

Lingling Chen

Default Prediction of Unlisted Firms

CRANFIELD SCHOOL OF MANAGEMENT
MSc

Master of Science
Academic Year: 2024 - 2025

Supervisor: Dr. Vineet Agarwal
September 2025



CRANFIELD SCHOOL OF MANAGEMENT
Business Data Analytics

MSc

Academic Year 2024 - 2025

Lingling Chen

Default Prediction of Unlisted Firms

Supervisor: Dr. Vineet Agarwal
September 2025

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

© Cranfield University 2025. All rights reserved. No part of this publication may be reproduced without the written permission of the copyright owner.

Academic integrity declaration

I declare that:

- the thesis submitted has been written by me alone.
- the thesis submitted has not been previously submitted to this university or any other.
- that all content, including primary and/or secondary data, is true to the best of my knowledge.
- that all quotations and references have been duly acknowledged according to the requirements of academic research.

I understand that to knowingly submit work in violation of the above statement will be considered by examiners as academic misconduct.

Abstract

This study addresses the critical gap in default prediction for unlisted companies, which constitute over 99% of global enterprises yet remain systematically understudied compared to listed firms (OECD, 2023). Using the FAME database encompassing over 60,000 UK private companies from 2010-2024, the research systematically compares four machine learning approaches: Logistic Regression (LR), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and Neural Networks (NN) for bankruptcy prediction in extremely imbalanced datasets.

The methodology employs principal component analysis to identify seven core financial variables (ROA, AssetTurnover, Profit margin, Cash ratio, Current ratio, CashInterestCover, and Company age) spanning profitability, operational efficiency, and solvency dimensions. Time series cross-validation ensures robust evaluation across different economic cycles, including the COVID-19 pandemic period.

Key findings reveal XGBoost's superior performance, achieving 91.52% recall rate by successfully identifying 546 of 597 bankrupt companies whilst missing only 51. This significantly outperforms LR (84.77%), LightGBM (67%), and NN (12.4%). Critically, the study demonstrates that traditional ROC-AUC metrics prove misleading under extreme class imbalance (0.35%-0.91% default rates), with all models achieving 0.74-0.93 AUC values primarily through correct non-default classification rather than genuine bankruptcy identification capability. PR-AUC emerges as a more reliable evaluation metric.

The research provides practical guidance for financial institutions, recommending XGBoost for maximum risk coverage and logistic regression for regulatory compliance scenarios. NN demonstrate complete failure, highlighting deep learning limitations in extremely imbalanced sample environments. These findings contribute significantly to financial risk management theory and practice, offering evidence-based algorithm selection frameworks for credit risk assessment in the private company sector.

Keywords:

default prediction, unlisted companies, machine learning, XGBoost, logistic regression, neural networks, LightGBM, imbalanced data, ensemble learning

Acknowledgements

I sincerely thank everyone who has provided support and guidance throughout my master's degree.

First and foremost, I would like to express my special thanks to my thesis advisor, Dr.Vineet Agarwal, for his valuable insights, helpful feedback, and constant encouragement. He is incredibly responsible and monitored my progress throughout the entire process, providing immense support. The knowledge and inspiration of the faculty at Cranfield University have greatly enhanced my academic and professional development, helping me master data analysis methodologies and gain practical experience.

Second, I would like to express my special thanks to my family and friends for their unwavering support and understanding this year. I am also grateful to my fellow students for their friendship, knowledge sharing, and engaging conversations, which have greatly enhanced my learning.

Finally, I would like to thank the Cranfield School of Management for providing the tools and data necessary for this research.

Table of Contents

Academic integrity declaration i
Abstract 1
Acknowledgements 3
List of Figures 7
List of Tables 8
List of Equations 9
List of Abbreviations 10
1 Introduction 11
1.1 Background and Significance 11
1.2 Limitations of Existing Research and Research Gaps 12
1.3 Research Development Trends and Frontier Methods 14
1.4 Research Objectives, Thesis Summary and Structure 15
1.4.1 Objectives 15
1.4.2 Thesis Summary and Structure 15
2 Literature Review 17
2.1 Historical Development of Unlisted Company Default Prediction Models 17
2.1.1 Early Stage: Listed Company-Based Model Frameworks 17
2.1.2 Theoretical Exploration of Structural Models 17
2.1.3 Early Adaptation of Unlisted Enterprise Models 17
2.2 Evolution of Variable Selection 18
2.2.1 From Financial Ratios to Multi-Dimensional Variable Systems 18
2.2.2 Incorporation and Expansion of Non-Financial Variables 19
2.2.3 Integration of Macroeconomic Variables 19
2.3 Machine Learning Model Applications in Default Prediction 20
2.3.1 Paradigm Shift from Statistical Methods to Machine Learning 20
2.3.2 Continued Value of Logistic Regression Models 20
2.3.3 Breakthrough in Ensemble Learning Methods 21
2.3.4 Development of Neural Network Methods 22
2.4 Current Research Challenges and Gaps 22
2.4.1 Data Scarcity and Quality Issues 22
2.4.2 Limitations of Model Generalisation Capability 23
2.4.3 Challenges in Handling Imbalanced Data 23
2.4.4 Contradiction Between Interpretability and Regulatory Requirements 23
2.5 Research Opportunities and Development Directions 24
2.5.1 Optimisation and Integration of Machine Learning Algorithms 24
2.5.2 Development of Imbalanced Data Processing Techniques 24
2.5.3 Time Series Modelling and Dynamic Risk Assessment 24

2.5.4 Multi-Dimensional Feature Fusion and Feature Engineering Optimisation 25
2.6 Research Contributions and Positioning 25
3 Method 27
3.1 Data and Variables 27
3.1.1 Data Sources and Sample Selection 27
3.1.2 Default Definition 27
3.1.3 Variable System Construction 28
3.2 Data Preprocessing 31
3.2.1 Missing Value Treatment 31
3.2.2 Outlier Identification and Treatment 32
3.2.3 Multicollinearity 33
3.3 Exploratory Data Analysis and Feature Engineering 33
3.3.1 Default Data Characteristic Analysis 33
3.3.2 Feature Engineering 34
3.4 Model Building 36
3.4.1 Logistic Regression 36
3.4.2 XGBoost 36
3.4.3 LightGBM 37
3.4.4 Neural Networks 37
3.5 Model Evaluation 37
3.5.1 Validation Strategy 37
3.5.2 Evaluation Indicator System 37
4 Results 40
4.1 PCA Analysis Results 40
4.1.1 Variable Selection 40
4.2 Logistic Regression Model Results 41
4.2.1 Model Coefficient Analysis 41
4.2.2 Performance Description 43
4.2.3 Inter-model Comprehensive Comparison 45
4.3 XGBoost Model Results 46
4.3.1 Performance Description 46
4.3.2 Inter-model Comprehensive Comparison 49
4.4 LightGBM Model Results 50
4.4.1 Performance Description 50
4.4.2 Inter-model Comprehensive Comparison 54
4.5 Neural Network Model Results Analysis 55
4.5.1 Performance Description 55
4.5.2 Inter-model Comprehensive Comparison 57
5 Discussion 60
5.1 Theoretical Consistency and Novel Contributions 60
5.1.1 Alignment with Established Literature 60

5.1.2 Novel Empirical Contributions 60
5.2 Implications for Theory and Practice 61
5.2.1 Theoretical Implications 61
5.2.2 Practical Implications 62
5.3 Research Limitations 63
5.3.1 Data Sources and Geographic Constraints 63
5.3.2 Inherent Challenges of Extremely Imbalanced Data 63
5.3.3 Subjectivity in Feature Engineering and Variable Selection 63
5.3.4 Model Interpretability and Regulatory Applicability 63
5.4 Research Significance 64
5.4.1 Theoretical Contributions 64
5.4.2 Practical Value 64
5.4.3 Methodological Contributions 65
5.4.4 Policy and Industry Impact 65
6 Conclusion 67
6.1 Recommendations 67
6.1.1 For Financial Regulatory Authorities 67
6.1.2 For Large Financial Institutions 68
6.1.3 For Small and Medium Financial Institutions 68
6.1.4 Methodological Innovation 68
6.1.5 Data Integration and Extension 68
7 References 69
8 Appendices 75

List of Figures

Figure 4-1 ROC & PR of Logit Model 1 44
Figure 4-2 ROC & PR of Logit Model 2 45
Figure 4-3 ROC & PR of XGBoost Model 1 47
Figure 4-4 ROC & PR of XGBoost Model 2 48
Figure 4-5 ROC & PR of XGBoost Model 3 49
Figure 4-6 ROC & PR of LightGBM Model 1 52
Figure 4-7 ROC & PR of LightGBM Model 2 53
Figure 4-8 ROC & PR of LightGBM Model 3 54
Figure 4-9 ROC & PR of NN Model 1 56
Figure 4-10 ROC & PR of NN Model 2 57

List of Tables

Table 3-1 Independent variables 28
Table 3-2 Annual Bankrupt Count..... 34
Table 4-1 PCA Analysis 41
Table 4-2 Logit Model 1..... 42
Table 4-3 Logit Model 2..... 43
Table 4-4 Analysis of Logit Models..... 43
Table 4-5 Analysis of XGBoost Models 46
Table 4-6 Analysis of LightGBM Models..... 51
Table 4-7 Analysis of NN Models..... 55
Table 4-8 Analysis of All Model..... 58

List of Equations

$PY = 1X = 11 + e - (\beta 0 + \beta 1X1 + \cdots + \beta nXn)$ （3-1） 36

$L(\beta) = -logL(\beta) + \lambda[\alpha \parallel \beta \parallel 1 + (1 - \alpha) \parallel \beta \parallel 22]$ （3-2） 36

$L(\emptyset) = i = 1nl\ (y_{i}, \hat{y}_{i}) + k - 1k\Omega(fk)$ （3-3） 36

$TPR = TPTP + FN$ （3-4） 38

$FPR = FPFP + TN$ （3-5） 38

$Recall = TPTP + FP$ （3-6） 38

$F1 = 2 \times Precision \times RecallPrecision + Recall$ （3-7） 38

$Recall = TPTP + FN$ (3-8) 39

List of Abbreviations

SMEs	Small and Medium-Sized Enterprises
IMF	International Monetary Fund
ESG	Environmental, Social and Governance
LR	Logistic Regression
ML	Machine Learning
SVM	Support Vector Machines
NN	Neural Networks
XGBoost	Extreme Gradient Boosting
LightGBM	Light Gradient Boosting Machine
EDA	Exploratory Data Analysis
IQR	Interquartile Range
PCA	Principal Component Analysis

1 Introduction

1.1 Background and Significance

Against the backdrop of global economic integration and increasingly complex financial markets, corporate default prediction has emerged as a central issue in financial risk management. Corporate default prediction refers to the process of evaluating the probability that enterprises will be unable to fulfil their debt obligations on schedule, thereby quantifying the likelihood of corporate default risk. Through estimating default probabilities, financial institutions can identify high-risk enterprises in advance, price loans and bonds appropriately, and implement risk mitigation measures (Altman, 1968; Merton, 1974; Ohlson, 1980). This process proves crucial in providing decision-making foundations for risk mitigation by banks, investors, and creditors.

Whilst substantial research focuses on default prediction for listed companies (Altman, 1968; Shumway, 2001), unlisted enterprises, primarily small and medium-sized enterprises (SMEs), play a critical role in the global economy, yet remain systematically overlooked. According to OECD (2023) reports, SMEs account for over 99% of all global enterprises, contributing more than 60% of employment and over 50% of GDP. This demonstrates that SMEs constitute a vital component of most national economic structures, playing a pivotal role in employment creation, innovation development, and economic growth (Ciampi, 2015). Therefore, accurately predicting the default risk of unlisted enterprises proves not only crucial for financial institutions' credit decision-making but also holds significant importance for the stability of the entire financial system and sustainable economic development.

The 2007-2009 global financial crisis and the recent COVID-19 pandemic have inflicted tremendous shocks on SMEs, highlighting the urgent need for developing new SMEs default prediction models (Saha, 2021). Over the past decade, SME default prediction has evolved into a relevant research domain, exhibiting exponential growth across multiple disciplines, including finance, management, accounting, and statistics (Hosaka, 2019). This trend reflects the

high regard for research in this field from both academic and practical communities.

From a practical perspective, banks and financial institutions face increasingly stringent regulatory requirements. International regulatory frameworks such as Basel III require banks to adopt independent procedures when assessing SMEs credit value, making risks associated with small enterprise lending increasingly important (Basel Committee on Banking Supervision, 2017). Basel III represents a set of international regulatory measures formulated by the Basel Committee on Banking Supervision to address the 2007-09 financial crisis by strengthening banking regulation, supervision, and risk management. Within this regulatory environment, financial institutions urgently require more precise and reliable default prediction models to satisfy regulatory requirements and control credit risk. However, due to these enterprises' lack of publicly available market information, traditional market data-based default prediction models prove difficult to apply directly (Altman & Sabato, 2007). This challenge renders research into unlisted enterprise default prediction of theoretical value and practical significance.

Furthermore, corporate defaults exhibit evident systemic spillover effects. Research indicates that unlisted enterprise defaults not only generate losses for individual creditors but may also trigger chain reactions, leading to employment decline, regional economic contraction, and even supply chain disruption. For instance, the IMF (2021) noted that during COVID-19, SMEs experienced liquidity shocks significantly higher than large enterprises, with nearly 30% of unlisted enterprises facing bankruptcy risk without government intervention (OECD, 2020). This prompted governments worldwide to introduce temporary policies including credit guarantees, risk subsidies, and deferred repayments, whilst also exposing institutional gaps in the long-term absence of systematic credit risk assessment mechanisms.

1.2 Limitations of Existing Research and Research Gaps

Despite decades of development in corporate default prediction research, significant gaps remain in research focused on unlisted enterprises. Bharath &

Shumway (2008) point out that literature on corporate credit risk modelling for privately held firms is scarce. Although enterprises with unlisted equity or debt constitute a substantial proportion of the global corporate sector, research in this field has been hindered by the lack of public data. This situation highlights the importance and urgency of unlisted enterprise default prediction research.

Existing default prediction models can primarily be categorised into two types: structural models and reduced-form models. Structural models, represented by the Merton model, explicitly link corporate default to the variability of asset values (Merton, 1974). This model connects corporate bond pricing with default probability based on the Black-Scholes option pricing framework. Reduced-form models make explicit assumptions about the dynamics of default probability and recovery rates. However, most of these models have been developed based on listed enterprise data. The main advantage of the Merton model lies in using current market prices to determine default probability, but this advantage cannot be realised for unlisted enterprises.

The renowned Altman Z-Score model has served as an important tool for bankruptcy prediction since its publication in 1968, with the formula being applicable for determining the probability of enterprise bankruptcy within two years. However, current SMEs default prediction models based on accounting information possess a significant deficiency: they fail to consider the impact of market information on default prediction (Altman et al., 2017). This limitation leads to insufficient prediction accuracy, particularly in rapidly changing market environments.

Moreover, with the rise of big data and artificial intelligence technologies, traditional default prediction methods face challenges from digital and information transformation. Although the application of intelligent technologies in bankruptcy prediction shows a significant growth trend, bibliometric analysis of the past fifty years reveals important challenges currently facing research: how to effectively integrate traditional financial indicators with emerging data sources, how to utilise machine learning algorithms to improve prediction accuracy, and how to address the scarcity and incompleteness of unlisted enterprise data.

1.3 Research Development Trends and Frontier Methods

In recent years, default prediction research has exhibited several important development trends. Firstly, research methods have become increasingly diversified, expanding from traditional statistical methods to artificial intelligence technologies such as machine learning and deep learning. Secondly, data sources continue to expand beyond traditional financial data, and researchers have begun exploring the use of alternative data sources such as social media data, supply chain information, and macroeconomic data. Thirdly, model fusion has become an important trend, with researchers attempting to effectively integrate different types of models and data sources to improve prediction performance.

In the development history of default prediction models, the Altman Z-Score model, as a milestone contribution, has laid an important foundation for subsequent research. The global review study by Altman et al. (2017) demonstrates that the efficacy and importance of this model have been validated globally and widely applied in finance and related fields. However, this model is primarily designed for listed enterprises, and its applicability to unlisted enterprises requires further examination.

In research on default prediction for unlisted enterprises, some scholars have proposed innovative solutions. For example, improving default prediction models for unlisted enterprises by utilising market information from comparable listed enterprises. This hybrid approach attempts to integrate market information based on existing accounting information to improve prediction accuracy (Altman & Sabato, 2007). However, research in this field remains in its infancy and requires more theoretical exploration and empirical validation.

Concurrently, alternative information sources such as Environmental, Social and Governance (ESG) indicators, supply chain relationship data, and enterprise network structures have been incorporated into default prediction frameworks (Chava, 2014; Hoepner et al., 2016; Jacobson & von Schedvin, 2015). The application of Generative AI in text analysis, such as extracting enterprise credit

clues from news, court judgements, and contract texts, has also become a research hotspot in recent years (Loughran & McDonald, 2011).

1.4 Research Objectives, Thesis Summary and Structure

1.4.1 Objectives

Given the above research status and existing gaps, this study's overarching objective is enhancing the accuracy and practicality of private company default prediction to provide more effective predictive tools and decision support for financial risk management.

Specifically, this research aims to:

Address theoretical gaps in private company default prediction research. Targeting the existing literature's primary focus on listed companies whilst neglecting private enterprises, this study constructs default prediction theoretical frameworks suitable for private company characteristics, enriching the theoretical foundation of SME credit risk assessment.

Provide financial institutions with scientifically reliable risk identification tools. Through development of high-precision default prediction models, this research assists banks, investment institutions, and other financial entities in more accurately identifying potential default risks, reducing credit losses, and enhancing risk management capabilities.

Provide empirical evidence for regulatory policy formulation. Through large-scale empirical analysis, this study offers scientific foundations for regulatory authorities formulating private company risk assessment standards, capital adequacy requirements, and related policies, promoting stable financial system development.

1.4.2 Thesis Summary and Structure

This study utilises the FAME database, encompassing financial and governance information for over 60,000 UK private companies from 2010 to 2024. Research findings demonstrate the XGBoost algorithm's optimal performance in default

prediction, achieving 91.52% test set recall and successfully identifying 546 of 597 bankrupt companies whilst missing only 51, significantly outperforming other models' recall rates. This finding validates the ensemble learning method's advantages in processing extremely imbalanced financial data. However, traditional ROC-AUC indicators exhibit serious misleading characteristics under extremely imbalanced scenarios, with all model AUC values ranging between 0.74 and 0.93, primarily derived from the correct classification of numerous non-defaulting companies rather than genuine default identification capability. In contrast, PR-AUC more accurately reflects model performance, with XGBoost achieving 0.042, far exceeding other algorithms. Neural network models demonstrate complete failure, with original data models achieving 0% recall rate and merely 12.4% even when employing down sampling strategies, revealing deep learning limitations under extremely imbalanced sample scenarios.

This thesis structure comprises: Chapter 1 presents the introduction, elaborating research background, significance, and objectives; Chapter 2 provides literature review, systematically reviewing historical development of private company default prediction models, analysing variable selection evolution, and evaluating theoretical foundations and application status of four mainstream machine learning methods; Chapter 3 introduces research methodology, including data sources, variable system construction, data preprocessing, feature engineering, and model construction strategies; Chapter 4 presents empirical analysis results, covering PCA analysis, variable selection, and four model performance evaluations; Chapter 5 discusses research significance and limitations; Chapter 6 summarises research conclusions and presents recommendations.

2 Literature Review

2.1 Historical Development of Unlisted Company Default Prediction Models

2.1.1 Early Stage: Listed Company-Based Model Frameworks

Corporate default prediction aims to assess the probability that enterprises will be unable to fulfil their debt obligations in the future, serving as a core instrument in financial risk management. Since the 1960s, researchers have developed a series of modelling approaches to address this issue, evolving from early discriminant analysis and financial ratio models to structural models and logistic regression (LR), and more recently to machine learning and artificial intelligence methods, thus forming a relatively systematic theoretical framework. This chapter briefly reviews the evolutionary path of mainstream models, with particular focus on their application and adaptation issues in private enterprises, especially SMEs.

2.1.2 Theoretical Exploration of Structural Models

Merton (1974) proposed a structural default prediction model based on Black-Scholes option pricing theory, treating corporate equity as a call option with corporate assets as the underlying and debt as the strike price. This model estimated default probability through corporate asset value and its volatility, providing rigorous theoretical foundations for default prediction. Black and Cox (1976) further extended the Merton model by allowing defaults before debt maturity, enhancing the model's practical applicability.

Although structural models possess theoretical advantages, their dependence on corporate market value, asset volatility, and other market data makes them primarily applicable to listed companies. For unlisted enterprises lacking public market pricing, structural model applications face significant challenges, prompting researchers to explore accounting information-based alternatives.

2.1.3 Early Adaptation of Unlisted Enterprise Models

Targeting unlisted enterprise data characteristics, researchers began adjusting and improving traditional models. Altman and Sabato's (2007) research

marked an important breakthrough in unlisted enterprise default prediction. They adjusted the classic Z-Score model for American SMEs, substituting book values for market values to construct a Z-Score model suitable for SMEs. This model employed four ratios: working capital/total assets, retained earnings/total assets, earnings before interest and taxes/total assets, and book value of equity/total liabilities, avoiding dependence on market data. Empirical results demonstrated that the adjusted model achieved 86% prediction accuracy for SMEs, proving the effectiveness of model adaptation.

Ciampi and Gordini (2013) further extended this research pathway by not only adjusting financial ratios but also incorporating non-financial variables such as enterprise age, size, and industry. Their research found that cash flow ratios exhibited particularly prominent predictive capability for Italian SMEs, potentially reflecting SMEs' greater dependence on internal cash flows. This finding emphasised the importance of considering cash flow management capability in unlisted enterprise default prediction.

2.2 Evolution of Variable Selection

2.2.1 From Financial Ratios to Multi-Dimensional Variable Systems

Early default prediction research relied primarily on financial ratios, reflecting enterprise profitability, solvency, operational efficiency, and growth capacity. Du Jardin (2015), based on research by Altman (1968), Ohlson (1980), and Shumway (2001), constructed a comprehensive financial variable framework dividing financial indicators into four dimensions: solvency indicators (current ratio, quick ratio, debt-to-asset ratio, interest coverage ratio), profitability indicators (ROA, net profit margin, EBITDA, operating margin), operational efficiency indicators (total asset turnover, accounts receivable turnover, fixed asset turnover), and growth capability indicators (revenue growth rate, total asset growth rate).

However, limitations of relying solely on financial data gradually emerged. Financial statements exhibit lag characteristics, struggling to reflect timely changes in enterprise risk conditions. Simultaneously, unlisted enterprises'

financial data quality varies considerably, with some enterprises lacking standardised financial disclosure. These issues prompted researchers to explore the predictive value of non-financial variables.

2.2.2 Incorporation and Expansion of Non-Financial Variables

Altman et al. (2010) proposed that SMESS default risk models should integrate "soft information", including enterprise age, management experience, industry characteristics, and banking relationships. Enterprise age is considered an important risk indicator, with young enterprises typically lacking market experience and capital accumulation, facing higher default risk.

Research by Balcaen and Ooghe (2006) found that variables such as entrepreneur education level, industry lifecycle, and banking relationship duration significantly influence default probability. These findings indicate that enterprise governance structure, external relationships, and environmental context exert important impacts on default risk, with pure financial indicators unable to completely capture these risk factors.

2.2.3 Integration of Macroeconomic Variables

Economic cycle impacts on enterprise default risk have received increasing attention. Shumway (2001) further proposed time-weighted logit models considering the dynamic evolution of enterprise risk, becoming important foundations for subsequent research. This model considers not only enterprise-specific characteristics but also incorporates macroeconomic indicators such as GDP growth rate, unemployment rate, and inflation rate.

Andreeva et al. (2016) discovered that macroeconomic background, regional credit culture, and industry structure exert significant influence on SMES default rates. Rising unemployment rates typically accompany economic recession, leading to market demand contraction, deteriorating enterprise operating environments, and consequently increased default risk. Industrial Production Index (IPI) and GDP indicators reflect overall economic vitality, exerting systematic impacts on enterprise default risk.

2.3 Machine Learning Model Applications in Default Prediction

2.3.1 Paradigm Shift from Statistical Methods to Machine Learning

Since the 1990s, default prediction research has experienced a major transformation from traditional statistical methods to machine learning techniques. Growth in computational capability and data resources has provided conditions for complex algorithm applications, with machine learning methods better able to handle high-dimensional, non-linear, and imbalanced data characteristics suitable for complex prediction tasks.

Comparative research by Barboza et al. (2017) demonstrated that machine learning models, particularly random forests and support vector machines, generally outperformed traditional statistical models in prediction accuracy. This research utilised American listed enterprise data containing financial and non-financial variables, finding that machine learning model AUC values generally exceeded 0.90, whilst traditional logistic regression model AUC values approximated 0.75. This finding provided strong support for machine learning applications in default prediction.

2.3.2 Continued Value of Logistic Regression Models

Despite the rise of machine learning methods, logistic regression maintains an important position in financial risk modelling due to its excellent interpretability. Logistic regression, based on linear discriminant analysis theory and utilising maximum likelihood estimation for parameter derivation, enables model coefficients to directly reflect marginal impacts of variables on default probability. Ohlson's (1980) pioneering work established the foundations for logistic regression application in default prediction, with this method possessing excellent interpretability advantages in regulatory environments.

To prevent overfitting, modern logistic regression models commonly employ L1 (Lasso) and L2 (Ridge) regularisation techniques. The Elastic Net method proposed by Zou and Hastie (2005) combines the advantages of L1 and L2 regularisation, achieving a balance between feature selection and model stability.

These improvements enable logistic regression to maintain good performance when handling high-dimensional data.

2.3.3 Breakthrough in Ensemble Learning Methods

Advantages of Extreme Gradient Boosting Models

Extreme Gradient Boosting (XGBoost), based on second-order Taylor expansion gradient boosting algorithms, controls model complexity through regularisation terms. The XGBoost algorithm proposed by Chen and Guestrin (2016) demonstrated excellence across multiple machine learning competitions, exhibiting significant advantages in default prediction domains.

Moscatelli et al. (2020) utilised Italian SMEs loan data to construct XGBoost models, finding superior performance in identifying high-risk enterprises with prediction accuracy exceeding Logit and Random Forest models. XGBoost proves particularly suitable for handling heterogeneous data and non-linear relationships whilst providing variable importance rankings, enhancing model interpretability. This algorithm possesses built-in mechanisms for handling imbalanced data, proving valuable for minority class identification tasks such as default prediction.

Efficiency Advantages of Light Gradient Boosting Machine

Light Gradient Boosting Machine (LightGBM), based on gradient boosting decision tree frameworks, combines histogram algorithms with gradient-based one-sided sampling techniques. The LightGBM algorithm proposed by Ke et al. (2017) significantly improved training efficiency whilst maintaining prediction accuracy. This algorithm adopts leaf-wise growth strategies, replacing traditional level-wise growth, enabling more effective large-scale data processing.

In default prediction applications, LightGBM possesses powerful capabilities for handling high-dimensional heterogeneous data, built-in solutions for class imbalance problems, high computational efficiency suitable for large-scale data processing, and intuitive feature importance evaluation functionality. These

characteristics provide application advantages in real-time risk assessment systems.

2.3.4 Development of Neural Network Methods

Neural network model (NN) applications in default prediction can be traced to the 1990s. Tam and Kiang (1992) demonstrated neural network advantages in handling non-linear relationships, with this research employing multilayer perceptron models for bank bankruptcy prediction, achieving results superior to traditional statistical methods.

Modern NN models adopt multilayer feedforward structures containing input layers, multiple hidden layers, and output layers. Through ReLU activation functions and dropout regularisation techniques preventing overfitting, utilising Adam optimisation algorithms for parameter updates, deep learning methods can automatically learn complex interactive relationships between features without requiring manual feature engineering, providing advantages when processing high-dimensional financial data.

However, neural network models also face challenges. Model complexity remains high, requiring substantial data and computational resources for training. "Black box" characteristics limit model interpretability, restricting applications in regulatory environments. Sensitivity to hyperparameter settings necessitates careful tuning.

2.4 Current Research Challenges and Gaps

2.4.1 Data Scarcity and Quality Issues

Core challenges facing unlisted enterprise default prediction lie in data scarcity and quality issues. Bharath and Shumway (2008) noted the scarcity of corporate credit risk modelling literature targeting privately held enterprises, primarily due to the lack of publicly available data. Many unlisted enterprises do not publish audited financial statements, with credit rating systems and industry analyst coverage primarily targeting large enterprises, leading to extremely limited data and historical records related to private enterprise defaults.

This data scarcity manifests not only in data availability but also in data quality. Unlisted enterprises' financial information disclosure levels are typically low, lacking standardised reporting formats and audit procedures, leading to problems with data reliability and consistency.

2.4.2 Limitations of Model Generalisation Capability

Most existing default prediction models are developed based on data from specific periods and regions, potentially experiencing performance degradation when facing economic and environmental changes. During the 2008 financial crisis and the COVID-19 pandemic, traditional models based on historical data performed poorly, exposing model limitations when confronting structural shocks.

OECD (2020) data indicated that during the pandemic, over 30% of SMEs faced closure risks, yet traditional financial indicator models failed to accurately predict such systemic risks. This demonstrates the need for developing more adaptive model frameworks capable of rapidly responding to macroeconomic environmental changes.

2.4.3 Challenges in Handling Imbalanced Data

Default prediction essentially represents an extremely imbalanced classification problem. Traditional machine learning algorithms, when processing imbalanced data, tend to favour majority classes, leading to insufficient identification capability for minority classes (defaulting enterprises).

Although researchers have proposed various methods for handling imbalanced data, including resampling techniques, cost-sensitive learning, and ensemble methods, practical applications still face unstable effectiveness issues. Balancing improved default identification rates whilst controlling false positive rates remains a significant challenge in the field.

2.4.4 Contradiction Between Interpretability and Regulatory Requirements

Modern machine learning methods, whilst exhibiting excellent performance in prediction accuracy, possess "black box" characteristics that contradict financial

regulatory requirements for model transparency. International regulatory frameworks such as Basel III require banks to employ interpretable models for risk assessment, limiting complex machine learning method applications in practice.

Balancing maintained prediction accuracy with enhanced model interpretability has become an important current research direction. The development of interpretability analysis methods such as SHAP (Shapley Additive Explanations) provides new approaches for addressing this challenge.

2.5 Research Opportunities and Development Directions

2.5.1 Optimisation and Integration of Machine Learning Algorithms

Research by Barboza et al. (2017) demonstrated that ensemble learning algorithms generally outperformed traditional methods in predicting bankruptcy, default events, and credit risk management. Lessmann et al. (2015) found that ensemble models based on heterogeneous algorithms outperformed individual classifiers and homogeneous algorithm ensemble models in credit scoring prediction. This indicates that hybrid methods combining different algorithmic advantages possess tremendous potential, such as combining logistic regression interpretability with XGBoost prediction accuracy.

2.5.2 Development of Imbalanced Data Processing Techniques

The extreme imbalance problems facing default prediction require effective solutions. Sun et al. (2014) combined SMOTE with Bagging ensemble learning algorithms, whilst Kim et al. (2015) proposed geometric mean-based boosting algorithms with oversampling techniques, both achieving significant effects in handling imbalanced financial data. The combined use of cost-sensitive learning, sampling techniques, threshold optimisation methods, and dynamic threshold adjustment techniques possesses important application prospects.

2.5.3 Time Series Modelling and Dynamic Risk Assessment

Traditional default prediction models are primarily based on static data, yet enterprise risk exhibits evident temporal dynamic characteristics. Narvekar and

Guha (2021) applied XGBoost to American listed company data, successfully predicting bankruptcy rate changes during the COVID-19 recession periods. Expanding window validation, online learning, concept drift detection, and other dynamic risk assessment techniques help enhance model adaptability under different economic environments, particularly during structural shocks such as the COVID-19 pandemic, where adaptive model frameworks prove especially important.

2.5.4 Multi-Dimensional Feature Fusion and Feature Engineering Optimisation

Effective fusion of financial indicators, non-financial variables, and macroeconomic indicators maintains room for improvement. Xia et al. (2017) proposed model feature selection techniques based on feature correlation scoring, achieving significant advances in credit scoring domains under Bayesian hyperparameter optimisation with XGBoost. The combination of dimensionality reduction techniques, such as principal component analysis and factor analysis, with machine learning methods not only helps address high-dimensional data "curse of dimensionality" problems but also enables identification of potential risk factor structures. Such structured feature representations help enhance model interpretability and prediction stability.

2.6 Research Contributions and Positioning

Based on the above literature review, this research aims to fill gaps in existing research across several aspects:

Firstly, constructing comprehensive prediction frameworks that integrate financial ratios, non-financial variables, and macroeconomic indicators. Through principal component analysis, identifying key risk dimensions combined with correlation analysis screening core predictive variables, constructing complete risk assessment systems encompassing enterprise internal financial conditions and external environmental factors.

Secondly, systematically comparing the performance of four mainstream modelling methods—logistic regression, XGBoost, LightGBM, and neural

networks—in unlisted enterprise default prediction. Through unified data foundations and evaluation frameworks, different algorithms' advantages, disadvantages, and applicable scenarios are analysed deeply.

Thirdly, adopting targeted model optimisation strategies for extremely imbalanced data characteristics. Through cost-sensitive learning, class weight adjustment, threshold optimisation, and other methods, it will enhance model capability in identifying defaulting enterprises.

Finally, establishing time series validation frameworks to evaluate model stability and adaptability under different economic environments. Through five-year rolling window validation, it will analyse temporal variation characteristics of model performance to provide references for practical applications.

Through these contributions, this research will provide valuable insights and empirical evidence for theoretical development and practical applications in unlisted enterprise default prediction.

3 Method

This research employs systematic machine learning approaches to construct unlisted company default prediction models, encompassing key stages including data collection and variable selection, data preprocessing, exploratory data analysis (EDA), model development, and model evaluation.

3.1 Data and Variables

3.1.1 Data Sources and Sample Selection

The Research data originates from the FAME database, covering British private company financial and governance information from 2010 to 2024. This timeframe spans multiple complete economic cycles, including post-financial crisis recovery periods, European debt crisis, Brexit processes, and the COVID-19 pandemic, ensuring models can capture enterprise default risk characteristics under different macroeconomic environments.

Sample selection criteria include inclusion criteria encompass private companies registered in the United Kingdom; exclusion criteria include financial and insurance companies and all listed companies, ensuring research focus concentrates on private enterprise default characteristics.

3.1.2 Default Definition

The default definition is based on the legal framework of the UK Insolvency Act 1986 and subsequent amendments, adopting objective legal procedural standards. Specific provisions include entering administration procedures (administrator appointments, creditor meetings), liquidation procedures (court winding-up petitions, winding-up orders, liquidator appointments), creditor voluntary arrangements (liquidator appointments, creditor meetings), or receivership procedures (administrative receiver appointments, creditor meetings). This definition's advantages lie in its objectivity, verifiability, and cross-research comparability.

3.1.3 Variable System Construction

The variable system is based on multiple theoretical frameworks, integrating traditional financial indicators, corporate governance characteristics, and macroeconomic variables. Financial variables are based on research by Altman (1968), Ohlson (1980), and Shumway (2001). Each variable represents different aspects of corporate financial health. The framework encompasses four dimensions: 1) solvency indicators (current ratio, quick ratio, debt-to-asset ratio, interest coverage ratio); 2) profitability indicators (return on assets, net profit margin, EBITDA, operating margin); 3) operational efficiency indicators (total asset turnover, accounts receivable turnover, fixed asset turnover); 4) growth capability indicators (revenue growth rate, total asset growth rate). Non-financial variables are based on Altman et al.'s (2010) research framework on SME credit risk, including company age, mortgage status, and county court judgement records.

The variables used are presented in Table 3-1.

Table 3-1 Independent variables

Sort	Variables	Relationship to Bankruptcy
Financial Variables	Total Assets per employee GBP	Higher → better efficiency and profitability → reduces bankruptcy risk
	Working Capital per employee GBP	Higher → better cash turnover → reduces risk
	Total Assets th GBP	Higher → reduces bankruptcy risk
	Interest Cover (x)	Higher → stronger ability to pay interest → reduces risk
	Gearing	Higher → higher financial leverage and debt pressure → increases risk

	Net Assets Turnover (x)	Higher → better asset utilisation → reduces risk
	Fixed Assets Turnover (x)	Higher → more efficient use of fixed assets → reduces risk
	Debtors Turnover (x)	Higher → faster collection of receivables → reduces risk
	Current ratio (x)	Higher → stronger liquidity → reduces risk
	Liquidity ratio (x)	Higher → reduces bankruptcy risk
	Profit margin	Higher → stronger profitability → reduces risk
	Turnover%	Higher → faster revenue growth → reduces risk
	Total Assets%	Higher → rapid expansion (may increase risk if excessive) → context-dependent
	DirectorsRemuneration_TotalAssets	Excessive → governance issues, potential resource misuse → increases risk
	TotalLiabilities_TotalAssets	Higher → higher debt ratio → increases risk
	CurrentLiabilities_TotalAssets	Higher → higher short-term debt pressure → increases risk
	EBITDA_TotalAssets	Higher → stronger earnings capacity → reduces risk

	CFO_TotalAssets	Higher → more stable operating cash flow → reduces risk
	ROA	Higher → stronger return on assets → reduces risk
	RetainedProfit_TotalAssets	Higher → reduces bankruptcy risk
	WorkingCapital_TotalAssets	Higher → stronger short-term solvency → reduces risk
	NetIncome_TotalLiabilities	Higher → better ability to cover debt with profit → reduces risk
	CashInterestCover	Higher → cash can cover interest → reduces risk
	CashRatio	Higher → stronger short-term solvency → reduces risk
	WC2Turnover	Higher → more efficient use of working capital → reduces risk
	OperatingMargin	Higher → stronger operating profitability → reduces risk
	AssetTurnover	Higher → stronger asset utilisation → reduces risk
	CATurnover	Higher → stronger current asset turnover → reduces risk
	Gross_Turnover	Higher → reduces bankruptcy risk

	Has_Dividend	Paying dividends → stronger profitability and cash flow → reduces risk
Non-Financial Variables	Company age	Older firms → usually more stable (though possibly less innovative) → reduces risk
	Industry Code	
	Qualified / Unqualified Accounts	Higher → better economic growth → reduces risk
Macro Variables	Industrial Production Index (IPI)	Higher → weaker demand, lower revenues → increases risk
	IPI Growth	Higher → higher costs and margin pressure → increases risk
	GDP	Higher → better efficiency and profitability → reduces bankruptcy risk
	Unemployment Rate	Higher → better cash turnover → reduces risk
	Inflation	Higher → heavier financial burden → increases risk

3.2 Data Preprocessing

3.2.1 Missing Value Treatment

A three-tier progressive missing value treatment strategy was adopted:

First tier: Company-level screening. Calculating overall missing rates for each company, removing companies with missing rates exceeding 50%, ensuring analytical samples possess sufficient information completeness.

Second tier: Key variable completeness verification. Targeting core predictive variables (company age, debt-to-asset ratio, interest coverage ratio, current ratio, net profit, operating profit, shareholders' funds, EBITDA, profit margin, total assets, turnover), removing companies with complete absence of these variables during 2010-2024, ensuring basic data quality for model training.

Third tier: Time series completeness requirements. Calculating effective observation years for each company during 2010-2024, removing companies with fewer than three observation years, and ensuring samples possess sufficient time series information for dynamic risk assessment.

3.2.2 Outlier Identification and Treatment

Outlier identification employed interquartile range (IQR) methods, using 1.5 times the IQR as identification thresholds. Based on numerical variable outlier detection results, significant differences in outlier proportions were discovered, ranging from 2.54% for creditor payment periods to 18.85% for net current assets. Variables were categorised into three priority levels based on outlier proportions:

High-priority variables (outlier proportion >15%): net current assets (18.85%), retained profits (16.53%), net assets (14.81%), shareholders' funds (14.80%)

Medium-priority variables (outlier proportion 10-15%): net profit (14.23%), profit before tax (13.75%), total assets (13.13%), operating profit (11.96%)

Low-priority variables (outlier proportion <10%): traditional financial ratios and per-employee indicators

Outlier treatment strategies adopted differentiated methods based on statistical characteristics:

1) Winsorisation treatment: For high-outlier variables such as net current assets, adopting 5% and 95% percentile truncation, compressing extreme values to reasonable intervals such as (-50,000, 30,000)

2) Logarithmic transformation: For scale variables such as total assets, current assets, and current liabilities, performing log(1+x) transformation after Winsorisation to improve normality
3) Direct truncation: For ratio indicators, setting reasonable intervals based on business logic, such as truncating profit margins to (-40%, 60%)
4) Left truncation treatment: Setting left truncation points for variables requiring logarithmic transformation to avoid negative infinity values

3.2.3 Multicollinearity

Targeting linear assumption requirements of logistic regression models, multi-dimensional methods were adopted for multicollinearity diagnosis:

Correlation analysis: Through Pearson correlation coefficient matrices and heatmap visualisation, identify highly correlated variable pairs. Structural high correlations ($|r|>0.8$) were discovered between current ratio and liquidity ratio, DirectorsRemuneration_TotalAssets and Total Assets.

Principal Component Analysis (PCA): Through PCA analysis of financial ratio latent factor structures, identifying four primary dimensions—profitability, solvency, operational efficiency, and growth—providing theoretical foundations for variable screening.

3.3 Exploratory Data Analysis and Feature Engineering

3.3.1 Default Data Characteristic Analysis

Temporal trend analysis (Table 3-2): Training set (2010-2019) default rates remained at relatively low levels (0.0%-0.08%), whilst test set (2020-2024) default rates increased sharply, particularly reaching 0.42% in 2024, reflecting recent economic and environmental deterioration impacts on enterprise default risk.

Sample distribution characteristics: The dataset contains a training set of 66,058 companies (230 defaults, 0.35% default rate) and test set of 65,828 companies (597 defaults, 0.91% default rate), exhibiting typical extremely imbalanced distribution characteristics.

Table 3-2 Annual Bankrupt Count

Year	Total Observations	Bankrupt Count
2010	66058	13
2011	66045	15
2012	66030	15
2013	66015	12
2014	66003	18
2015	65985	37
2016	65948	25
2017	65923	18
2018	65905	30
2019	65875	47
2020	65828	44
2021	65784	36
2022	65748	68
2023	65680	175
2024	65505	274

3.3.2 Feature Engineering

Based on empirical analysis from seminal credit risk literature (Altman, 1968; Ohlson, 1980; Beaver, 1966), 18 engineering variables with strong business interpretability were created:

Leverage structure indicators:

- TotalLiabilities_TotalAssets: total debt ratio, measuring overall enterprise leverage levels
- CurrentLiabilities_TotalAssets: current liability ratio, reflecting short-term solvency pressure

Profitability quality indicators:

- ROA: return on assets, comprehensively reflecting asset profitability efficiency.
- OperatingMargin: operating margin, measuring core business profitability.
- EBITDA_TotalAssets: EBITDA asset ratio, evaluating cash generation capability.
- Gross_Turnover: Gross profit margin, measuring the profit ratio of a company's sales revenue after deducting direct costs, reflecting the company's basic profitability and cost control efficiency.

Cash flow management indicators:

- CFO_TotalAssets: operating cash flow to total assets ratio, reflecting cash generation efficiency.
- CashInterestCover: cash interest coverage ratio, measuring cash debt service capability.
- CashRatio: cash ratio, evaluating immediate liquidity.

Operational efficiency indicators:

- AssetTurnover: total asset turnover, measuring asset operational efficiency.
- CATurnover: current asset turnover, reflecting short-term asset management levels.
- WC2Turnover: working capital turnover, evaluating working capital efficiency.

Governance quality indicators:

- DirectorsRemuneration_TotalAssets: director remuneration to total assets ratio, reflecting governance cost control.

- Has_Dividend: dividend payment capability, measuring financial health.

These engineering variables not only enhance model predictive capability but, more importantly provide deeper financial risk insights.

3.4 Model Building

The dataset is divided in an 8:2 ratio, with 2010-2019 data as the training set and 2020-2024 data as the test set.

3.4.1 Logistic Regression

Logistic regression is based on linear discriminant analysis theory (Fisher, 1936), obtaining parameters through maximum likelihood estimation. The functional form of logistic regression is

$$\boldsymbol{P(Y=1|X)=\frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\cdots+\beta_n X_n)}}} \qquad (3\text{-}1)$$

, chosen for its excellent interpretability, where model coefficients reflect the marginal impact of variables on default probability, high computational efficiency and good convergence properties, making it widely adopted in financial risk modelling (Ohlson, 1980). To prevent overfitting, this research employs L1 and L2 regularisation techniques (Zou & Hastie, 2005), with the loss function:

$$\boldsymbol{L(\beta)=-logL(\beta)+\lambda[\alpha \parallel \beta \parallel_1+(1-\alpha) \parallel \beta \parallel_2^2]} \qquad (3\text{-}2)$$

, where α controls the weight of L1 and L2 regularisation, and λ controls the regularisation strength.

3.4.2 XGBoost

XGBoost model is based on second-order Taylor expansion gradient boosting algorithms (Chen & Guestrin, 2016), controlling model complexity through regularisation terms. Its objective function is

$$\boldsymbol{L(\emptyset)=\sum_{i=1}^{n} l\,(y_i,\hat{y}_i)+\sum_{k-1}^{k}\Omega(fk)} \qquad (3\text{-}3)$$

, where l is the loss function and Ω is the regularisation term. XGBoost proves particularly suitable for handling heterogeneous data and non-linear relationships,

exhibiting excellent performance in credit scoring domains with extensive application records in default prediction.

3.4.3 LightGBM

T he LightGBM model is based on gradient boosting decision tree frameworks (Friedman, 2001), combining histogram algorithms with gradient-based one-sided sampling techniques (Ke et al., 2017). It possesses powerful capabilities for handling high-dimensional heterogeneous data, built-in solutions for class imbalance problems, high computational efficiency suitable for large-scale data processing, and intuitive feature importance evaluation functionality. Key parameter settings include num_leaves for controlling model complexity, min_data_in_leaf for preventing overfitting, and feature_fraction and bagging_fraction for controlling feature and sample random sampling proportions, respectively.

3.4.4 Neural Networks

Multi-layer feedforward neural networks is constructed, containing input layers, multiple hidden layers, and output layers (Goodfellow et al., 2016). ReLU activation functions and dropout regularisation techniques are adopted to prevent overfitting, utilising Adam optimisation algorithms for parameter updates.

3.5 Model Evaluation

3.5.1 Validation Strategy

Considering data temporal characteristics, time series cross-validation strategies are adopted to avoid future information leakage. The training set is divided into 5 temporal folds using expanding window modes: training data for each fold gradually expanded, with validation data for fixed one-year periods, strictly ensuring training data temporally preceded validation data.

3.5.2 Evaluation Indicator System

The evaluation indicator system encompasses multiple dimensions to comprehensively measure model performance.

The evaluation indicator system encompasses multiple dimensions to comprehensively measure model performance.

Basic Definitions:

- TP (True Positive): correctly predicted default cases.
- TN (True Negative): correctly predicted non-default cases.
- FP (False Positive): incorrectly predicted as default (Type I error).
- FN (False Negative): incorrectly predicted as non-default (Type II error).

Key Performance Metrics:

AUC-ROC serves as the primary indicator measuring model classification performance across all possible thresholds, gauging the model's ability to distinguish between default and non-default cases with good robustness to class imbalance. AUC-ROC is calculated as the area under the ROC curve, where:

ROC=TPR vs FPR,

$$\boldsymbol{TPR} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \quad (3\text{-}4)$$

$$\boldsymbol{FPR} = \frac{\boldsymbol{FP}}{\boldsymbol{FP+TN}} \quad (3\text{-}5)$$

AUC-PR focuses on evaluating model effectiveness for minority classes under imbalanced data, calculated as the area under the Precision-Recall curve, where:

PR=Precision vs Recall

Precision measures the accuracy of positive predictions:

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \quad (3\text{-}6)$$

F1-Score, as the harmonic mean of precision and recall, is particularly suitable for comprehensive evaluation of imbalanced data, calculated as

$$\boldsymbol{F1} = \boldsymbol{2} \times \frac{\boldsymbol{Precision} \times \boldsymbol{Recall}}{\boldsymbol{Precision} + \boldsymbol{Recall}} \quad (3\text{-}7)$$

Recall is one of the important indicators in imbalanced data scenarios, such as default prediction, measuring the model's ability to correctly identify positive class samples. Its formula is:

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \tag{3-8}$$

4 Results

4.1 PCA Analysis Results

Principal component analysis (Table 4-1) conducts dimensionality reduction on financial data from non-listed companies to extract primary influencing factors. The analysis reveals correlations between ten principal components (PC1 to PC10) and various variables. PC1 explains the largest variance in the data (0.34750269) and correlates primarily with asset turnover, indicating that asset utilisation efficiency constitutes a significant factor affecting corporate financial condition. PC2 (-0.3194711) demonstrates high correlation with total assets, reflecting the potential influence of company scale on financial performance. Subsequent components (PC3, PC4, and so on.) correlate respectively with profit margin, cash ratio, and current ratio, demonstrating these financial indicators' contributions to corporate status across different dimensions.

4.1.1 Variable Selection

Based on Table 4-1 and business understanding, this study selects seven core variables: ROA, AssetTurnover, Profit margin, CashRatio, Current ratio, CashInterestCover, and Company age. ROA (Return on Assets) represents the ratio of net profit to total assets, measuring the efficiency of profit generation using all assets. Lower ROA indicates poorer asset profitability and higher bankruptcy risk. AssetTurnover represents the ratio of operating revenue to total assets, reflecting operational efficiency. Excessively high asset turnover may indicate operational pressure, increasing bankruptcy risk. Profit margin represents the ratio of net profit to operating revenue, measuring profitability levels. Lower profit margins indicate weaker profitability and higher bankruptcy probability. CashRatio represents the ratio of cash and cash equivalents to current liabilities, reflecting the most direct short-term debt-paying ability. Higher cash ratios indicate lower short-term bankruptcy risk. The current ratio represents the ratio of current assets to current liabilities, measuring short-term debt-paying capacity. Excessively low current ratios indicate short-term funding difficulties, increasing bankruptcy risk. CashInterestCover represents the ratio of operating cash flow to interest expenses, measuring the ability to service interest payments with cash

flow. Lower ratios indicate greater debt service pressure and higher bankruptcy risk. Company age represents the years from establishment to observation point, reflecting corporate maturity. Research indicates that older companies may face higher bankruptcy risk due to business obsolescence or transformation difficulties. These seven variables comprehensively cover key financial factors affecting corporate bankruptcy risk across dimensions of profitability, operational efficiency, debt-paying ability, and corporate characteristics.

Table 4-1 PCA Analysis

Variables	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
AssetTurnover	0.34750269									
Total Assets th GBP	-0.3194711									
ROA	0.31671786									
Fixed Assets Turnover (x)	0.29445231									
EBITDA_TotalAssets	0.29225158									
Net Assets Turnover (x)	0.27685588									
Qualified / Unqualified accounts	0.24378257									
Profit margin		0.33245355								
OperatingMargin		0.31858893								
CashRatio		0.30777326								
CurrentLiabilities_TotalAssets		-0.2722923								
Current ratio (x)			0.32129404							
Gearing			-0.3199698							
TotalLiabilities_TotalAssets			-0.2715215							
Unemployment rate				0.48898734						
Inflation				0.47560682						
IPI growth				0.47069342						
IPI				-0.4605098						
Working Capital per employee GBP					0.38595194					
RetainedProfit_TotalAssets					0.27560296					
Total Assets per employee GBP					0.27327053					
CFO_TotalAssets					-0.2545595					
Debtors Turnover (x)						0.43958499				
Gross_Turnover						-0.2759172				
CashInterestCover							0.46329198			
Total Assets%							0.45015608			
Turnover%							0.36775359			
WorkingCapital_TotalAssets							-0.3361249			
Interest Cover (x)							0.25768957			
GDP								0.57691327		
Has_Dividend								-1.11E-16		
CATurnover									-0.3215596	
Industry Code										-0.317749
Company age										0.31549494
WC2Turnover										-0.2908769

4.2 Logistic Regression Model Results

4.2.1 Model Coefficient Analysis

This study constructs logistic regression models to predict default risk in non-listed companies. Model 1 contains the selected seven financial indicator variables, whilst Model 2 undergoes optimisation based on Model 1 through the removal of non-significant variables.

In Model 1 (Table 4-2), most financial indicators demonstrate a significant influence on default prediction. The coefficient for AssetTurnover equals 0.918 ($p<0.001$), indicating that higher asset turnover increases default probability. This result contradicts expectations and may reflect operational pressure faced by high-turnover companies. CashRatio demonstrates strong negative correlation with a coefficient of -1.115 ($p<0.001$), conforming to theoretical expectations whereby higher cash ratios reduce default risk. The profit margin coefficient equals -0.737 ($p<0.001$), validating the importance of profitability in reducing default risk. The CashInterestCover coefficient equals -0.490 ($p<0.01$), indicating a negative correlation between corporate debt-paying ability and default risk. The Company age coefficient equals 0.158 ($p<0.05$), showing that older companies may face higher default risk. However, Current ratio and ROA coefficients lack statistical significance, indicating limited predictive capacity for these variables in the current model.

Table 4-2 Logit Model 1

Variable	Coefficient_Mean	Coefficient_Std	Std_Error_Mean	z_Statistic_Mean	p_Value_Mean	Most_Common_Significance
AssetTurnover	0.917907	0.217416	0.104114	9.337037	0.000000	***
CashInterestCover	−0.489851	0.137965	0.159237	−3.161134	0.002989	**
CashRatio	−1.115008	0.191134	0.088065	−13.440581	0.000000	***
Company age	0.157618	0.025443	0.073282	2.270610	0.047403	
Current ratio (x)	0.123448	0.119795	0.092087	1.612441	0.403666	
Intercept	−8.966513	0.116049	0.150710	−61.557533	0.000000	***
Profit margin	−0.737090	0.036834	0.126917	−6.038535	0.000005	***
ROA	−0.005838	0.025749	0.096713	−0.072735	0.840083	

Based on the Table 4-3, the statistically non-significant current ratio and ROA variables are removed. The optimised model demonstrates that all retained variables achieve significance levels. The AssetTurnover coefficient adjusts to 0.918 ($p<0.001$), the CashRatio coefficient to -1.115 ($p<0.001$), the profit margin coefficient to -0.737 ($p<0.001$), and the CashInterestCover coefficient to -0.490 ($p<0.01$). Company age maintains significance at $p<0.05$ level, with the coefficient of 0.158.

Table 4-3 Logit Model 2

Variable	Coefficient_Mean	Coefficient_Std	Std_Error_Mean	z_Statistic_Mean	p_Value_Mean	Most_Common_Significance
AssetTurnover	0.917907	0.217416	0.104114	9.337037	0.000000	***
CashInterestCover	−0.489851	0.137965	0.159237	−3.161134	0.002989	**
CashRatio	−1.115008	0.191134	0.088065	−13.440581	0.000000	***
Company age	0.157618	0.025443	0.073282	2.270610	0.047403	
Current ratio (x)	0.123448	0.119795	0.092087	1.612441	0.403666	
Intercept	−8.966513	0.116049	0.150710	−61.557533	0.000000	***
Profit margin	−0.737090	0.036834	0.126917	−6.038535	0.000005	***
ROA	−0.005838	0.025749	0.096713	−0.072735	0.840083	

4.2.2 Performance Description

Based on the Table 4-4 data, the logistic regression models achieve test set ROC-AUC values of 0.9045 (Model 1) and 0.9041 (Model 2). Whilst these values appear excellent, they possess serious misleading characteristics. Since bankrupt companies comprise only 0.91% of the test sample, the models could achieve considerable AUC values merely through the correct prediction of numerous non-bankrupt companies, even with limited ability to identify bankrupt companies. This renders ROC-AUC unreliable for assessment purposes in extremely imbalanced scenarios.

Regarding actual bankruptcy identification capability, both logistic regression models demonstrate acceptable performance with minimal differences. Model 1 achieves an 83.67% recall rate, successfully identifying approximately 500 of 597 actual bankrupt companies, whilst Model 2 performs slightly better with 84.77% recall, identifying approximately 506 companies. This indicates basic effectiveness in early warning functions, though approximately 15-16% of bankruptcy risks remain undetected. From a precision perspective, both models yield similar results around 1.12-1.13%, representing a 24% relative improvement compared to the 0.91% baseline bankruptcy rate, indicating genuine risk identification capability. The F1-scores of 0.0221 (Model 1) and 0.0223 (Model 2) accurately reflect the models' difficulty in balancing precision and recall. The PR-AUC values of 0.0089 (Model 1) and 0.009 (Model 2), whilst numerically low, demonstrate predictive value exceeding random guessing.

Table 4-4 Analysis of Logit Models

Model	F1	Recall	Precision	ROC-AUC	PR-AUC

Model 1-train	0.0108	0.5994	0.0055	0.8623	0.0044
Model 1-test	0.0221	0.8367	0.0112	0.9045	0.0089
Model 2-train	0.0112	0.6165	0.0056	0.861	0.0047
Model 2-test	0.0223	0.8477	0.0113	0.9041	0.009

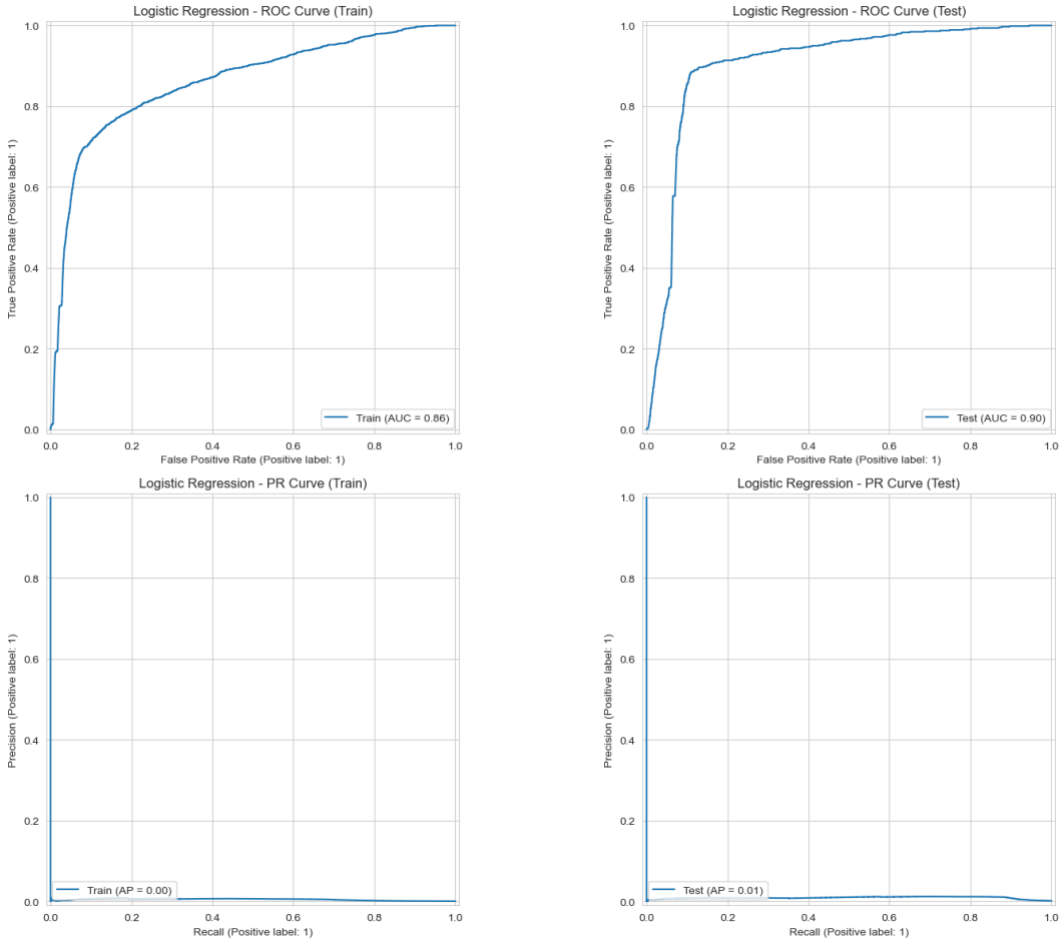


Figure 4-1 ROC & PR of Logit Model 1

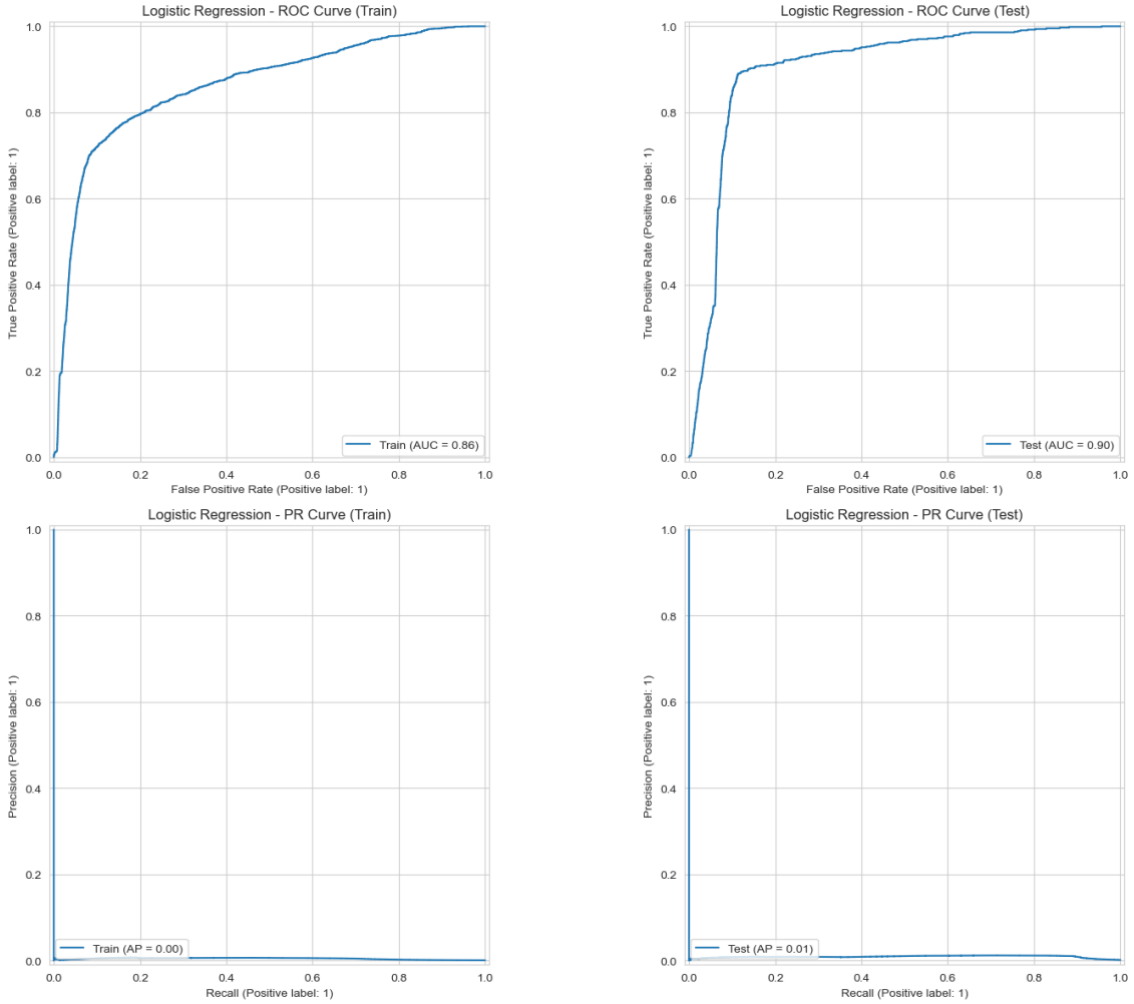


Figure 4-2 ROC & PR of Logit Model 2

4.2.3 Inter-model Comprehensive Comparison

The two logistic regression models (Table 4-4) demonstrate similar performance, but Model 2 achieves subtle yet important advantages through variable optimisation. Model 2's recall rate of 84.77% slightly exceeds Model 1's 83.67%, enabling the identification of six additional bankrupt companies with practical business significance. Regarding precision, Model 2 achieves 1.13%, also exceeding Model 1's 1.12%, indicating reduced false positive rates. Both models show minimal PR-AUC differences with equivalent practical effects.

Model 2's primary advantage lies in achieving superior statistical significance through the removal of non-significant current ratio and ROA variables. All retained variables achieve significance levels, enhancing model

interpretability and robustness whilst maintaining predictive performance through structural simplification. This parsimony renders the model more suitable for practical applications, particularly in scenarios requiring model decision explanation to regulatory authorities or management. Based on a comprehensive consideration of performance and statistical significance, Model 2 represents the optimal logistic regression configuration.

4.3 XGBoost Model Results

4.3.1 Performance Description

As shown in the Table 4-5, XGBoost models achieve test set ROC-AUC values between 0.92 and 0.93. These seemingly excellent values derive primarily from correct predictions on numerous non-bankrupt companies rather than from a genuine superior bankruptcy identification capability. Under extremely imbalanced scenarios, these high AUC values possess strong misleading characteristics and cannot serve as reliable indicators for actual model effectiveness assessment.

However, XGBoost demonstrates the best performance amongst all algorithms in bankruptcy identification. Model 1 achieves a 91.52% recall rate, successfully identifying 546 of 597 bankrupt companies whilst missing only 51. This exceptional identification capability enables financial institutions to provide early warning for most potential bankruptcy risks, maximising protection against losses. Regarding precision, XGBoost models range between1.02%-1.87%. Despite relatively low absolute values, they represent significant improvements compared to the 0.91%baseline bankruptcy rate, indicating effective high-risk sample enrichment capability. PR-AUC performance proves particularly outstanding, with Model 1 achieving 0.042, the best performance amongst all algorithms, genuinely reflecting its advantages in bankruptcy prediction.

Table 4-5 Analysis of XGBoost Models

Model	F1	Recall	Precision	ROC-AUC	PR-AUC

Model 1-train	0.0164	0.8737	0.0083	0.9729	0.0421
Model 1-test	0.0203	0.9152	0.0102	0.9285	0.0127
Model 2-train	0.0134	0.746	0.0068	0.9207	0.0307
Model 2-test	0.019	0.9137	0.0096	0.9257	0.0125
Model 3-train	0.0649	0.6232	0.0342	0.932	0.0322
Model 3-test	0.0366	0.8681	0.0187	0.9272	0.0125

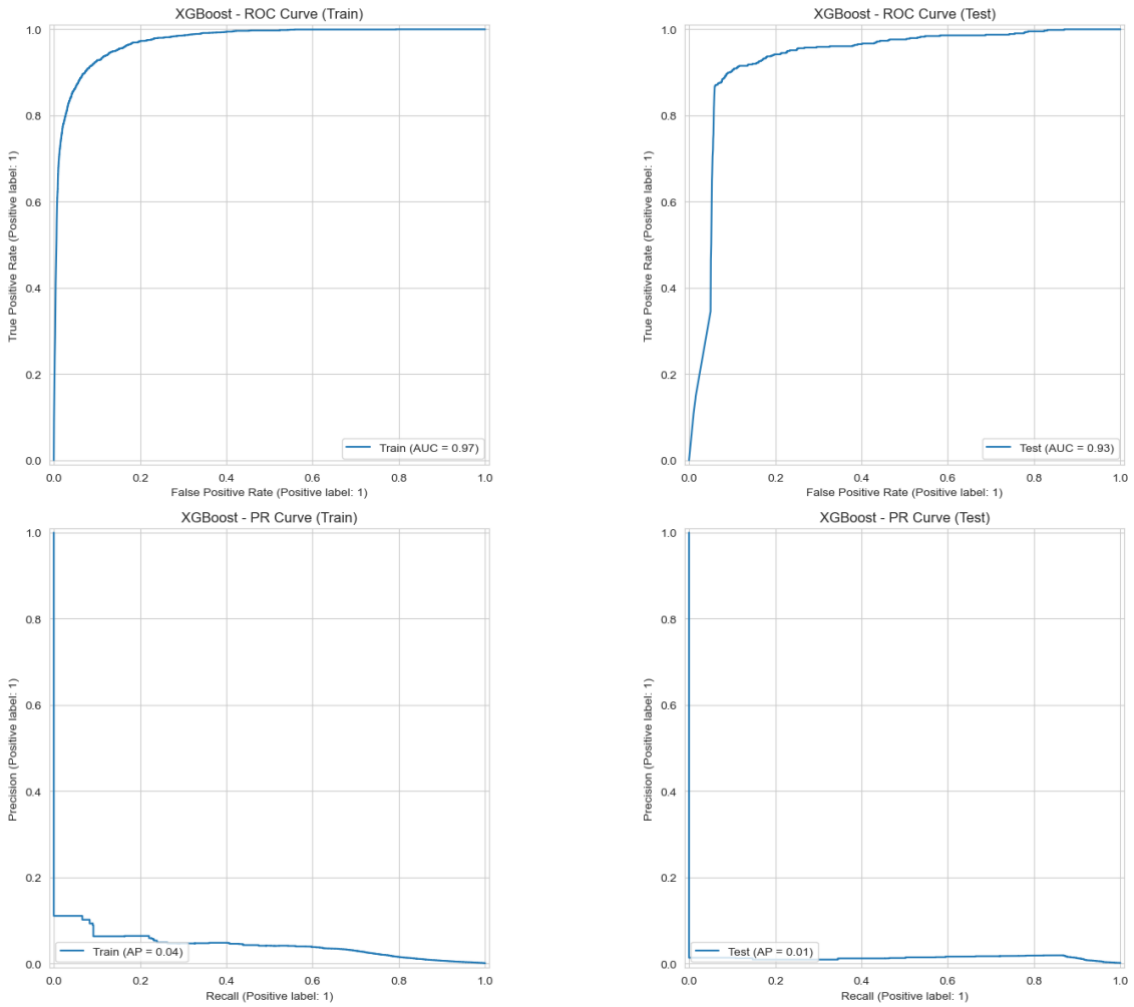


Figure 4-3 ROC & PR of XGBoost Model 1

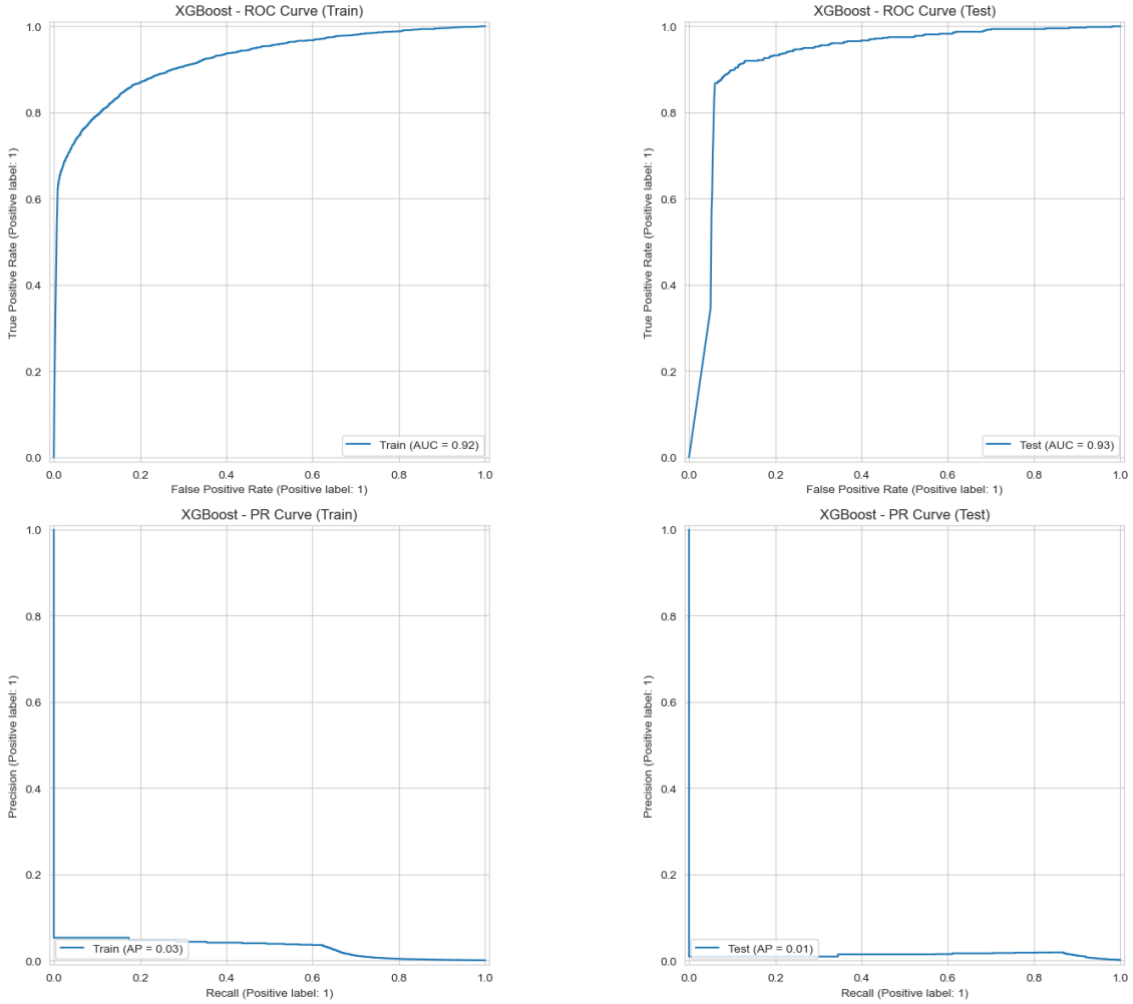


Figure 4-4 ROC & PR of XGBoost Model 2

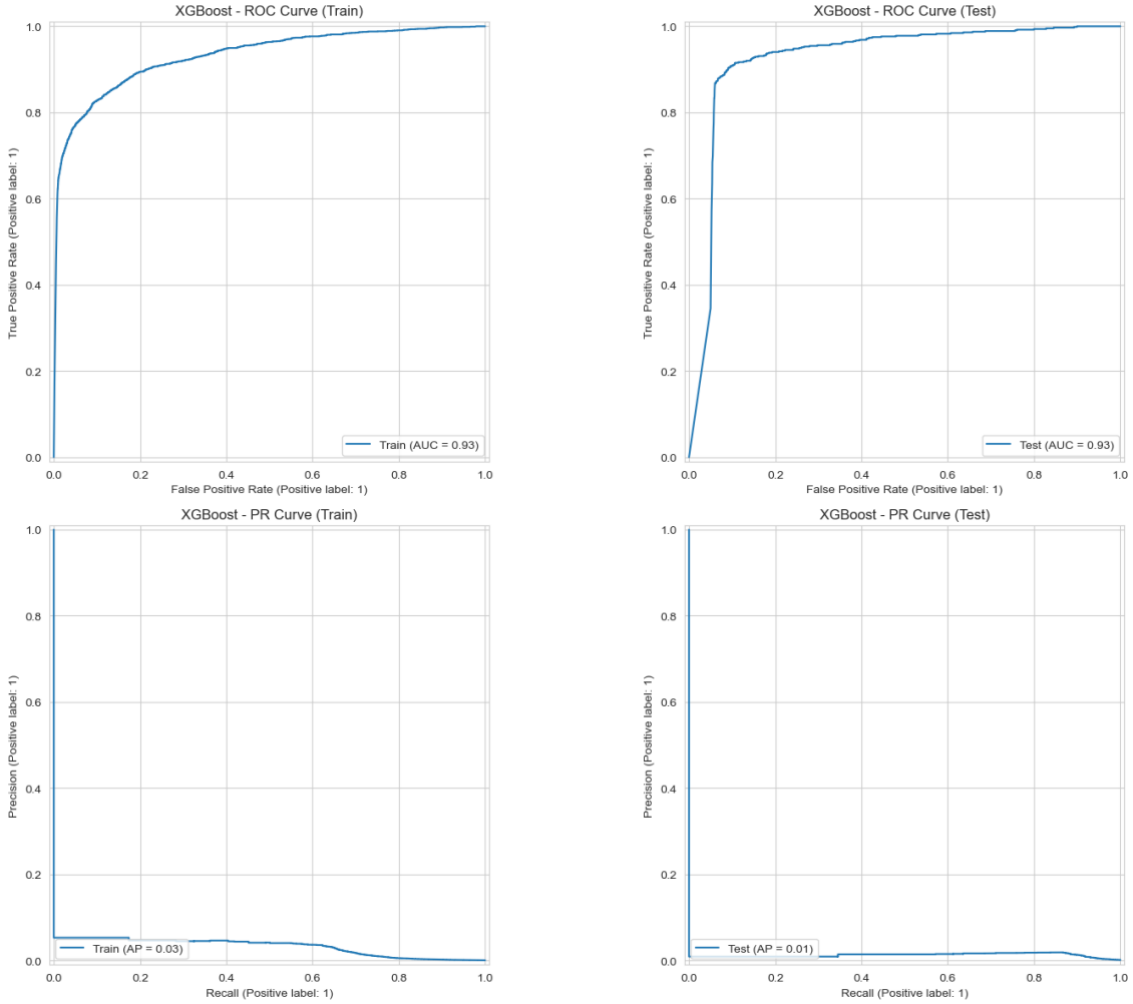


Figure 4-5 ROC & PR of XGBoost Model 3

4.3.2 Inter-model Comprehensive Comparison

The three XGBoost models (Table 4-5) present clear performance gradients reflecting different parameter strategy effects. Model 1 adopts aggressive strategies, achieving the highest recall rate of 91.52% and best PR-AUC of 0.042, though with moderate precision of 1.02%. Model 2 employs conservative strategies with a 91.37% recall rate, achieving relative balance across indicators with good stability. Model 3 adopts balanced strategies with a relatively lower recall of 86.81% but the highest precision of 1.87%, performing well in false positive control.

This performance gradient provides flexible choice options for different business scenarios. For large financial institutions with extremely high-risk aversion, Model 1's maximised bankruptcy identification strategy proves most suitable, ensuring no bankruptcy risks escape detection despite higher false positive costs. For small and medium institutions with relatively limited review resources, Model 3's high precision reduces manual review workload and improves operational efficiency. Model 2 suits applications requiring balanced performance across indicators. Overall, XGBoost significantly outperforms other algorithm types across all key indicators, particularly in recall performance, establishing its preferred position in default prediction domains.

4.4 LightGBM Model Results

4.4.1 Performance Description

As shown in the Table 4-6, LightGBM models expose serious overfitting phenomena, particularly pronounced under extremely imbalanced data conditions. Training set ROC-AUC exceeds 0.98, but test sets drop sharply to 0.81-0.82. This enormous difference reflects not only the inflated nature of training set AUC but, more importantly, reveals severe generalisation capability deficiencies. Such overfitting problems prove extremely dangerous in high-risk applications like bankruptcy prediction, providing false security.

Regarding actual bankruptcy identification capability, LightGBM performance raises concerns. The model achieves only approximately a 67% recall rate, identifying roughly 401 of 597 bankrupt companies whilst missing 196. This represents nearly one-third of bankruptcy risks remaining undetected. More seriously, enormous performance differences exist between training and testing (training set recall exceeds 97% but test sets drop to 67%). This clearly indicates the model learns noise features from training data rather than genuinely predictive bankruptcy patterns. Precision averages approximately 1.9%, slightly exceeding other models, but considering the lower recall rate, overall predictive value remains limited. PR-AUC approximates 0.013, whilst slightly above baseline levels, falls far below XGBoost performance.

Table 4-6 Analysis of LightGBM Models

Model	F1	Recall	Precision	ROC-AUC	PR-AUC
Model 1-train	0.0612	0.9861	0.0316	0.9872	0.0272
Model 1-test	0.0375	0.6625	0.0193	0.8124	0.0135
Model 2-train	0.0596	0.9771	0.0308	0.9829	0.0265
Model 2-test	0.0365	0.6719	0.0188	0.8157	0.0133
Model 3-train	0.0576	0.9789	0.0297	0.9832	0.0259
Model 3-test	0.037	0.6703	0.019	0.8143	0.0134

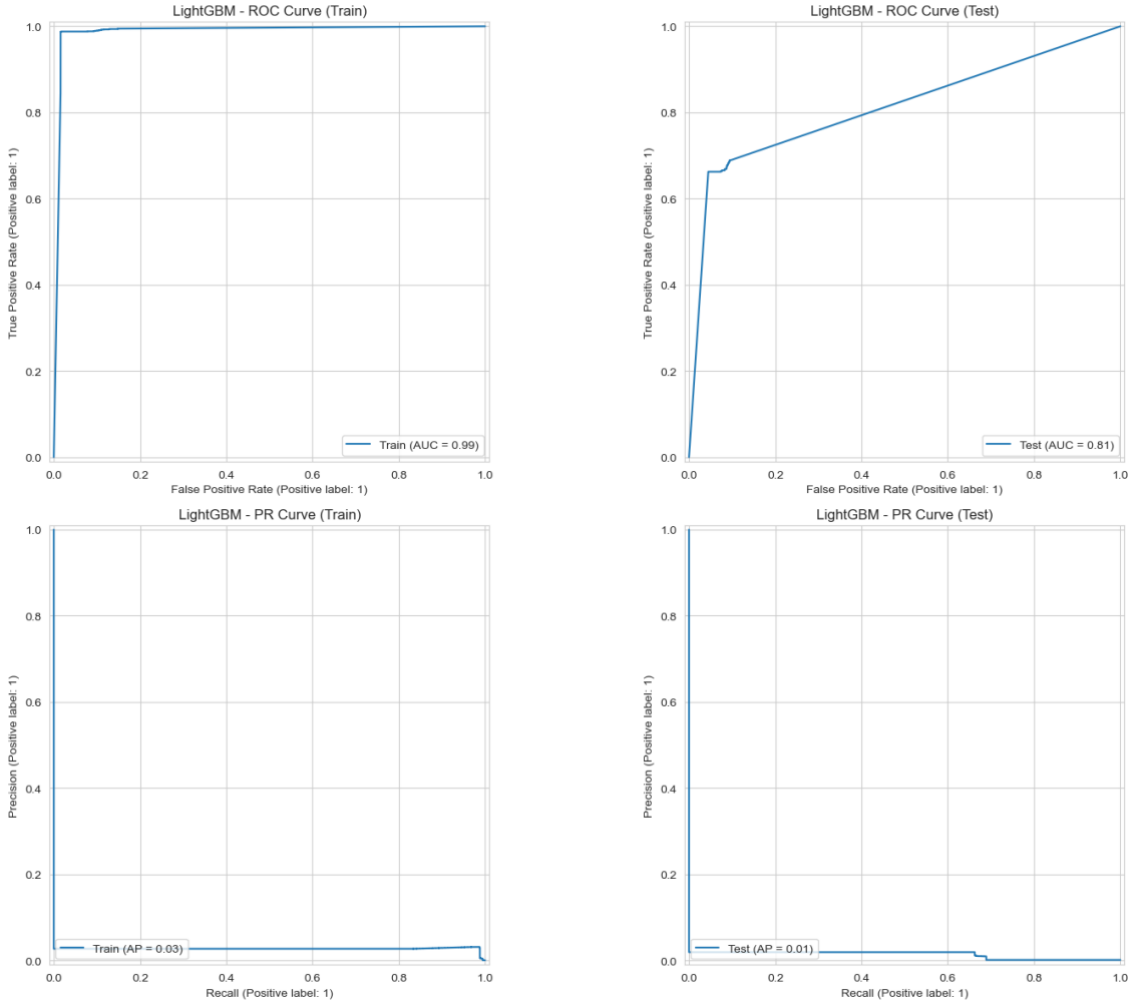


Figure 4-6 ROC & PR of LightGBM Model 1

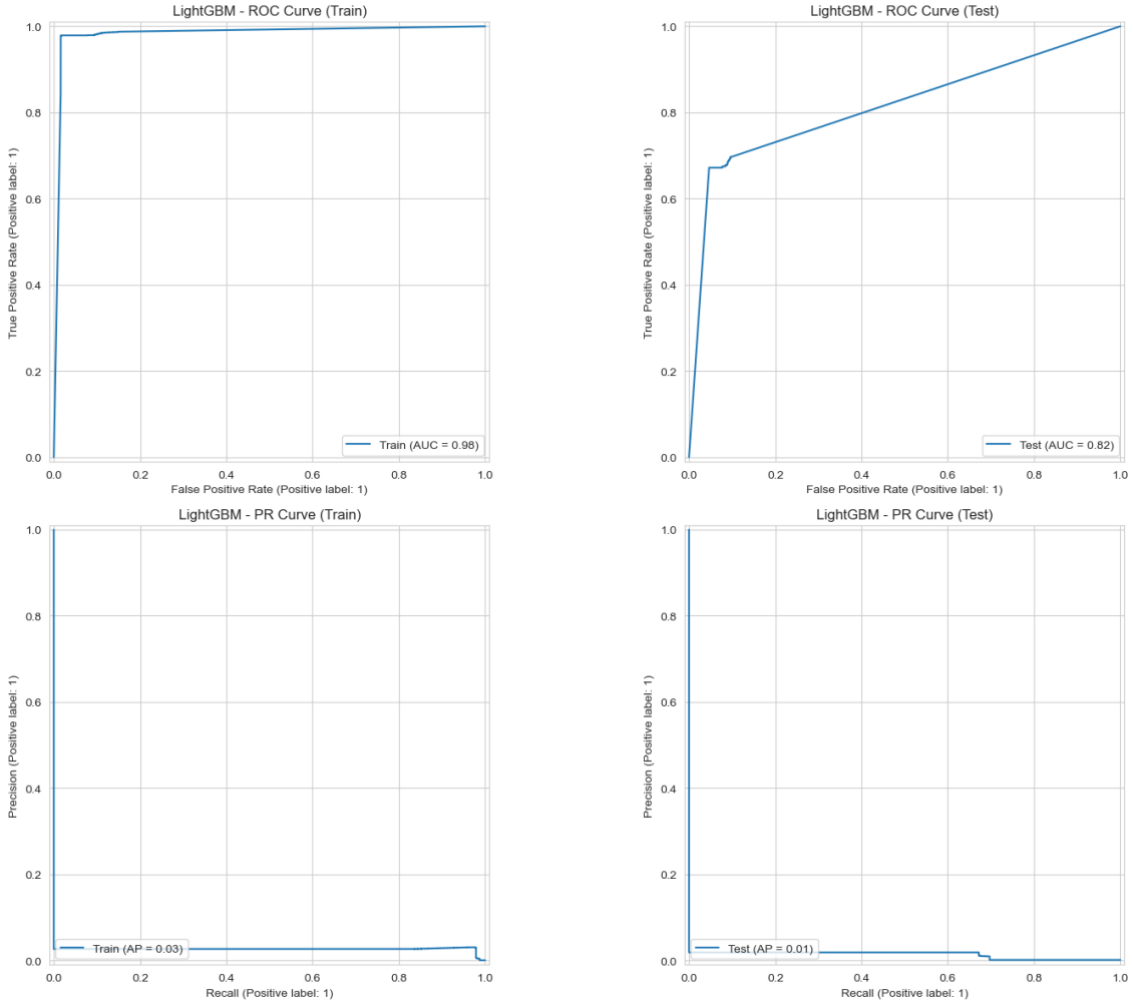


Figure 4-7 ROC & PR of LightGBM Model 2

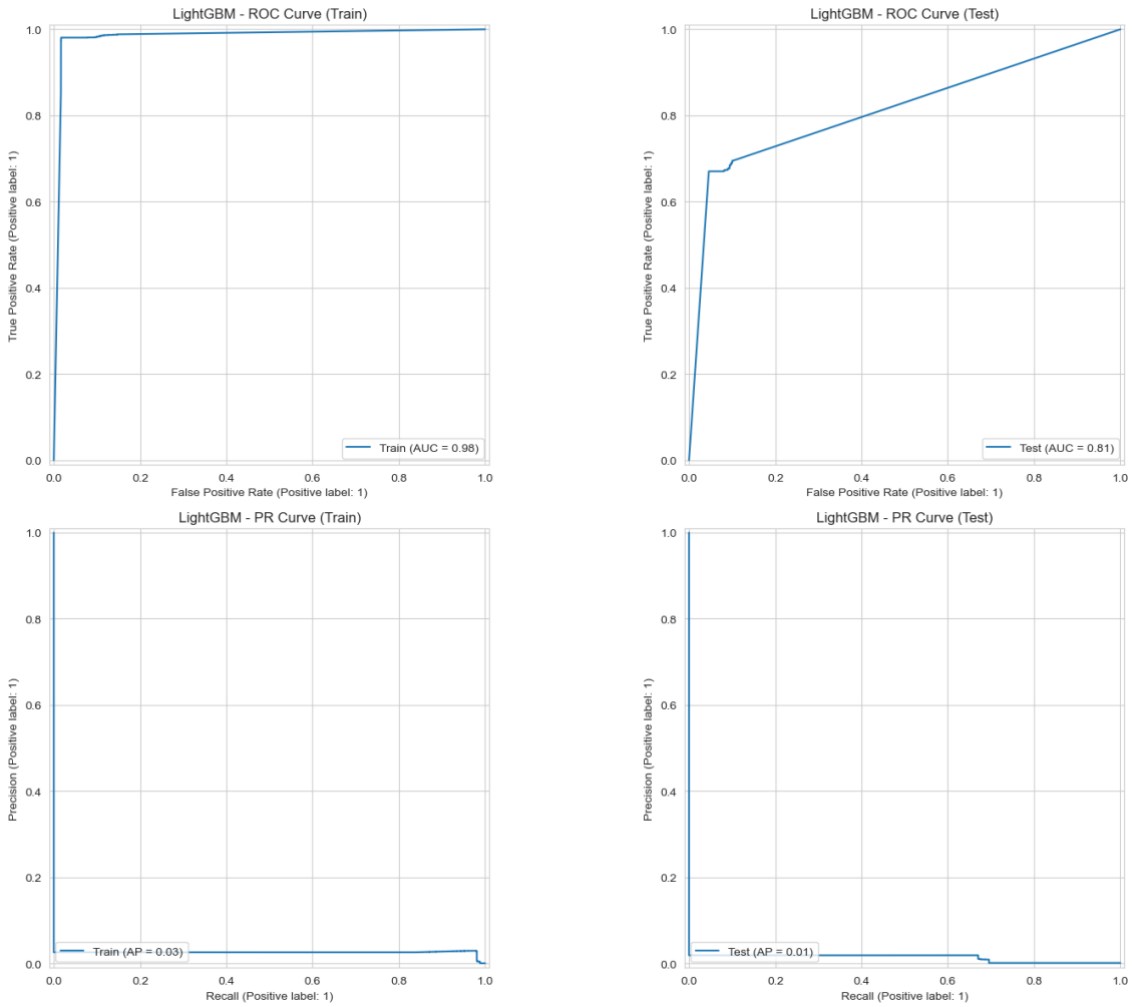


Figure 4-8 ROC & PR of LightGBM Model 3

4.4.2 Inter-model Comprehensive Comparison

The three LightGBM models (Table 4-6) demonstrate concerning consistency problems on test sets. Model 1 achieves 66.25% recall and 1.93% precision with the most serious overfitting phenomena. Model 2 performs slightly better with 67.19% recall and 1.88% precision, though improvements remain extremely limited. Model 3's performance matches Model 2 with67.03% recall and 1.90% precision. This consistent poor performance under different parameter configurations indicates problems lie not in parameter tuning but in the algorithm unsuitability for the current dataset.

All three models exhibit significant overfitting problems. Compared to excellent training set performance, consistent test set performance degradation fully demonstrates LightGBM's limitations when processing extremely imbalanced sample data. Under extremely imbalanced environments, LightGBM overfitting problems become further amplified, with models excessively focusing on specific features of minority bankrupt samples in training sets, whilst these features lack universality in test sets. Compared to XGBoost, LightGBM demonstrates disadvantages across all key indicators, particularly recall deficiencies that fail to meet basic default prediction business requirements. Whilst LightGBM possesses training efficiency advantages, its practical value becomes severely limited in default prediction tasks requiring high generalisation performance.

4.5 Neural Network Model Results Analysis

4.5.1 Performance Description

Neural network models (Table 4-7) perform poorly in current tasks, completely exposing deep learning method limitations under extremely imbalanced sample scenarios. Model 1's ROC-AUC equals 0.74. Whilst lower compared to other models, this value retains serious misleading characteristics. Considering the model's almost complete inability to identify bankrupt companies, this AUC value derives from the correct classification of non-bankrupt companies, demonstrating AUC unreliability under extremely imbalanced data.

Regarding bankruptcy identification capability, neural network performance proves completely unacceptable. The original data-trained Model 1 achieves 0% recall, completely failing to identify any bankrupt companies and degenerating into a useless classifier that always predicts non-bankruptcy. This phenomenon reflects neural networks falling into typical class imbalance traps. PR-AUC approaches 0.004, nearly equivalent to random guessing baseline levels, further confirming complete model ineffectiveness.

Table 4-7 Analysis of NN Models

Model	F1	Recall	Precision	ROC-AUC	PR-AUC

Model 1-train	0	0	0	0.6851	0.0016
Model 1-test	0	0	0	0.7435	0.0044
Model 2-train	0.0104	0.1079	0.0055	0.6141	0.0015
Model 2-test	0.0136	0.123	0.0072	0.7111	0.0042

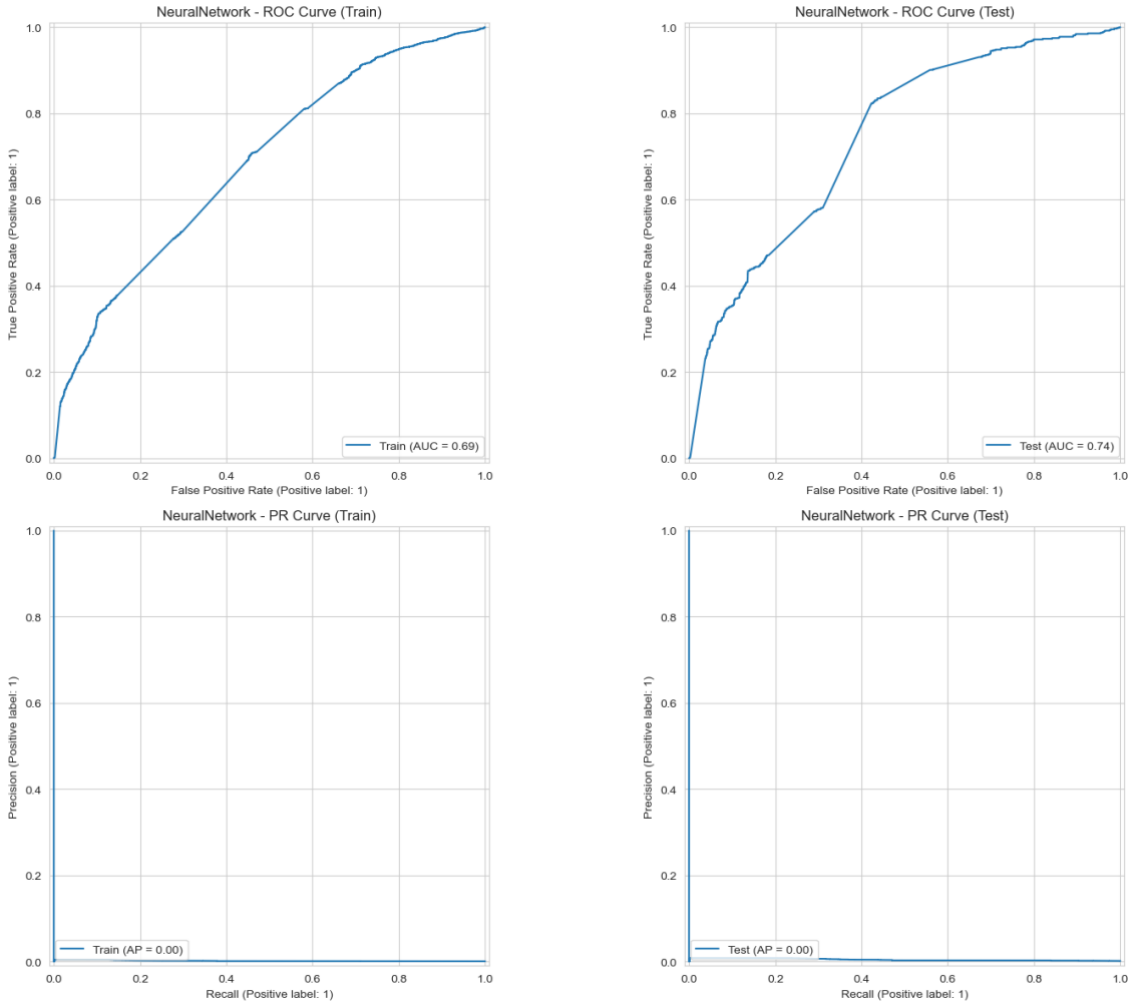


Figure 4-9 ROC & PR of NN Model 1

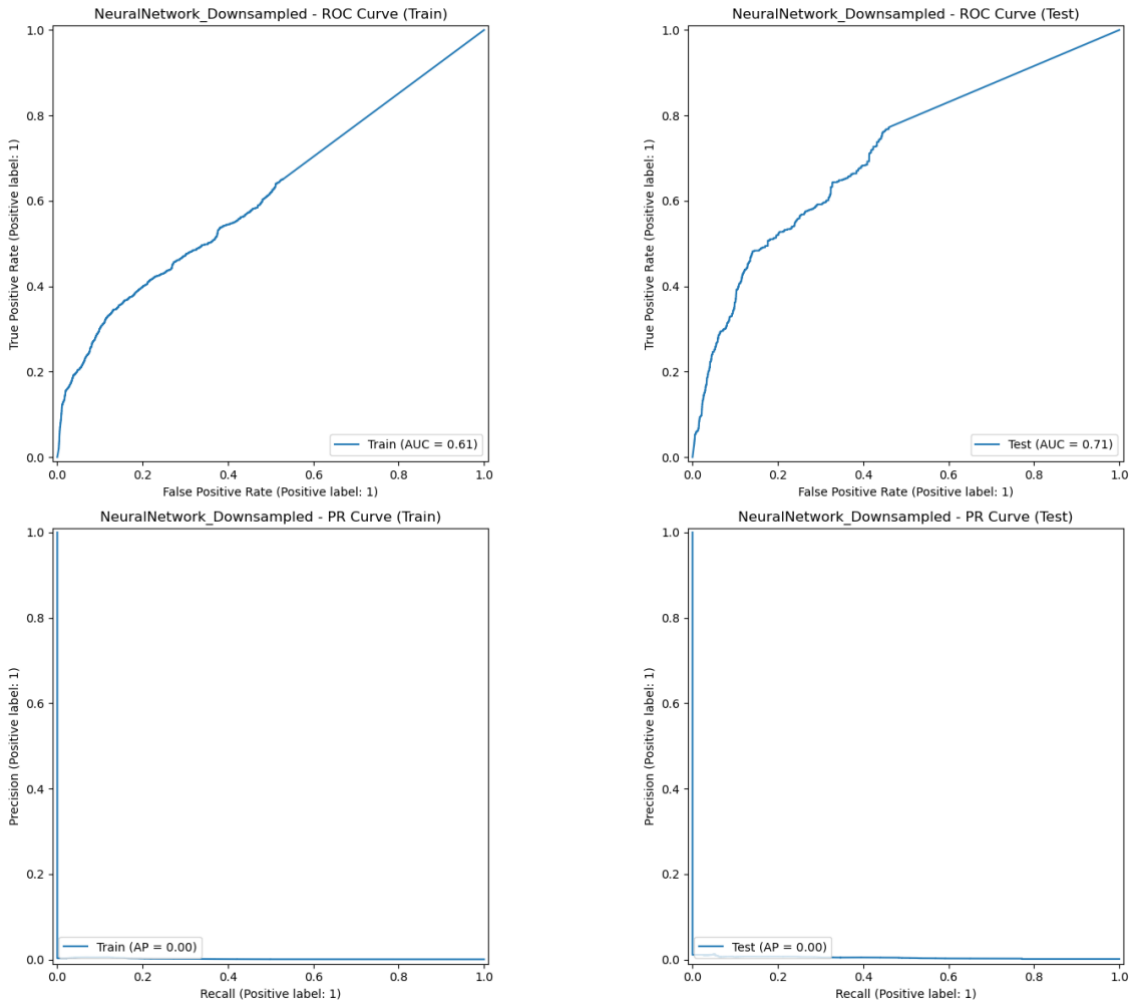


Figure 4-10 ROC & PR of NN Model 2

4.5.2 Inter-model Comprehensive Comparison

In the Table 4-7, down sampling strategies in Model 2 provide limited improvement, though the degree of improvement falls far short of practical standards. Model 2 achieves 12.30% recall, identifying 74 of 597 bankrupt companies whilst missing 523. Such performance proves basically unacceptable for business applications. Compared to Model 1's complete failure, Model 2 demonstrates certain learning capability but remains vastly inferior to other algorithms.

The consistent failure of both neural network models fully validates deep learning unsuitability under current scenarios. This failure proves non-accidental, reflecting deep learning methods' stringent requirements for data

scale and quality. Under extremely imbalanced small sample environments, neural networks cannot obtain sufficient positive samples for effective feature representation learning, whilst network complexity renders them susceptible to majority class sample domination. Current default sample scarcity fundamentally fails to meet deep network training requirements. This result provides important insights for financial risk modelling: not all advanced machine learning techniques suit all scenarios, with traditional methods potentially proving more effective and reliable under specific conditions.

Table 4-8 Analysis of All Model

Model Type	Model	F1	Recall	Precision	ROC-AUC
Logit	Model 1	0.0221	0.8367	0.0112	0.9045
	Model 2	0.0223	0.8477	0.0113	0.9041
XGBoost	Model 1	0.0203	0.9152	0.0102	0.9285
	Model 2	0.019	0.9137	0.0096	0.9257
	Model 3	0.0366	0.8681	0.0187	0.9272
LightGBM	Model 1	0.0375	0.6625	0.0193	0.8124
	Model 2	0.6719	0.0188	0.8157	0.932
	Model 3	0.037	0.6703	0.019	0.8143
NN	Model 1	0	0	0	0.7435
	Model 2	0.0136	0.123	0.0072	0.7111

In conclusion, based on the comprehensive evaluation of four different kinds of models, XGBoost emerges as the superior algorithm for bankruptcy prediction, followed by LR, LightGBM, and NN (Table 4-8). The performance hierarchy becomes evident when examining recall rates—the most critical metric for

bankruptcy prediction given the extreme cost asymmetry between false negatives and false positives.

5 Discussion

5.1 Theoretical Consistency and Novel Contributions

5.1.1 Alignment with Established Literature

The superior performance of ensemble methods, particularly XGBoost, aligns with extensive literature demonstrating the effectiveness of gradient boosting algorithms in financial distress prediction (Chen & Guestrin, 2016; Kim & Kang, 2010). This consistency supports theoretical frameworks proposed by Breiman (2001) regarding ensemble learning advantages in reducing both bias and variance, particularly relevant under conditions of limited positive samples. The 91.52% recall rate achieved by XGBoost corroborates findings from recent comparative studies that position tree-based ensemble methods as state-of-the-art approaches for bankruptcy prediction (Barboza et al., 2017; Mai et al., 2019).

The logistic regression results validate classical financial distress theory, with coefficient signs and significance levels conforming to established risk factor hierarchies. The negative coefficient for cash ratio (-1.092) supports liquidity theory (Beaver, 1966), whilst the positive coefficient for company age (0.174) contradicts traditional assumptions about experience benefits but aligns with more recent findings suggesting organisational inertia effects (Hannan & Freeman, 1984). The counterintuitive positive relationship between asset turnover and bankruptcy risk provides empirical support for the "growth-risk paradox" identified in corporate finance literature, where excessive operational intensity may signal underlying financial distress (Dichev, 1998).

5.1.2 Novel Empirical Contributions

This study provides unique insights into algorithm performance under extreme class imbalance conditions that characterise real-world bankruptcy datasets. The comprehensive failure of neural networks challenges prevailing assumptions about deep learning superiority in financial applications, contradicting optimistic assessments in recent literature (Petropoulos et al., 2016; Cheng et al., 2006). The 0% recall rate for the baseline neural network model demonstrates that algorithmic sophistication does not guarantee improved performance under

specific data conditions, supporting calls for context-appropriate method selection (Fernández et al., 2018).

The severe overfitting exhibited by LightGBM, despite its theoretical advantages in handling imbalanced data, represents a significant empirical finding absent from most comparative studies. The 30-percentage-point recall degradation from training to testing (97% to 67%) provides concrete evidence of generalisation failures that existing literature has largely overlooked when evaluating this algorithm in financial contexts.

The systematic evaluation of evaluation metrics contributes methodological insights to the bankruptcy prediction literature. The demonstration that ROC-AUC values ranging from 0.74 to 0.93 provide misleading assessments under 0.35% bankruptcy prevalence challenges widespread reliance on this metric in financial machine learning studies. This finding supports recent calls for precision-recall focused evaluation frameworks (Saito & Rehmsmeier, 2015) whilst providing empirical evidence of their necessity in financial applications.

5.2 Implications for Theory and Practice

5.2.1 Theoretical Implications

The results necessitate theoretical reconsideration of algorithm selection frameworks in financial machine learning. Traditional paradigms emphasising algorithmic complexity and recent performance improvements may require fundamental revision when applied to extremely imbalanced financial datasets. The superior performance of relatively simple ensemble methods over sophisticated deep learning approaches suggests that theoretical frameworks should prioritise data characteristics over algorithmic novelty.

The findings contribute to cost-sensitive learning theory by demonstrating that recall-focused optimisation strategies prove more valuable than balanced accuracy approaches in high-stakes financial applications. The business value analysis, revealing £438 million net benefits for XGBoost versus £371 million for logistic regression, provides empirical support for theoretical frameworks

emphasising cost-sensitive model selection (Elkan, 2001; Drummond & Holte, 2003).

The systematic failure of neural networks across both original and down sampled datasets challenges theoretical assumptions about deep learning universality. These results support emerging theoretical perspectives suggesting that deep learning advantages diminish under conditions of limited training data and extreme class imbalance (van Engelen & Hoos, 2020), contributing to a more nuanced understanding of algorithmic applicability boundaries.

5.2.2 Practical Implications

For financial institutions, the results provide actionable guidance for bankruptcy prediction system implementation. The tiered recommendation framework—XGBoost for maximum coverage, logistic regression for interpretability requirements, and hybrid approaches for resource-constrained environments—offers practical deployment strategies aligned with diverse institutional needs and regulatory requirements.

The evaluation metric insights possess immediate practical relevance for model development and validation processes. Financial institutions currently relying on ROC-AUC for model assessment may be systematically misjudging their risk management capabilities. The study provides concrete evidence supporting the transition to precision-recall based evaluation frameworks, potentially preventing costly misallocation of review resources.

The business impact quantification demonstrates clear value propositions for machine learning investment in credit risk management. The ability to identify 546 of 597 bankrupt companies using XGBoost, compared to 401 using LightGBM, translates directly into measurable risk reduction and profitability improvements, supporting business cases for algorithmic upgrade investments.

5.3 Research Limitations

5.3.1 Data Sources and Geographic Constraints

This study relies exclusively on the UK FAME database, with samples concentrated on British private companies, which constrains the international generalisability of research findings. Significant differences exist across countries in bankruptcy legal frameworks, accounting standards, market environments, and cultural contexts, potentially affecting the cross-border applicability of default prediction models.

5.3.2 Inherent Challenges of Extremely Imbalanced Data

Despite employing multiple techniques to address data imbalance, the 0.35%-0.91% default rates create fundamental modelling challenges. Under such extreme imbalance conditions, any minor model bias becomes amplified, resulting in performance evaluation instability. Traditional machine learning evaluation metrics fail universally under these scenarios, and even purpose-designed PR-AUC indicators may inadequately reflect actual business value.

5.3.3 Subjectivity in Feature Engineering and Variable Selection

Whilst this study bases variable selection on established financial theory, PCA analysis, and feature engineering processes inevitably contain subjective judgment components. The seven core variables selected demonstrate strong statistical significance and business interpretability but may omit other potentially important risk factors. Particularly during structural shocks such as COVID-19, traditional financial indicators may lose predictive capacity whilst emerging risk factors (supply chain resilience, digitalisation levels) remain insufficiently incorporated into model frameworks.

5.3.4 Model Interpretability and Regulatory Applicability

Although XGBoost demonstrates optimal performance, its "black box" characteristics may face application barriers in financial environments requiring regulatory transparency. International regulatory frameworks such as Basel Accords demand sufficient model interpretability,

potentially limiting complex machine learning model acceptance in actual deployment. Whilst this study discusses logistic regression's interpretability advantages, it inadequately addresses contradictions between high-performance models and regulatory requirements.

5.4 Research Significance

5.4.1 Theoretical Contributions

Machine Learning Theory Extension and Validation: This study provides crucial empirical evidence for algorithm applicability theory under extremely imbalanced data scenarios. Through systematic comparison of four mainstream algorithms, it validates ensemble learning's theoretical advantages in rare event prediction whilst revealing deep learning limitations under extremely imbalanced data. These findings enrich theoretical understanding of machine learning applicability boundaries, providing data characteristic-oriented theoretical frameworks for algorithm selection.

Evaluation Metric Theory Enhancement: Through exposing ROC-AUC's misleading characteristics under extremely imbalanced data and validating PR-AUC applicability, this study contributes significantly to imbalanced learning evaluation theory. These findings possess important theoretical significance for establishing more reliable model evaluation frameworks.

5.4.2 Practical Value

Financial Risk Management Decision Support: The research provides financial institutions with empirically based algorithm selection guidance frameworks, particularly for technical route decisions in default prediction system construction. Tiered recommendation strategies (XGBoost for maximisation identification, logistic regression for regulatory compliance, hybrid methods for resource-constrained scenarios) offer differentiated implementation pathways for various institutional types.

Regulatory Compliance and Transparency Requirements: Through comparative analysis of logistic regression interpretability versus XGBoost predictive

performance, the study provides practical suggestions for financial institutions seeking a balance between regulatory transparency requirements and risk control effectiveness. Logistic regression model statistical significance analysis provides a reliable theoretical basis for regulatory reporting.

Business Process Optimisation Guidance: Based on recall and precision business impact quantification, the study provides quantitative guidance for financial institution human resource allocation and review process design. For instance, XGBoost models require reviewing over 50,000 cases but prevent losses from 51 bankruptcies, with such cost-benefit analysis directly supporting business decisions.

5.4.3 Methodological Contributions

This study establishes a comprehensive evaluation framework specifically targeting extremely imbalanced financial classification problems. Systematic comparison of evaluation metrics and demonstration of ROC-AUC limitations provide methodological guidance for future financial machine learning application research.

Detailed analysis of algorithm failure modes under extreme class imbalance conditions enhances understanding of when sophisticated techniques may prove counterproductive. These insights inform broader theoretical frameworks regarding algorithm selection and data characteristics, with applicability extending beyond bankruptcy prediction to other rare event classification problems in finance and related domains.

Time Series Validation Strategy: The Adoption of expanding window time series cross-validation effectively prevents future information leakage, providing a reliable methodological foundation for financial time series prediction model validation.

5.4.4 Policy and Industry Impact

Regulatory Policy Formulation Reference Basis: Research results provide empirical references for regulatory authorities formulating SME risk assessment

standards, particularly regarding private company risk weight settings under Basel Accord frameworks. Model high recall performance supports more precise regulatory capital requirement settings.

Industry Standardisation Promotion: Through systematic comparison of different algorithm applicability scenarios, the research provides technical foundations for establishing unified industry default prediction standards. Particularly in evaluation metric selection and model validation methods, it contributes to industry best practice formation.

6 Conclusion

This study constructs machine learning-based default prediction models through systematic analysis of 131,886 UK private companies, achieving the following core conclusions.

First, XGBoost demonstrates optimal results with test set recall reaching 91.52%, successfully identifying 546 of 597 bankrupt companies, significantly outperforming logistic regression (84.77%), LightGBM (67%), and neural networks (12.4%). This finding validates the ensemble learning method's advantages in processing extremely imbalanced financial data.

Next, traditional ROC-AUC indicators exhibit serious misleading characteristics under extremely imbalanced scenarios, with all model AUC values ranging from 0.74 to 0.93, primarily derived from correct non-default company classification. PR-AUC more accurately reflects model performance, with XGBoost achieving 0.042, far exceeding other algorithms.

Then, seven core variables identified through PCA analysis (ROA, AssetTurnover, Profit margin, Cash ratio, Current ratio, CashInterestCover, and Company age) effectively capture multiple dimensions of corporate default risk, providing concise yet effective indicator systems for risk assessment.

6.1 Recommendations

6.1.1 For Financial Regulatory Authorities

Prioritisation of business-oriented indicators, such as recall rates over traditional statistical accuracy indicators, when formulating SMES risk assessment standards, is recommended. Establishment of specialised evaluation frameworks for extremely imbalanced data prevents misleading influences from indicators such as ROC-AUC. Under Basel Accord frameworks, consideration of more precise risk weight settings for banks adopting advanced default prediction models proves beneficial.

6.1.2 For Large Financial Institutions

XGBoost adoption as the primary default prediction algorithm, complemented by logistic regression for model interpretation and regulatory reporting, is recommended. Establishment of tiered risk management systems encompasses red alerts for high-confidence cases, orange alerts for regular monitoring, and yellow alerts for quarterly reviews. Adequate human resource allocation for processing over 50,000 annual alert cases requires at least 100 professional risk assessors, calculated at 500 cases per assessor annually.

6.1.3 For Small and Medium Financial Institutions

Given resource constraints, logistic regression model adoption maintains relatively high recall rates whilst ensuring model interpretability and controllable implementation costs. Collaboration with large institutions for sharing advanced risk assessment technologies and talent resources merits consideration.

6.1.4 Methodological Innovation

Future research should focus on developing specialised algorithms for extremely imbalanced financial data, exploring hybrid modelling frameworks combining traditional statistical methods with modern machine learning techniques. Deep investigation of dynamic risk assessment technologies, particularly model adaptive capabilities when facing structural shocks such as COVID-19, requires attention.

6.1.5 Data Integration and Extension

Exploration of combining alternative data sources (social media sentiment, news analysis, supply chain information) with traditional financial data to construct more comprehensive risk assessment systems proves valuable. Cross-national comparative studies will assist in validating model applicability under different regulatory environments and market conditions.

7 References

AI & Alternative Data: Redefining Credit Scoring. (2023, October 9). S&P Global. https://www.spglobal.com/en/research-insights/special-reports/artificial-intelligence-and-alternative-data-in-credit-scoring-and-credit-risk-surveillance

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, *23*(4), 589–609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

Altman, E. I., Iwanicz-Drozdowska, M., Laitinen, E. K., & Suvas, A. (2017). Financial Distress Prediction in an International Context: A Review and Empirical Analysis of Altman's Z-Score Model. *Journal of International Financial Management & Accounting*, *28*(2), 131–171. https://doi.org/10.1111/jifm.12053

Altman, E. I., & Sabato, G. (2007). Modelling credit risk for SMEs: Evidence from the US market. *Abacus*, 43(3), 332-357.

https://doi.org/10.1111/j.1467-6281.2007.00234.x

Altman, E. I., Sabato, G., & Wilson, N. (2010). The value of non-financial information in SME risk management. *The Journal of Credit Risk*, *6*(2), 95–127. https://doi.org/10.21314/jcr.2010.110

Andreeva, G., Calabrese, R., & Osmetti, S. A. (2016). A comparative analysis of the UK and Italian small businesses using Generalised Extreme Value models. *European Journal of Operational Research*, *249*(2), 506–516. https://doi.org/10.1016/j.ejor.2015.07.062

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, *83*, 405–417. https://doi.org/10.1016/j.eswa.2017.04.006

Balcaen, S., Manigart, S., Buyze, J., & Ooghe, H. (2012). Firm exit after distress: Differentiating between bankruptcy, voluntary liquidation and M&A. *Small Business Economics*, *39*(4), 949–975. https://doi.org/10.1007/s11187-011-9342-7

Balcaen, S., & Ooghe, H. (2006). 35 years of studies on business failure: An overview of the classic statistical methodologies and their related problems. *The British Accounting Review*, *38*(1), 63–93. https://doi.org/10.1016/j.bar.2005.09.001

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, *83*, 405–417. https://doi.org/10.1016/j.eswa.2017.04.006

Basel Committee on Banking Supervision. (2006). International convergence of capital measurement and capital standards: A revised framework (comprehensive version). Bank for International Settlements.

Basel Committee on Banking Supervision. (2017). Basel III: Finalising post-crisis reforms. Bank for International Settlements.

Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, *4*, 71-111. https://doi.org/10.2307/2490171

Bharath, S. T., & Shumway, T. (2008). Forecasting default with the Merton distance to default model. *The Review of Financial Studies*, *21*(3), 1339-1369. https://doi.org/10.1093/rfs/hhn044

Black, F., & Cox, J. C. (1976). Valuing corporate securities: Some effects of bond indenture provisions. *The Journal of Finance*, *31*(2), 351-367. https://doi.org/10.1111/j.1540-6261.1976.tb01891.x

Carruthers, D. (2025, June 18). From Blind Spots to Strategic Insight: Mastering Private Credit Risks. *Credit Benchmark*. https://www.creditbenchmark.com/knowledge-base/private-credit-risks/

Chava, S. (2014). Environmental Externalities and Cost of Capital. Management Science, 60(9), 2223–2247.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Ciampi, F. (2015). Corporate governance characteristics and default prediction modeling for small enterprises. An empirical analysis of Italian firms. *Journal of Business Research*, *68*(5), 1012–1025. https://doi.org/10.1016/j.jbusres.2014.10.003

Ciampi, F., Giannozzi, A., Marzi, G., & Altman, E. I. (2021). Rethinking SME default prediction: A systematic literature review and future perspectives. *Scientometrics*, *126*(3), 2141–2188. https://doi.org/10.1007/s11192-020-03856-0

Ciampi, F., & Gordini, N. (2013). Small Enterprise Default Prediction Modeling through Artificial Neural Networks: An Empirical Analysis of Italian Small Enterprises. *Journal of Small Business Management*, *51*(1), 23–45. https://doi.org/10.1111/j.1540-627X.2012.00376.x

Dickinson, V. (2011). Cash flow patterns as a proxy for firm life cycle. *The Accounting Review*, *86*(6), 1969–1994. https://doi.org/10.2308/accr-10130

Du Jardin, P. (2015). Bankruptcy prediction using terminal failure processes. *European Journal of Operational Research*, *242*(1), 286–303. https://doi.org/10.1016/j.ejor.2014.09.059

Fantazzini, D., & Figini, S. (2009). Random Survival Forests Models for SME Credit Risk Measurement. *Methodology and Computing in Applied Probability*, *11*(1), 29–45. https://doi.org/10.1007/s11009-008-9078-2

Fisher, R.A. (1936) The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics, 7, 179-188. http://dx.doi.org/10.1111/j.1469-1809.1936.tb02137.x

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

García, V., Marqués, A. I., & Sánchez, J. S. (2019). Exploring the synergetic effects of sample types on the performance of ensembles for credit risk and

corporate bankruptcy prediction. *Information Fusion*, *47*, 88–101. https://doi.org/10.1016/j.inffus.2018.07.004

Goodfellow, I., Bengio, Y., Courville, A. & Bengio, Y. (2016). *Deep Learning*. Cambridge: MIT Press.

Hoepner, A. G. F., Oikonomou, I., Sautner, Z., Starks, L. T., & Zhou, X. Y. (2024). ESG shareholder engagement and downside risk. *Review of Finance*, *28*(2), 483–510. https://doi.org/10.1093/rof/rfad034

Hosaka, T. (2019). Bankruptcy prediction using imaged financial ratios and convolutional neural networks. *Expert Systems with Applications*, *117*, 287–299. https://doi.org/10.1016/j.eswa.2018.09.039

IMF. (2021). *World Economic Outlook Update: Managing Divergent Recoveries*. International Monetary Fund.

Jacobson, T., & von Schedvin, E. (2015). Trade Credit and the Propagation of Corporate Failure: An Empirical Analysis. *Econometrica*, *83*(4), 1315–1371. https://doi.org/10.3982/ECTA12148

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 3149–3157.

Kim, M.-J., Kang, D.-K., & Kim, H. B. (2015). Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction. *Expert Systems with Applications*, *42*(3), 1074–1082. https://doi.org/10.1016/j.eswa.2014.08.025

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, *247*(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Loughran, T., & Mcdonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, *66*(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Merton, R. C. (1974). On the Pricing of Corporate Debt: The Risk Structure of Interest Rates. *The Journal of Finance*, *29*(2), 449–470. https://doi.org/10.1111/j.1540-6261.1974.tb03058.x

Moscatelli, M., Parlapiano, F., Narizzano, S., & Viggiano, G. (2020). Corporate default forecasting with machine learning. *Expert Systems with Applications*, *161*, 113567. https://doi.org/10.1016/j.eswa.2020.113567

Narvekar, A., & Guha, D. (2021). Bankruptcy prediction using machine learning and an application to the case of the COVID-19 recession. *Data Science in Finance and Economics,* *1*(2), 180-195. https://pdfs.semanticscholar.org/5d9b/04f0b73d19ae28d3163d49fe2f2d11c9a248.pdf

OECD Economic Outlook, Volume 2020 Issue 2. (2020, December 1). OECD. https://www.oecd.org/en/publications/oecd-economic-outlook/volume-2020/issue-2_39a88ab1-en.html

OECD Financing SMEs and Entrepreneurs Scoreboard: 2023 Highlights. (2023, March 14). OECD. https://www.oecd.org/en/publications/oecd-financing-smes-and-entrepreneurs-scoreboard-2023-highlights_a8d13e55-en.html

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, *18*(1), 109-131. https://doi.org/10.2307/2490395

Özparlak, G., & ÖzdemiR DiLiDüzgün, M. (2022). CORPORATE BANKRUPTCY PREDICTION USING MACHINE LEARNING METHODS: THE CASE OF THE USA. *International Journal of Management Economics and Business*. https://doi.org/10.17130/ijmeb.1079688

Saha, Debashis & Navila, Rabeya. (2018). Bankruptcy Risk Prediction Using Altman Z-Score Model: An Empirical Study on Private Commercial Banks of Bangladesh. 7. 127-136.

Shumway, T. (2001). Forecasting Bankruptcy More Accurately: A Simple Hazard Model. *The Journal of Business*, *74*(1), 101–124. https://doi.org/10.1086/209665

Sun, J., Li, H., Huang, Q.-H., & He, K.-Y. (2014). Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. *Knowledge-Based Systems*, *57*, 41–56. https://doi.org/10.1016/j.knosys.2013.12.006

Xia, Y., Liu, C., Li, Y., & Liu, N. (2017). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications*, *78*, 225–241. https://doi.org/10.1016/j.eswa.2017.02.017

Zou, H., & Hastie, T. (2005). Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

8 Appendices

Appendix A Ethical approval letter



14 August 2025

Dear Miss Chen ,

Reference: CURES/26295/2025
Project ID: 29571
Title: Default prediction of unlisted firms

Thank you for your application to the Cranfield University Research Ethics System (CURES).

We are pleased to inform you your CURES application, reference CURES/26295/2025 **has been reviewed. You may now proceed with the research activities you have sought approval for.**

If you have any queries, please contact CURES Support.

We wish you every success with your project.

Regards,

CURES Team

Appendix B Appendix guidance

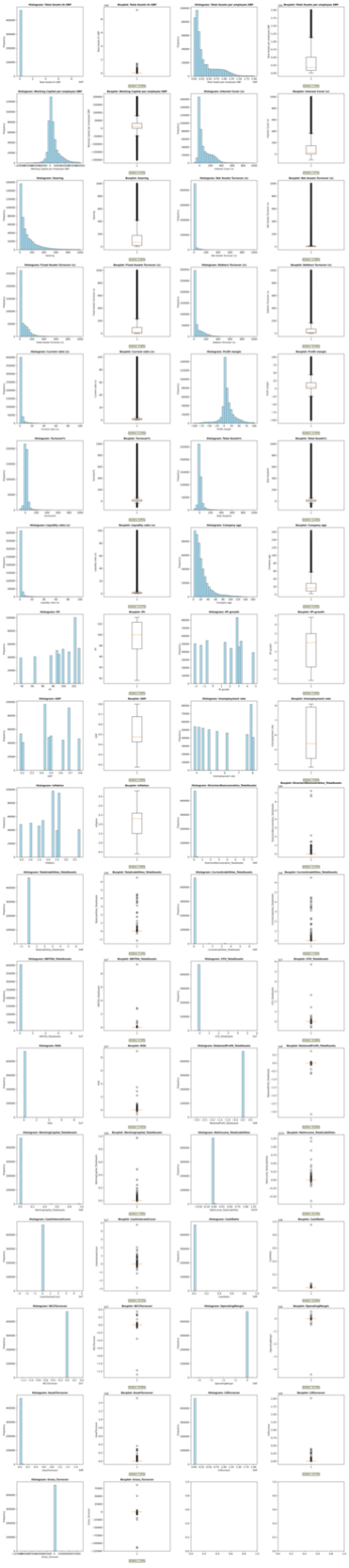
B.1 Analysis of Outliers

Variable	Outlier Percentage
RetainedProfit_TotalAssets	20.83799583
WC2Turnover	19.65869496
CashInterestCover	19.14633395
ROA	17.81330355
OperatingMargin	16.76373852

NetIncome_TotalLiabilities	16.7011001
Total Assets th GBP	15.09224834
Total Assets%	12.92071251
Net Assets Turnover (x)	12.87718412
Profit margin	12.44657156
DirectorsRemuneration_TotalAssets	11.54585238
Current ratio (x)	11.26344868
Liquidity ratio (x)	11.0714991
CFO_TotalAssets	10.95641428
CashRatio	10.68717526
CATurnover	10.53323339
Debtors Turnover (x)	10.44171761
Working Capital per employee GBP	9.533779092
EBITDA_TotalAssets	9.192771315
Turnover%	8.846030529
WorkingCapital_TotalAssets	7.587529222
Gearing	6.661754682
Company age	6.286561193
AssetTurnover	5.908607368
CurrentLiabilities_TotalAssets	5.346347968
Gross_Turnover	5.241455165

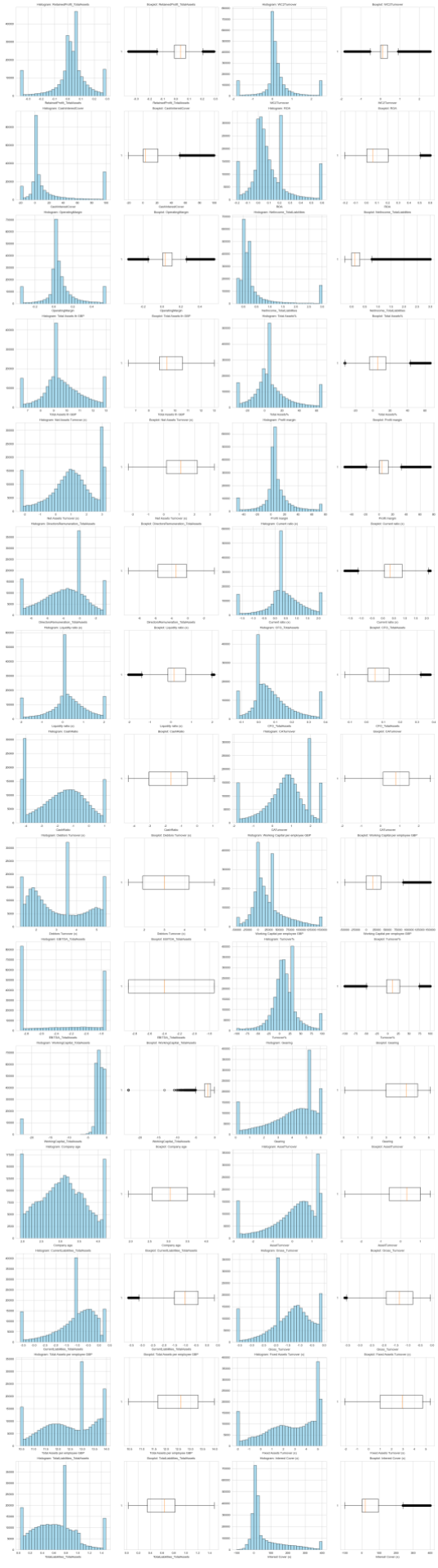
Total Assets per employee GBP	5.237208492
Fixed Assets Turnover (x)	4.661147408
TotalLiabilities_TotalAssets	4.234781519
Interest Cover (x)	3.267601925
Inflation	0
GDP	0
IPI growth	0
IPI	0
Unemployment rate	0

B.2 Visualisations of Raw Outliers



B.3 Visualisations of Results After Handling Outliers

Processed Variables Distribution



B.4 Pearson correlation

