - All Active Customers	Customer ID Year and month Monthly transfers (counts and values) Monthly payments (counts and values)	65.5
CBK7_POS POS Transactions 2014-2018 - All Active Customers	Customer ID Year and month Monthly purchases (count and values)	557.64
CBK8: Customer Data	Customer ID Education level	610.53
Total raw data size for bank 4		34,969

Table A-6-3: Bank 3 raw data

Name	Consists of (column heading)	Size (MB)
CBK1: Daily Account Transactions	Customer ID	6,922

Name	Consists of (column heading)	Size (MB)
	Transaction Date Billing Currency Type of Card Transaction Type Value of Transactions	
CBK2: Customer Data, Customer General Information	Customer ID Age Gender Address Date of relationship	4
CBK3: Income Customer Income 2008-2018	Customer ID The income from January until December for each year	111
CBK4: Accounts counts and balances 2008-2018	Customer ID Type of account (current, saving or deposit) The Account counts from January until December for each month in 8 years	482
CBK5: Loans 2008-2018	Customer ID Loan type Loans counts Total loan amount Principle paid Interest paid Outstanding balance from January until December for each year	298
CBK6: Credit Cards	Customer ID The Credit Card Limit The Credit Card Outstanding Balance	697

Name	Consists of (column heading)	Size (MB)
	from January until December for each year	
CBK7: Electronic banking transactions (ATM, Call centre, Mobile, Internet, POS) 2008-2018	Customer ID Year and month Monthly Withdrawals (counts & values) Monthly deposits (counts and values) Monthly transfers (counts and values) Monthly payments (counts and values)	95.2
CBK8: Customer information 2008-2018	Customer ID Nationality address Gender	8.74
Total raw data size for bank 2		8,609

Table A-6-4: Bank 4 raw data

Name	Consists of (column heading)	Size (MB)
CBK1: Daily Account Transactions	Customer ID Dates Business Date Number of Transactions Value of Transactions Value of Transaction in Foreign Currency Currency Transaction Type Type of Card	16,015.94
CBK2: Customer Data, Customer General Information - All Active Customers	Customer ID Age Gender	35.95

	Address Education	
CBK3: Income Customer Income 2008-2018 - All Active Customers	Customer ID The income from January until December for each 11 years	137.53
CBK4: Accounts counts and balances 2008-2018 – All Customers	Customer 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	834.83
CBK5: Loans 2008-2018 – All Active Customers	Customer 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	285.54
CBK6: Credit Cards 2008-2018 – All Active Customers	Customer ID The Credit Card Limit The Credit Card Outstanding Balance from January until December for each month in the 11 Years	42.37
CBK6: Credit Card - Credit Score -2008-2018 – all active Customers	Customer ID Credit card credit scoring	118.15
	Customer ID	1,351.51

CBK7: Electronic banking transactions (ATM, Call centre, Mobile Internet, POS) 2008-2018 - All Active Customers	Year and month	
	Monthly Withdrawals (counts & values)	
	Monthly deposits (counts and values)	
	Monthly transfers (counts and values)	
	Monthly payments (counts and values)	
CBK8: Customer information 2008-2018 - All Active Customers	Customer ID	528.78
	Education level	
Total raw data size for bank 3		19,350.60

Table A-6-5: Bank 5 raw data

Name	Consists of (column heading)	Size (MB)
CBK1: Daily Account Transactions	Customer ID Transaction Date Billing Currency Type of Card Transaction Type Value of Transactions	488.37
CBK2: Customer Data, Customer General Information	Customer ID Age Gender Address	13,420.98
CBK3: Income Customer Income 2012- 2018 -	Customer ID The income from January until December for each year	7.6
CBK4: Accounts counts and balances 2012-2018	Customer ID Type of account (current, saving or deposit) The Account counts from January until December for each month in 11 years	13,420.98

Name	Consists of (column heading)	Size (MB)
	The account balance from January until December for each year	
CBK5: Loans_ 2012-2018	Customer ID Loan type Loans counts Total loan amount Principle paid Interest paid Outstanding balance from January until December for each year	77.33
CBK6: Credit Cards	Customer ID The Credit Card Limit The Credit Card Outstanding Balance from January until December for each year	6.29
CBK7: Electronic banking transactions (ATM, Call centre, Mobile, Internet, POS) 2012-2018	Customer ID Year and month Monthly Withdrawals (counts & values) Monthly deposits (counts and values) Monthly transfers (counts and values) Monthly payments (counts and values)	69.2
CBK8: Customer information 2012-2018	Customer ID Customer nationality Date of relationship	1.83
Total raw data size for bank 5		27,493

Table A-6-6: Bank 6 raw data

Name	Consists of (column heading)	Size (MB)
CBK1: Daily Account Transactions	Customer ID Transaction Date Value of Transactions Billing Currency Type of Card Transaction Type	91,399.90
CBK2: Customer Data, Customer General Information	Customer ID Age Gender Address Date of relationship Nationality	63,823.67
CBK3: Income Customer Income 2008-2018	Customer ID The income from January until December for each year	490.47
CBK4: Accounts counts and balances 2008-2018	Customer 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	2,197.98
CBK5: Loans_ 2008-2018	Customer ID Loan type Loans counts Total loan amount Principle paid Interest paid Outstanding balance	4,660.12

Name	Consists of (column heading)	Size (MB)
	from January until December for each year	
CBK6: Credit Cards	Customer ID The Credit Card Limit The Credit Card Outstanding Balance from January until December for each year	165.93
CBK7: Electronic banking transactions (ATM, Call centre, Mobile, Internet, POS) 2008-2018	Customer ID Year and month Monthly Withdrawals (counts & values) Monthly deposits (counts and values) Monthly transfers (counts and values) Monthly payments (counts and values)	631.49
CBK8: Customer information 2008-2018	Customer General information Customer Income Customer education history	2,500.32
Total raw data size for bank 6		165,869.88

A.2 Data conversion, number of observations and execution time

Table A-6-7: Bank 1 transferred data from CSV to MATLAB

Name	Consists of (column heading)	Format	Number of observations	Size (MB)
	Customer ID	Table	40,483,159	

Name	Consists of (column heading)		Format	Number of observations		Size (MB)
CBK1: Branch	Transaction Date Value of Transaction Currency of Transaction Card Type Transaction Type Year					50,851
CBK1: Cards (2009-2018)	Customer ID Transaction Date Value of Transaction Currency of Transaction Card Type Transaction Type Year					
CBK2: Customer Data, Customer General Information	A_Customer ID AGE GENDER ADDRESS Relationship Date Nationality		Table	295,801		249.45
CBK3: Income	Customer ID The Monthly income from January until December for each customer dates		Table	8,345,352		17.37
CBK4: Accounts	TABLE 1 Customer ID Dates	TABLE 2 Customer ID Dates	Structure (table 1: counts,	TABLE 1 12,552,552	TABLE 2 12,552,552	49.44

Name	Consists of (column heading)				Format	Number of observations				Size (MB)
	CA. counts DA counts JA counts SA counts		CA. balances DA balances JA balances SA balances		table 2: balance s					
CBK5: Loans	Table 1	Table 2	Table 3	Table 4	Structur e (Table 1: counts, Table 2: interest, Table 3: outstand ing, Table 4: principle)	Table 1	Table 2	Table 3	Table 4	33.5 7
	Custom er ID	Custo mer ID	Custo mer ID	Custo mer ID						
	Dates	Dates	Dates	Dates						
	Consu mer counts	Consu mer interes t	Consu mer outsta nd.	Consu mer principl e		2,639, 784	2,198, 508	2,639, 784	2,147, 100	
	Instalm ents counts	Instalm ent interes t	Instalm ent outsta nd.	Instalm ent principl e						
CBK6: Credit cards and Risk factor	Customer ID				Table	1,659,112				686. 36
	Dates									
	Credit card limit									
	Credit card outstanding balances									
	Risk score									
	Risk rating									
CBK7: Electronic channels	Table 1	Table 2	Table 3	Table 4	Structur e (Table 1: ATM, Table 2: call centre, Table 3:	Table 1: 735,356				11.8 5
	Table 5					Table 2: 63,792				
	For each table:					Table 3: 129,215				
	Customer ID					Table 4: 83,597				
	Year and month									

Name	Consists of (column heading)	Format	Number of observations	Size (MB)
	Monthly Withdrawals (counts & values) Monthly deposits (counts and values) Monthly transfers (counts and values) Monthly purchases (count and values) Monthly payments (counts and values)	internet, Table 4: mobile, Table 5: P.O.S)	Table 5: 751,168	
CBK8: Education Level	Customer ID Age Gender Address Education level	Table	236,927	295
Total MATLAB data				52,194

Table A-6-8: Bank 2 transferred data from CSV to MATLAB

Name	Consists of (column heading)	Format	Number of observations	Size (MB)
CBK1: Monthly Average Account Transactions	Customer ID Dates Number of Transactions Value of Transactions	Table	400,781,903	12,825
CBK2: Customer Data, Customer General Information	Customer ID Age Gender Address Relationship date	Table	875,191	1,146
CBK3: Income	Customer ID	Table	52,511,520	

Name	Consists of (column heading)				Format	Number of observations				Size (MB)
	The Monthly income from January until December for each customer Dates									65.15
CBK4: Accounts	Table 1 Customer ID Dates CA. counts		Table 2 Customer ID Dates CA. balances		Structur e (table 1: counts, table 2: balances)	Table 1 5,370,264		Table 2 5,370,264		19.15
	Table 1		Table 2			Table 1		Table 2		
	Customer ID		Customer ID			5,370,264		5,370,264		
	Dates		Dates							
CBK5: Loans	Table 1	Table 2	Table 3	Table 4	Structur e (table 1: counts, table 2: interest, table 3: outstanding, table 4: principle)	Table 1	Table 2	Table 3	Table 4	16.72
	Custom er ID	Custo mer ID	Custo mer ID	Custo mer ID						
	Dates	Dates	Dates	Dates						
	Consum er counts	Consu mer interest	Consu mer outstan d.	Consu mer principl e		1238088	1207296	1238088	1211628	
	Instalm ents counts	Instalm ent interest	Instalm ent outstan d.	Instalm ent principl e						
CBK6: Credit cards and Credit Card - Credit Score	Customer ID				Table	2,140,200				6.1
	Dates									
	Credit card limit									
	Credit card outstanding balances									
	Credit cards credit score									
CBK7: Electronic channels	Table 1	Table 2	Table 3	Table 4	Structur e (table 1: ATM, table 2: call	Table 1: 2,682,302				33.31
	Table 5									
	For each table:									
	Customer ID									

Name	Consists of (column heading)	Format	Number of observations	Size (MB)
	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)	centre, table 3: internet, table 4: mobile, table 5: P.O.S)	Table 2: 139,701 Table 3: 327,794 Table 4: 263,586 Table 5: 2,512,264	
CBK8: Customer Data	Customer ID Education level	Table	1,750,384	3.73
Total MATLAB Data				14,1 15

Table A-6-9: Bank 2 transferred data from CSV to MATLAB

Name	Consist of (column heading)	Format	Number of observations	Size (MB)
CBK1: Monthly Average Account Transactions	Customer ID Dates Number of Transactions Value of Transactions	Table	7,853,183	160
CBK2: Customer Data, Customer General Information	Customer ID Age Gender Address Date of relationship	Table	95,429	0.9 2
CBK3: Income	Customer ID The Monthly income from January until December for each customer Dates	Table	3,357,771	10. 26

Name	Consist of (column heading)				Format	Number of observations				Size (MB)
CBK4: Accounts	Table 1 Customer ID Dates CA. counts JA. Count	Table 2 Customer ID Dates CA. balances JA balances			Structure (table 1: counts, table 2: balances)	Table 1 4,103,174	Table 2 4,110,543			18.68
CBK5: Loans	Table 1 Customer ID Dates Consumer counts Instalments counts	Table 2 Customer ID Dates Consumer interest Instalment interest	Table 3 Customer ID Dates Consumer outstanding. Instalment outstanding	Table 4 Customer ID Dates Consumer principle Instalment principle	Structure (table 1: counts, table 2: interest, table 3: outstanding, table 4: principle)	Table 1 976,156	Table 2 882,911	Table 3 950,305	Table 4 875,343	13.63
CBK6: Credit cards	Customer ID Dates Credit card limit Credit card outstanding balances Credit Score				Table	1,213,261				2.46
CBK7: Electronic channels	Table 1 For each table: Customer ID	Table 2	Table 3	Table 4	Table 5 Structure (table 1: ATM, table 2: call)	Table 1: 1,431,388				7.45

Name	Consist of (column heading)	Format	Number of observations	Size (MB)
	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)	centre, table 3: internet, table 4: mobile, table 5: P.O.S)	Table 2: 33,048 Table 3: 120,989 Table 4: 2,017 Table 5: 101,190	
CBK8: Customer Data	Customer ID Nationality Address Gender	Table	240,993	1.23
Total MATLAB Data				214

Table A-6-10: Bank 3 transferred data from CSV to MATLAB

Name	Consists of (column heading)	Format	Number of observations	Size (MB)
CBK1: Monthly Average Account Transactions	Customer ID Dates Number of Transactions Value of Transactions	Table	3,796,794	21.02
CBK2: Customer Data, Customer General Information	Customer ID Age Gender Address Relationship date Education	Table	392,995	3.88

Name	Consists of (column heading)				Format	Number of observations				Size (MB)
CBK3: Income	Customer ID The Monthly income from January until December for each customer Dates				Table	4,950,996				24.2
CBK4: Accounts	Table 1 Customer ID Dates CA. counts	Table 2 Customer ID Dates CA. balances	Structure (table 1: counts, table 2: balances)		Table 1 15,201,378	Table 2 15,201,378				71.69
CBK5: Loans	Table 1 Customer ID Dates Consumer counts Instalment counts	Table 2 Customer ID Dates Consumer interest Instalment interest	Table 3 Customer ID Dates Consumer outstanding. Instalment outstanding.	Table 4 Customer ID Dates Consumer principle Instalment principle	Structure (table 1: counts, table 2: interest, table 3: outstanding, table 4: principle)	Table 1 2,878,089	Table 2 2,878,089	Table 3 2,878,089	Table 4 2,878,089	62.21
CBK6: Credit cards	Customer ID Dates Credit card limit Credit card outstanding balances				Table	1,079,603				3.48
CBK7: Electronic channels	Table 1 Table 5 For each table:	Table 2	Table 3	Table 4	Structure (Table 1: ATM,					250.56

Name	Consists of (column heading)	Format	Number of observations	Size (MB)
	Customer 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)	table 2: call centre, table 3: internet, table 4: mobile, table 5: P.O.S)	Table 1: 7,734,642 Table 2: 22,450 Table 3: 1,072,257 Table 4: 813,364 Table 5: 7,547,495	
CBK8: Customer Data	Customer ID Education level	Table	392,995	0.73
Total MATLAB Data				437.77

Table A-6-11: Bank 5 transferred data from CSV to MATLAB

Name	Consists of (column heading)	Format	Number of observations	Size (MB)
CBK1: Monthly Average Account Transactions	Customer ID Dates Number of Transactions Value of Transactions	Table	663,667	4.98
CBK2: Customer Data, Customer General Information	Customer ID Age Gender Address	Table	70,501	0.92
	Customer ID	Table	599,028	1.1

Name	Consists of (column heading)				Format	Number of observations				Size (MB)
CBK3: Income	The Monthly income from January until December for each customer Dates									
CBK4: Accounts	Table 1 Customer ID Dates CA. counts JA. Count	Table 2 Customer ID Dates CA. balances JA balances	Structure (table 1: counts, table 2: balance s)			Table 1 1,327,392	Table 2 1,327,392			5.06
CBK5: Loans	Table 1 Customer ID Dates Consumer counts Instalments counts	Table 2 Customer ID Dates Consumer interest Instalment interest	Table 3 Customer ID Dates Consumer outstanding Instalment outstanding.	Table 4 Customer ID Dates Consumer principle Instalment principle	Structure (table 1: counts, table 2: interest, table 3: outstanding, table 4: principle)	Table 1 2,146,620	Table 2 2,146,620	Table 3 2,146,620	Table 4 2,146,620	9.37
CBK6: Credit cards	Customer ID Dates Credit card limit Credit card outstanding balances				Table	279,600				0.62
CBK7: Electronic channels	Table 1 Table 5 For each table:	Table 2	Table 3	Table 4	Structure (Table 1: ATM,					9.13

Name	Consists of (column heading)	Format	Number of observations	Size (MB)
	Customer 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)	table 2: call centre, table 3: internet, table 4: mobile, table 5: P.O.S)	Table 1: 892,089 Table 2: 1,100,740 Table 3: 412,332 Table 4: 278,556 Table 5: 1,005,757	
CBK8: Customer Data	Customer ID Nationality Date of relationship	Table	70,501	28.98
Total MATLAB Data				55.18

Table A-6-12: Bank 6 transferred data from CSV to MATLAB

Name	Consist of (column heading)	Format	Number of observations	Size (MB)
CBK1: Monthly Average Account Transactions	Customer ID Dates Number of Transactions Value of Transactions	Table	74,632,245	515.37
CBK2: Customer Data, Customer General Information	Customer ID Age Gender Address Date of relationship Nationality	Table	1,382,942	9.13

Name	Consist of (column heading)				Format	Number of observations				Size (MB)
CBK3: Income	Customer ID The Monthly income from January until December for each customer Dates				Table	34,566,719				93.95
CBK4: Accounts	Table 1 Customer ID Dates CA. counts JA. Count	Table 2 Customer ID Dates CA. balances JA balances	Structure (table 1: counts, table 2: balances)			Table 1 106,701,077	Table 2 117,078,628			453.38
CBK5: Loans	Table 1 Customer ID Dates Consumer counts Instalment counts	Table 2 Customer ID Dates Consumer interest Instalment interest	Table 3 Customer ID Dates Consumer outstanding. Instalment outstanding.	Table 4 Customer ID Dates Consumer principle Instalment principle	Structure (table 1: counts, table 2: interest, table 3: outstanding, table 4: principle)	Table 1 3,024,481	Table 2 3,024,481	Table 3 3,024,481	Table 4 3,024,481	44.82
CBK6: Credit cards	Customer ID Dates Credit card limit Credit card outstanding balances Risk factor number				Table	2,393,127				7.42
	Table 1 Table 5	Table 2	Table 3	Table 4	Structure (Table					9.13

Name	Consist of (column heading)	Format	Number of observations	Size (MB)
CBK7: Electronic channels	For each table: Customer 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)	1: ATM, table 2: call centre, table 3: internet, table 4: mobile, table 5: P.O.S)	Table 1: 931,605 Table 2: 22,822 Table 3: 38,722 Table 4: 381,589 Table 5: 1,047,685	
CBK8: Customer Data	Customer ID Age Gender Nationality Date of relationship	Table	1,382,942	584.08
Total MATLAB Data				1,192.78

Table A-6-13: Banks' code execution times

Transfer from CVS to MATLAB (cleaning)	Bank 1	Bank 2	Bank 3	Bank 4	Bank 5	Bank 6
CBK1	5:23 hrs	3.40 hrs *	28 mins*	7 mins*	29 mins*	6.43 hrs*
CBK2		5 min	1 min	1 min	1 min	2 min
CBK3		48.48 hrs	1 min	2 mins	4 mins	17.58 hrs
CBK4	47.43 hrs	42 mins	2 mins	3 mins	10 mins	97.34 hrs
CBK5	15.20 hrs	10 mins	1 min	1 min	29 mins	1 min
CBK6	15 mins	6 mins	1 min	1 min	2 mins	2 mins
CBK7	2 mins	1 min	1 min	7 mins	42 mins	1 min
CBK8	4 mins	2 mins	1 min	2 mins	1 min	2 mins

*After running the code for two weeks and experiencing memory failures the failure of memory capacity for running and saving the table required, we agreed to gather the data into monthly periods rather than daily and do not save it. Bank 2 and 6 CBK1 is gathered on Monthly basis.