PhD-FSTM-2025-006
The Faculty of Science, Technology and Medicine

DISSERTATION

Defense held on 10 February 2025 in Luxembourg

to obtain the degree of

DOCTEUR DE L´UNIVERSITÉ DU LUXEMBOURG
EN INFORMATIQUE

by

XINLIN WANG

Born on 2nd January 1992 in Anhui, China

From Data to Decision: Enhancing SME Financial Health Prediction with Advanced Machine Learning Techniques

Dissertation defense committee

Dr. Mats BRORSSON, Dissertation supervisor
Research Scientist, Université du Luxembourg

Dr. Jacques KLEIN, Chair
Professor, Université du Luxembourg

Dr. Zsófia KRÄUSSL
Senior Lecturer, City University of London - Bayes Business School

Dr. Manxing DU
Senior Researcher, HOOTSUITE, Luxembourg

Dr. Radu STATE, Vice Chair
Professor, Université du Luxembourg

Dedicated

to

My Families

ACKNOWLEDGMENTS

This doctoral thesis has been performed at *The Interdisciplinary Centre for Security, Reliability and Trust (SnT) at the University of Luxembourg.* This research was funded by Yoba S.A. and Luxembourg National Research Fund (FNR), grant reference 15403349.

I would like to express my deepest gratitude to everyone who has contributed to the success of this project. First and foremost, I am incredibly thankful for the opportunity to work on this challenging and inspiring project, which has significantly enriched my academic and personal growth.

I owe my sincere thanks to my supervisor, Dr. Mats Brorsson, whose support, insightful guidance, and patience have been indispensable throughout this journey. His expertise and encouragement have not only shaped my research but have also motivated me to strive for excellence.

Appreciation is extended to my colleagues and friends, particularly my PhD pal, Braulio Blanco. Thank you for your collaborative spirit, constructive feedback, and the stimulating discussions shared, which have greatly enhanced the quality of this project. It has been a privilege to work alongside such talented and dedicated individuals. Profound gratitude is expressed to the individuals met in Luxembourg, especially Yidi Yang, whose constant support, warmth, and insight have guided this journey. Her gentle encouragement and unwavering belief in my potential have been a quiet, yet powerful force during the highs and lows of these years.

Finally, I am forever grateful to my parents and grandparents. Their endless love, understanding, and sacrifices have been the cornerstone of my success. Although we are thousands of miles apart, our daily video calls, where we share our everyday lives, are a constant source of strength for me. Without their unwavering support, this achievement would not have been possible.

Thank you all for being an essential part of this journey.

ABSTRACT

Small and medium-sized enterprises (SMEs) are critical to global economic stability; however, they are particularly vulnerable to financial risks and bankruptcy. This dissertation focuses on enhancing SME financial risk prediction through advanced data-driven methods. Leveraging financial and non-financial datasets, this research aims to address the limitations of traditional bankruptcy prediction models and to develop a robust, automated credit reporting system tailored to SMEs. The research begins with a thorough literature review that establishes a taxonomy of datasets used in bankruptcy prediction and highlights key challenges related to data quality and integration. It then introduces an automatic feature engineering (AFE) framework to extract meaningful features from financial data, outperforming traditional financial ratio-based approaches. Further exploration of large language models (LLMs) for financial analysis demonstrates their potential in calculating financial ratios, conducting the Altman Z-score model and DuPont analysis, and predicting bankruptcy risk and key financial indicators with enhanced accuracy under optimized settings. Expanding beyond financial data, this dissertation integrates company adjustment behavioral data into hybrid datasets. Through uplift modeling and machine learning techniques, it reveals how non-financial factors significantly influence financial health. Considering the dynamic nature of company adjustments, MTDnet is proposed to estimate the uplift with multiple time-dependent treatments. It outperforms other uplift models, establishing the necessity of considering the sequence of treatments. These findings underscore the importance of combining financial and non-financial data for comprehensive financial risk assessment. The culmination of this research is the design of an automated credit reporting system that synthesizes financial ratios, company adjustments, and predictive analytics into actionable insights. This system offers SMEs and stakeholders a scalable, data-driven tool for real-time analysis of financial health and bankruptcy risk, fostering informed decision-making and proactive risk management. By advancing methods in feature engineering, hybrid datasets, uplift modeling, and the application of LLMs, this dissertation contributes to the interdisciplinary field of data science and financial risk management. It highlights the transformative potential of integrating diverse data sources and cutting-edge technologies, paving the way for more accurate, transparent, and equitable financial systems for SMEs worldwide.

Keywords: Data-driven financial health assessment, feature engineering, uplift modeling, large language model, automated credit reporting system, bankruptcy prediction

Contents

Acknowledgments ii
Abstract iii

1 Introduction **1**
1.1 Research Context and Motivation 1
1.1.1 SMEs in the Economy and Importance of Their Financial Health 1
1.1.2 Developmental Overview of Bankruptcy Risk Prediction and Corresponding Challenges 3
1.2 Research Questions and Significance 5
1.3 Research Objectives and Scope 7
1.4 Contributions 9
1.5 Dissertation Structure 10

2 Literature Review **14**
2.1 Existing Datasets and Taxonomy 14
2.1.1 Survey of Public Available Datasets Used by Advanced Bankruptcy Prediction 15
2.1.1.1 American Dataset 16
2.1.1.2 Polish Dataset 16
2.1.1.3 Taiwanese Dataset 17
2.1.1.4 SMEsD 17
2.1.1.5 HAT 17
2.1.1.6 Other Publicly Available Datasets with A Subscription Fee 18
2.1.2 Taxonomy of Datasets Used for Advanced Bankruptcy Prediction 20
2.2 Current Methods for SMEs' Financial Health and Bankruptcy Prediction 24
2.2.1 Traditional Financial Ratio-based Methods 24
2.2.2 Deep Learning and Machine Learning Methods 26
2.2.3 Other Advanced Methods 29
2.2.3.1 Uplift Modeling 29

2.2.3.2 Large Language Models in Financial Statement Analysis 31
2.3 Summary of Findings and Research Gaps 32

3 Data collection **35**
3.1 Introduction of Raw Data from LBR 35
3.2 Financial Data 37
3.3 Non Financial Data 39

4 Automatic Feature Engineering: Enhancing the Use of Financial Statements for Bankruptcy Prediction **40**
4.1 Problem Statement 41
4.2 Methodology 43
4.2.1 Overview for Automatic Feature Engineering 43
4.2.2 Feature Generation 45
4.2.3 Termination Condition 46
4.3 Experiments 46
4.3.1 Comparison with Feature Sets Generated by Other Methods . 49
4.3.2 Models for Evaluating Different Feature Generation Approaches 50
4.3.3 Feature Performance Indicators for Comparing Feature Contribution 51
4.4 Results and Discussion 52
4.4.1 Comparison with Financial Ratios 52
4.4.1.1 Model Performance 52
4.4.1.2 Feature Performance 52
4.4.2 Comparison with DeepFM 53
4.4.3 Comparison with Deep Feature Synthesis 55
4.4.4 Comparison with Raw Data 56
4.4.5 Explainability and Extensibility 57
4.4.5.1 Explainability 57
4.4.5.2 Extensibility 57

5 Leveraging Large Language Models to Analyze Financial Statements **59**
5.1 Problem Statement 59
5.2 Experimental Design 60
5.2.1 Dataset and Data Preprocessing 60
5.2.2 Methodology 63
5.2.3 Evaluation Metrics 66
5.3 Results analysis 67

5.3.1 Answers completion . . . 67
5.3.2 Evaluation on calculation steps . . . 67
5.3.3 Financial Metric Calculation Accuracy . . . 69
5.3.4 Bankruptcy Prediction . . . 70
5.3.5 EBITDA and Sales Forecasting . . . 71
5.3.6 Resources Consumption . . . 72

6 Incorporating Company Adjustments: Hybrid Datasets to Improve the Bankruptcy Risk Prediction 73
6.1 Problem Statement . . . 73
6.2 Methodology . . . 74
6.2.1 Conceptual Framework . . . 75
6.2.2 Variables . . . 77
6.3 Experimental Setup . . . 79
6.3.1 Dataset Description . . . 79
6.3.2 Models . . . 80
6.3.3 Performance Evaluation . . . 83
6.4 Results & Discussion . . . 84
6.4.1 Features Evaluation . . . 84
6.4.2 Ablation Experimental results . . . 85
6.4.3 Performance about Covid Period . . . 86

7 Causal Insights: Uplift Modeling of Company Adjustments and Financial Health 89
7.1 Problem Statement . . . 89
7.2 Methodology . . . 91
7.2.1 Research Questions and Data Mapping . . . 91
7.2.2 MTDnet Framework . . . 92
7.2.3 Metrics . . . 94
7.2.4 Uplift Models . . . 95
7.2.5 Implementation Details . . . 96
7.3 Results and Discussion . . . 97
7.3.1 RQ1: Identifying Effective Company Adjustment . . . 97
7.3.2 RQ2: Necessity of Considering the Sequence of Treatments . . . 98

8 Towards Automation: Designing a Credit Reporting System for SMEs 100
8.1 Problem statement . . . 100
8.2 Methodology and Workflow . . . 101
8.3 Key Features of the Automated Credit Report . . . 102

8.4 Potential Use Cases 103
8.5 Challenges and Limitations 104

9 Conclusion and Future Work **105**
9.1 Summary of Key Findings 105
9.2 Limitations and Future Work 108
9.2.1 Limitations 108
9.2.2 Future work 108

10 Appendix **110**
10.1 A 110
10.2 B 116
10.3 C 117
10.4 D 120
10.5 E 122

Bibliography **149**

List of Figures

1.1 Roadmap of this dissertation 12

2.1 Distribution of survey papers according to taxonomy 23

3.1 Example of raw data webpage from LBR 36
3.2 Distribution of different types of reported files 37
3.3 Distribution of financial statement disclosures among companies. . . . 38
3.4 Distribution of company adjustment acts 39

4.1 Pipeline for automatic feature engineering process 44
4.2 AUC improvement of AFE compared to FR on each dataset 53
4.3 Feature importance of AFE and FR on each dataset 54
4.4 Information value of features in AFE and FR on each dataset 54
4.5 AUC improvement of AFE compared to DFS with all the features on each dataset 56
4.6 AUC improvement of AFE compared to DFS with the selected features on each dataset 57
4.7 AUC improvement of AFE compared to raw data on each dataset . . 58

5.1 Workflow of constructing datasets for training and testing. 61
5.2 Workflow of experimental structure. 63
5.3 Distribution of completion rate over different combinations of LLMs and optimization techniques. 68
5.4 sMAPE for financial ratios by over different combinations of LLMs and optimization techniques. 70

6.1 Conceptual framework of experimental design 75
6.2 Bankruptcy rate of three datasets from 2012 to 2021 78
6.3 Definition of time period 78
6.4 Structure of MLP 82
6.5 IV for features created from AFE, FR and RB 85
6.6 ROC curve of lightGBM and LSTM on 1-year datasets 86

7.1 Framework of multiple time-dependent uplift modelling 94
7.2 Uplift comparison for different adjustment 98

10.1 Average time consumption for each answer. 117
10.2 Average CPU memory consumption for inference. 117
10.3 Average GPU consumption for inference. 118
10.4 The rest results for ROC curve of 6 models on 3 datasets 121

List of Tables

2.1 Summary of related studies on bankruptcy prediction 27
2.2 Description of various treatments and studies. 30

4.1 Summary of sub-datasets sampled according to n consecutive years . 48
4.2 The set of financial ratios used in this chapter 49
4.3 AUC of models trained on AFE and DeepFM 55

5.1 Summary of constructed datasets for LLMs' experiments 63
5.2 Settings of LLMs 65
5.3 ROUGE-L comparison of different combinations of LLMs and optimization techniques 69
5.4 Performance evaluation for bankruptcy prediction by LLMs 71
5.5 Comparison of the forecasting ability of LLMs and financial expert. . 72

6.1 Summary of three datasets sampled by sliding time window approach 78
6.2 Description of variables in Chapter 4 79
6.3 Description of training, testing, Pre-Covid and Post-Covid datasets . 80
6.4 Information of training machine 80
6.5 Range of parameters for LR models 81
6.6 Range of parameters for RF models 81
6.7 Range of parameters for LGB models 81
6.8 AUC of models on testing, pre-Covid and post-Covid sets 88

7.1 Description of the four datasets for uplift models 92
7.2 Scope of hyperparameters used in optimizing the model performance 97
7.3 Results of the experiments for evaluating the necessary of considering treatment sequence. 99

10.1 Summary for related studies 111
10.2 Comparison of the forecasting ability of LLMs and financial expert. . 116

Chapter 1

Introduction

Contents

1.1 Research Context and Motivation **1**
1.1.1 SMEs in the Economy and Importance of Their Financial Health 1
1.1.2 Developmental Overview of Bankruptcy Risk Prediction and Corresponding Challenges 3
1.2 Research Questions and Significance **5**
1.3 Research Objectives and Scope **7**
1.4 Contributions **9**
1.5 Dissertation Structure **10**

1.1 Research Context and Motivation

1.1.1 SMEs in the Economy and Importance of Their Financial Health

Small and medium-sized enterprises (SMEs) hold significant importance in the economy, making significant contributions to employment, innovation, and economic development. Their importance is evident across diverse economies. SMEs contribute significantly to the GDP of many countries. They are often regarded as the backbone of local communities and play a crucial role in job creation and innovation. According to the World Bank, SMEs constitute approximately 90% of all companies and play a substantial role in generating over 50% of global employment [15]. In Europe, SMEs represent 99.8% of all businesses in the non-financial sectors and employ approximately 88.7 million people, making them a cornerstone of the EU's economic structure [76]. Similarly, in the

United States, SMEs represent 99.9% of all businesses and employ approximately 61.6 million people, which equates to nearly 45.9% of the US workforce [125]. In developing economies, SMEs are equally pivotal. They not only create job opportunities but also act as vehicles for poverty reduction and economic inclusion by enabling individuals from underprivileged backgrounds to participate in entrepreneurial activities.

SMEs promote competition, drive innovation, and increase productivity within the market, which is essential for economic growth and development. These businesses are often agile, making them capable of rapidly adopting new technologies and business models so they play a leading role in sectors such as software development, renewable energy, and digital platforms. SMEs operate predominantly within localized communities, providing employment opportunities to local residents and acting as drivers of economic advancement by fostering competition, innovation, and increased productivity. In addition, SMEs demonstrate a strong sense of social responsibility and commitment to sustainability, frequently prioritizing community engagement and environmental preservation [163].

The financial health of SMEs is an essential determinant of their ability to sustain operations, scale, and contribute effectively to the broader economy. Predicting SMEs' financial health, especially bankruptcy risk, can help enterprises to proactively identify potential risks, adapt their business strategies, and improve their overall competitiveness and stability in a timely manner. Furthermore, bankruptcy risk prediction can facilitate the assignment of credit ratings for SMEs, providing credit endorsements, and enabling them to access better financial services as limited access to affordable financing remains a critical bottleneck.

The financial health of SMEs is not just a concern for the businesses themselves; it has far-reaching implications for the overall economy. Financially healthy SMEs can invest in infrastructure, adopt innovative technologies, and expand their market reach. This, in turn, improves productivity, increases employment opportunities, and enhances national economic performance. During economic downturns, the collapse of SMEs can trigger widespread unemployment and disrupt supply chains, amplifying the impact of economic crises. This was evident during the COVID-19 pandemic, during which the global SME sector faced unprecedented disruptions. In response, governments worldwide implemented relief measures, such as low-interest loans, wage subsidies, and grant programs, to safeguard SME operations and preserve employment [2]. Given their critical role, ensuring the financial health of SMEs should remain a top priority for policymakers, financial institutions, and international organizations.

1.1.2 Developmental Overview of Bankruptcy Risk Prediction and Corresponding Challenges

The assessment of bankruptcy risk represents a critical area of research and practice within finance and business management. This process evaluates the likelihood of a company failing to meet its financial obligations, resulting in bankruptcy or liquidation. The accurate assessment of bankruptcy risk is essential for various stakeholders, including investors, creditors, regulators, and the companies themselves, as it informs decision-making processes related to credit risk management, investment strategies, regulatory oversight, and corporate governance. Over the decades, numerous models have been developed to assess bankruptcy risk, ranging from traditional financial ratios to sophisticated machine learning algorithms. The development of these models has been driven by the need for more accurate and reliable predictions, especially in the wake of financial crises that have exposed the limitations of conventional methods.

The earliest models for bankruptcy risk assessment, developed in the late 1960s, relied heavily on financial ratios derived from balance sheets and income statements. Financial ratios are accounting-based ratios used to assess the financial health of a company, typically derived from its financial statements [150]. Pioneering work in this area includes the Logit model proposed by Beaver in 1966 [18] and the Z-score model proposed by Altman in 1968 [8]. Both models were formulated using financial ratios and played a significant role during their respective periods, establishing a fundamental framework for subsequent investigations into the prediction of bankruptcy. Despite their simplicity and ease of use, these models often lack predictive accuracy and fail to capture the dynamic and multifaceted nature of financial distress.

With the development of the financial industry and the field of data science, numerous studies have attempted to make a breakthrough by applying new models into bankruptcy prediction, and directly use well-calculated financial ratios to find the most predictive model and make predictions, moving from basic financial ratio analysis to more complex models [8, 156, 155, 165, 36, 85, 6, 27]. Models based on decision trees, neural networks, and support vector machines have emerged, providing researchers with greater flexibility and higher accuracy in prediction.

These methods, whether based on financial experts or machine learning methods, improved predictive power but were still limited by their reliance on historical financial data and full financial statements. Not all companies are able to provide their complete historical data for financial calculations. This phenomenon often holds for SMEs. Due to the nature or maturity of their business, SMEs often lack complete financial statements, making it difficult to calculate financial ratios and conduct comprehensive risk assessments. This frequently results in high rates of declined credit applications. The credit risk assessment of SMEs is often conducted using financial ratios, which might not fully reflect the company's operational dynamics and potential for restructuring or recovery. There is a need for more comprehensive and holistic approaches to bankruptcy prediction that consider a wider range of data, like company adjustment acts and other non-financial indicators, instead of purely relying on financial ratios to encompass all significant aspects of a company's financial health.

As the economic landscape becomes increasingly complex, the traditional reliance on financial ratios to assess the financial health of businesses has been supplemented by a broader spectrum of data types. Also, advanced models can process vast amounts of data, including non-financial indicators such as market sentiment, macroeconomic factors, and corporate governance practices. More recent advances have seen the development of ensemble methods and hybrid models that combine multiple algorithms to enhance predictive performance. These approaches leverage the strengths of different models to address their individual weaknesses, resulting in more robust and reliable insolvency risk assessments. This dissertation aims to analyze and predict the financial health and bankruptcy risk of SMEs using both financial and non-financial data.

1.2 Research Questions and Significance

This dissertation focuses on the application of advanced techniques to comprehensively improve the understanding and prediction of SMEs' financial health and bankruptcy using both financial and non-financial data. To achieve this, a series of research questions have been formulated from a data perspective to guide the investigation and ensure a structured approach to problem-solving.

This dissertation begins by focusing on data, which serves as the foundation for this research topic, and formulates three research questions.

RQ1.1: What types of datasets can be used in bankruptcy prediction and what are their characteristics?

RQ1.2: How does the diversity of datasets affect the efficiency and accuracy of bankruptcy prediction?

RQ1.3: What are the challenges and limitations associated with different types of datasets in bankruptcy prediction?

With the development of data science, researchers have begun to use different types of data, from traditional financial indicators to unstructured data, to construct and optimize bankruptcy prediction models. Understanding and enhancing the taxonomy and quality of datasets are critical for building robust and accurate models. By identifying the effective datasets and addressing integration challenges, this research can provide better predictions for SMEs, which often lack a comprehensive financial history and financial statements.

Then, this dissertation continues to raise research questions from the perspectives of both financial and non-financial data. Due to the limitations of SMEs' financial data, the following research questions are explored.

RQ2.1: For SMEs with limited or incomplete financial data, how can existing data be fully utilized to assess the financial health or bankruptcy risk of the company?

RQ2.2: How effective is automatic feature engineering (AFE) in predicting bankruptcy from financial statements, especially when compared to domain knowledge-based financial ratio methods and other machine learning methods?

AFE methods can generate high-quality, explainable features that enhance prediction accuracy, particularly for SMEs with limited financial histories. This addresses a key limitation of traditional models that rely heavily on expert-defined financial ratios.

With the advancement of large language models (LLMs), several explorations have been undertaken in this field.

RQ3.1: Can LLMs accurately process financial data and calculate financial ratios based on the provided financial statement data?

RQ3.2: How effectively can LLMs predict critical financial indicators and bankruptcy risk using methodologies such as the Altman Z-score model and DuPont analysis?

RQ3.3: How do different optimization techniques (e.g., RAG, fine-tuning) affect the performance of LLMs in financial risk prediction?

LLMs represent a cutting-edge approach to financial analysis, offering the potential for automation and scalability. However, understanding their strengths, limitations, and optimization strategies is critical for their successful application in financial statement analysis.

Meanwhile, non-financial data is also an indispensable part of comprehensively assessing the financial health of a company. This dissertation focuses on the aspect of company adjustment acts.

RQ4.1:How can a hybrid dataset that combines financial ratios and company adjustment acts improve the accuracy of bankruptcy prediction?

RQ4.2:Which machine learning model performs best when using a hybrid dataset, and what are the implications for practical adoption?

This dissertation employs causal inference to further investigate the impact of company adjustment acts on financial health.

RQ5.1: How can uplift modeling be utilized to estimate the causal effects of company adjustments on financial health? What types of company adjustment acts help improve financial status and prevent bankruptcy?

RQ5.2: Should the sequence and timing of company adjustment acts be considered when measuring their impact on financial health? How does the proposed MTDnet framework compare to existing uplift modeling techniques in capturing the time-dependent nature of company adjustments?

Company adjustments are the key to preventing financial distress, but their effectiveness depends on timing and sequence. Uplift modeling provides actionable insights into these adjustments, enabling SMEs to implement targeted strategies for financial resilience.

These research questions form the foundation of the dissertation, guiding the exploration of the problems and the development of solutions. They encompass the theoretical foundations, methodological approaches, empirical investigations, and practical applications of financial health assessment for SMEs. By addressing these questions, the dissertation aims to contribute to the body of knowledge in financial risk management and data science.

1.3 Research Objectives and Scope

This dissertation addresses a critical gap in the current financial health and bankruptcy prediction practices for SMEs. Financial health focuses on a broad evaluation of a company's performance and stability, while bankruptcy risk is a narrower concept that serves as a binary outcome prediction. They are closely related and interconnected concepts in corporate finance and risk management, so in this dissertation, both cases will be discussed. Traditional methods of bankruptcy prediction, which rely heavily on financial ratios and historical data, may not fully capture the complexity and dynamic nature of SMEs' financial health. Integrating diverse data sources and applying advanced machine learning techniques present promising avenues for enhancing the predictive accuracy and robustness of bankruptcy risk models. The primary research samples are SMEs in Luxembourg, with data sourced exclusively from the Luxembourg business register[1](LBR). Based on SMEs in Luxembourg, this dissertation focuses on developing a general holistic framework to enhance the prediction of financial health and bankruptcy for SMEs that integrates diverse data sources and leverages advanced machine learning techniques.

Specifically, the dissertation aims to:

- To provide a taxonomy that categorizes the datasets used in bankruptcy

[1] www.lbr.lu

prediction research.

- To provide insights that help researchers and practitioners select the appropriate datasets for their bankruptcy prediction studies.
- To develop an automated feature engineering approach for any financial statement quality level, that generates effective, explainable, and extensible features for bankruptcy prediction models.
- To assess the effectiveness of LLMs in analyzing financial statements and identify the most effective combination of LLMs and optimization techniques for financial statement analysis.
- To improve bankruptcy prediction models by combining financial statements with company adjustment acts.
- To uncover the causal relationship between company adjustment acts and their effects on company financial health.
- To estimate the effect of what an adjustment act may lead to in terms of financial health and then provide insights into corporate governance to prevent financial distress.
- To propose an efficient framework (MTDnet) to estimate the individual effect with multiple time-dependent treatments.
- To disseminate the findings and insights from this research to the academic community, industry professionals, and policy makers, thereby contributing to the body of knowledge in financial risk management and data science.

By addressing these objectives, the research seeks to contribute to the field of bankruptcy risk assessment by providing a more nuanced and data-driven approach that can better serve the needs of SMEs in navigating financial challenges and preventing bankruptcy. This research extends the application of data science in the field of financial risk prediction, in particular by providing new theoretical perspectives and methodological support for financial prediction models for SMEs. Furthermore, the research results provide practical tools and techniques for SMEs in Luxembourg, assisting them in predicting financial risks more accurately. This, in turn, enhances their viability and competitiveness in the market.

1.4 Contributions

This research contributes significantly to the field of financial health and bankruptcy prediction, particularly for SMEs, by addressing existing gaps and proposing innovative solutions. The first contribution lies in the creation of a comprehensive taxonomy and evaluation framework for datasets used in bankruptcy prediction. By systematically categorizing and assessing accounting-based, market-based, macroeconomic, relational, and non-financial data, the research highlights the importance of integrating diverse datasets to improve prediction accuracy, particularly for SMEs with sparse financial data.

A second contribution is the development and validation of an automatic feature engineering approach tailored to SMEs. Unlike traditional methods that rely on predefined financial ratios, the AFE approach dynamically generates explainable, high-quality features, enabling more accurate predictions even in cases of incomplete financial records. This contribution not only enhances prediction models but also provides a scalable methodology applicable across diverse datasets and markets.

This research further advances the use of LLMs in financial analysis by systematically evaluating their capabilities in processing financial data, calculating key metrics, and predicting financial risks. By identifying the strengths and limitations of different LLM models and optimization strategies, the research contributes to the growing body of knowledge on how data science can be effectively applied in financial contexts.

The research also introduces the concept of hybrid datasets that combine financial ratios with company adjustment behaviors. By demonstrating the superior performance of these datasets in bankruptcy prediction, particularly during economic disruptions, the study offers a novel perspective on how behavioral data can complement traditional financial indicators to provide a holistic view of company health.

Another significant contribution is the application of uplift modeling to assess the impact of company adjustments on financial health. The development of the MTDnet framework, which incorporates temporal dynamics into uplift modeling, represents a methodological advancement. This contribution enables a more nuanced understanding of how the timing and sequence of company adjustments influence financial outcomes, providing actionable insights for proactive risk management.

Lastly, the research culminates in the design and implementation of an automated credit reporting system that integrates all these findings into a unified framework. This system provides a practical tool for generating comprehensive and actionable credit reports for SMEs, facilitating better risk assessment and decision-making by lenders and policymakers. By bridging theoretical advancements with real-world applications, the research offers a holistic solution for improving financial risk prediction and management for SMEs.

This dissertation contains the following papers published as first authors. The papers are listed in the order in which they appear in this dissertation.

- *Datasets for Advanced Bankruptcy Prediction: A survey and Taxonomy* Published in arXiv. [169](Included in Chapter 2)
- *Effective Automatic Feature Engineering on Financial Statements for Bankruptcy Prediction* Presented and published in 2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME) [170] (Included in Chapter 4)
- *Can Large language model analyze financial statements well?* Presented and published in the 31st International Conference on Computational Linguistics [168] (Included in Chapter 5)
- *Augmenting Bankruptcy Prediction Using Reported Behavior of Corporate Restructuring* Presented and published in BenchCouncil International Symposium on Intelligent Computers, Algorithms, and Applications [166] (Included in Chapter 6)
- *Which Company Adjustment Matter? Insights from Uplift Modeling on Financial Health* Presented and published in Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2024) [167](Included in Chapter 7)

1.5 Dissertation Structure

This dissertation discusses a series of models and methodologies designed to adapt or enhance available datasets for assessing the financial health of SMEs, particularly in Luxembourg. A novel approach is introduced to improve the feature engineering process for data from financial statements. Several well-known models for bankruptcy prediction are compared and analyzed, not only on the financial datasets but also on the non-financial datasets and the hybrid datasets. Finally, the causal relationship between company adjustment acts and its financial health is explored by uplift modeling. Fig. 1.1 shows the structure of the remaining chapters in this dissertation. It starts with Chapter 2 and Chapter 3 to introduce the background and data-related information. Then Chapter 4 and Chapter 5 explain how the advanced techniques are proposed and applied to make the best use of financial data. Chapter 6 and Chapter 7 describe the exploration of non-financial data. In Chapter 8, an automated system is introduced to generate the credit report, which contains the research results derived from both financial data and non-financial data. Finally, this dissertation ends with Chapter 9 reaching to the conclusions.

Chapter 2 reviews existing research related to financial risk prediction, particularly focusing on the application of machine learning and deep learning models in this area. It provides a comprehensive review of datasets used in bankruptcy prediction, proposing a taxonomy to categorize them and metrics to assess their quality and informativeness. It highlights the importance of data quality for accurate bankruptcy forecasting and discusses the challenges of using different types of data, including accounting-based, market-based, macroeconomic, relational, and non-financial datasets. This study emphasizes the necessity of careful data integration and model evaluation in bankruptcy prediction research.

Chapter 3 introduces the raw data and constructed datasets that are used in this dissertation. It also explains the characteristics of the data and the data preprocessing methods.

Chapter 4 introduces an automated feature engineering (AFE) approach for bankruptcy prediction that outperforms traditional methods. It uses financial data to generate effective features for model training, with better performance compared to financial ratios and other machine learning techniques. This approach aims to improve bankruptcy prediction for professionals who may not possess the necessary engineering expertise or efficient data.

Chapter 5 evaluates the effectiveness of LLMs in financial analy-

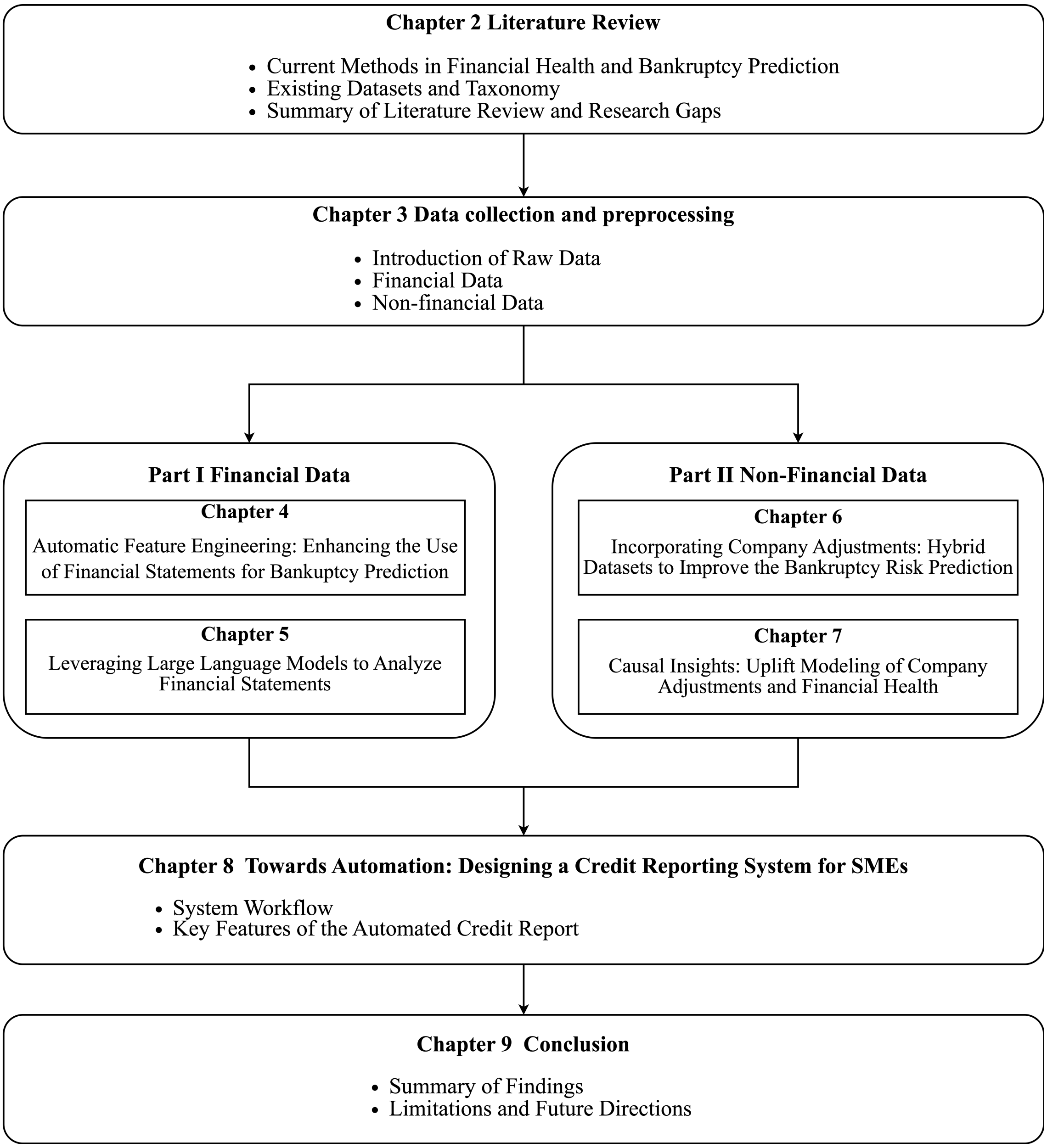


Figure 1.1: Roadmap of this dissertation

sis, including calculating financial ratios, predicting bankruptcy risks, and forecasting financial indicators. It compares combinations of LLMs (Llama 3.2 3B, Llama 3.1 8B, and Mistral 7B) and optimization techniques (zero-shot, few-shot, Retrieval Augmented Generation, and fine-tuning) against expert predictions and ground truth. The study finds that while LLMs show potential in financial analysis, they still lag behind human experts in forecasting complex financial metrics. It highlights the need for selecting the right model and optimization strategy based on task requirements and resource constraints.

Chapter 6 presents a new method for predicting the bankruptcy by combining financial statements with data on company adjustment behaviors. Utilizing a hybrid data set provides a more comprehensive and holistic perspective on a company's financial position and the dynamics of its business operations. The study demonstrates that using machine learning models with a hybrid dataset improves prediction accuracy by 4%-13% compared to financial data alone.

Chapter 7 applies uplift modeling to analyze how company adjustments affect financial health and bankruptcy risk. They introduce a new framework, MTDnet, to handle the complexity of time-dependent treatments. The study finds that information and business-related adjustments are most effective in improving financial health, while basic binary adjustments have the least impact. The research highlights the importance of considering the sequence and timing of company adjustments for accurate effect estimation.

Chapter 8 describes a system to automatically generate the credit report for SMEs. All the research fruits from previous chapters are included in this report, including the better way for credit scoring, risk alert from company adjustments, human-readable text generated by LLMs to analyze the financial statements.

Chapter 9 summarizes and concludes the results of the findings in this dissertation. Then it analyzes the limitations of this dissertation and makes recommendations for future research.

Chapter 2

Literature Review

Contents

2.1 Existing Datasets and Taxonomy **14**
2.1.1 Survey of Public Available Datasets Used by Advanced Bankruptcy Prediction 15
2.1.2 Taxonomy of Datasets Used for Advanced Bankruptcy Prediction 20
2.2 Current Methods for SMEs' Financial Health and Bankruptcy Prediction **24**
2.2.1 Traditional Financial Ratio-based Methods 24
2.2.2 Deep Learning and Machine Learning Methods 26
2.2.3 Other Advanced Methods 29
2.3 Summary of Findings and Research Gaps **32**

2.1 Existing Datasets and Taxonomy

With the emergence of advanced analytical techniques powered by machine learning and vast data availability, increasingly diverse datasets have been applied to bankruptcy prediction. To thoroughly investigate the area of bankruptcy prediction, this research first assesses and collects datasets comprehensively covering the scientific domain of this area. To ensure an unbiased environment for data collection, Google Scholar[1] was utilized. The search was conducted using 'bankruptcy prediction' and 'insolvency prediction' as primary query terms to identify and collect relevant datasets, drawing motivation from the seminal work of Bellovary et al. [19], which provides an in-depth survey of various methods for bankruptcy prediction. The term 'financial distress' was deliberately

[1] https://scholar.google.com/

excluded to avoid conflating it with bankruptcy, as financial distress pertains to operational challenges, whereas bankruptcy involves deliberate asset protection decisions by companies [134].

Relevant papers were identified using Google Scholar based on the following criteria: (1) relevance to bankruptcy or insolvency prediction, (2) detailed dataset descriptions, (3) the adoption of machine learning or deep learning methods for model training, and (4) publication within the past decade (2013–2023) or the inclusion of representative datasets predating 2013. This ten-year timeframe was selected to focus on recent developments in big data and machine learning and their implications for bankruptcy research.

To maximize the retention of diverse datasets, the investigation focused on significant works assessed by (1) citation count and (2) the quality of the publication venue, such as ranked conferences and journals. To maintain the diversity of the datasets that were selected for bankruptcy research, thus to create a rather robust environment for this assessment, we also included papers that were operating unique, problem-specific datasets yet achieved fewer citations. Ultimately, we searched for and included all papers that employed machine learning and deep learning techniques for bankruptcy prediction, independent of the type of dataset that was instrumentalized for the analysis. This research resulted, therefore, in a diverse selection of 47 papers in total, including both public and private datasets that combined both quantitative (e.g., numeric) and qualitative (e.g., text) variables. To assess and analyze the used data, we further needed to restrict our analysis to research papers that instrumentalized publicly available. The findings of the survey are summarized in the following subsections.

2.1.1 Survey of Public Available Datasets Used by Advanced Bankruptcy Prediction

After manually reviewing the papers found, we collected information to describe the characteristics of the instrumentalized datasets for bankruptcy prediction in as much detail as possible. Table 10.1 in the Appendix 10.1 structures the datasets of the collected 47 papers along the following variables: (1) publication year, (2) number of samples that the study uses, (3) bankruptcy rate, (4) number of features that are applied in the model, (5) data type, (6) data source, (7) publicly available? (8) fee applied?. The slash symbol "/" was used to indicate where information is not specifically retrievable from a particular paper. While many studies claim the datasets they employ are publicly available, some datasets are only conditionally accessible, such as with a subscription fee. It should be noted that it may actually cause a barrier for reviewers to revisit a particular study and/or to use a particular dataset. In the following, the publicly available datasets with free access identified in the reviewed studies are introduced and described.

2.1.1.1 American Dataset

This dataset was published by Lombardo et al. [106]on GitHub[2] in 2022. The authors collected accounting data from 8262 different companies in the period between 1999 and 2018 related to the public companies in the US stock market. A company is labeled "Bankruptcy" (1) if it filed for bankruptcy under Chapter 11 or Chapter 7 of the Bankruptcy Code in the subsequent fiscal year; otherwise, it is labeled "Alive" (0). The dataset comprises 78,682 firm-year observations without missing or synthetic values, divided into training (1999-2011), validation (2012-2014), and test (2015-2018) sets.

2.1.1.2 Polish Dataset

This dataset is published in the UCI machine learning repository, which can be accessed and downloaded[3] [161]. The dataset focuses on the bankruptcy prediction of Polish companies, collected from the Emerging Markets Information Service[4] (EMIS) covering bankruptcies from 2000 to 2012 and active companies from 2007 to 2013. This dataset doesn't show the original value of the financial statements but contains 64 financial ratios as model training features calculated from the financial statements. The response variable Y uses 0 to represent that the company does not go bankrupt and uses 1 to represent that the company is going bankrupt. Five classification cases are based on the forecasting period. The nth year means the data contains financial rates from the nth year of the forecasting period and the corresponding class label that indicates bankruptcy status after (5-n) years.

[2] https://github.com/sowide/bankruptcy_dataset
[3] https://archive.ics.uci.edu/dataset/365/polish+companies+bankruptcy+data
[4] http://www.securities.com

2.1.1.3 Taiwanese Dataset

This dataset is published in the UCI machine learning repository, and it can also be found in Kaggle[5] [1]. The dataset, spanning from 1999 to 2009, was collected from the Taiwan Economic Journal. Company bankruptcy is defined according to the business regulations of the Taiwan Stock Exchange. The dataset comprises 6,819 instances and 95 variables, with the response variable being binary: 220 companies are labeled as bankrupt and 6,599 as non-bankrupt, resulting in a bankruptcy rate of approximately 3.23%. A notable advantage of this dataset is the absence of missing values and the high quality of the variables, which comprehensively cover important financial ratios. However, a significant limitation is its relatively small size, which may be insufficient for training machine learning or deep learning models effectively.

2.1.1.4 SMEsD

This dataset was published by Zhao et al. [182] in 2024 on Github[6]. This dataset specially contains the information of lawsuits, which is very rarely public available data. SMEsD covers 3,976 SMEs and affiliated individuals in China, covering the period from 2014 to 2021. This database forms a comprehensive enterprise knowledge graph linking all enterprises and their related individuals by their basic information and lawsuit records from 2000 to 2021. Basic information for each company includes registered capital, paid-in capital, and establishment date. Lawsuit records provide details such as the involved plaintiff, defendant, subjects, court level, outcome, and timestamp. It can be used to split the out-of-time dataset for testing.

2.1.1.5 HAT

This dataset was collected in early October 2020 and published in 2021 [184], also available on GitHub[7]. The authors curated a real-world dataset encompassing the board member network and the shareholder network of 13,489 companies in China. Data was sourced from various public platforms. Specifically, they randomly sampled 1,000 companies located

[5] https://archive.ics.uci.edu/dataset/572/taiwanese+bankruptcy+prediction
[6] https://github.com/shaopengw/comrisk
[7] https://github.com/hetergraphforbankruptcypredict/HAT

in the same city that experienced bankruptcy in 2018. They then expanded the network by capturing information on all shareholders and board members associated with these firms. This process was repeated twice, resulting in the augmentation of the original 1,000-node network to a more extensive network comprising 13,489 nodes.

2.1.1.6 Other Publicly Available Datasets with A Subscription Fee

Bureau Van Dijk Bureau Van Dijk, a Moody's Analytics company, provides a range of business intelligence and financial datasets that are widely used in various industries for decision-making, risk assessment, and research. Some notable datasets offered by Bureau van Dijk include:

- Orbis: A comprehensive global company database featuring information on private and public companies worldwide, including detailed financial statements, ownership structures, and key business details.
- Amadeus: Focuses on European companies, this datasets provides detailed financial information and company profiles, which is useful for financial analysis, market research, and assessing business relationships.
- Fame: A UK and Irish-based database offering financial information on companies operating in the United Kingdom and Ireland, including financial statements, directors' details, and ownership structures.
- Zephyr: A global mergers and acquisitions (M&A) database, offering information on deals, rumors, and market rumors, valuable for analyzing M&A trends, deal structures, and market dynamics.
- Mint Global: Offers company information and financial for businesses globally, combining data from various sources. Useful for assessing business risk, conducting due diligence, and market research.
- TP Catalyst: Specializes in transfer pricing documentation and compliance data. Helps businesses navigate international transfer pricing regulations.

- Osiris: Focuses on emerging markets, providing financial information on companies in Asia, Latin America, the Middle East, and Africa. Supports financial analysis and risk assessment in these regions.
- Belfirst: A Belgium and Luxembourg-based database offering financial information on companies operating in Belgium and Luxembourg, and its datasource is National Bank of Belgium and Creditreform Luxembourg.

In addition, Belfirst and Fame datasets include relational data for bankruptcy prediction, as introduced by Tobback et al. [160]. They collected data from 2011 to 2014 covering more than 400,000 Belgian SMEs and 2,000,000 UK SMEs. This relational data, based on shared directors/managers, creates a directed graph where nodes represent companies and edges represent links between them.

These datasets from Bureau van Dijk are utilized by financial institutions, corporations, researchers, and analysts to gain insights into company performance, conduct market research, and manage risks associated with business operations and investments.

Compustat Compustat datasets are comprehensive financial databases that provide detailed information on publicly traded companies. These datasets are widely used by researchers, analysts, and financial professionals to conduct financial analysis, modeling, and research. Compustat is a product of Standard & Poor's (S&P) and is considered one of the leading sources of financial data.

Key features and components of Compustat datasets include:

Financial Statements: Compustat provides detailed financial statements, including income statements, balance sheets, and cash flow statements. These statements offer a comprehensive view of a company's financial performance over time.

Ratios and Metrics: The datasets include various financial ratios and metrics that help in evaluating a company's liquidity, profitability, and overall financial health. Common ratios such as return on equity (ROE), debt-to-equity ratio, and earnings per share (EPS) are included.

Segment Data: Compustat datasets often include segment-level information, allowing users to analyze the performance of different business segments within a company.

Stock Market Data: Information related to stock prices, trading volumes, and market capitalization is available, enabling the analysis of a company's stock performance and market trends.

Ownership Data: Compustat provides data on institutional ownership, allowing users to understand the distribution of a company's shares among different institutional investors.

Corporate Governance: Some Compustat datasets include information on corporate governance, executive compensation, and board composition, providing insights into the management structure of companies.

Global Coverage: While the primary focus is on U.S. companies, Compustat also includes data on international companies, expanding its coverage to a global scale.

Researchers and analysts use Compustat datasets to perform financial modeling, conduct valuation analyses, and gain insights into industry trends. The data is valuable for academic research, investment analysis, and strategic decision-making within the business and financial sectors.

Thomson Reuters Datastream Database Thomson Reuters Datastream is a global financial data platform developed and delivered by Thomson Reuters, one of the world's leading information service providers. The platform is focused on providing a wide range of economic, financial, and market data to professional investors, financial professionals, and researchers. This dataset can be publicly accessed with a subscription fee. The key indicators of macroeconomic data from Thomson Reuters Datastream are *Gross Domestic Product*, *Inflation Rate*, *Interest Rates*, *Trade Balance*, *Money Supply (M1, M2, M3)*, *Consumer Confidence Index* and so on.

2.1.2 Taxonomy of Datasets Used for Advanced Bankruptcy Prediction

It is a well-known fact in data science that the quality of the data determines the upper boundary of the model performance. Still, despite the importance of recognizing and acknowledging quality measures of the used data, as we argue, bankruptcy prediction research using machine learning has been dominated by a strong, model-driven focus. Numerous scholarly papers have been reviewing and addressing the performance of bankruptcy prediction models [13, 5, 148, 39, 73]. Both Alaka et al. [5] and Clement et al. [39] found that no single tool or model consistently outperforms others, suggesting that a hybrid approach may be more effective. The other paper points out that advanced machine learning methods appear to have the greatest promise for future research on firm failure [73]. In contrast to these scientific works, we analyze the various datasets that are instrumentalized for bankruptcy prediction. To our knowledge, this is the first study in the data science literature devoted to this subject.

After manually reviewing the papers, we detected a feature-driven pattern of the applied datasets, allowing us to classify the research papers on bankruptcy prediction. We conceptualized and summarized our findings by a *taxonomy*. Based on the keywords and the title of the surveyed papers, combined with the leading variables of the applied methodology for bankruptcy prediction, we classified the employed datasets into the following five categories: (1) accounting-based, (2) market-based, (3) macroeconomic, (4) relational, and (5) non-financial. Below we summarize and define the taxonomy in detail.

Accounting-based Datasets Traditional financial data is one of the most commonly used data sources in bankruptcy forecasting. These data include financial statements such as income statements, balance sheets, and cash flow statements, as well as data related to financial metrics such as operating income, net profit, and return on assets. These data sources provide important clues about the financial health of a company and are often used to build bankruptcy prediction models based on ratio analysis and statistical modeling.

Market-based Datasets Market-based indicators reflect the sentiment of investors and the overall market perception of a company's value. The key indicators include stock prices, credit ratings, and bond yields, which are examples of market-based indicators. Changes in these indicators may signal financial distress. Analysts incorporate market data to capture external perceptions and reactions to a company's financial health, which may not be fully reflected in financial statements alone.

Macroeconomic Factors Macroeconomic data sources include data related to the industry and macroeconomic environment, such as industry growth rates, gross domestic product (GDP), interest rates, and inflation rates. These data can help bankruptcy forecasting models to better consider the impact of the external environment on the business, thus improving the accuracy of the forecast.

Relational Data Enterprises are an important part of society and have a network of relationships in society that encompasses a wide range of relationships, including with other enterprises, government agencies, non-profit organizations, customers, suppliers, employees, and communities. These relationships have a significant impact on the success and sustainability of the enterprise, and they also reflect the state of the enterprise. We also found the transactional data is an important part to form relations between companies[87, 93], however, unfortunately there is no publicly available data upon that.

Non-financial Data This category is defined to encompass all other types of typically descriptive data, as noted in [9]. Based on the surveyed papers, the majority of instrumentalized datasets are private data sources; however, more studies instrumentalize annual reports of publicly listed companies. As a consequence of the further development of different language models, more studies started paying attention to the annexes of such annual reports. The information that is extracted from these annexes can serve as sentiment variables for bankruptcy prediction, as explained by Mai et al. [112].

By reviewing the features of the applied methodologies of the papers, we can indirectly refer to the characteristics of non-financial datasets. According to our review, fundamental information about a company, such as registration date, location, legal form, ownership structure, product information, or the service portfolio, is the most frequently used non-financial data for bankruptcy prediction. Furthermore, company governance indicators, such as the company's management ability, business feasibility, and technical ability, are also crucial elements for bankruptcy assessment. As the reviewed papers [37, 98] argue, governance indicators can reflect the business status and signalize future development of a company. There are also some studies that use tax-based variables or legal (i.e. lawsuit-related) data to predict bankruptcy. Due to data sensitivity, such data is typically owned by private entities.

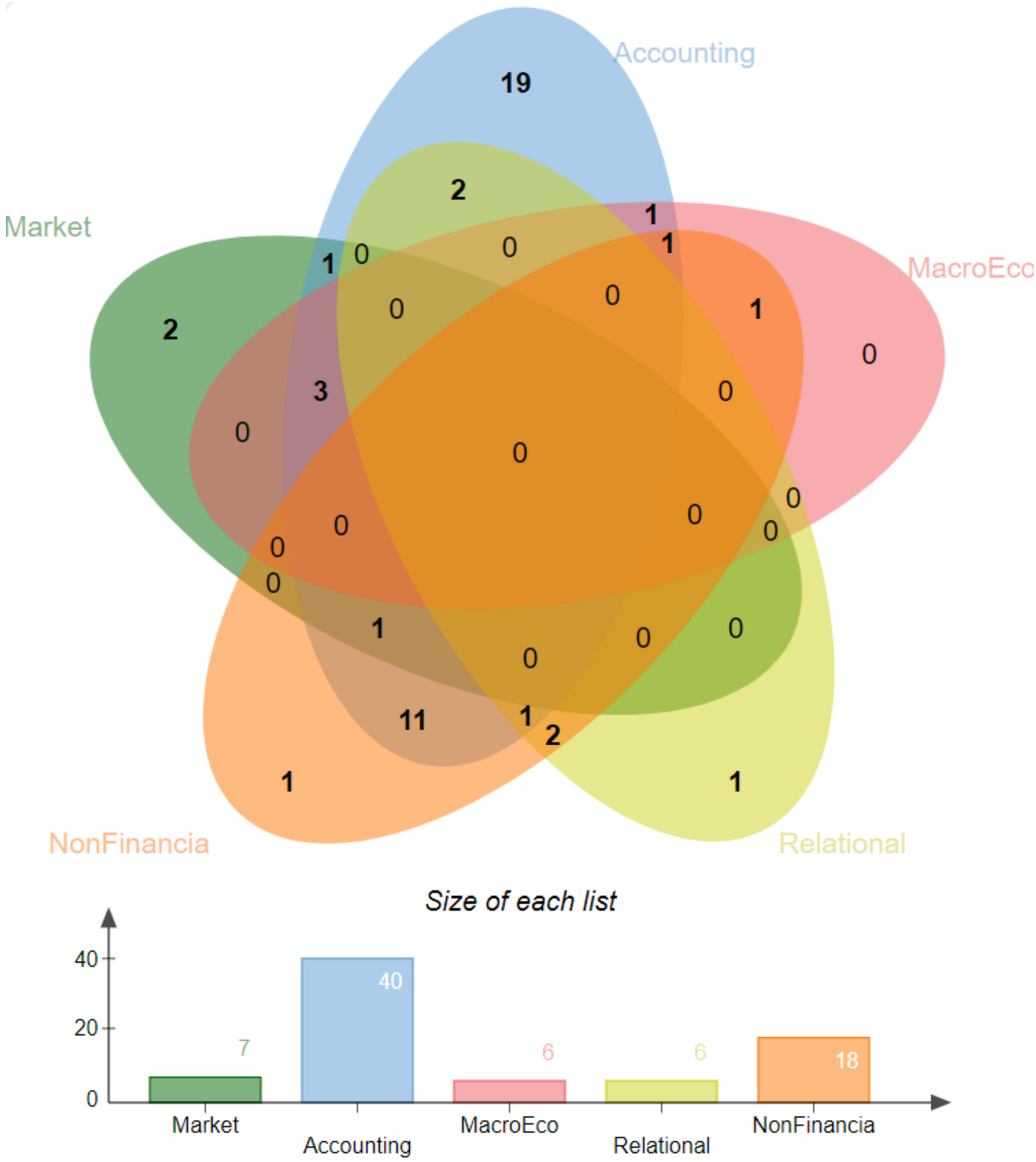


Figure 2.1: Distribution of survey papers according to taxonomy

Applying our taxonomy on the total set of surveyed papers, Fig. 2.1 depicts the distribution of different dataset types. As our findings show, accounting-based data is still the most commonly utilized data source, which accounts for the vast majority of the reviewed papers. From the in total 47 reviewed studies, 40 papers used accounting-based data. What indicates, however, the emerging trend of using mixed datasets for bankruptcy prediction is that 19 papers out of 40 instrumentalized accounting-based data as the only resource. The remaining 21 studies used other types of datasets to supplement the observations to predict bankruptcy. Only two studies relied purely on market-based data for modeling, and five papers combined market-based data with other data types. From this Venn graph, we know that most papers rely on a single data type to do the research. It may be because of the difficulty in doing the data fusion.

With the development of graph neural networks, relational data has been gradually applied to bankruptcy prediction models in recent years. Non-financial data has become a very broad category, including not only company information on board and management structure and on annual reports, but also on regulatory and compliance data, such as taxation and litigation, as discussed earlier. Our findings indirectly confirm that research methodologies for bankruptcy prediction have indeed adopted the available data variables and applied them accordingly. In total, we found 24 papers that use either relational (6 papers) or non-financial (18 papers) data. Most frequently, these datasets were applied in combination with accounting data. There are four papers that exclusively use these datasets for bankruptcy prediction.

2.2 Current Methods for SMEs' Financial Health and Bankruptcy Prediction

2.2.1 Traditional Financial Ratio-based Methods

As a sign of the beginning of applying financial ratios to predict bankruptcy, Beaver [18] proposed and implemented a univariate analysis on the bankruptcy predictive ability of six financial ratios in 1966. The best result of the examination was that the ratio of cash flow to total debt had only a 13% misclassification rate, which was quite good for bankruptcy prediction using a single feature. Two years later, Altman [8] introduced the famous Z-score model, which used five variables based on financial ratios, to predict bankruptcy. This model combines multiple financial ratios into a single score that predicts the likelihood of a company facing bankruptcy. The original Z-Score formula is as follows:

$$Z = 1.2 * X_1 + 1.4 * X_2 + 3.3 * X_3 + 0.6 * X_4 + 1.0 * X_5 \tag{2.1}$$

Where:
X_1= Working Capital to Total Assets

X_2 = Retained Earnings to Total Assets
X_3 = Earnings Before Interest and Taxes to Total Assets
X_4 = Market Value of Equity to Book Value of Total Liabilities
X_5= Sales to Total Assets

The Z-Score classifies companies into three distinct zones:
Safe Zone: Z-Score > 2.99, indicating a low risk of bankruptcy.
Grey Zone: $1.81 <$ Z-Score < 2.99, indicating moderate risk.
Distress Zone: Z-Score < 1.81, indicating a high risk of bankruptcy.

The model was shown to be successful in providing bankruptcy predictions and served as a foundation for subsequent researchers to further refine bankruptcy models. Numerous scholars and researchers have continued to refine and develop various types of bankruptcy prediction models since then, narrowing the general focus between different industries and countries [35, 12, 63, 38, 19, 89].

Financial ratio analysis is highly valued for its simplicity. By performing basic calculations on readily available financial data, a company's financial condition can be quickly assessed [120]. Its straightforward nature allows even nonfinance professionals to use it. These methods do not require advanced technical skills or complex software, highlighting their practicality in resource-limited environments. Financial ratio analysis offers a high degree of interpretability. Ratios such as liquidity, profitability, and leverage ratios clearly reflect the financial condition of a company. For example, the current ratio can be used to measure a company's ability to repay short-term debt. This clarity makes financial ratios effective tools for forecasting. Financial ratio analysis is widely applicable across various industries and organizations and has been a primary tool of financial analysis since its inception. It provides a standardized approach for assessment that is suitable for both large corporations and SMEs, offering a common language to compare the financial performance of different entities. This broad applicability makes it a foundational element of financial analysis, ensuring its continued relevance even with the emergence of more advanced methods [118]. Due to its practicality and ease of use, financial ratio analysis has held a dominant position in the early stages of financial analysis and bankruptcy prediction. In the early days, financial ratios were among the few systematic tools available for assessing financial risk and performance, supported by the limited availability of data and computational resources. Even today, despite the advent of more advanced techniques, financial ratio analysis remains fundamental to financial risk assessment and is often the first step in a comprehensive analysis.

2.2.2 Deep Learning and Machine Learning Methods

With the development of machine learning and deep learning, plenty of studies have attempted to make a breakthrough by applying new models into bankruptcy prediction, and directly use well-calculated financial ratios to find the most predictive model and make predictions [14, 16, 66, 122, 135, 162]. In the early stage of applying machine learning methods for bankruptcy prediction, logistic regression [126] was once the most widely-used model in predicting bankruptcy. Even today, many financial institutions still adopt the logistic regression as the primary approach for building the credit scorecards because of its interpretability and stability [148, 20]. DL and ML methods are particularly well-suited for processing and analyzing large-scale datasets. Techniques such as neural networks and ensemble learning can efficiently handle vast amounts of financial data, including high-dimensional variables and extensive historical records [28]. Their scalability allows for the integration of diverse data sources, such as transaction records, market data, and economic indicators, providing a comprehensive basis for financial risk assessment. In Table 2.1, a list of recent studies is provided that focus on input data and selected representative works that are based on financial ratios. Unlike traditional methods, which may struggle with data volume and complexity, these techniques leverage advanced computational power to derive insights from big data.

One of the key advantages of DL and ML methods is their ability to model complex non-linear relationships within financial data [40]. Neural networks, with their multiple layers and non-linear activation functions, can learn intricate patterns and interactions between variables that traditional linear models might miss. This capability is crucial in financial risk prediction, where relationships between financial indicators and risk factors are often non-linear and dynamic. By accurately modeling these complex interactions, DL and ML methods provide a more nuanced understanding of risk and performance. As demonstrated in Table 2.1, many studies have been dedicated to improving prediction accuracy using various models. Bankruptcy prediction models must be

Table 2.1: Summary of related studies on bankruptcy prediction

Study	Data Category	Data Type	Prediction Models	Evaluation Approaches	Sample Size	Publish Year
[10]	Financial ratios,basic firm information, reported and compliance, operational risk	Numerical data	Altman's Z-score, generic model	AUC, roc curve	3,462,619	2008
[4]	Financial ratios, market-based variables	Numerical data	Black and Scholes models, Altman's Z-score	ROC curve, information content tests	15,384	2008
[29]	Financial ratios, market-based variables	Numerical data	MLP, CART, LR, RF, SVM, ensemble, boosting	Accuracy, sensitivity, specificity	16816	2009
[83]	Financial ratios	Numerical data	MLP, boosting, bagging	Accuracy ratio, AUC	1458	2009
[101]	Financial ratios, corporate governance indicators	Numerical data	Altman's Z-score, SVM	Type I error, Type II error, average accuracy, brier score	108	2010
[129]	Financial ratios	Numerical data	DT, LR, MLP, RBFN, SVM	Correct classification rate	1321	2012
[159]	Accounting, market and macroeconomic data	Numerical data	LR, Altman's Z-score, MLP	AUC, Gini rank coefficient,Kolmogorov–Smirnov	23,218	2013
[98]	Financial ratios, corporate governance indicators	Numerical data	SVM, KNN, NB,CART, MLP	ROC curve	478	2016
[66]	Financial ratios	Image data	CNN	Identification rates, ROC curve	7520	2019
[160]	Financial ratios, relational data	Numerical data, graph data	SVM,GNN	AUC	60,000	2017
[119]	Financial ratios	Numerical data	LR, ANN, SVM, PLS-DA, SVM-PLS	Confusion matrix, accuracy, sensitivity, specificity, AUC	212	2017
[80]	Financial ratios, macroeconomic indicators, industrial factors	Numerical data	MDS	/	165	2019
[112]	Accounting-based ratio, market-based variables, textual discolures	numerical data, text data	Word embedding, CNN, DNN	Accuracy ratio, AUC	11,827	2019
[152]	Financial ratios, basic firm information	numerical data, categorical data	LR, RF, XGBoost, LightGBM, ANN	AUC	977,940	2019
[87]	Basic firm informaion, SME network-based variables, transactional data	Numerical data, graph data, categorical data	LDA, LR, SVM, DT, RF, XGB, NN	AUC	340,531	2021
[68]	Textual sentiment	text	SVM, Bayes, KNN, DT, CNN, LSTM	AUC	10,034	2022

applied practically in the financial industry, necessitating both model accuracy and explainability. The work [129] compares the accuracy and explainability of different data mining methods for predicting bankruptcy. The study compares algorithms such as neural networks, support vector machines, and decision trees, and concluded that decision trees are both more accurate and easier to interpret compared to neural networks and support vector machines. In this study [95], it was demonstrated that LightGBM achieved the highest performance, with fast and cost-effective training, and the model's results could be interpreted using SHAP value analysis. In contrast, authors [112] argue that simple deep learning models outperform other data mining models. Therefore, there are always controversial results concluded under different experimental data and models.

DL and ML methods also excel in automatic feature extraction, a critical step in the data preprocessing phase [121]. When faced with non-intuitive numbers, credit analysts typically calculate financial ratios for analysis, but SMEs often cannot provide complete financial statements, which leads to the inability to calculate financial ratios and complete credit assessments. Neural Networks and Ensemble Learning algorithms can automatically identify and extract relevant features from raw financial data, reducing the need for manual intervention and domain expertise. This automation not only streamlines the data preparation process but also enhances model performance by discovering hidden patterns and relationships that may not be evident through manual feature engineering. There are plenty of explorations on automatic feature engineering based on feature interaction to improve the performance of predicting click-through rate in recommendation systems, such as the Factorization Machine (FM) [140], DeepFM [164], AutoInt [153], and AutoFIS [102] each of which builds the extensive feature interactions to obtain a good result of the model. Deep feature synthesis (DFS) [75] is another well-known algorithm, which is built on feature combination. As a result, these feature engineering methods can improve the overall efficiency and effectiveness of financial risk prediction.

2.2.3 Other Advanced Methods

2.2.3.1 Uplift Modeling

Uplift modeling is dedicated to estimate the causal impact of a treatment at an individual level. It has found widespread application in domains such as marketing, healthcare, and personalized recommendations. Rubin's seminal work introduced the potential outcome framework, which forms the basis for causal inference by controlling for all variables except the treatment [141]. However, this framework relies on stringent assumptions like stable unit treatment value, ignorability, and positivity that can be challenging to satisfy in real-world scenarios.

Recent research [57] has proposed new estimations defining treatment and interference effects using observational data. This sets the groundwork for the use of industrial observational data to estimate uplift. Subsequent studies have increasingly applied deep learning models in causal inference following the publication of Shalit et al. [147]. Künzel et al. [91] introduced meta-learners for uplift modeling, dividing the learning process into two stages: one for predicting outcomes under treatment and another for predicting outcomes under control. While effective for binary treatments, this approach encounters challenges with more complex treatment scenarios. Numerous studies have been conducted on various types of treatments in subsequent years; these studies and their target treatments are summarized in Table 2.2. Despite research on five different types of treatments, multiple time-dependent treatments remain unexplored. Liu et al.[103] addresses the challenge of fully exploiting the interaction between treatment and context information. We designed the multiple time-dependent treatment module based in part on the basis of this study.

In related academic studies [10, 98, 101], it has been fully verified that there is a significant correlation between company adjustment acts and the company's financial status. Another study suggests that there are significant relationships between financial distress and corporate governance practices within variables of the number of members on the board of directors of the company, institutional ownership, managerial ownership, CEO duality and financial leverage [157]. The study [43] identifies the close relationship between governance structure and the possibility of filing for bankruptcy, and highlights the complex interaction between preinsolvency financial, managerial, and governance factors. An-

Table 2.2: Description of various treatments and studies. T refers to treatment. $\mathbb{N}$ denotes the set of nonnegative integers. $\mathbb{R}$ denotes the nonnegative set. k refers to the k_{th} treatment.

Type of treatments	Mathematic definition	Studies
Binary	$T \in \{0, 1\}$	[96, 91, 69, 142]
Multiple	$T \in \mathbb{N}$	[127, 180, 108]
Continuous	$T \in \mathbb{R}$	[74, 26, 24, 78]
Multi-cause	$(T_1, T_2, ..., T_k), k \in \mathbb{N}$	[138, 111, 50, 86]
Time-dependent	$(T \mid time)$	[158, 22, 17]

other study [44] finds that larger boards reduce the risk of bankruptcy in complex firms, while a higher proportion of inside directors lowers bankruptcy risk in firms needing specialist knowledge but increases it in technically unsophisticated firms. Furthermore, the influence of corporate governance variables grows stronger as the time to bankruptcy lengthens, indicating that governance changes may be too late to save firms on the brink of bankruptcy.

However, while the correlation between company adjustments and bankruptcy prediction has been extensively studied, the causal impact of such company adjustment acts on bankruptcy has not yet been fully investigated. When using machine learning or deep learning models to predict bankruptcy, it is common practice to incorporate various relevant factors into the model [98, 166]. In doing so, these models can make binary classification predictions about whether or not a company will go bankrupt. Although this approach considers the correlation between multiple features, including company adjustment acts and prediction targets, it focuses solely on identifying patterns without considering causality.

In practical applications, however, it is not only important to understand the probability of bankruptcy but also essential to gain insight into how to avoid such situations altogether. Understanding this causal relationship is crucial for developing effective strategies to prevent or mitigate the risk of bankruptcy. When a company faces credit risk, the people in charge may take some actions such as adjusting the people on board or relocating the store to prevent it. If we can know beforehand whether these actions will be effective or not, we can act more confidently.

2.2.3.2 Large Language Models in Financial Statement Analysis

Financial analysis is a cornerstone of corporate finance, supporting decision-making in areas such as investment, risk management, and corporate governance. Traditional approaches rely on financial metrics derived from balance sheets, income statements, and cash flow statements, with ratios such as *profitability, liquidity, leverage*, and *efficiency* serving as essential indicators [41]. These ratios form the basis for advanced analytical frameworks like DuPont analysis and the Altman Z-score model. DuPont analysis decomposes return on equity (ROE) into three components: *profit margin*, *asset turnover*, and *financial leverage*, allowing analysts to identify sources of financial performance [151]. Similarly, the Altman Z-score model predicts bankruptcy risk through a weighted combination of financial ratios [8]. However, these methods are labor-intensive, prone to human error, and constrained in their ability to process large datasets or deliver real-time insights.

Advances in artificial intelligence (AI) and machine learning (ML) offer opportunities to automate financial analysis. While these methods improve efficiency and consistency, they often focus on pure numerical predictions [186, 7] or textual sentiment analysis [104], falling short of replicating traditional frameworks like DuPont and Z-score [51]. LLMs represent a transformative technology in this space, demonstrating exceptional abilities in natural language understanding and complex problem-solving [3, 116]. By mastering complex linguistic patterns, LLMs excel in various domains, including customer support automation, content generation, and coding assistance [33].

In financial contexts, however, LLMs face unique challenges. Financial documents often contain jargon, numerical data, and intricate relationships that demand both linguistic and mathematical precision [62]. While LLMs like GPT-3.5 and GPT-4 have shown promise in tasks such as sentiment analysis [104], their numerical reasoning abilities are limited, particularly in multi-step calculations or exact numerical tasks [25, 181]. Studies highlight that even state-of-the-art LLMs often miscalculate or misinterpret numerical contexts, leading to inaccurate financial projections [64, 177]. This limitation underscores the critical importance of precise numerical reasoning in financial decision-making, where even minor errors can lead to flawed conclusions.

Efforts to enhance LLMs' numerical reasoning have explored hybrid approaches, such as *Retrieval-Augmented Generation* (RAG), which integrates external databases for improved factual accuracy [60, 130]. Fine-tuning on domain-specific datasets [154] and techniques like Chain-of-Thought prompting have also been proposed to improve performance on complex financial tasks [81]. These methods have demonstrated the potential to bridge gaps between LLM capabilities and traditional financial analysis. For instance, GPT-4 has been shown to outperform human analysts in predicting earnings changes [81], while few-shot learning has proven effective for text classification in finance with minimal labeled data [109].

Despite these advances, no consensus exists on the optimal strategies for enhancing LLMs in numerical and domain-specific tasks. This paper seeks to address this gap by systematically benchmarking various methods, including zero-shot, few-shot, RAG, and fine-tuning, to evaluate their efficacy in financial applications. The findings aim to establish a clearer framework for leveraging LLMs in finance and identify trade-offs between performance and computational efficiency.

2.3 Summary of Findings and Research Gaps

Although traditional methods based on financial ratios have played a significant role in financial analysis and bankruptcy prediction, their limitations have become increasingly evident with the continuous improvement of analytical and forecasting techniques. Traditional financial ratio analysis primarily relies on data from financial statements, usually focusing solely on financial metrics while neglecting the impact of non-financial factors such as the market environment, competitive landscape, and management quality [53]. This limitation may result in analysis that is not comprehensive enough to fully reflect a company's actual situation. Financial ratio analysis typically assumes that the relationships between financial metrics are linear. However, in reality, the relationship between financial data and company performance is often nonlinear [53]. For instance, the relationship between profitability and debt levels may vary depending on the company's size and industry characteristics. Traditional methods are not well-equipped to identify and model these complex nonlinear relationships. Financial ratio analysis often relies on historical data for forecasting and evaluation, which limits its ability to respond to rapidly changing market environments and economic conditions [58]. When faced with economic fluctuations, market trends, or company strategy adjustments, traditional methods may struggle to adjust predictions in a timely manner, leading to delayed or inaccurate risk assessments. Therefore, it's necessary to develop algorithms to generate useful features rather than only relying on financial ratios.

The methods based on feature interaction are more suitable for the highly sparse categorical data and the business scenario of recommendation systems. Moreover, these features are deeply integrated with the deep learning models and therefore it is not meaningful to just take out the features as such. Furthermore, these features typically lack interpretability. As financial business use-cases usually pay great attention to explainability. Although DFS is interpretable which can give users a clearer insight into the business and is good at handling relational data, it will generate all the features according to the manual setting of the hyperparameters regardless of whether they are useful or not. It just implements the aggregation functions on the features so it cannot generate the features with the necessary depth. Additionally, modeling with redundant features can be costly and may lead to unfavorable outcomes. In Chapter 4, an effective feature engineering algorithm is proposed to address these issues.

Above mentioned studies have shown the benefits of incorporating diverse input data into bankruptcy prediction models. These studies have expanded beyond traditional data, such as financial ratios and market-based variables, to explore various types of input data. This work [146] confirms that financial ratios are predictive indicators of firm failure. The study also suggests that non-financial variables, such as localization and economic conditions, are drivers of SMEs failure. The study [80] combines financial ratios and macroeconomic data to analyze their impact on firms, providing evidence for the reliability of macroeconomic data. Another study [87] focused on using SMEs' transaction data for the prediction of bankruptcy, without relying on accounting data. The results showed that this approach outperformed the benchmark method. Some research pairs also include studies on different types of data. The authors [160] used shared directors and managers to establish a connection between two companies and developed a model using relational data to identify the companies with the highest risk. In contrast, Mai et al. [112] focuses on using a deep learning model to extract textual information as a complementary variable to accounting and market data to improve prediction accuracy. There are two pieces of work [10] and [98] that use similar indicators related to company adjustments, but they are static and cannot reflect dynamic behavior. Therefore, the correlation between company adjustments and bankruptcy prediction needs to be studied.

When it comes to identifying the causal relationship between company adjustments and financial health, the application of uplift modeling to company adjustment analysis is relatively nascent. Most existing studies on uplift modeling focus on simpler binary treatment scenarios and do not consider the complexities of multiple time-dependent treatments typical in company adjustment. This gap presents an opportunity for the development and application of more sophisticated models that can handle these complexities.

Recent advancements in machine learning and deep learning provide promising avenues for addressing these challenges. Long short-term memory (LSTM) networks and attention mechanisms, for example, have shown great potential for capturing temporal dependencies and interactions in sequential data. These techniques can be leveraged to develop uplift models that account for the dynamic nature of company adjustment and their impact on financial outcomes. MTDnet proposed in Chapter 7 aims to fill this gap by integrating these advanced techniques to handle multiple time-dependent treatments in the company adjustment analysis. By doing so, it seeks to provide more accurate and actionable insights into how different company actions influence financial stability and the likelihood of bankruptcy.

Chapter 3

Data collection

Contents

3.1 Introduction of Raw Data from LBR 35
3.2 Financial Data 37
3.3 Non Financial Data 39

This research focuses on SMEs in Luxembourg, and the methodologies proposed in this dissertation are mainly trained and validated within this scope. This chapter will give an overall and brief description of the dataset that is used in this dissertation. The experiments from later chapters will resample the data according to the research questions and include other data if necessary.

3.1 Introduction of Raw Data from LBR

The dataset utilized for the experiments is a publicly available dataset obtained from the portal of Registre de Commerce et des Sociétés (RCS) of LBR[1]. Companies are mandated to disclose files pertaining to their registered information, financial reports, managerial changes, and other essential details in Luxembourg. Fig. 3.1 is an example of the raw data in LBR website. It contains the company name, registration data, the files uploaded by the company, and so on. This data is up-to-date, so it dynamically depicts the operational disclosures of a company.

Apart from the information showed in the webpage, the files that uploaded by the company are very important for data extraction. The files include the financial statements, the registration form, the court

[1] www.lbr.com

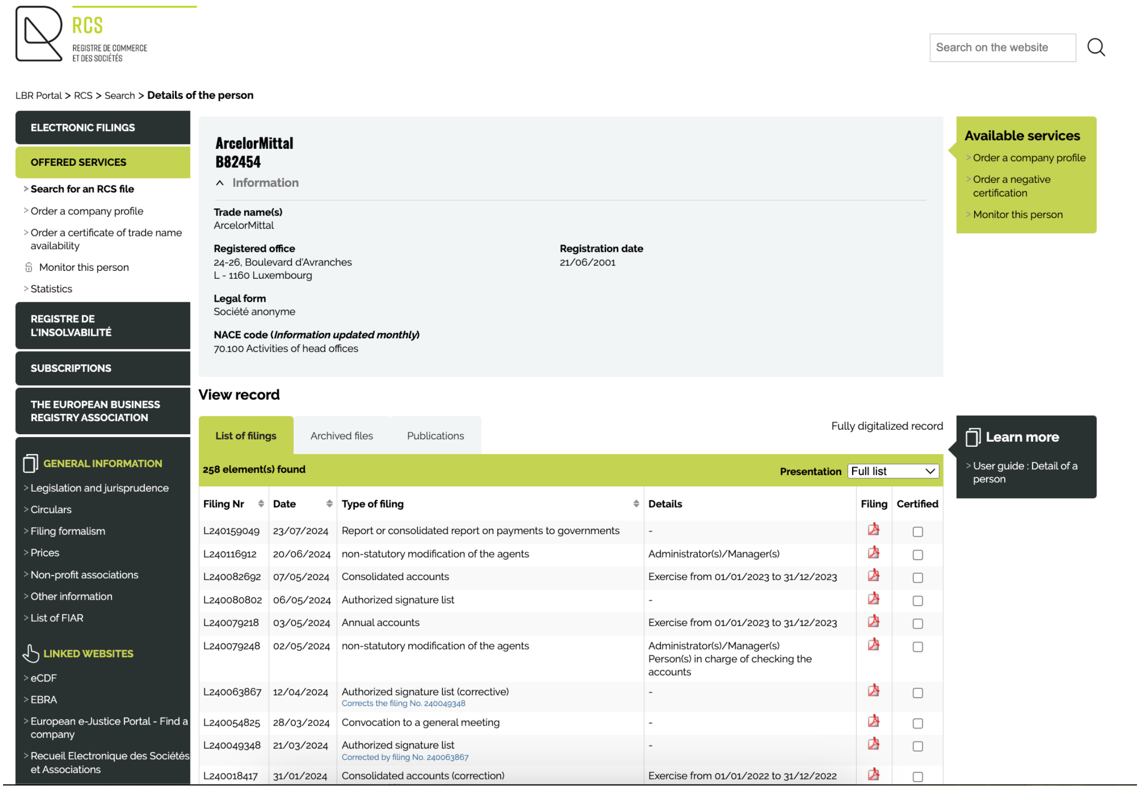


Figure 3.1: Example of raw data webpage from LBR

orders regarding bankruptcy, and so on. Pie chart 3.2 is the proportion of file types that account for more than 1%. As can be seen from this figure, the largest proportion of documents is financial statements, followed by acts of company adjustment. Other documents include information on bankruptcy and exit, which are the labels for model training. These three types are the focus of this dissertation. The data collected from LBR spans from 2011 to 2023, with a subsequent focus on SMEs within this research scope. The research of Chapter 4 and Chapter 6 uses the data from 2011 to 2021 because these two studies finished at an earlier time. The research of Chapter 5 and Chapter 7 uses the data from 2011 to 2023.

Although Luxembourg thrives on the financial institutions and companies, we excluded them from our dataset and focused on non-financial limited companies. We omit financial institutions not only because they have inherently complex organizational forms but also because the aim of opening and closing of financial institutions in Luxembourg is usually to benefit from the regulatory environment rather than to really operate a business. This means their business status may not directly relate to their financial statements. In addition, due to favorable tax policies,

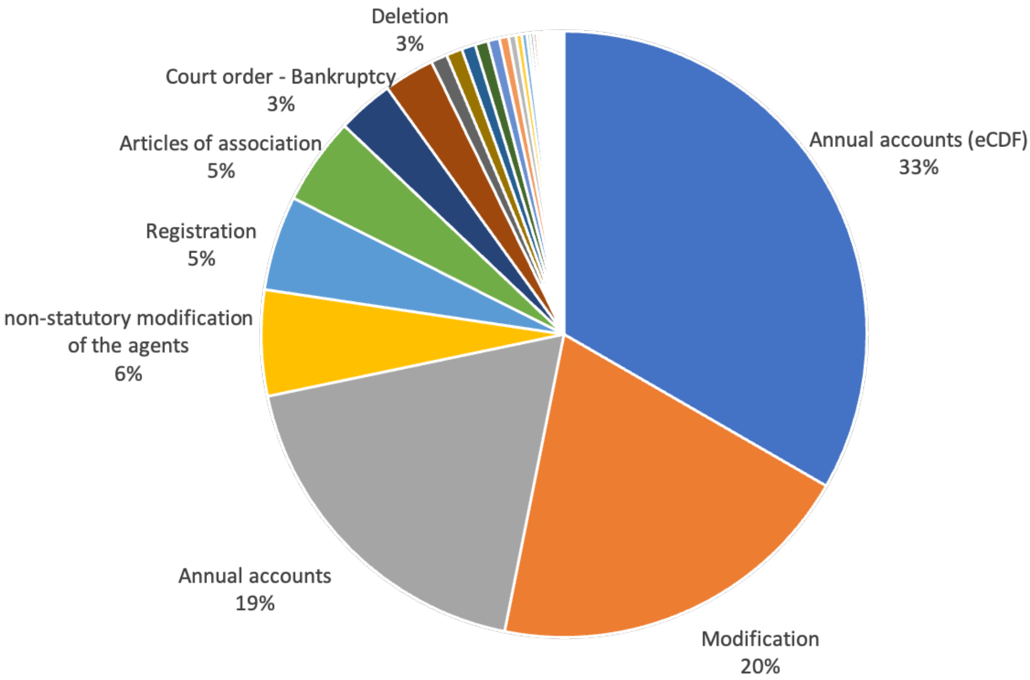


Figure 3.2: Distribution of different types of reported files

there is a notable presence of financial companies and large corporate headquarters in Luxembourg; however, this may not be representative of the general market conditions in other countries, so we exclude the financial-related companies.

3.2 Financial Data

The dataset used in this study consists of financial statements submitted as part of their annual reports to the authorities in Luxembourg and which have been made publicly available by LBR. We find that less than half of SMEs have submitted their profit & loss statements and few SMEs have submitted the cash flow statements as submitting these two statements is not mandatory for unlisted companies. At the minimum, these statements consist of the balance sheets, which are submitted in one of two fixed formats making them amenable for data processing, and an annex describing the company results in a free text format.

Fig 3.3 shows the number of companies on the y-axis against the number of balance sheet or profit and loss statements they disclosed on the x-axis. The number of companies decreases as the number of disclosed statements increases for both balance sheet and profit and loss

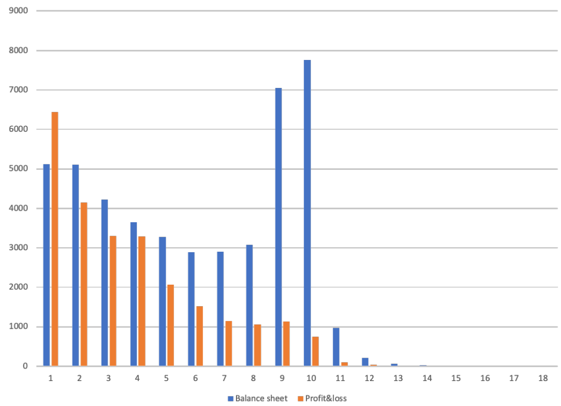


Figure 3.3: Distribution of Financial Statement Disclosures Among Companies. y-axis refers to the number of companies, x-axis refers to the number of financial statements disclosed with separate counts for balance sheet and profit & loss statements as indicated in the legend.

statements. Few companies disclosed more than 10 statements, as indicated by the sharp decline in counts after the 9th disclosure point. Across all disclosure levels, the count of companies disclosing balance sheet data, represented by blue bars, is consistently higher than those disclosing profit and loss data, represented by orange bars. This suggests that companies are more likely to disclose balance sheet information compared to profit and loss statements. A noticeable peak is observed at the 9th disclosure point for balance sheet data, where the count of companies rises significantly, potentially indicating a reporting or regulatory requirement for companies to disclose a specific number of balance sheets, such as 9 years. For companies that disclosed only one or two statements, the gap between balance sheet and profit and loss disclosures is more pronounced, suggesting that companies with minimal disclosure prioritize balance sheets. This pattern might have implications for research, as companies with fewer disclosures could pose challenges for accurate analysis due to limited data, and the prevalence of balance sheet disclosures could be leveraged in models that depend on historical financial data.

Therefore, we only choose the balance sheet as the raw data in order to cover all of the target companies. But even just for the balance sheet, we still face several challenges. Luxembourg is a multilingual country that has four official languages. So there are three language versions for each form. The format of these forms changes from time to time, and also the accounting subjects from the financial statements are not the

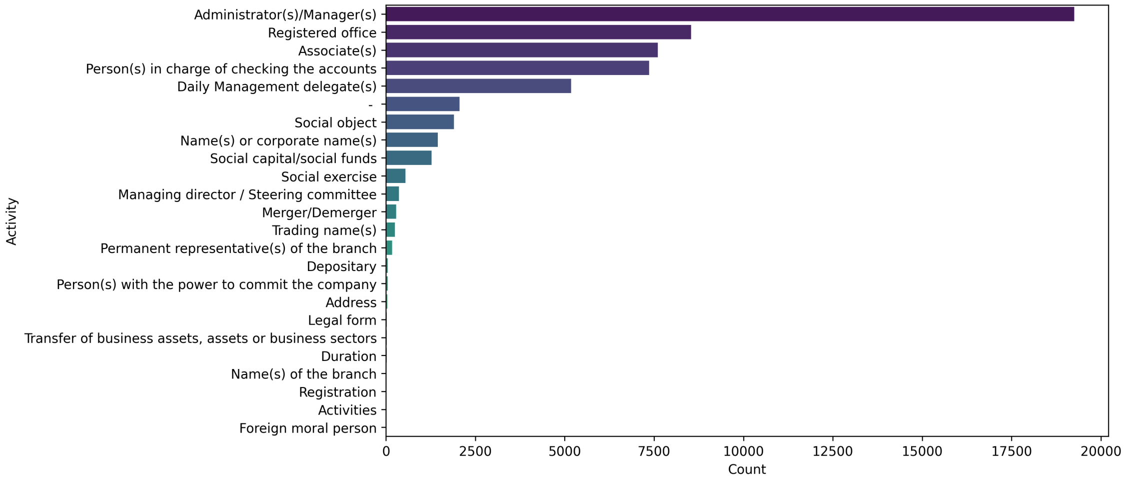


Figure 3.4: Distribution of company adjustment acts

same in international accounting.

3.3 Non Financial Data

Non-financial data in this dissertation mainly refers to the company's information excluding the financial statements. The information about a company from LBR comprises basic information, financial statements, and company adjustment acts derived from company-reported files. In chapter 6, we analyze if the bankruptcy prediction can be improved by considering the company adjustment acts.

For analysis and uplift modeling purposes, we have selected the most recent financial statements of each company along with their adjustment acts occurring between the release date of these statements and their current financial status (active or bankrupt). Basic information encompasses legal form, operational duration, and sector classification. Figure 3.4 illustrates the distribution of different adjustment acts observed in this study. The Luxembourg Business Register classifies company adjustments into 24 distinct categories; however, it should be noted that these acts exhibit a significant imbalance in distribution. The most frequent acts are related to the changing of managerial people.

Chapter 4

Automatic Feature Engineering: Enhancing the Use of Financial Statements for Bankruptcy Prediction

Contents

4.1 Problem Statement **41**

4.2 Methodology **43**

4.2.1 Overview for Automatic Feature Engineering 43

4.2.2 Feature Generation 45

4.2.3 Termination Condition 46

4.3 Experiments **46**

4.3.1 Comparison with Feature Sets Generated by Other Methods 49

4.3.2 Models for Evaluating Different Feature Generation Approaches 50

4.3.3 Feature Performance Indicators for Comparing Feature Contribution 51

4.4 Results and Discussion **52**

4.4.1 Comparison with Financial Ratios 52

4.4.2 Comparison with DeepFM 53

4.4.3 Comparison with Deep Feature Synthesis 55

4.4.4 Comparison with Raw Data 56

4.4.5 Explainability and Extensibility 57

4.1 Problem Statement

Bankruptcy prediction represents a critical aspect of financial risk management, with significant implications for businesses, employees, creditors, and the broader economy. Accurately predicting the likelihood of a company's bankruptcy facilitates the early identification of financial distress, enabling timely interventions that could prevent economic losses. Traditional bankruptcy prediction models primarily rely on financial ratios derived from comprehensive financial statements, such as the balance sheet, cash flow statement, and profit and loss statement. These financial ratios have demonstrated strong predictive power across various studies, offering valuable insights into a company's financial health by summarizing its business activities and performance over a defined period.

However, the effectiveness of these traditional models is contingent upon the availability and completeness of financial statements. While listed companies are generally required to disclose detailed financial statements, many unlisted companies, including SMEs, are not obligated to publicly disclose this information. The absence of publicly available financial statements presents a significant challenge for bankruptcy prediction models that rely on financial ratios. For instance, without access to cash flow statements and profit and loss statements, a substantial portion of the necessary financial ratios cannot be calculated. In some cases, this may lead to the loss of up to 73% of the required input features for accurate prediction, thereby severely undermining the model's reliability and effectiveness.

This issue is particularly pronounced among SMEs, which play a vital role in local economies by generating employment and contributing to economic growth. SMEs often face difficulties in providing the necessary financial history due to their smaller scale, limited resources, or the nascent stage of their business operations. Consequently, these enterprises encounter higher risks of credit declines and financial instability, which can trigger cascading effects throughout the broader economy. The lack of comprehensive financial data not only hinders the calculation of traditional financial ratios but also complicates the credit risk assessment process, making it challenging for analysts to make informed decisions.

To address these challenges, the research proposes a novel automatic feature engineering (AFE) approach designed to overcome the limitations associated with incomplete financial statements. The AFE approach aims to generate a robust set of features that can be employed for bankruptcy prediction, even when key financial statements are missing. Unlike traditional financial ratios, which are developed based on domain expertise and require specific data inputs, the AFE approach is flexible and capable of generating meaningful features irrespective of which financial statements are available. This method is particularly advantageous in real-world scenarios in which data is often incomplete or inconsistent.

The AFE approach offers several key benefits. First, it reduces the dependency on complete financial statements, making it possible to develop effective bankruptcy prediction models with limited data. Second, the features generated by AFE are fully transparent, which enables a detailed investigation into the importance and contribution of each feature to the prediction model. This transparency is essential for understanding and explaining the underlying factors that drive the model's predictions, which is critical for gaining stakeholders' trust and making informed decisions.

To validate the effectiveness of the AFE approach, a series of experiments was conducted using manually collected data from real-world companies. The results from these experiments demonstrate that models trained on features generated by the AFE approach perform comparably to, and in many cases better than, models based on traditional financial ratios. This finding is significant, as it suggests that the AFE approach not only compensates for the lack of financial data but also enhances the overall quality of bankruptcy prediction. Furthermore, the AFE approach outperforms other feature generation methods in most cases, highlighting its potential as a superior alternative for feature engineering in bankruptcy prediction models.

In summary, the research presented in this dissertation addresses a critical gap in the field of bankruptcy prediction by proposing and validating an automatic feature engineering approach. This approach holds particular value for SMEs and other companies lacking comprehensive financial disclosures, offering a new avenue for improving bankruptcy prediction and, by extension, financial risk management. The contributions of this research include the development of a novel AFE algorithm, the implementation of this algorithm in a real-world context, and a comparative analysis of its performance against traditional financial ratios and other feature generation methods. The results underscore the potential of the AFE approach to transform the way bankruptcy prediction models are constructed and utilized, ultimately contributing to more accurate and reliable risk assessments in the financial industry.

4.2 Methodology

4.2.1 Overview for Automatic Feature Engineering

Fig. 4.1 illustrates the complete process of automatic feature engineering (AFE) from raw financial records to the derived feature set used for bankruptcy prediction. The first step involves preprocessing the raw data for feature construction. During this phase, extreme values are addressed by replacing infinity with finite extremes, and missing values are treated as zero, in accordance with accounting subjects. Prior to advancing to the next step, the hyperparameters k_1 and k_2 are determined. These represent the number of features selected from feature aggregation and feature crossing, respectively. Additionally, $batch_size$ is specified to indicate the number of feature pairs resulting from one iteration of the first feature crossing round.

Feature generation process consists of two independent parts: aggregation and crossing. In the feature aggregation process, features are generated using the aggregation method in a single step, after which the most valuable features are selected. Conversely, the feature crossing process involves a loop of feature crossing and feature selection, where new features are generated through the crossing method, followed by the same selection criteria employed in the feature aggregation. Subsequently, if the newly generated feature set does not meet the termination condition, feature crossing continues. Otherwise, the process halts, and the derived feature set is obtained by combining the outcomes from the crossing process with the features generated during the aggregation process.

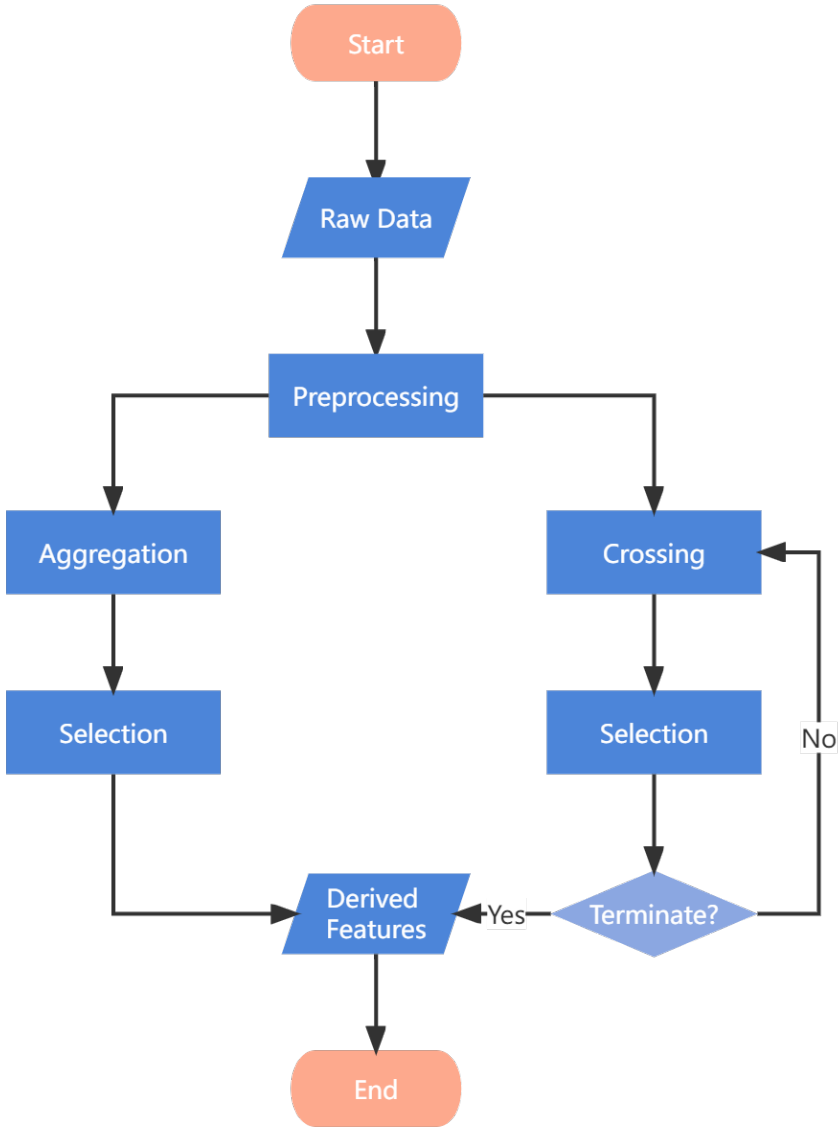


Figure 4.1: Pipeline for automatic feature engineering process

4.2.2 Feature Generation

For feature aggregation, statistical descriptive indicators are calculated for the features of each company over n years and utilized as new features. Specifically, the maximum value (max), the minimum value (min), the sum (sum), the average ($mean$), the standard deviation (std), and the percentage change (pct_change) between the current and the previous year are employed as descriptive indicators. The feature importance from lightGBM is adopted to evaluate a feature's contribution in identifying the targets due to its fast and efficient computation, high reliability, and strong interpretability [77]. Top k_1 features are kept as part of the final feature set.

For the first round $i = 0$ of feature crossing, as in (4.1), $(f_1 f_2 \cdots f_n)$ represents the n features from the input dataset, while S_0 denotes the derived feature set that follows the initial round of feature combination. The symbol $\odot$ indicates four basic operands: addition $(+)$, subtraction $(-)$, multiplication $(*)$, and division $(/)$. These operands are intended to simulate the calculations conducted by experts when determining financial ratios for each feature pair. Taking one element $(f_1 \odot f_2)$ as an example, this element signifies four new feature values $(f_1 + f_2), (f_1 - f_2), (f_1 * f_2), (f_1/f_2)$.

$$
\begin{aligned}
S_0 &= \begin{pmatrix} f_1 & f_2 & \cdots & f_n \end{pmatrix}^T \odot \begin{pmatrix} f_1 & f_2 & \cdots & f_n \end{pmatrix} \\
&= \begin{pmatrix} (f_1 \odot f_1) & (f_1 \odot f_2) & \cdots & (f_1 \odot f_n) \\ (f_2 \odot f_1) & (f_2 \odot f_2) & \cdots & (f_2 \odot f_n) \\ \vdots & \vdots & \ddots & \vdots \\ (f_n \odot f_1) & (f_n \odot f_2) & \cdots & (f_n \odot f_n) \end{pmatrix}
\end{aligned}
\tag{4.1}
$$

Following this, the feature selection process is initiated. The feature set S_0 is analyzed, and the newly derived features are incorporated into a LightGBM classification model to ascertain their feature importance. The top k_2 features, consistent with the results, are selected as input features for the next round of feature generation. Simultaneously, these k_2 features are added to the final feature set for the prediction model.

For the subsequent rounds $i > 0$, similar to (4.2), the new derived feature set is S_i. The steps outlined previously are then repeated, resulting in k_2 new generated features for each order until the termination condition is satisfied. The detailed steps of automatic feature engineering are presented in algorithm 1.

$$
\begin{aligned}
S_i &= \overbrace{\begin{pmatrix} f_a & f_b & \cdots & f_k \end{pmatrix}}^{k\ features\ from\ S_0}{}^{T} \odot \begin{pmatrix} f_1 & f_2 & \cdots & f_n \end{pmatrix} \\
&= \begin{pmatrix} (f_a \odot f_1) & (f_a \odot f_2) & \cdots & (f_a \odot f_n) \\ (f_b \odot f_1) & (f_b \odot f_2) & \cdots & (f_b \odot f_n) \\ \vdots & \vdots & \ddots & \vdots \\ (f_k \odot f_1) & (f_k \odot f_2) & \cdots & (f_k \odot f_n) \end{pmatrix}
\end{aligned} \tag{4.2}
$$

4.2.3 Termination Condition

The algorithm for feature engineering offers two termination options. The first option involves a maximum iteration limit for the feature generation loop, which prevents infinite feature generation. This limit can be manually defined and is set to the number of input features by default, enabling each feature to combine with every other feature once.

The other method is automatic termination. As shown in algorithm 1, for each round of feature generation, a LightGBM model is trained to make predictions on the newly generated features, followed by a comparison of the AUC value of the current round with that of the previous round. If the AUC of the current round is greater than that of the previous round, it indicates that the newly generated features enhance the performance of the prediction model. Therefore, these features should be retained and added to the final feature set. Conversely, if the AUC of the current round is less than that of the previous round, it suggests that the newly generated features do not enhance the prediction model, leading to the conclusion that the feature generation process should be terminated.

In this case, the default number for the maximum feature generation loop is adopted, along with automatic termination, to achieve minimal manual intervention and obtain the derived feature set.

4.3 Experiments

The time period of the data spans from 2011 (the earliest annual data available via LBR) to 2021, encompassing records of both bankrupt and well-operating companies. Nine datasets are constructed with different

Algorithm 1 The Automatic Feature Generation Process

Definitions:

F_{raw}: Input parameter. A vector with the original features in the raw data,
D: Input parameter. A random selection of rows from 70% of raw pre-processed data, also as the training data for the prediction models,
k = the number of highest ranked features,
$batch\_size$ = the number of feature pairs in each batch,
$f_x \odot f_y$ = operation that yields a set of four values: $f_x + f_y, f_x - f_y, f_x * f_y, f_x/f_y$,
AUC = area under receiver operating characteristic curve,
$AggIndicators(data, *arg)$ is a function to generate the descriptive indicators $*arg$ of each feature in $data$,
$TopFeatures(model, FC, k)$ is a function to rank the Feature Candidate set (FC) based on the feature importance calculated from $model$, and keep the top k.
lgb: LightGBM model.

Returns:

F = Constructed feature set

function AFE(F_{raw}, D)
  $F_{agg} \leftarrow AggIndicators(F_{raw}, *arg)$ ▷ $*arg$: a set of operations max, min, sum, mean, std, pct_change
  $FC \leftarrow F_{agg}$
  $FS_{agg} \leftarrow TopFeatures(lgb(D[FC]), FC, k_1)$
  $n_0 \leftarrow |F_{raw}|$
  $n\_batches \leftarrow n_0^2/batch\_size$
  $F_{cross} \leftarrow F_{raw}^T \odot F_{raw}$ ▷ See (4.1)
  $FS_0 \leftarrow \emptyset$
  for $b \leftarrow 0, n\_batches - 1$ **do**
    $FC \leftarrow F_{crossed}[batch]$
    $FS_b \leftarrow TopFeatures(lgb(D[FC]), FC, k_2)$
    ▷ Selecting all features in a batch and keep the top k
    $FS_0 \leftarrow FS_0 \cup FS_b$
  end for
  $FS_1 \leftarrow TopFeatures(lgb(D[FS_0]), FS_0, k_2)$ ▷ Select the k top most important feature pairs
  $AUC_0 \leftarrow 0$
  for $i \leftarrow 2, n_0 - 1$ **do**
    $F_{cross} \leftarrow F_{raw}^T \odot FS_{i-1}$
    $FC \leftarrow F_{crossed}$
    $FS_i \leftarrow FS_i \cup TopFeatures(lgb(D[FC]), FC, k)$
    $AUC_i \leftarrow Calculate\_AUC(lgb(D, FS_i))$
    if $AUC_i <= AUC_{i-1} | reached\_limit(i)$ **then**
      return FS_i
    end if
  end for
  $F \leftarrow FS_{agg} \cup FS_i$
end function

historical periods (1-year, 2-year,..., and 9-year) to cover all time spans from 2011 to 2021. The n ranges from 1 to 9 years, indicating that the data contains balance sheets from n consecutive years for each company. The target label corresponds to the business status (bankrupt or non-bankrupt) of each company in the year following n consecutive years. The descriptive statistics of the datasets, which list the number of positive samples (bankrupt companies), negative samples (non-bankrupt companies), positive rate (bankruptcy rate), feature sizes of the AFE approach, and raw data for each historical period, are presented in Table 4.1.

Table 4.1: Summary of sub-datasets sampled according to n consecutive years

Dataset	**Negative**	**Positive**	**Positive Rate**	**Num of AFE Features**	**Num of Input Features**
1-year	31528	4351	12%	42	64
2-year	28570	3271	10%	58	128
3-year	26029	2348	8%	50	192
4-year	23783	1675	7%	34	256
5-year	21664	1133	5%	66	320
6-year	19689	737	4%	50	384
7-year	17560	440	2%	50	448
8-year	15036	229	2%	58	512
9-year	7913	88	1%	50	576

Hyperparameters of the AFE approach can be selected to align with the characteristics of different input data. The selection of hyperparameters involves a trade-off between model performance, training time, and resource consumption. Hyperparameters for the 1-year, 2-year, and 3-year datasets are explored to optimize performance. It is necessary to establish $k1$ and $k2$ for selecting top features from the feature aggregation and feature crossing processes, as well as $batch\_size$ to prevent memory exhaustion during feature generation. $k1$ and $k2$ directly impact the number of features selected for the bankruptcy prediction model. After conducting experiments on $k1$, varying from 5 to 50, $k2$, varying from 2 to 20, and $batch\_size$, varying from 100 to 512,800, $k1$ was set to 15, $k2$ to 8, and $batch\_size$ to 30,000 for optimal model performance.

4.3.1 Comparison with Feature Sets Generated by Other Methods

Financial Ratios Financial ratios are developed and formulated based on the expertise of domain experts. Due to the limitations of cash flow and profit & loss statements, only 10 financial ratios can be replicated based on the top 20 most frequently used financial ratios in bankruptcy prediction, as discussed in [23], along with 8 financial ratios from other studies [174, 187, 188]. The considered financial ratios are depicted in Table 4.2. The variables marked ⋆ are derived from [23] and the ones marked † come from [174, 187, 188].

Table 4.2: The set of financial ratios used in this chapter

Variable	Financial Ratios	Description
f1⋆	current ratio	current assets ÷ current liabilities
f2⋆	debt to equity	debt ÷ equity
f3⋆	working capital to total assets	(current assets-current liabilities) ÷ total assets
f4⋆	total liabilities to total assets	total liabilities ÷ total assets
f5⋆	equity to total assets	equity ÷ total assets
f6⋆	quick ratio	(cash + marketable securities + accounts receivable) ÷ current liabilities
f7⋆	current assets to total assets	current assets ÷ total assets
f8⋆	cash to total assets	cash ÷ total assets
f9⋆	cash to current liabilities	cash ÷ current liabilities
f10⋆	long term debt to equity	long term debt ÷ equity
f11†	total assets growth rate	(total assets of current year - total assets of previous year) ÷ total assets of previous year
f12†	quick assets to total assets	(current assets-inventory-prepaid expenses) ÷ total assets
f13†	current assets to current liabilities	current assets ÷ current liabilities
f14†	(cash or marketable securities) to total assets	(cash + marketable securities) ÷ total assets
f15†	total debt to total assets	debt ÷ total assets
f16†	equity to fixed assets	equity ÷ fixed assets
f17†	current assets to total liabilities	current assets ÷ total liabilities
f18†	short-term liabilities to total assets	short-term liabilities ÷ total assets

First, we compare the performance of the representative bankruptcy prediction models, which are trained by the features generated by financial ratios and AFE; then we evaluate the prediction contribution of the features from these two different approaches.

DeepFM DeepFM is a prominent approach in the field of recommendation systems [164]. It was developed from the factorization machine(FM). The DeepFM model consists of two parts: FM and DNN. The FM model extracts low-order features, while the DNN model extracts high-order features, allowing for the simultaneous learning of low- and high-order feature interactions. The output is the sum of the FM part and the DNN part, as shown in (4.3). Because the input consists of raw features and both FM and DNN share these features, training the DeepFM model is efficient. This chapter focus on the results obtained using the default setting of DeepFM, rather than fine-tuning the model. As this is a black box model, the exact features generated by DeepFM cannot be determined, and it returns the prediction as the model result.

$$\hat{y} = sigmoid(y_{FM} + y_{DNN}) \tag{4.3}$$

Deep Feature Synthesis We also compare our approach with deep feature synthesis(DFS) mentioned in Section 2.2.2. We deploy the `featuretools` package and set seven primitives: ”sum”,”std”,”max”,”skew”,”min”,”mean” and ”trend”. These primitives are what we can have according to the input data. We compare DFS with the AFE approach in two ways. One is to keep all the features that are generated by DFS. The other is to select the same number of features from DFS as in the AFE approach. This two-way comparison is to show the redundant features generated by DFS and the necessity of feature selection.

4.3.2 Models for Evaluating Different Feature Generation Approaches

We evaluate the effectiveness of the features generated by two approaches by comparing the performance of the representative models trained from the features generated by these two approaches. Each of the nine datasets (1-year, 2-year,...and 9-year) is divided randomly into training and testing sets with the ratio of 7:3. Subsequently, we train the models mentioned below on these nine datasets. The following are the representative bankruptcy prediction models and their settings:

Logistic Regression (LR) The sigmoid function is utilized to convert the output from a linear model into a classification result. Ohlson [126] applied this model to make bankruptcy predictions, and it has gained significant traction in this area. In this chapter, we use the LR function of the Scikit-learn package (version 1.0.1), setting the $C = 0.1$ and *class_weight='balanced'*.

Random Forest (RF) This chapter discusses a typical bagging ensemble model that is effective in classification tasks. Kruppa et al. [90] compared RF with LR in individual credit risk prediction, and the results showed that RF outperformed LR. In this chapter, we use the RF function of the Scikit-learn package (version 1.0.1), setting the *max_depth=2* and *n_estimators = 10*.

LightGBM (LGB) It is an improved model based on extreme gradient boosting (XGBoost) and also a representative model of a boosting ensemble model [77]. Son et al. [152] compared several models for predicting bankruptcy, finding that LightGBM performed the best among all evaluated models. In this chapter, we use the LightGBM package (version 3.2.2) and the GridSearch method to find the best parameters for learning rate, max depth, and number of leaves. We keep other parameters in default settings.

Multilayer Perceptron (MLP) Inspired by the study [106, 136], we train a MLP model with four hidden layers. We use the ReLU function as the activation function, and the dropout rate is 0.3. The loss function chosen is cross entropy, with a learning rate is set at 0.01. We implemente the MLP model using PyTorch (version 1.10.0).

4.3.3 Feature Performance Indicators for Comparing Feature Contribution

We evaluate the performance of features by feature importance and information value. We conduct feature importance ranking and calculate the information value by the `LGBMClassifier()` function with default parameters of `lightgbm` package on the combination dataset of features created by different approaches. We set the option of "importance_type" to "split" to calculate the importance, which is a split-based method. Feature importance is generally used to assess the contribution of each feature during model training [11]. A higher rank corresponds to a greater effect on the model. Information value serves as an indicator for measure the predictive power of an independent variable [67].

A higher information value indicates that the feature possesses more predictive power. We adopt this indicator to evaluate features' performance due to its widespread application in feature selection for credit risk assessment within the financial industry. The calculation can be described as follows [149]:

$$IV = \sum_{i=1}^{n} (\frac{G_i}{G} - \frac{B_i}{B}) * \ln \frac{G_i/G}{B_i/B} \tag{4.4}$$

where n denotes the number of bins for each feature, G_i and B_i represent the counts of negative and positive samples within bin i, while G and B indicate the total counts of negative and positive samples in the population.

4.4 Results and Discussion

4.4.1 Comparison with Financial Ratios

4.4.1.1 Model Performance

Fig. 4.2 shows the performance of mentioned models trained on automatic feature engineering (AFE) and financial ratios (FR) from Table 4.2. The x-axis of the figure represents nine datasets. The y-axis represents the improvement of AUC from models trained by AFE compared to AUC from models trained by FR. It can be observed that the models trained by AFE exhibit significant advantages over the models trained by FR. In total, AFE outperforms FR in 35 out of 36 cases. Consequently, the models using the automatic feature engineering approach demonstrate a superior ability to predict bankruptcy under these scenarios.

4.4.1.2 Feature Performance

We evaluate the contribution of each feature by putting them in the same bankruptcy prediction model. Fig. 4.3 illustrates a comparison of the feature importance between the features identified by the AFE algorithm and the financial ratios. It is evident that the features generated by AFE consistently achieve a higher rank than those produced by FR across all

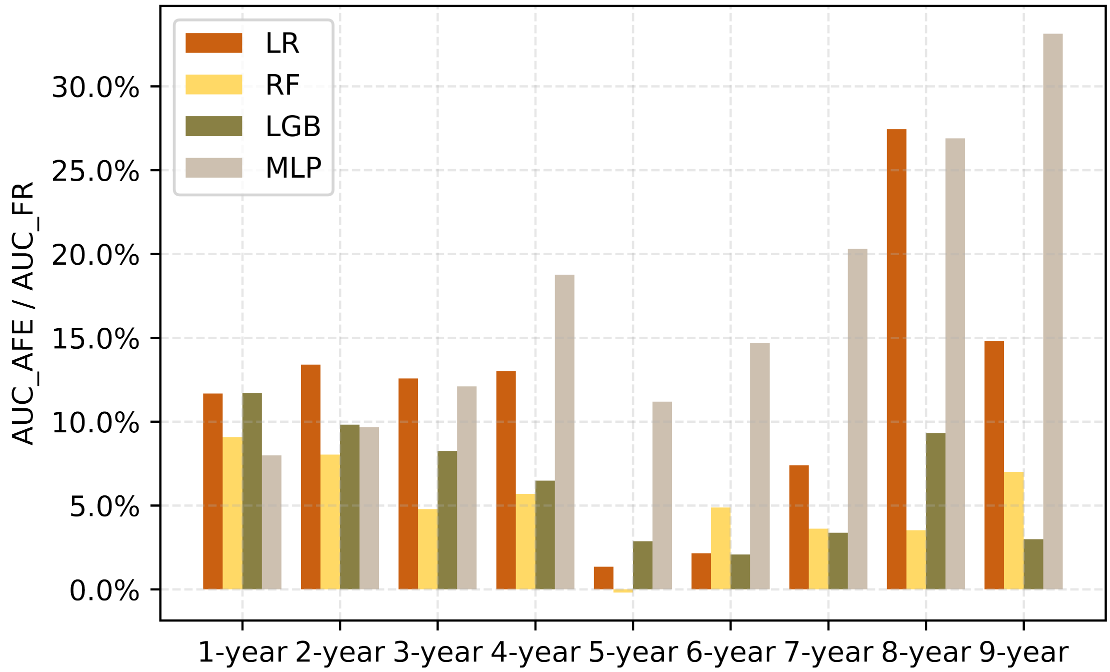


Figure 4.2: AUC improvement of AFE compared to FR on each dataset

nine datasets, which further suggests that models utilizing AFE tend to exhibit superior performance.

Fig. 4.4 illustrates the information value, indicating that most features created by AFE possess higher information values than those generated by the FR approach across all datasets, except for the 3-year dataset. Although the median IV of 3-year features from AFE is slightly lower than that of 3-year features from FR, the first quartile of features from AFE remains larger than that of 3-year features from FR. This suggests that AFE essentially contributes features with higher IV, which could lead to improved model performance. Therefore, it can be concluded that features created by AFE demonstrate better performance than those created by FR, based on the results of both feature importance and information value.

4.4.2 Comparison with DeepFM

We compare the performance of the four models with the AUC of DeepFM result. From Table 4.3, it shows that automatic feature engineering has the absolute advantages over all the nine datasets by all the four models. AUC of DeepFM for nine datasets is all between 0.6 and 0.7. But for LightGBM model on AFE, the AUC could reach more than 0.85 on the

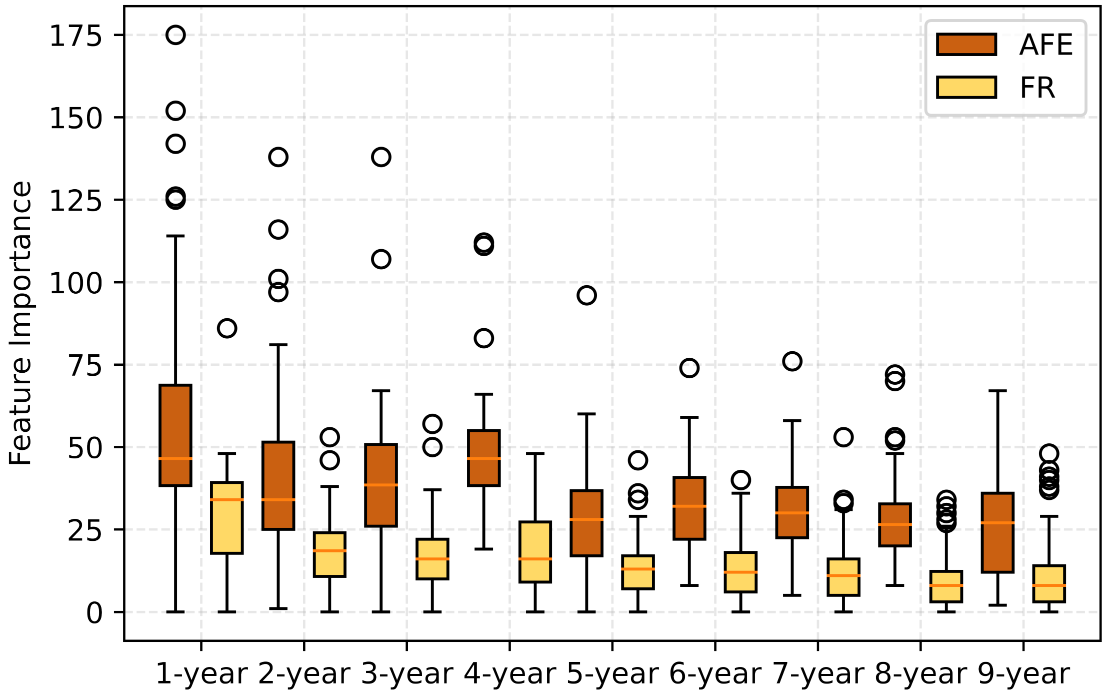


Figure 4.3: Feature importance of AFE and FR on each dataset

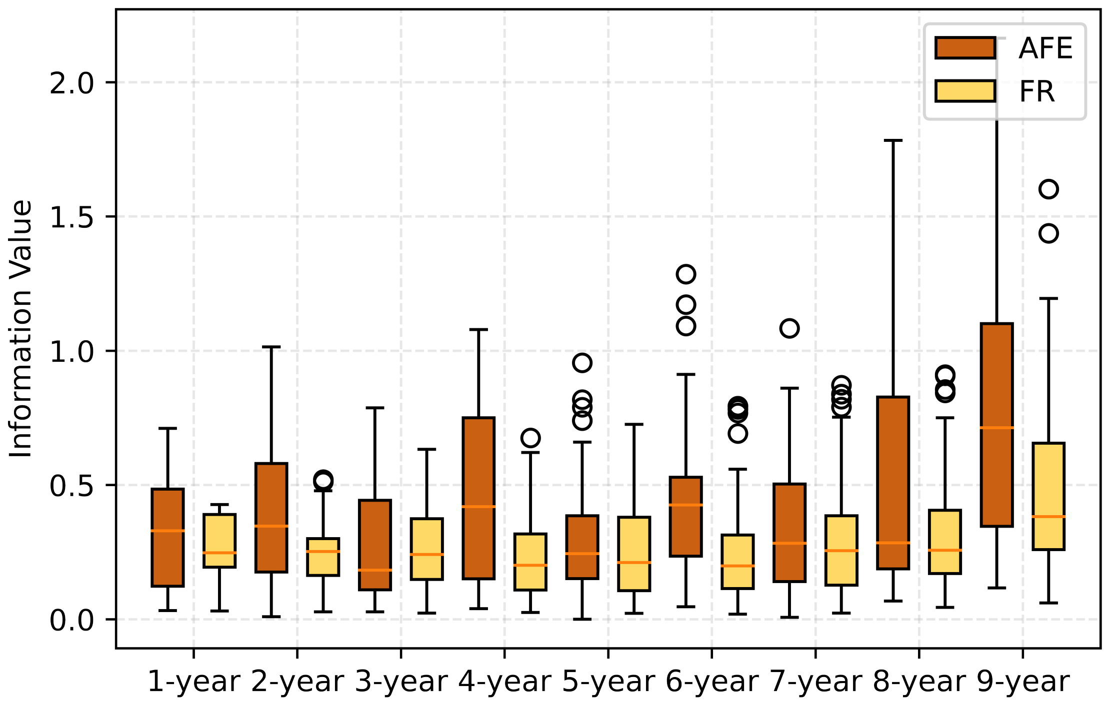


Figure 4.4: Information value of features in AFE and FR on each dataset

1-year and 2-year datasets and around 0.8 on other datasets, which is an impressive improvement compared to DeepFM. We also notice that the LightGBM model on AFE has a better performance than the logistic regression model and random forest on AFE. It is also observed that the LightGBM model with automatic feature engineering outperforms both the logistic regression model and the random forest model when employing automatic feature engineering.

Table 4.3: AUC of models trained on AFE and DeepFM

Dataset	**AFE**				**DeepFM**
	LR	**RF**	**LGB**	**MLP**	
1-year	0.7466	0.7940	**0.8713**	0.7658	0.6352
2-year	0.7542	0.7867	**0.8589**	0.7833	0.6502
3-year	0.7574	0.7725	**0.8458**	0.7872	0.6443
4-year	0.7797	0.7964	**0.8474**	0.8090	0.6560
5-year	0.6640	0.7517	**0.8206**	0.7292	0.6256
6-year	0.6734	0.8103	**0.8257**	0.7360	0.6303
7-year	0.6444	0.7693	**0.7997**	0.7120	0.6062
8-year	0.7722	0.7636	**0.8038**	0.7141	0.6262
9-year	0.7693	0.8207	**0.8301**	0.7993	0.6762

4.4.3 Comparison with Deep Feature Synthesis

We compare automatic feature engineering with deep feature synthesis in two ways. Fig. 4.5 illustrates the comparison, where all features generated by DFS are retained. The x-axis of the figure represents nine datasets. The y-axis depicts the improvement in AUC of models trained on AFE in comparison to the AUC of models trained on DFS. From this figure, it is evident that models trained on AFE display a clear advantage over those trained on DFS using all features. We also identified that the logistic regression model trained on AFE lacks the ability to provide favorable results compared to the random forest model and LightGBM model. This observation suggests that features generated by AFE perform more effectively within tree models than in linear models. We can consider this as the result of adopting the tree model to select features during the process of the automatic feature generation approach. In a nutshell, AFE outperforms DFS with all features in 28 out of 36 cases, indicating that it is still advantageous to use AFE features than DFS when training models.

Fig. 4.6 presents a comparison between the number of features se-

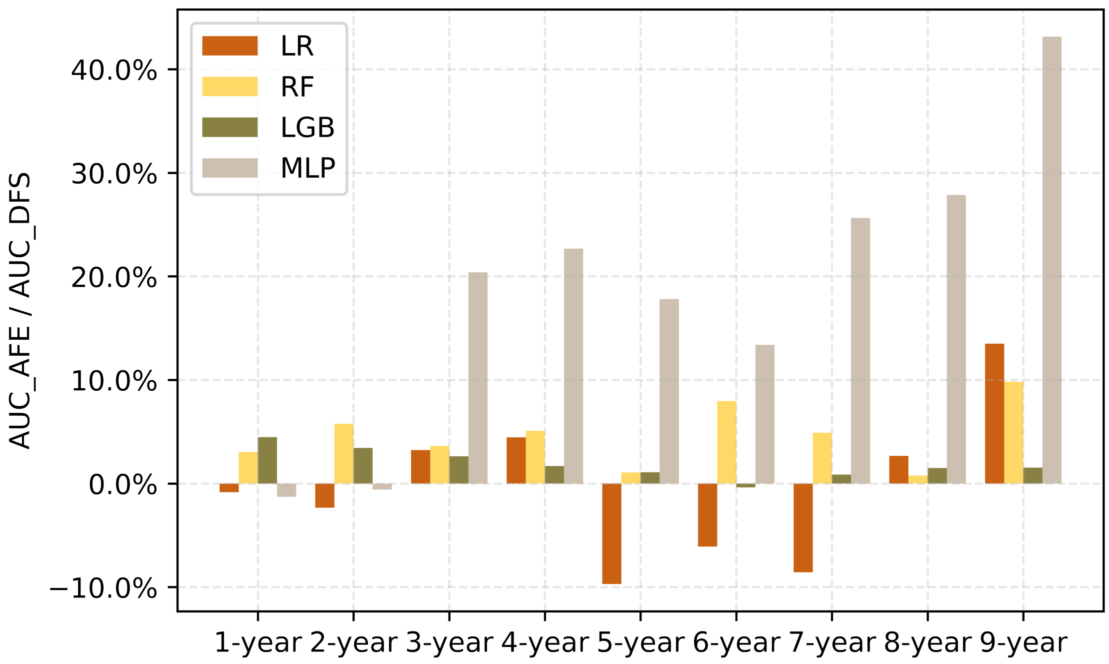


Figure 4.5: AUC improvement of AFE compared to DFS with all the features on each dataset

lected from the DFS and those selected in the AFE approach. The results are similar to those displayed in Fig. 4.5. In summary, AFE outperforms DFS with selected features in 28 out of 36 cases. In some instances, DFS with selected features yields better results than DFS with all features, indicating that eliminating redundant features during model training can enhance performance.

4.4.4 Comparison with Raw Data

To prove the necessity of feature engineering for the financial statements, we compare our approach with the raw data from financial statements. Since the data from financial statements consists entirely of numerical values, we implement the same data preprocessing steps to handle the extreme values and the missing values as we adopt in automatic feature engineering.

The comparison of raw data and automatic feature engineering can be found in Fig. 4.7. The AFE has a higher AUC in most cases and at the same time, AFE has less advantage in the logistic regression models but in total AFE outperforms raw data with selected features in 26 out of 36 cases, which means feature engineering indeed improves the

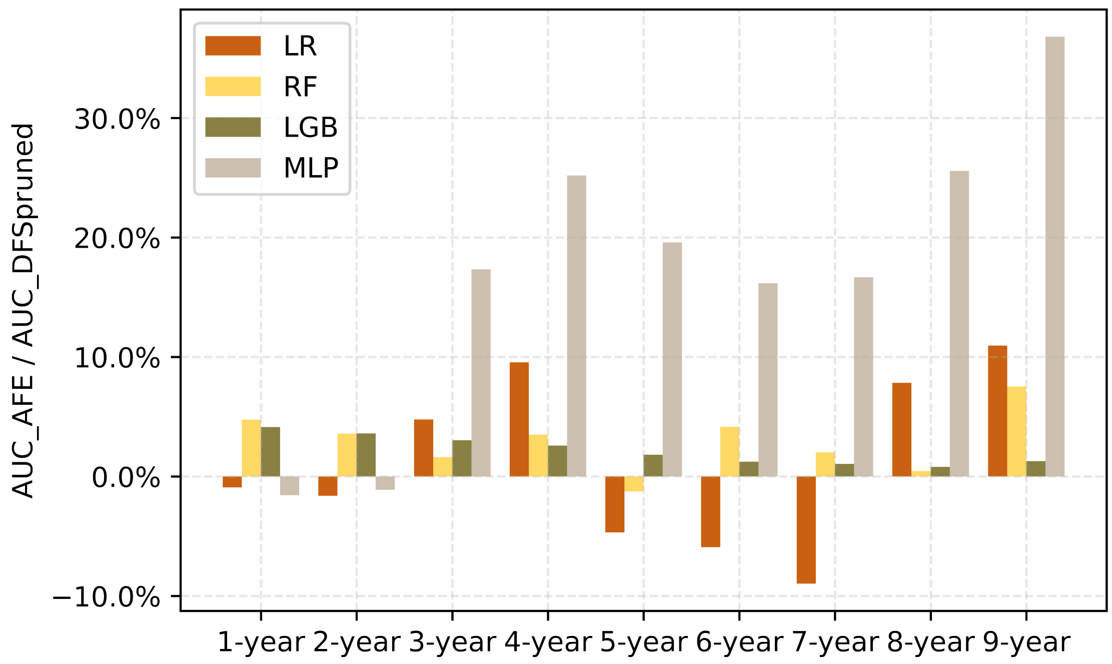


Figure 4.6: AUC improvement of AFE compared to DFS with the selected features on each dataset

predictive ability of the features for model training.

4.4.5 Explainability and Extensibility

4.4.5.1 Explainability

All features created by the AFE algorithm take the form of simple arithmetic expressions. Taking the AFE feature with the highest feature importance from the 1-year data set as an example: $(fid\_321 + fid\_322)$, where each term fid_x represents the name of one of the original numerical features derived from the financial statement of a company. fid_321 and fid_322 denote the profit or loss of the current and previous years, respectively. The sum of these two values indicates the profit or loss over the recent two years, which can be regarded as an important factor related to the business status of a company.

4.4.5.2 Extensibility

For this particular solution design and experiment, we adopted seven operands for the aggregation process and four operands for the crossing process. The automatic feature engineering process, however, could be

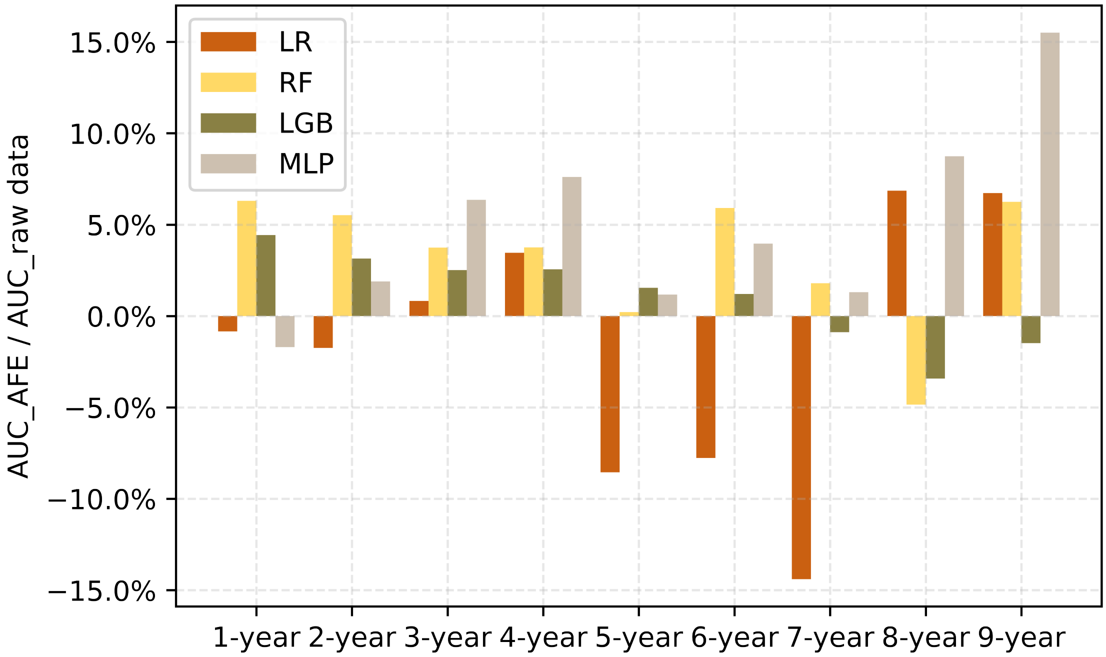


Figure 4.7: AUC improvement of AFE compared to raw data on each dataset

extended by adding more operands for both the aggregation and crossing. It depends on the users to decide the number of operands based on the available data set.

Chapter 5

Leveraging Large Language Models to Analyze Financial Statements

5.1 Problem Statement

The core objective of this research is to assess the effectiveness of small open-source LLMs in analyzing financial statements and making financial projections, compared to traditional methods and expert forecasts. Building on previous research (Section 2.2.3.2), which highlights the potential and limitations of small open-source LLMs in financial statement analysis and numerical reasoning, this chapter aims to identify the most effective models and methodologies for financial analysis tasks.

To achieve this, we address the following research questions in this chapter:

RQ1: How accurately can small open-source LLMs compute financial ratios based on provided financial statement data? LLMs must first identify relevant accounting subjects in financial statements and subsequently perform step-by-step calculations to derive key financial ratios. The accurate computation of financial ratios serves as a foundational step for subsequent analyses. The results are benchmarked against manually calculated values to evaluate their accuracy.

RQ2: How effectively can small open-source LLMs predict bankruptcy risks using methodologies such as the Altman Z-score model and DuPont analysis? LLMs utilize financial ratios from RQ1 to apply specific equations for the Altman Z-score and the DuPont analysis. Inferred bankruptcy risks are compared with ground-truth values to assess the reliability of LLM-generated predictions.

RQ3: How capable are small open-source LLMs in forecasting critical financial indicators? This question evaluates the ability of LLMs to independently predict key metrics, such as EBITDA and sales, for the upcoming financial year. Using only the financial statements for the current year and the embedded knowledge of the model, the LLM forecasts are compared to expert predictions to determine their practical applicability.

RQ4: What is the optimal combination of models and approaches balancing efficiency and effectiveness? LLMs are computationally intensive, which requires an evaluation of resource usage (e.g., time, CPU, and GPU memory). By analyzing the trade-offs across various approaches (e.g., zero-shot, few-shot, RAG, and fine-tuning), we aim to identify strategies that optimize both performance and efficiency.

To better study these questions, we prepared a special dataset to simulate a qualified and experienced financial analyst, allowing LLMs to acquire knowledge from this dataset through RAG or fine-tuning.

5.2 Experimental Design

5.2.1 Dataset and Data Preprocessing

For this research, data preparation involves selecting both training and validation datasets. Fig. 5.1 shows the process of constructing the training set and testing set. We have five raw data sources, including a question-answer pair dataset, raw PDF files, and publicly available accessible databases. By combining Compustat with the Institutional Brokers' Estimate System (IBES) using the company's stock ticker, a hybrid dataset of Compustat and IBES is constructed. The FinQA and CFA-QA datasets are only involved in the training set, the other three datasets are used in both training set and testing set. The details of these datasets will be introduced in the following sections.

FinQA Dataset : This dataset is constructed from financial experts' annotations on earning reports of S&P 500 companies and comprises unstructured documents and tables from financial reports that reflect the real-world finance context [32]. It provides a strong foundation for

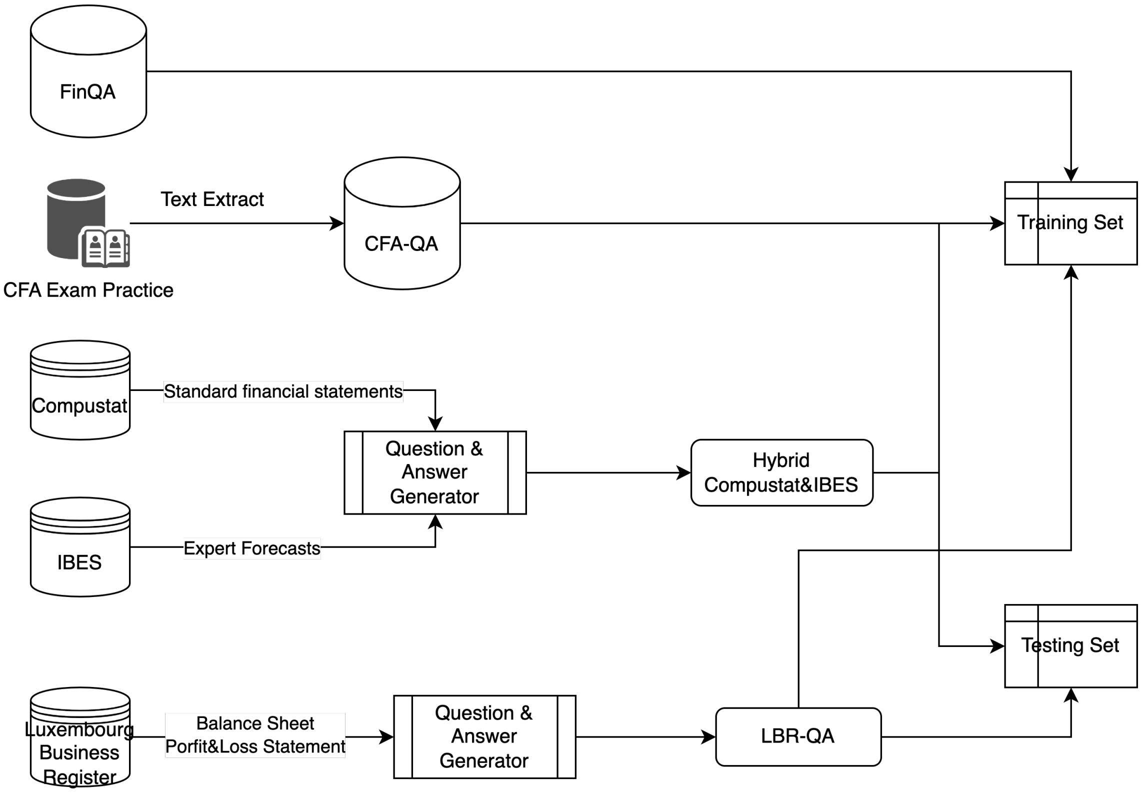


Figure 5.1: Workflow of constructing datasets for training and testing.

training the model to address both financial comprehension and quantitative reasoning questions. The dataset consists of training, validation, and testing sets, which can be accessed from GitHub[1]. In this chapter, we use its training set with 6251 samples for our training set and generated the question-answer pairs from its features. We then splice together the text from the features named `post_text`, `pre_text`, `table`, and `question` as the question, and spliced together the text from the features `answer` and `gold_evidence` as the answer.

CFA-QA Dataset : This dataset consists of 208 question-answer pairs derived from Level I CFA exam practice documents[2]. A study has demonstrated that, through few-shot learning, ChatGPT can pass all sections of accounting certification exams [52], indicating that LLMs may possess the ability to function as certified experts. As the Level I CFA exam covers various topics in financial statement analysis, this dataset is particularly valuable for LLMs with RAG and fine-tuning to align with expert-level financial analysis standards. We first extracted the text and recorded the table in Markdown format from the original

[1] https://github.com/czyssrs/FinQA/tree/main
[2] https://www.cfainstitute.org/

PDF files, then manually checked the correctness of extraction, and finally compiled the questions and answers as question-answer pairs of the CFA-QA dataset for further processing. The CFA-QA dataset aims to emulate the reasoning of a skilled financial analyst by incorporating challenging questions that require not only financial knowledge but also contextual understanding and judgment.

Compustat : This is a comprehensive database that provides standardized financial statements, company filings, market data, and other publicly available documents about North American and global companies. It is widely used in academic research due to its detailed financial disclosures, which enable robust calculations of metrics required for bankruptcy prediction and financial performance analysis. We used 50 accounting subjects from the standardized financial statements of North American companies. Since the performance evaluation concerns bankruptcy prediction, the focus is on the fiscal years 2014 to 2019, extracting 50 accounting subjects and excluding pandemic-related anomalies.

Institutional Brokers' Estimate System (IBES) : IBES provides forecast results on publicly traded American companies from expert analysts, including predictions for EBITDA and sales, which are essential for the study's focus on evaluating the forecasting performance of the LLMs. It provides a reliable benchmark against which model predictions can be compared. We chose the median value of the analyst-predicted EBTIDA and the median value of analyst-predicted sales from IBES as the baseline of human expert forecasting. In conjunction with the samples selected from Compustat, a total of 4,957 companies with 21,496 fiscal years were considered. To accommodate the experimental time for LLMs inference, 1,000 samples were randomly chosen for the training set and 1,000 samples for the testing set.

Luxembourg Business Register : LBR offers publicly available balance sheets and profit and loss statements of Luxembourg-based companies. These financial reports do not follow the standard accounting format and subjects, which is beneficial for LLMs to train and test the model's understanding of complex real-world financial statements. To ensure a more universal sample set, we exclude the finance-related companies as many financial institutions start companies in Luxembourg because of favorable policy. Since submitting the profit and loss statement is not mandatory in Luxembourg, to standardize the data, companies are included only if they had submitted both a balance sheet and profit and loss statement for the same fiscal year. This criterion ensures that all necessary financial ratios and metrics can be accurately computed for each sample, facilitating a consistent comparison between models. In total we have 15908 company's fiscal year as samples, then we randomly choose 1000 samples for training set and 1000 samples for testing set considering the experimental time of LLMs inference.

In summary, the number of samples included in the training and testing sets, along with their sources, is presented in Table 5.1.

Table 5.1: Summary of constructed datasets for LLMs' experiments

	Dataset	# samples
Training set	FinQA	6251
	CFA-QA	208
	Hybrid Compustat& IBES	1000
	LBR-QA	1000
Testing set	Hybrid Compustat& IBES	1000
	LBR-QA	1000

5.2.2 Methodology

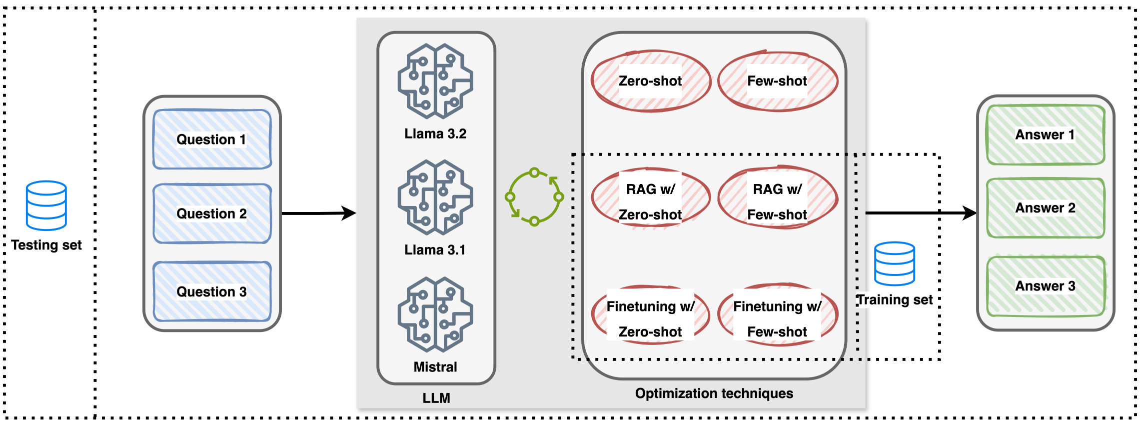


Figure 5.2: Workflow of experimental structure.

To understand which models and methods are most effective for analyzing financial statements, we chose three state-of-the-art small open-source LLMs: Llama 3.2 3B[3], Llama 3.1 8B[4], Mistral 7B[5]. These three models are all open-source, enabling us to have complete control over the model's architecture, parameters, and training data without dependence on third-party platforms. In contrast to closed-source models such as GPT-4, this transparency and controllability allow for flexible adjustments and optimizations. Furthermore, the research requires the exploration of diverse combinations of models and methods. Compared to the cost of using the GPT-4 API, the Llama and Mistral models can easily be run locally on a single GPU machine and thereby significantly reducing experimental costs. The capability of researching open-source models could offer enterprises or research institutions solutions rather than relying solely on commercial models.

Llama 3 models, particularly the latest version, exhibit competitive capabilities compared to leading models such as GPT-4, especially in multilingual support and complex reasoning tasks [49]. Llama 3.2, as the latest version, incorporates higher parameter optimization and knowledge updates, and holds the potential to perform exceptionally well in understanding complex language tasks and mathematical reasoning. In contrast, Llama 3.1, the previous version, can be utilized for comparisons to assist in analyzing whether version iterations yield significant improvements. Mistral emphasizes efficient parameter utilization, excelling in minimizing hallucinations and achieving performance that approaches leading models while using fewer parameters [70]. This characteristic makes it suitable for contrast experiments sensitive to resource efficiency, particularly for analyzing the actual performance of the model under limited computing power. We use the same setting for LLMs in this chapter considering the needs of comparison: max_new_tokens is set to 2048 to ensure a complete answer, the temperature is set to 0 or 1e-5 to maintain a consistent answer set, load_in_4bit is true to facilitate the smooth deployment of LLMs.

The settings for these LLMs can be found in Table 5.2.

To optimize the performance of these LLMs, this research employed three primary strategies: prompt engineering, retrieval-augmented generation (RAG), and fine-tuning. Prompt engineering involved zero-shot and few-shot learning. In zero-shot learning, no previous examples were

[3]https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct
[4]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[5]https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

Table 5.2: Settings of LLMs

Parameter	**Value**
max_new_tokens	2048
do_sample	True
temperature	0 or 1e-5
load_in_4bit	True
bnb_4bit_use_double_quant	True
bnb_4bit_quant_type	True
torch_dtype	bfloat16

provided, allowing the evaluation of the model's baseline capabilities. Few-shot learning was conducted by presenting the model with a limited number of question-answer pairs, testing its ability to generalize from minimal context in financial tasks. For RAG, a vector database was incorporated to retrieve domain-specific financial knowledge, which the models used to enhance accuracy in question answering and financial ratio computations. Fine-tuning was performed using supervised training on domain-specific question-answer pairs, allowing the models to align more closely with the requirements of financial statement analysis.

Fig 5.2 illustrates the overall experimental design, where the training set is exclusively used for RAG and fine-tuning, while the testing set evaluates all combinations of models and optimization techniques. We designs three categories of questions according to the RQs. Question 1 focused on computing financial ratios, Z-score values, and bankruptcy risks using the Altman Z-score model. Question 2 involved calculating financial ratios, return on equity (ROE), and bankruptcy risks by DuPont analysis. Question 3 is to ask for the predicted EBIDTA and sales based on provided financial statements and its own knowledge. Combining the financial statements from hybrid Compustat/IBES and LBR, we can have the full text of questions. For the answers, we populate the manually calculated financial ratios, Z-score value and ROE value into the fixed-format text as the ground truth.

With zero-shot learning and few-shot learning, LLMs will directly return the answers. We deploy RAG and fine-tuning in conjunction with the same prompts as used in zero-shot learning and few-shot learning for the questions. Therefore, there are six techniques in the optimization techniques part. Considering the LLMs, in total, we have 18 different combinations of LLMs and optimization techniques, which constitute a comprehensive evaluation of how LLMs can be adapted to tackle financial analysis tasks.

5.2.3 Evaluation Metrics

The inference tasks of this research not only emphasize text generation, but also highlight the importance of the correctness of mathematical calculations related to financial ratios. Therefore, to fully evaluate the effectiveness of the model, we apply a set of evaluation metrics across the four research questions.

Completion rate : In this chapter, we particularly define a metric named completion rate for the research questions 5.1. For Question 1 to Question 3, we require the LLMs to summarize the required values in JSON format. Therefore, it it vital for a qualified answer to have this complete JSON to present the required calculated or forecasted values of corresponding questions. The completion rate is defined in equation 5.1.

$$R = \frac{\sum_{i=1}^{N}(\mathrm{A}_i \cdot \mathrm{B}_i \cdot \mathrm{C}_i)}{N} \tag{5.1}$$

where, N means the total number of generated answers, A_i represents whether the i-th answer contains a valid JSON format. It is 1 if valid, otherwise 0. B_i indicates whether the JSON contains all the required fields. It is 1 if all fields are present, otherwise 0. C_i checks if the values of the fields in the JSON are numbers (either integers or floats). It is valued at 1 if all values are numeric; otherwise, it is 0.

Recall-Oriented Understudy for Gisting Evaluation(ROUGE) : ROUGE can measure the degree of overlap between the generated answers and the reference answers in terms of n-grams or the longest common subsequence, with particular emphasis on coverage [100]. In this chapter, we employed ROUGE-L to evaluate the calculation steps of financial ratios or the reasoning behind predictions, as it not only assesses whether the generated text covers the reference content but also pays special attention to whether the answers are presented in sequence. In this chapter, we used the rouge-score package in Python for calculation[6].

[6] https://pypi.org/project/rouge-score/

Symmetric Mean Absolute Percentage Error (sMAPE) : MAPE measures the percentage error relative to the actual value, which means it is scale-independent among all the financial ratios that LLMs need to forecast. However, when actual values are near zero, MAPE can become excessively large or undefined. sMAPE (see equation 5.2) avoids the problem of infinite values when actual values are zero, making it more reliable in such cases.

$$\text{sMAPE} = \frac{1}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{\frac{|y_i| + |\hat{y}_i|}{2}} \times 100 \tag{5.2}$$

where, y_i is the actual value for the i-th data point, $\hat{y}_i$ is the predicted value for the i-th data point, n is the total number of data points.

5.3 Results analysis

5.3.1 Answers completion

Fig. 5.3 highlights clear distinctions in the performance of the three LLMs across optimization strategies. Llama 3.1 8B outperforms its counterparts in 4 scenarios, particularly excelling in zero-shot learning and finetuning with few-shot learning. Llama 3.2 3B, while demonstrating strong general performance, exhibits minor declines in completion rates under specific fine-tuning and RAG scenarios, suggesting some sensitivity to the optimization approach. Mistral 7B, although competitive in RAG with zero-shot learning, significantly lags behind in other settings, indicating potential architectural or pre-training limitations in handling structured output requirements.

These results underscore the significance of aligning model selection and optimization strategies with specific task requirements. Llama 3.1 8B and Llama 3.2 3B emerge as reliable choices for tasks demanding consistent and complete outputs, while Mistral 7B's use may be more suited to resource-constrained scenarios or specific RAG applications.

5.3.2 Evaluation on calculation steps

Table 5.3 reveals distinct performance patterns among the three LLMs across the Altman Z-score model and DuPont analysis. Llama 3.1 8B

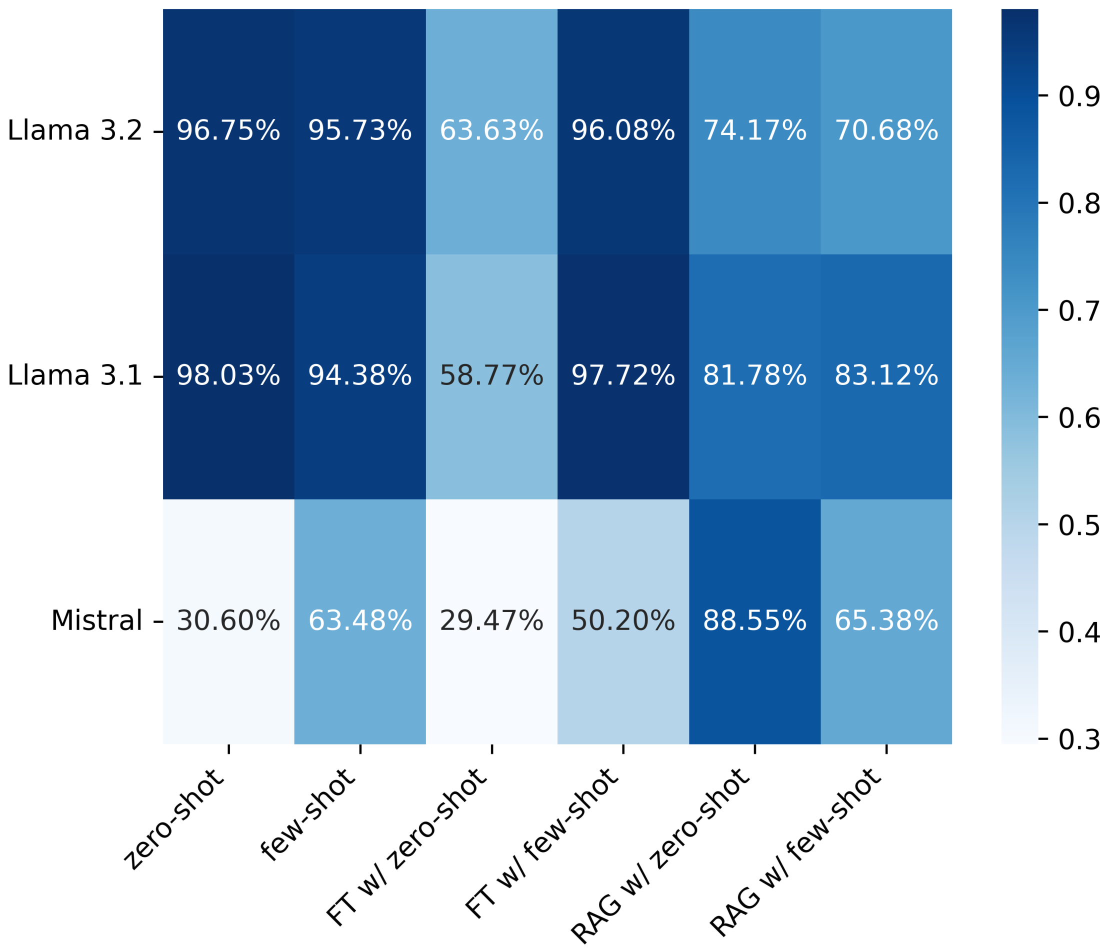


Figure 5.3: Distribution of completion rate over different combinations of LLMs and optimization techniques.

consistently achieves the highest overall performance, excelling particularly in fine-tuning tasks, where it demonstrates superior F1 scores for both analysis methods. Llama 3.2 3B performs well in structured optimization tasks but underperforms in certain retrieval-augmented generation (RAG) scenarios. Mistral 7B, while generally less effective, exhibits competitive results in RAG-based tasks, particularly with the DuPont analysis.

For the Altman Z-score model, Llama 3.1 8B dominates in fine-tuning (87.60% F1), while Mistral 7B performs better in zero-shot RAG tasks (75.82%). In the DuPont analysis, Llama 3.1 8B also leads in fine-tuning scenarios, while Mistral 7B achieves its highest performance in RAG with zero-shot learning (89.33%), surpassing both Llama models. Across both methods, introducing few-shot examples in RAG results in slight performance declines for most models; however, Llama 3.1 8B maintains its lead.

Table 5.3: ROUGE-L comparison of different combinations of LLMs and optimization techniques

		Altman Zscore Model			DuPont analysis		
		Recall	Precision	F1 score	Recall	Precision	F1 score
Llama 3.2	zero-shot	31.80%	35.90%	33.06%	30.90%	39.17%	34.34%
	few-shot	12.27%	62.93%	19.21%	8.14%	70.58%	13.10%
	FT w/ zero-shot	79.70%	90.50%	84.30%	88.18%	92.48%	89.92%
	FT w/ few-shot	50.08%	80.87%	56.53%	85.16%	88.79%	86.89%
	RAG w/ zero-shot	75.27%	69.59%	69.69%	59.15%	58.61%	56.29%
	RAG w/ few-shot	43.94%	52.46%	46.51%	59.29%	63.48%	58.96%
Llama 3.1	zero-shot	29.50%	41.73%	31.36%	31.99%	47.08%	35.73%
	few-shot	60.36%	79.94%	68.49%	48.19%	68.97%	55.79%
	FT w/ zero-shot	82.77%	93.70%	**87.60%**	88.01%	**95.58%**	**91.50%**
	FT w/ few-shot	70.46%	90.18%	78.78%	**89.27%**	93.45%	91.26%
	RAG w/ zero-shot	**83.00%**	87.57%	84.73%	88.93%	67.81%	75.35%
	RAG w/ few-shot	68.44%	79.77%	73.57%	80.36%	85.65%	82.02%
Mistral	zero-shot	30.62%	52.89%	37.74%	24.74%	36.94%	29.28%
	few-shot	36.73%	41.92%	38.43%	85.14%	78.33%	80.74%
	FT w/ zero-shot	66.12%	**96.94%**	78.04%	86.08%	95.32%	90.32%
	FT w/ few-shot	34.64%	54.25%	42.10%	85.09%	88.11%	86.08%
	RAG w/ zero-shot	73.31%	80.40%	75.82%	88.10%	90.82%	89.33%
	RAG w/ few-shot	52.13%	79.41%	62.77%	84.72%	91.60%	87.99%

5.3.3 Financial Metric Calculation Accuracy

Fig 5.4 shows significant variation in model performance across datasets, ratios, and optimization configurations. Llama 3.2 3B demonstrates the most notable improvement in the Altman Z-score Model, reducing sMAPE from 186.8 (zero-shot) to 135.0 (RAG with few-shot). Similarly, Llama 3.1 8B shows effective enhancement in the Working Capital/Total Assets ratio, where sMAPE improves from 96.1 to 75.9 with few-shot learning. In contrast, Mistral 7B displays inconsistencies, particularly in ratios such as Earnings Before Interest and Tax/Total Assets, where RAG with zero-shot leads to a high sMAPE of 191.1, indicating limited benefit from additional vector database information.

RAG with few-shot consistently emerges as the most reliable method, particularly for complex financial prediction tasks. However, ratios involving equity and earnings, such as the Market Value of Equity/Total Liabilities and Earnings Before Interest and Tax/Total Assets, remain challenging due to their sensitivity to financial volatility. High sMAPE values, such as 196.3 (Llama 3.1 8B) and 161.7 (Mistral 7B) for equity-related ratios, highlight the need for improved approaches.

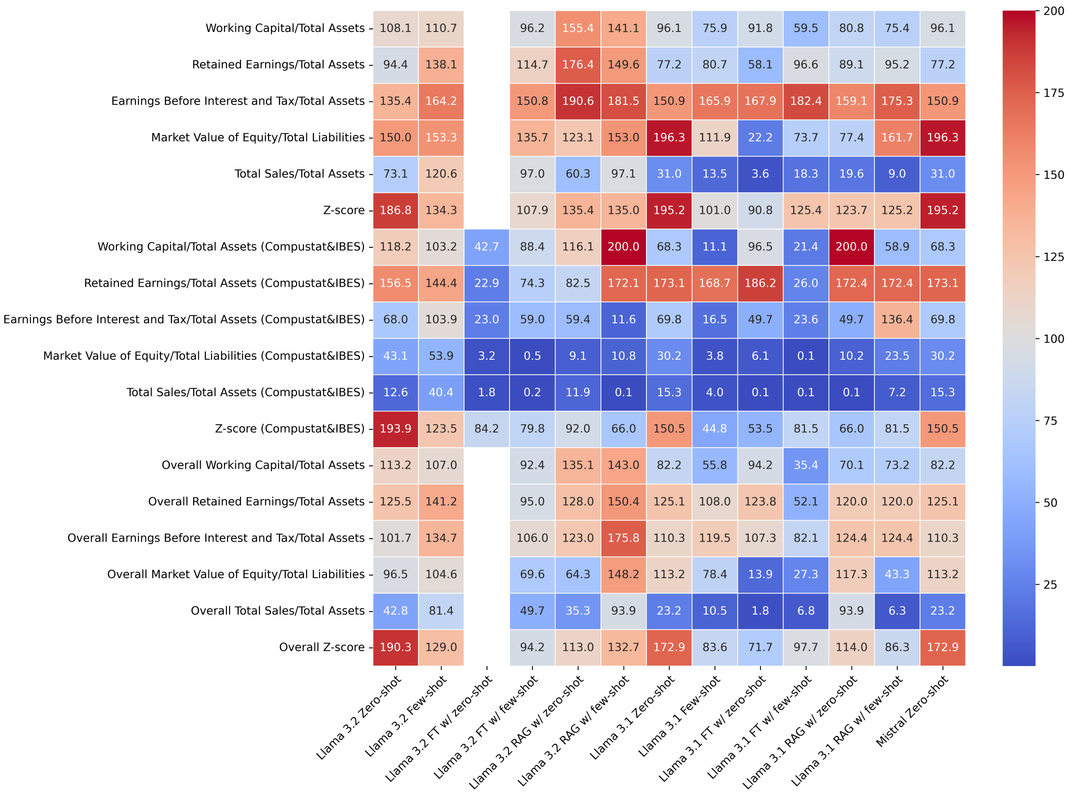


Figure 5.4: sMAPE for financial ratios by over different combinations of LLMs and optimization techniques.

While the overall sMAPE is high, for certain ratios, such as total sales to total assets (Compustat&IBES), all the LLMs perform well, which means LLMs indeed have the potential to anaylze the financial statements.

5.3.4 Bankruptcy Prediction

Table 5.4 reveals significant variability in LLM performance for bankruptcy prediction, with results heavily influenced by the optimization strategy. Llama 3.2 3B shows the most consistent performance in bankruptcy prediction, particularly with zero-shot learning, achieving up to 82% accuracy and 0.62 AUC for DuPont analysis. However, its performance declines in few-shot learning and fine-tuning, highlighting the limitations of these methods. Llama 3.1 8B underperforms overall but demonstrates potential in combining retrieval-based techniques with few-shot training, achieving an AUC of 0.76 for the Altman Z-score model. Mistral 7B delivers mixed results, showing competitive zero-shot accuracy but poor performance in fine-tuning, particularly for DuPont analysis.

Table 5.4: Performance evaluation for bankruptcy prediction by LLMs

		Altman Zscore Model		DuPont Analysis	
		Accuracy	AUC	Accuracy	AUC
Llama 3.2	zero-shot	79%	0.61	82%	0.62
	few-shot	78%	0.36	74%	0.44
	FT w/ zero-shot	/	/	46%	0.59
	FT w/ few-shot	79%	0.52	77%	0.53
	RAG w/ zero-shot	63%	0.56	35%	0.49
	RAG w/ few-shot	64%	0.50	57%	0.30
Llama 3.1	zero-shot	66%	0.65	66%	0.59
	few-shot	61%	0.58	53%	0.61
	FT w/ zero-shot	/	/	44%	0.48
	FT w/ few-shot	73%	0.62	69%	0.58
	RAG w/ zero-shot	60%	0.65	47%	0.46
	RAG w/ few-shot	66%	0.76	51%	0.58
Mistral	zero-shot	79%	0.67	67%	0.75
	few-shot	/	/	65%	0.62
	FT w/ zero-shot	/	/	22%	0.41
	FT w/ few-shot	/	/	69%	0.30
	RAG w/ zero-shot	65%	0.61	67%	0.63
	RAG w/ few-shot	/	/	53%	0.39

Overall, Llama 3.2 3B is the most reliable model for bankruptcy prediction, but its variability across optimization methods underscores the need for more robust strategies tailored to financial tasks.

5.3.5 EBITDA and Sales Forecasting

In Table 5.5, we only put the best forecasting from LLMs and compare it with the forecasts from human financial expert. The financial expert achieved exceptionally low sMAPE values of 25.1 for "Next Year Sales" and 44.9 for "Next Year EBITDA," far surpassing the results obtained by all LLM configurations(Table 10.2 in the Appendix 10.2). This significant gap in accuracy indicates that, despite advances in machine learning and natural language processing, LLMs are not yet capable of matching the forecasting precision of experienced financial analysts, particularly regarding complex financial metrics that require nuanced judgment and domain expertise.

Table 5.5: Comparison of the forecasting ability of LLMs and financial expert.

	Next Year Sales Prediction	Next Year EBITDA Prediction
Llama 3.2|zero-shot	/	129.6
Llama 3.1|few-shot	123.2	/
Expert Forecasting	**25.1**	**44.9**

5.3.6 Resources Consumption

In this chapter, we analyze the time, CPU memory, and GPU memory consumption across different models and optimization methods and reveal key performance trade-offs. The detailed records can be seen from Fig. 10.1, Fig. 10.3 and Fig. 10.2 in the Appendix 10.3. Llama 3.1 8B offers the most consistent performance, particularly in few-shot optimization, with the fastest response times (50 seconds). Mistral 7B also excels in few-shot scenarios but is less effective in more complex methods. Llama 3.2 3B, while delivering high performance, requires significantly more computational resources, especially for RAG-based tasks, with response times reaching up to 600 seconds.

Regarding CPU consumption, all models exhibit similar usage, with slight increases under RAG methods, particularly for Llama 3.2 3B. However, CPU requirements are not a major constraint for any model, with usage staying below 2.5GB in most cases. GPU consumption shows more variation, with Llama 3.1 8B consuming the most GPU memory (over 5GB), while Llama 3.2 3B is the most resource-efficient, particularly in zero-shot and few-shot learning scenarios.

In conclusion, Llama 3.1 8B offers the best balance of efficiency and performance for low-latency tasks, Mistral 7B is suitable for few-shot optimization in resource-constrained settings, and Llama 3.2 3B excels in high-quality tasks but requires more computational power, especially for complex optimization strategies like RAG.

Chapter 6

Incorporating Company Adjustments: Hybrid Datasets to Improve the Bankruptcy Risk Prediction

Contents

6.1 Problem Statement **73**

6.2 Methodology **74**

6.2.1 Conceptual Framework 75

6.2.2 Variables 77

6.3 Experimental Setup **79**

6.3.1 Dataset Description 79

6.3.2 Models 80

6.3.3 Performance Evaluation 83

6.4 Results & Discussion **84**

6.4.1 Features Evaluation 84

6.4.2 Ablation Experimental results 85

6.4.3 Performance about Covid Period 86

6.1 Problem Statement

The development of bankruptcy prediction models has a long history, beginning with the introduction of the Logit model by Beaver in 1966 and the Z-score model by Altman in 1968. These models, which rely on financial ratios derived from a company's financial statements, have laid the groundwork for subsequent research in the field. Financial ratios, which assess a company's financial health based on accounting data, have traditionally been the primary input for bankruptcy prediction models. Over time, numerous studies have expanded on this approach, employing various models to enhance predictive accuracy. Some studies have also incorporated additional data types, such as market-based variables and macroeconomic indicators, to improve predictions. However, while there has been extensive research on the models themselves, the exploration of alternative input data or features has been relatively limited.

In this chapter, we aim to enhance the accuracy of bankruptcy prediction models by incorporating data on reported company adjustment behaviors alongside traditional accounting-based ratios. Using a publicly available dataset from LBR, which includes basic company information, business operations, and financial statements, we create a hybrid dataset. This dataset combines accounting-based ratios with features related to restructuring behavior. To evaluate the effectiveness of this approach, we compare the bankruptcy prediction results of six machine learning models: logistic regression (LR), random forest (RF), lightGBM (LGB), multilayer perceptron (MLP), convolutional neural network (CNN), and long short-term memory (LSTM). We further assess the robustness of these models by comparing their performance on data from periods before and after the COVID-19 pandemic, which introduces significant economic challenges. By advancing the understanding of how company adjustments can be used in conjunction with traditional financial ratios, this chapter aims to improve bankruptcy prediction models, offering valuable tools for SMEs, financial institutions, and policymakers.

6.2 Methodology

The main focus of this chapter is to examine the effectiveness of reported company adjustment behaviors in predicting bankruptcy. We also aim to analyze the robustness of various bankruptcy models during the Covid-19 pandemic. In this section, we first present the overall framework for investigating these problems. Then, we focus on the details of the input data and explain the experimental design.

6.2.1 Conceptual Framework

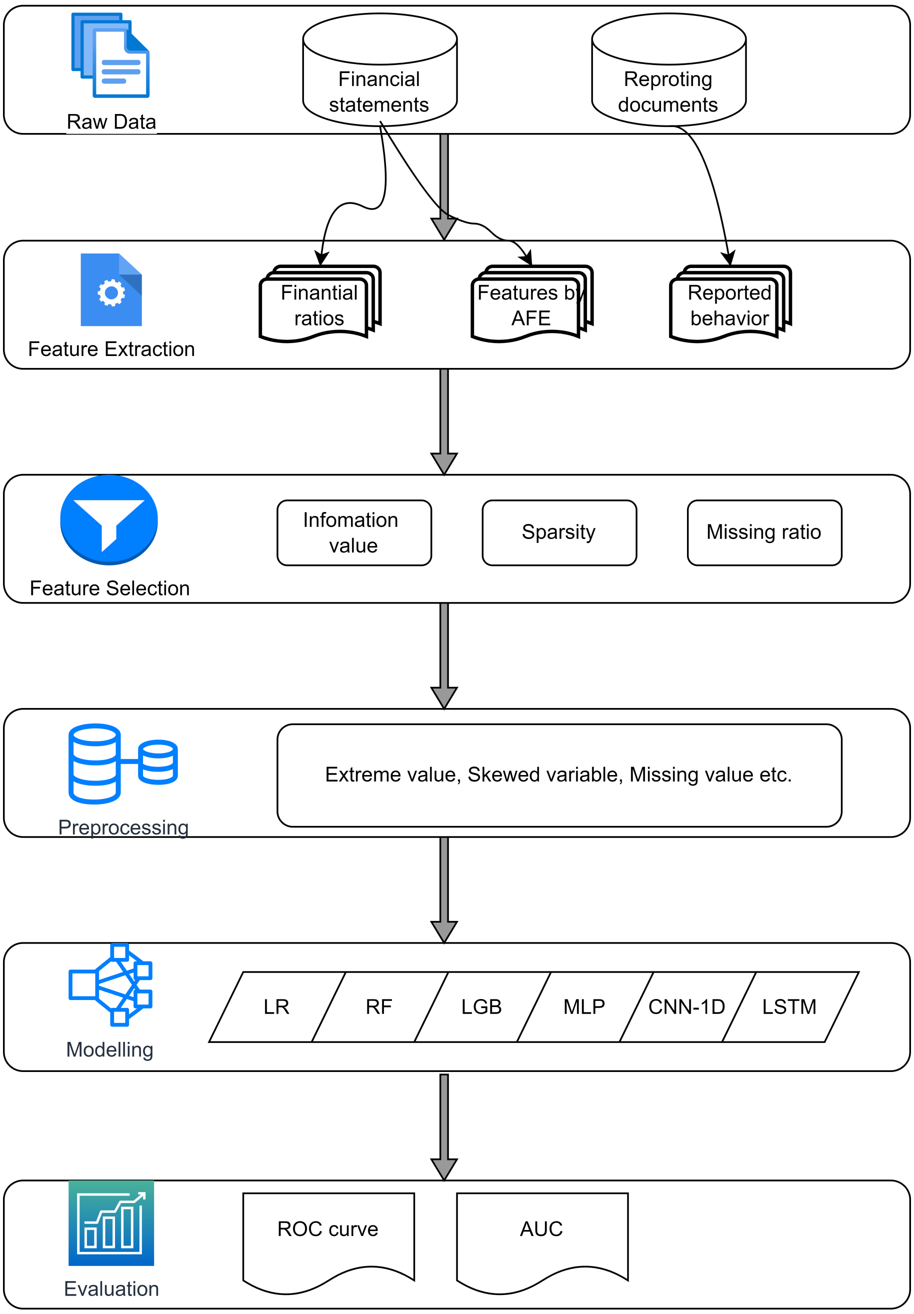


Figure 6.1: Conceptual framework of experimental design

Fig. 6.1 illustrates the six stages of a framework designed to conduct a comparative study of bankruptcy models using different input data. The data used in this chapter consists of financial statements and reporting documents. However, since the cash flow statement and profit & loss statement are not included, the financial statements only consisted of the balance sheet. The reporting documents that companies submit to disclose their operational behaviors are usually classified as textual files.

Different methods are used to extract three types of features from the raw data. The first is financial ratios. Since we do not have a cash flow statement or profit & loss statement, we create as many financial ratios as possible. The second type also includes accounting-based features. These variables are constructed using an automatic feature engineering method that we developed in our previous work. This method is capable of creating highly effective features even with limited data [171]. The final type of features is behavior-related features. We design these variables based on company adjustments, such as changes in registered addresses, manager resignations, and mergers and acquisitions.

The next stage of the framework involves selecting features from the current variables to eliminate unfavorable and redundant variables caused by sparsity, missing, and repetition. The information value (IV) is an indicator used to measure the predictive power of an independent feature [67]. A higher information value indicates that the feature has greater predictive power. The formula for calculating information value is as follows [149]:

$$IV = \sum_{i=1}^{n} \left(\frac{G_i}{G} - \frac{B_i}{B}\right) * \ln \frac{G_i/G}{B_i/B} \tag{6.1}$$

We select features with an IV value greater than 0.02 and a missing rate less than 0.7.

The data are pre-processed to address missing values, infinite values, and skewed variables, making it more suitable for modeling. We also exclude abnormal samples, such as companies that have submitted financial reports prior to the reference year. We replace infinite values with the highest finite value. In the fifth stage, we assess the effectiveness of behavior-related features by comparing the prediction performance of hybrid datasets that include both behavior-related features and accounting-based ratios with datasets that only contain accounting-based ratios. We train six popular models, including logistic regression (LR), random forest (RF), LightGBM (LGB), multiple perceptron (MLP), convolutional neural network (CNN) and long-short-term memory (LSTM) to compare their prediction results. We use the receiver operating characteristic curve (ROC curve) and Area under the ROC curve (AUC) as indicators of evaluating the performance of models, which are commonly used and discussed in Section 6.3.3.

6.2.2 Variables

The state of a company is not static and can change over time, either by being established or going bankrupt. This means that the company may enter or exit the sample set. We utilize a sliding time window approach to sample from the raw data. The sliding time window is a technique used to extract data from a time series dataset by defining a fixed period of time (window) and moving it forward by a certain interval (step size). This technique allows for continuous monitoring of system states [178].

As of June 2022, there are 74,611 companies in Luxembourg. The average lifespan of companies is approximately 3.5 years. Therefore, we have selected a timeframe of up to 3 years for predicting bankruptcy. We create datasets with three different windows (1-year, 2-year, and 3-year) to predict one step forward (one year). According to the timeline (see Fig. 6.3), the three datasets consist of 1-year data from t_{-1} to t_0, 2-year data from t_{-2} to t_0, and 3-year data from t_{-3} to t_0. The sample size of these three datasets, including solvent and bankrupted companies, was summarized in 6.1. However, there is another category of companies with an unknown status. Some companies have not uploaded annual reports or declared bankruptcy, which contributes to the variation in data from year to year. As depicted in the Fig. 6.2, the bankruptcy rate of SMEs in Luxembourg has decreased over the past decade. It may indicate that the business conditions of SMEs are improving or that the overall economic environment has improved, resulting in greater stability for SMEs. Additionally, other factors such as policy support or industry changes may also influence the bankruptcy rate of SMEs. It is surprising to find that the bankruptcy rate of SMEs has actually increased during the Covid-19 pandemic, suggesting that fewer SMEs are going bankrupt compared to previous periods. We hypothesize that this could be attributed to government financial assistance during the special period. Some companies may be technically bankrupt but have not yet filed for bankruptcy due to delays in filing.

As mentioned earlier, we derive three types of features from raw data: two accounting-based variables and one behavior-based variable. The statistics and descriptions of these features can be found in the table 6.2. SMEs are not required to prepare and disclose cash flow statements and income statements. Therefore, we can only calculate 18 financial indicators based on the available data [23, 174, 187, 188]. We develop an

Table 6.1: Summary of three datasets sampled by sliding time window approach

Year	**1-year**		**2-year**		**3-year**	
	Solvent	Bankrupt	Solvent	Bankrupt	Solvent	Bankrupt
2012	21738	621	/	/	/	/
2013	23804	687	17087	461	/	/
2014	25686	669	18790	451	16512	361
2015	27331	663	20301	436	18188	361
2016	28781	653	21475	461	19477	384
2017	30748	661	22789	449	20755	378
2018	32718	606	24419	392	22061	322
2019	34557	431	25793	319	23504	267
2020	36309	179	27034	138	24596	121
2021	22387	34	17195	28	15571	24

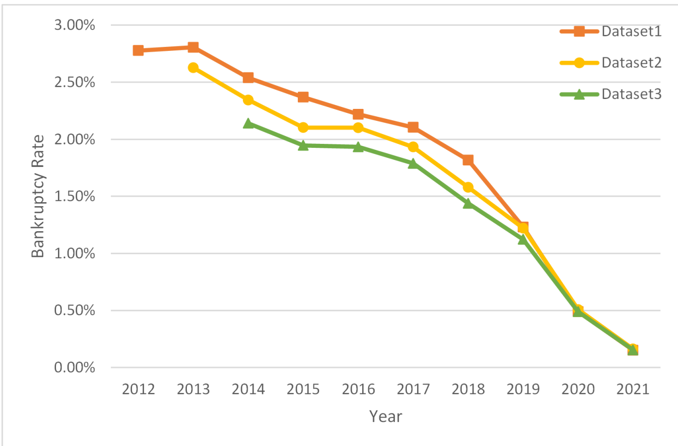


Figure 6.2: Bankruptcy rate of three datasets from 2012 to 2021

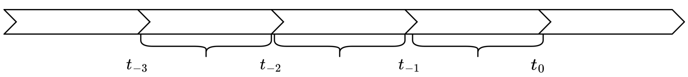


Figure 6.3: Definition of time period

algorithm for automatic feature engineering in Chapter 4 to derive as many useful features as possible from financial statements to address issues caused by the absence of certain financial statements or data quality problems. This algorithm maximizes data mining to generate high-quality features that enhance prediction accuracy. Behavior-based variables are derived from information reported by SMEs regarding company adjustments, including both statistical and trend variables.

Table 6.2: Description of variables in Chapter 4

Variable	**Description**
Financial ratios (FR)	current ratio, debt to equity, working capital to total assets ,total liabilities to total assets, equity to total assets, quick ratio ,current assets to total assets ,cash to total assets,cash to current liabilities ,long term debt to equity,total assets growth rate,quick assets to total assets,current assets to current liabilities,(cash or marketable securities) to total assets,total debt to total assets,equity to fixed assets,current assets to total liabilities,short-term liabilities to total assets
Automatic feature engineering (AFE)	automatically generate features from financial statements, which can adapt to any kind of numerical data
Reported company adjustment behavior-related features (RB)	Modification of name or corporate name, registered office, social object, administrator/manager, daily management delegate, associate, person in charge of checking the accounts, Social capital/social funds, managing director / steering committee, duration, legal form, social exercise, permanent representative of the branch, merger / demerger, depositary, transfer of business assets, assets or business sectors, address, trading name, activities, manager, seat, reason, name, chairman / director, personne autorisée à gérer, administrer et signer, person with the power to commit the company, ministerial approval

6.3 Experimental Setup

6.3.1 Dataset Description

We divide the datasets into two parts: the training set and the testing set, as outlined in table 6.3. To maintain consistency between the training set and the testing set, we divide the data from 2012 to 2018 into a 70% training set and a 30% testing set. Additionally, we creat two additional testing sets: one using solvent and bankrupt SMEs from 2019 as a pre-Covid testing set, and another using solvent and bankrupt SMEs from 2020 and 2021 as a post-Covid testing set. To train our models, we utilize the 5-fold cross-validation method and do not set aside a separate validation set. Table 6.3 have a bankruptcy rate below 3%, making them highly imbalanced. The negative datasets are typically large in size, so we use the under-sampling method during data preprocessing to balance the rate to 25%.

Table 6.3: Description of training, testing, Pre-Covid and Post-Covid datasets

		1-year	**2-year**	**3-year**
Training	Solvent (Negative)	110805	87467	67920
	Bankrupt (Positive)	2625	1797	1244
	Bankruptcy Rate	2.31%	2.01%	1.80%
Testing	Solvent (Negative)	47458	37403	29081
	Bankrupt (Positive)	1155	853	562
	Bankruptcy Rate	2.38%	2.23%	1.90%
Pre-Covid	Solvent (Negative)	28730	25793	23504
	Bankrupt (Positive)	368	319	267
	Bankruptcy Rate	1.26%	1.22%	1.12%
Post-Covid	Solvent (Negative)	48846	44229	40167
	Bankrupt (Positive)	181	166	145
	Bankruptcy Rate	0.37%	0.37%	0.36%

6.3.2 Models

In this chapter, we choose six bankruptcy prediction models, which include statistical, machine learning, and deep learning models, by synthesizing the statistics from previous studies in Part II. We comprehensively evaluate the behavior-based features by comparing the performance of representative models. The table 6.4 displays the environmental information used for model training.

Table 6.4: Information of training machine

Device name	Tesla V100-SXM2-32GB
Linux version	Red Hat 8.5.0-10
Python version	3.8.6
Pytorch version	1.10.1+cu111
Cuda version	11.1
Cudnn version	8005
Sklearn version	1.2.1
Number of GPU	2
Number of CPU	16

Logistic Regression predicts the likelihood of a binary outcome using one or more predictor variables. Logistic regression models have advantages in bankruptcy prediction due to their simplicity, fast computation, and better results when dealing with smaller datasets. In this research, we adopt *LogisticRegression* from *sklearn* package and use *GridSearch* to determine the optimal parameters within a specific range 6.5.

Table 6.5: Range of parameters for LR models

Parameter	**Range**
C	[0.01,0.1,0.5,1]
max_iter	[500,1000,2000]
penalty	[l1,l2]

Random Forest creates a forest of decision trees, with each tree being trained on a random subset of the data and a random subset of predictor variables. Random forest models outperform single decision tree models and other classification models in terms of predictive performance and robustness, and can effectively handle high-dimensional and complex datasets. In this research, we utilize *RandomForestClassifier* from *sklearn* package and employ *GridSearch* to determine the optimal parameters within a specific range 6.6.

Table 6.6: Range of parameters for RF models

Parameter	**Range**
max_depth	[2,3,4, 5]
n_estimators	[10,20,35,50]

LightGBM prioritizes speed and efficiency, specifically for managing large datasets. The method utilizes a gradient-based approach to construct decision trees and incorporates various optimization techniques to accelerate the training process. In this research, we adopt *LGBMClassifier* from *lightgbm* package and use *GridSearch* to decide the best parameters from a specific range 6.7.

Table 6.7: Range of parameters for LGB models

Parameter	**Range**
max_depth	[3, 4, 5]
num_leaves	[5, 6, 7, 14, 15, 30, 31]
learning_rate	[0.01,0.05]
reg_alpha	[0,10,100,1000]
reg_lambda	[0,10,100,1000]

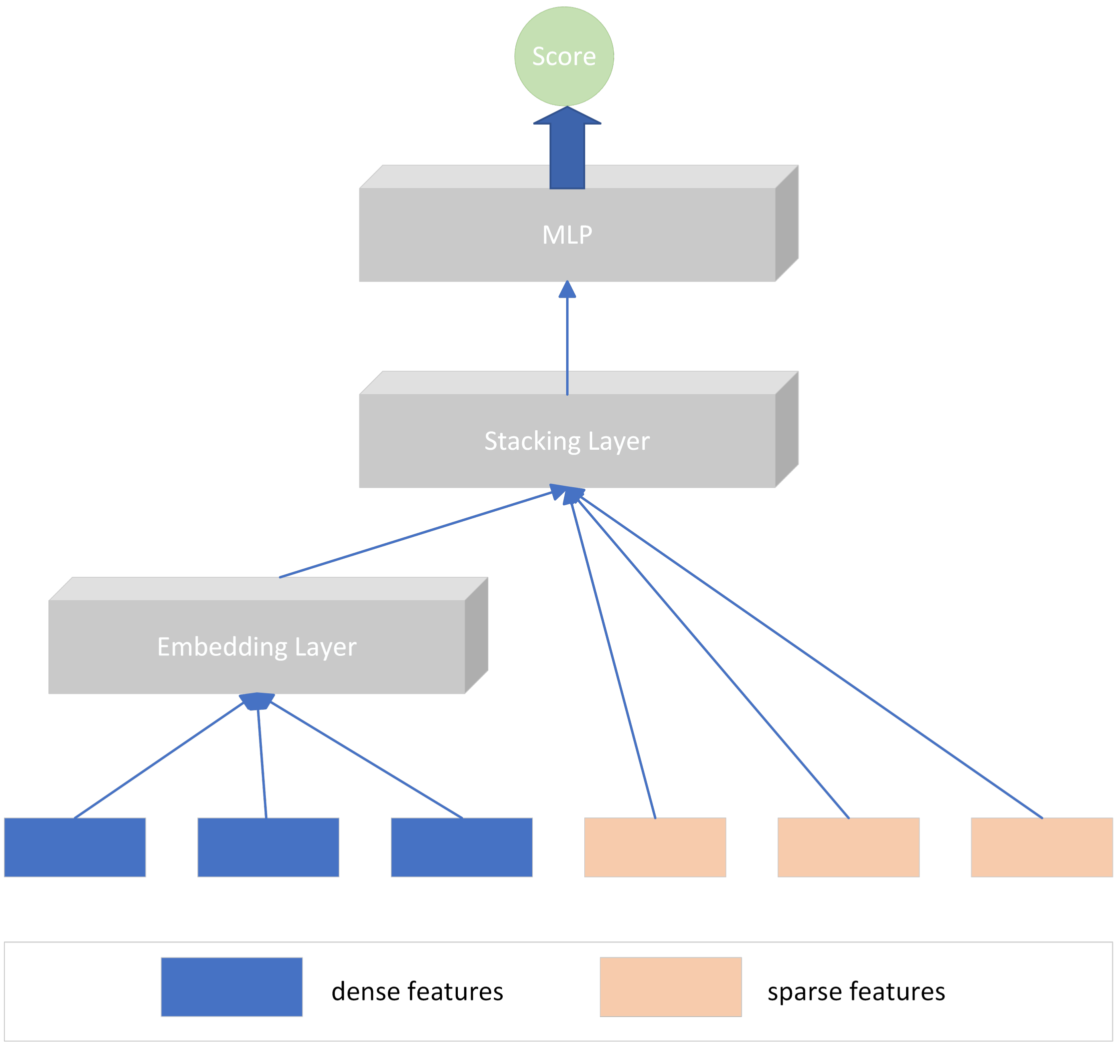


Figure 6.4: Structure of MLP

Multilayer Perceptron is commonly used for classification and regression tasks. It has the ability to learn complex non-linear relationships between inputs and outputs, making it a powerful tool for various applications. In this research, we incorporate embedding layers for sparse reported behavior features, as depicted in Fig. 6.4. We train the model by Pytorch. We choose *BCEWithLogitsLoss* as loss function, *Adam* as optimizer, and *auc* as metric function. We set batch size to 64, epoch to 50 and learning rate to 0.00001.

Convolutional Neural Network In this research, we only have tabular data, so we use a one-dimensional CNN (CNN-1D) for prediction. CNN-1D is more effective at capturing local features in the data and has a strong ability to adapt. The convolutional layer extracts features from input data, the pooling layer reduces the number of features and improves model robustness, and the fully connected layer maps the features to the output space for classification. We choose *BCEWithLogitsLoss* as loss function, *Adam* as optimizer, and *auc* as metric function. We set batch size to 1024, epoch equals to 50 and learning rate to 0.00005.

Long Short-term Memory aims to address the vanishing gradient problem commonly encountered in traditional recurrent neural networks. The model is capable of retaining long-term dependencies in the input data, making it suitable for various sequence prediction tasks. We reshape the data to fit the time step and features for LSTM in order to predict bankruptcy several years in advance.In this research, we choose *BCEWithLogitsLoss* as loss function, *Adam* as optimizer, and *auc* as metric function. We set batch size to 64, epochs to 50 and learning rate to 0.00001.

6.3.3 Performance Evaluation

In selecting the performance measures for the model, we refer to and synthesize previous studies in Section II and select two metrics, AUC and ROC curve, to assess the effectiveness of the model.

Area under the Receiver Operating Characteristic Curve (AUC) is a performance metric that assesses a classification model's ability to differentiate between positive and negative samples. AUC is not affected by sample imbalance or threshold selection, making it a more comprehensive measure of classifier performance compared to accuracy. The interpretation is straightforward as it summarizes the model's performance with a single scalar value. The formula for calculating AUC is:

$$\mathrm{AUC} = \int_0^1 \mathrm{TPR}(FPR^{-1}(t))\, dt \tag{6.2}$$

Receiver Operating Characteristic Curve (ROC curve) is a graphical representation of the True Positive Rate (TPR) plotted against the False Positive Rate (FPR) at various classification thresholds. TPR represents the proportion of positive samples correctly classified as positive. On the other hand, FPR represents the proportion of negative samples incorrectly classified as positive. ROC curve is a useful tool for visualizing the trade-off between TPR and FPR at various classification thresholds. The curve is created by plotting the TPR against FPR for every possible classification threshold. It offers a visual representation of the model's performance and helps in selecting the right classification threshold, considering the desired balance between TPR and FPR. The formula for calculating TPR and FPR is:

$$\text{FPR} = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}} \tag{6.3}$$

and

$$\text{FPR} = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}} \tag{6.4}$$

And we plot ROC curve by

$$\text{ROC curve: TPR vs FPR} \tag{6.5}$$

6.4 Results & Discussion

6.4.1 Features Evaluation

Information value (IV) is a widely used metric for selecting features in binary classification models. Assesses the ability of a feature to predict the target variable by analyzing its relationship. In essence, IV quantifies the amount of information that a feature provides about the target variable. It is commonly used to rank the importance of different features in a predictive model. We calculate the IV for AFE features, financial ratios, and behavior-related features. The results are displayed in Fig. 6.5. When performing feature selection using IV, features with high IV scores are generally considered more important and informative than those with low IV scores. By eliminating features with low IV scores, we can potentially simplify the model, enhance its performance, and identify the most significant predictors for a specific problem. Additionally, IV provides a standardized and interpretable measure of feature importance that can be easily communicated to stakeholders and decision-makers.

We observe that the number of AFE features is the highest, and most of these features have relatively high IV values. Financial ratios, while fewer in number compared to AFE features, have higher IV values and are less varied. On the other hand, behavior-related features exhibit

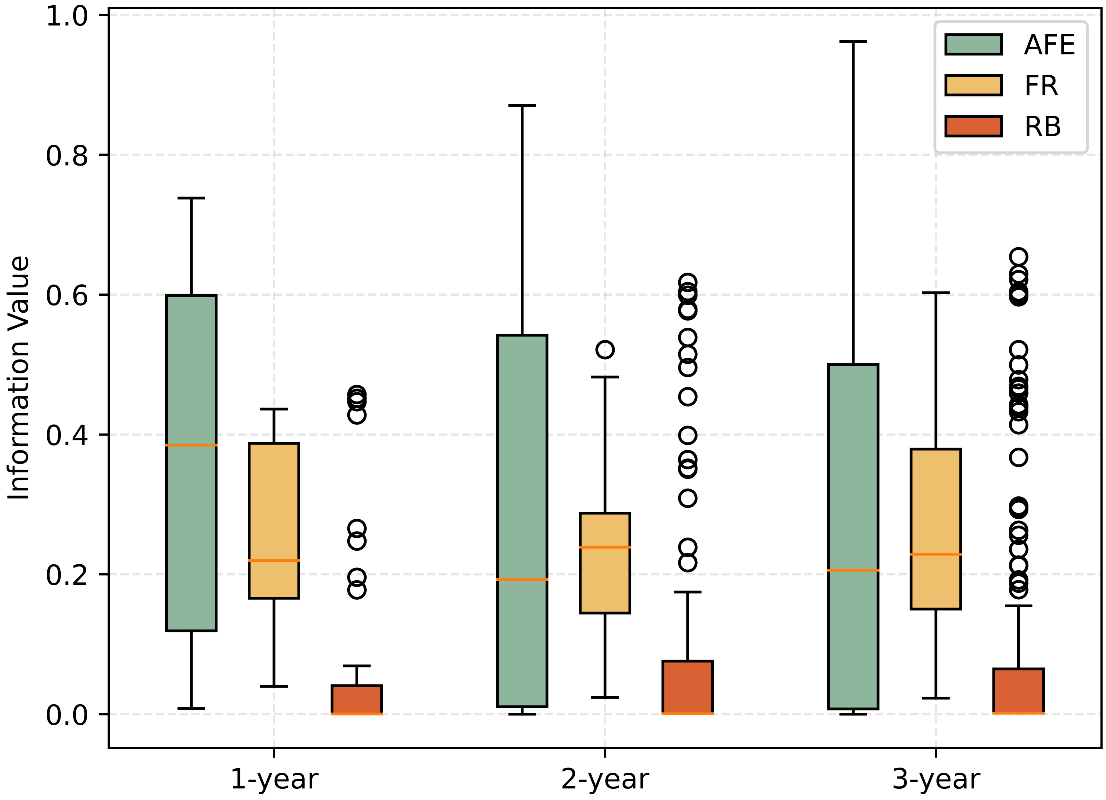
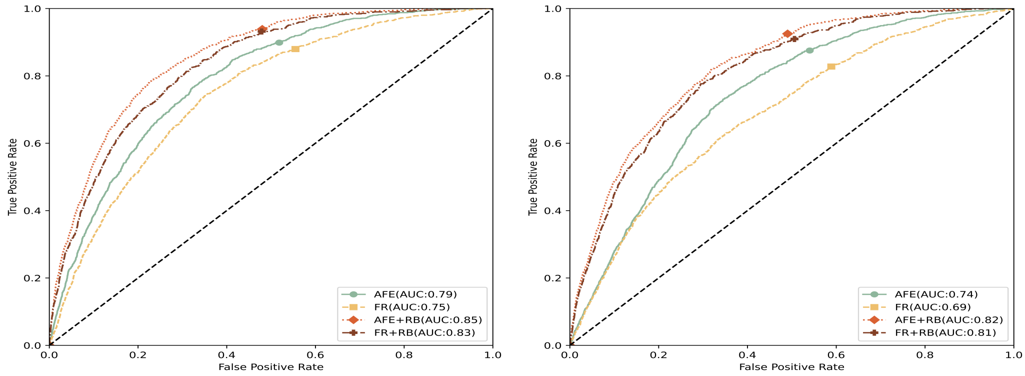


Figure 6.5: IV for features created from AFE, FR and RB

a wide range of IV values, with some having very high values and the majority clustered towards the lower end of the y-axis. This suggests that these features have little impact on predicting bankruptcy. Behavioral correlation features are often sparse matrices, with the majority of eigenvalues being 0. To mitigate the drawbacks of high coefficient matrices, we will employ feature filtering and summing techniques to maximize the utilization of the data.

6.4.2 Ablation Experimental results

We implement the ablation experiments to evaluate if the behaviour-related features can improve the model performance. We create four datasets: AFE, AFE+RB, FR and FR+RB to compare the model performance of with RB features and without RB features. The experiments were carry out on 6 models and 3 time periods. We select 2 out of the 18 results as the representative results and include all the other experimental results in the appendix for reference. Fig. 6.6 summarizes the performance of different features on lightGBM and LSTM by comparing their ROC curves. We use a green line to represent AFE features,



(a) LGB on 1-year dataset

(b) LSTM on 1-year dataset

Figure 6.6: ROC curve of lightGBM and LSTM on 1-year datasets

a yellow line to represent FR features, a red line to represent AFE and RB features, and a brown line to represent FR and RB features. From this figure, it can be clearly seen that models trained on hybrid datasets of financial and behavior-related features outperform datasets that only include financial features.

In the Appendix 10.4, the results of LR(Fig. 10.4a, Fig. 10.4b, Fig. 10.4c), LGB(Fig. 10.4g, Fig. 6.6b, Fig. 10.4h) and LSTM(Fig. 6.6b, Fig. 10.4n, Fig. 10.4p) very clearly show the advantages of hybrid datasets for bankruptcy prediction. Although the results of RF(Fig. 10.4d, Fig. 10.4d, Fig. 10.4f) and the results of MLP(Fig. 10.4i, Fig. 10.4j, Fig. 10.4k), we can still find the advantage of hybrid datasets, but not very obvious. The results of CNN-1D(Fig. 10.4l, Fig. 10.4m, Fig. 10.4n) are inconclusive, as the performance of financial-related features is comparable to random guessing. Additionally, the performance of models improves with longer training data periods. This suggests that using a larger data set can capture more accurate trends and patterns that indicate potential bankruptcy. Furthermore, it is worth noting that machine learning models such as LR, RF, and LGB outperform deep learning models such as MLP, CNN-1D, and LSTM. Overall, hybrid datasets offer significant advantages over single-source datasets for predicting bankruptcy. LightGBM model outperforms all other models in 3 time periods.

6.4.3 Performance about Covid Period

As described in Section III, the bankruptcy rate decreases significantly since 2019. It only has a 1% bankruptcy rate in 2019 and less than a 5% bankruptcy rate for 2020 and 2021. There are several reasons for the drop in the bankruptcy rate. First, the implementation of fiscal stimulus policies. Many countries adopt large-scale fiscal stimulus policies to ease the economic pressure caused by the epidemic, such as providing loans, tax cuts, and direct funding to businesses. Implementing these policies may help companies maintain cash flow and reduce the risk of bankruptcy. Second, debt moratorium and grace period. Many companies obtain debt moratorium and grace period arrangements during the pandemic, which allow them to delay debt repayment, thereby easing short-term financial stress and reducing the risk of bankruptcy.

However, this downward trend in bankruptcy rates may only be temporary, as these policies and arrangements may be unsustainable and companies are facing various uncertainties and challenges. For now, we can not see any evidence directly from the data but just observe that the distribution of both pre-Covid set and post-Covid set drift a lot from the training set. The experimental results verify this observation table 6.8.

More than 50% results show that the model performances of pre-Covid sets and post-Covid sets are better than those of testing sets, which is contrary to common sense. Furthermore, we find the hybrid datasets perform less favorable for both pre-Covid and post-Covid time period. We assume that this is because the reporting behavior of companies changed during the pandemic, which means companies may not submit or report their restructuring behavior in time due to the pandemic. This inconsistency on reported behavior data will confuse the model thus make the prediction performance not as good as testing set.

Table 6.8: AUC of models on testing, pre-Covid and post-Covid sets

	1-year			**2-year**			**3-year**		
	Testing	**Pre-Covid**	**Post-Covid**	**Testing**	**Pre-Covid**	**Post-Covid**	**Testing**	**Pre-Covid**	**Post-Covid**
LR									
AFE	0.7522	0.7539	**0.7669**	0.7494	0.7486	**0.7632**	**0.7702**	0.7599	0.7548
FR	0.7231	**0.7493**	0.7425	0.7384	**0.7621**	0.7605	0.7706	**0.7755**	0.7660
AFE+RB	**0.8234**	0.7433	0.7129	**0.8556**	0.7636	0.7124	**0.8767**	0.7755	0.7226
FR+RB	**0.8118**	0.7375	0.6948	**0.8642**	0.7758	0.7037	**0.8918**	0.7885	0.7178
RF									
AFE	0.7894	**0.8018**	0.7786	0.7860	**0.8077**	0.7775	0.8024	0.8052	**0.8078**
FR	0.7564	**0.7687**	0.7337	0.7676	**0.7838**	0.7454	0.7836	0.7810	**0.7873**
AFE+RB	0.7733	0.7739	**0.8195**	0.7733	0.7658	**0.8423**	0.7876	0.7711	**0.8589**
FR+RB	0.7294	0.7316	**0.7941**	0.745	0.7309	**0.8164**	0.7524	0.7286	**0.8418**
LGB									
AFE	0.7887	0.7980	**0.8092**	0.7925	0.7976	**0.8156**	0.8133	0.8108	**0.8147**
FR	0.7490	0.7629	**0.7741**	0.7634	0.7752	**0.7912**	**0.7990**	0.7865	0.7903
AFE+RB	**0.8542**	0.7732	0.7705	**0.8783**	0.8065	0.7623	**0.8930**	0.8168	0.7621
FR+RB	**0.8312**	0.7421	0.7226	**0.8706**	0.7759	0.7299	**0.8903**	0.7919	0.7313
MLP									
AFE	0.7408	0.7383	**0.7536**	0.7428	0.7447	**0.7752**	**0.7489**	0.7240	0.7263
FR	0.7282	**0.7482**	0.7299	0.7204	**0.7585**	0.7537	0.7301	**0.7462**	0.7329
AFE+RB	**0.8109**	0.6799	0.6775	**0.8014**	0.6841	0.7032	**0.7743**	0.6710	0.6643
FR+RB	**0.7980**	0.6904	0.674	**0.8145**	0.6946	0.6775	**0.8218**	0.7238	0.6902
CNN-1D									
AFE	0.7277	0.7288	**0.7335**	0.5607	0.5857	**0.6705**	0.4318	0.4629	**0.5317**
FR	0.6153	0.6495	**0.7055**	0.4922	0.4954	**0.5524**	0.7165	0.7196	**0.7333**
AFE+RB	**0.7442**	0.7090	0.7188	**0.7306**	0.6361	0.6229	**0.7243**	0.6298	0.5995
FR+RB	**0.7508**	0.6456	0.6504	**0.7139**	0.615	0.6141	**0.7217**	0.6339	0.6114
LSTM									
AFE	0.7404	0.7468	**0.7579**	0.7066	0.7036	**0.7406**	0.7193	0.7248	**0.7363**
FR	0.6879	0.7247	**0.7497**	0.7064	0.7369	**0.7628**	0.7419	0.7554	**0.7563**
AFE+RB	**0.8245**	0.7342	0.7145	**0.8158**	0.7094	0.6924	**0.8046**	0.7236	0.6951
FR+RB	**0.8087**	0.7159	0.6920	**0.8211**	0.7461	0.7146	**0.8187**	0.7470	0.7114

Chapter 7

Causal Insights: Uplift Modeling of Company Adjustments and Financial Health

Contents

7.1 Problem Statement 89

7.2 Methodology 91

7.2.1 Research Questions and Data Mapping 91

7.2.2 MTDnet Framework 92

7.2.3 Metrics 94

7.2.4 Uplift Models 95

7.2.5 Implementation Details 96

7.3 Results and Discussion 97

7.3.1 RQ1: Identifying Effective Company Adjustment 97

7.3.2 RQ2: Necessity of Considering the Sequence of Treatments 98

7.1 Problem Statement

Company adjustments, such as changes in management, registered addresses, or auditors, have been linked to financial distress in several studies, suggesting that these adjustments can be used as features in bankruptcy prediction models [105, 59, 173]. However, understanding which specific adjustment leads to particular financial outcomes remains a challenge. This chapter aims to uncover the causal relationships between different types of company adjustments and their financial effects, providing more precise estimates of how each adjustment might impact a company's financial health.

Accurately estimating the Individual Treatment Effect (ITE) is crucial in various fields. For instance, in healthcare, treatments can be personalized based on a patient's unique characteristics [79, 69], while in education, tailored interventions can be designed to prevent student dropouts [128]. In the business context, ITE helps companies understand their customers better and execute targeted marketing campaigns [183, 103]. In this chapter, we treat company adjustments as "treatments" to investigate their effects on financial outcomes.

The potential outcome model, proposed by Rubin in 1986 [141], provides a framework for estimating causal effects by controlling for all variables except the treatment. However, applying this model in practice, particularly in observational studies, is challenging due to the need to meet assumptions like stable unit treatment value and ignorability. Although Randomized Controlled Trials (RCTs) are considered the gold standard for estimating treatment effects, they are frequently impractical due to constraints related to time, cost, and ethical considerations. Furthermore, RCTs typically concentrate on average treatment effects rather than individual ones. Uplift modeling, which encompasses a set of machine learning techniques, presents a practical solution by estimating individual treatment effects from observational data [61, 179, 117].

Although uplift modeling has been applied in scenarios involving binary treatments [91, 69, 142], multiple treatments [143, 180, 145], and other complex cases [84, 131], the complexity increases significantly when dealing with multiple time-dependent treatments. Unlike static treatments that yield immediate effects, dynamic interventions unfold over time, rendering them more challenging to model.

In this chapter, we focus on the dynamic nature of company adjustments and their impact on bankruptcy prevention. Companies facing financial distress often undertake various strategic interventions, such as cost-cutting or restructuring, to avoid insolvency. We aim to estimate the effect of these adjustments on each company and provide insights into how corporate governance can prevent financial distress. To address the challenge of timing, we propose a novel uplift modeling framework for multiple time-dependent treatments. Our proposed framework leverages Long Short-Term Memory (LSTM) networks and attention mechanisms to capture the temporal dynamics of company adjustments and their impact on financial vulnerability.

Through empirical validation, we demonstrate the effectiveness of our framework and offer insights into its applicability in real-world scenarios. We propose an efficient framework for estimating individual effects with multiple time-dependent treatments. We estimate and analyze the effects of company adjustments on financial health. By exploring the dynamic interactions between company adjustments and financial vulnerability, this research aims to inform strategies for proactive risk management and enhance the resilience of enterprises against insolvency.

7.2 Methodology

7.2.1 Research Questions and Data Mapping

To investigate the effects of company adjustment acts on financial health, we examine the company's likelihood of bankruptcy as the dependent variable and company adjustment acts as treatments. Independent variables include a company's financial statements and basic information.

We apply the uplift model to estimate the ITE of each company under specific treatments. By conducting the counterfactual predictions $ITE = (y_i|(T = k_t) - y_i|(t = 0))$, we can determine whether a treatment has an effect on the outcome. In this context, y_i refers to company$i$, $T = k_t$ denotes that a company implements an adjustment act k at t moment and $T = 0$ indicates that a company does not take any action. A positive ITE suggests that adjustment acts have a positive effect on preventing bankruptcy, while a negative value indicates otherwise.

The research questions (RQs) defined in this chapter are as follows:

RQ1: What types of company adjustment acts help to improve financial status and prevent bankruptcy?

RQ2: Should the sequence and timing of company adjustment acts be considered when measuring their impact on financial health?

For RQ1, in order to study effective types of adjustment acts, we transform original treatments into binary treatments. Due to limited samples for some company adjustment, it is challenging for the uplift model to estimate each type individually. Therefore, we restructure and categorize these adjustment into four datasets for uplift models:

- Basic binary treatment: where $T = 0$ if no reported adjustment exists; otherwise $T = 1$

- Personnel-related treatment: where personnel-related acts are classified as $T = 1$; otherwise $T = 0$
- Information-related treatment: where changing information-related or business-related acts are categorized as $T = 1$; otherwise $T = 0$
- Other activities: all other unclassified adjustment acts are categorized as $T = 1$; otherwise $T = 0$

From Table 7.1, we observe the distribution across these four datasets with $y = 1$ indicating companies that went bankrupt within one year and $y = 0$ representing companies still operating. We implement the uplift models on these datasets and compare the uplift of these four different types of treatments.

For RQ2, we propose an uplift model to address the problem arising from the special data structure of company adjustment acts. We compare its performance with traditional uplift models used for binary or multiple treatments in order to assess whether considering different treatment sequences is necessary.

For RQ1, we randomly sample the training and test sets in the 7:3 ratio on the four datasets under the different treatment scenarios and measure the average uplift on the test set. For RQ2, we first split the dataset into training and test sets in the ratio of 7:3, and then get the binary treatments, multiple treatments, and original treatments by reshaping the dimension of the raw treatment value. We will explain the way of reshaping in Sect. 7.3.2. All the experimental results in Sect. 7.3 are the metrics evaluated in the test set.

Table 7.1: Description of the four datasets for uplift models

	Basic binary		Personnel		Information		Other	
	T=0	T=1	T=0	T=1	T=0	T=1	T=0	T=1
y=0	21820	17974	25600	14194	34417	5377	25600	14194
y=1	2281	4529	3589	3221	5347	1463	3589	3221

7.2.2 MTDnet Framework

The diagram in Fig. 7.1 illustrates the comprehensive framework to estimate the uplift with multiple time-dependent treatments. The input data shown in the figure represent contextual features, which are one-dimensional array-like data. The matrix-like data of multiple time-dependent treatments describe the time steps and multiple treatments. This framework consists of two main components: the representative module and the time-attention module. The uplift model is a two-head network model; one head estimates the label value when there is no treatment, while the other head estimates the label value when treatments are implemented. We use L_C to denote the loss of the control group ($t = 0$, where t denotes treatment) and L_T to denote the loss of treated group ($t = k$). In both cases, a lower loss indicates a more precise estimate. L_D measures the distance between the control group and the treated group using Kullback-Leibler (KLD) divergence. A smaller KLD signifies greater similarity between the control and treated groups, indicating that these experiments resemble randomized controlled trials (RCTs) more closely.

$$L_C = loss\{y(t = 0), \hat{y}(t = 0)\} \tag{7.1}$$

$$L_T = loss\{y(t = k), \hat{y}(t = k)\} \tag{7.2}$$

$$L_D = KLD\{dist.|control, dist.|treated\} \tag{7.3}$$

In general, the loss function of this model is

$$total\ loss = L_C + L_T + L_D \tag{7.4}$$

Representative Module For all the independent features, we put them in the representative module. We used multilayer perceptrons to capture the characteristics of both the control group and the treatment group. Independence is an important assumption in causal inference, which means that the potential causal variable should be independent of other potential causes of the outcome. This means that there should be no other factors that simultaneously affect both the potential cause and the outcome. When handling the observational data, we try to mimic the randomized controlled trials in order to have fewer selection biases. After sharing the same layers, we use the KLD to calculate the distance of the distribution of control group and treatment group. The lower KLD means that the distribution of the control group and the treatment group is more like, the selection bias will be less.

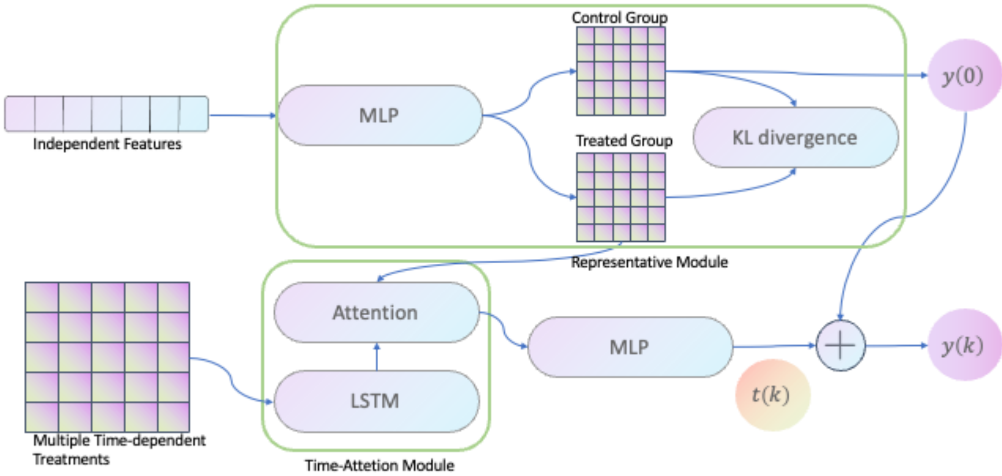


Figure 7.1: Framework of multiple time-dependent uplift modelling

Multiple Time-dependent Treatment Module In this module, we adopt the long and short-term networks (LSTM) to capture time-related characteristics of multiple time-dependent treatments. LSTM is flexible to handle time series data. After the LSTM layer, we apply the attention theory to optimize the treatment representation by considering the background features. We calculate the different attention weights of the treatments and multiply the weights by the treatments to get interactive results.

7.2.3 Metrics

We adopt four commonly used metrics to evaluate the performance of the proposed model [46, 45]: Average Uplift, Qini Score, AUUC, and Uplift at 30%.

Average Uplift refers to the average effect of a treatment or intervention across a population. It measures the difference in outcomes between those who received the treatment and those who did not, averaged across all individuals.

Qini Score computes normalized Area Under the Qini coefficient curve from prediction scores. By computing the area under the Qini curve, the curve information is summarized in one number. For binary outcomes, the ratio of the actual uplift gain curve above the diagonal to that of the optimum Qini curve.

AUUC computes normalized Area Under the uplift curve from prediction scores. There are many different ways to calculate the AUUC value, in this paper, we adopt the formulation from [61]. By computing the area under the uplift curve, the curve information is summarized in one number. For binary outcomes the ratio of the actual uplift gains curve above the diagonal to that of the optimum uplift curve.

Uplift at 30% computes uplift at first 30% observations by uplift of the total sample. After ordering the data by uplift prediction, the difference of conversions between control group and treatment group can be get.

7.2.4 Uplift Models

We have chosen two meta-learner models and three other models according to the study by Gutierrez and Gérardy [61].

S-Learner (Single Model Learner) is a fundamental model in uplift modeling that estimates the individual treatment effect by employing a single model. It is represented by the following equation:

$$\hat{Y}_i = \hat{Y}(X_i, 1) - \hat{Y}(X_i, 0)$$

Here, $\hat{Y}_i$denotes the uplift score for individual i, $\hat{Y}(X_i, 1)$ represents the predicted outcome for individual i under treatment, and $\hat{Y}(X_i, 0)$ represents the predicted outcome for individual i without treatment.

T-Learner (Two Model Learner) employs two separate models to estimate the expected outcomes for the treatment and control groups, respectively, and calculates the difference between them. It is expressed by the following equation:

$$\hat{Y}_i = \hat{Y}_1(X_i) - \hat{Y}_0(X_i)$$

Here, $\hat{Y}_i$ represents the uplift score for individual i, $\hat{Y}_1(X_i)$ denotes the predicted outcome for individual i under treatment, and $\hat{Y}_0(X_i)$ denotes the predicted outcome for individual $i$without treatment.

CEVAE (Causal Effect Variational Autoencoder) is a probabilistic graphical model-based uplift model that estimates individual uplift scores through a variational autoencoder (VAE). It involves complex equations related to the encoder and decoder structures for learning individual latent representations and causal relationships in the context of treatment effects. In essence, CEVAE aims to disentangle the latent factors that influence both treatment assignment and outcome, allowing for more accurate estimation of causal effects.

DragonNet consists of two neural networks: one to predict the outcome of the treatment group and the other to predict the outcome of the control group. The two networks share part of the structure (called the "shared representation"), and then each has its own output layer. By comparing the output of the two networks, Dragonnet can estimate individual processing effects. A key feature of Dragonnet is that it uses a special loss function designed to directly optimize the quality of the estimated individual processing effects. This allows Dragonnet to provide more accurate estimates than traditional causal inference methods when dealing with complex, high-dimensional data.

Ganite is a generative adversarial network (GAN)-based method for estimating ITE. It uses two neural networks: a generator and a discriminator. The generator's task is to generate possible therapeutic effects, while the discriminator's task is to distinguish between generated therapeutic effects and real therapeutic effects. In this way, the generator is trained to generate therapeutic effects that are closer to the real thing. This is achieved by having the generator generate a distribution of the therapeutic effects, rather than just a point estimate. This allows GAN-ITE to provide richer information to help decision-makers understand the uncertainty of treatment effects.

7.2.5 Implementation Details

We use PyTorch (version 2.2.1) to build the network. For uplift models Dragonnet and CEVAE, we use the CausalML package [30] with the default settings. For S-learner and T-learner, we use the lightgbm model as the estimator for the model with the default settings. For Ganite, we adopt the settings from the paper [172]. For tuning our proposed

Table 7.2: Scope of hyperparameters used in optimizing the model performance

Hyperparameter	Range
Batch size	32,64,128
Number of Epoch	50,100,150
Learning rate	0.0001, 0.00001,0.000005
L2	$1e^{-}4$,$1e^{-}5$,$1e^{-}6$
Hidden size	2^5,2^6,2^7,2^8
Output size	2^3,2^4,2^5

model, we use the GridSearch method and the early stop condition that patience of 8. The hyperparameters can be seen from Table 7.2.

7.3 Results and Discussion

In this section, the results and findings of the previously defined research questions are discussed.

7.3.1 RQ1: Identifying Effective Company Adjustment

We evaluate the uplift for different categories of treatments by the uplift models mentioned in Sect. 7.2.4. Fig. 7.2 indicates that the overall results for different uplift models show the same conclusion. The dataset of information and business-related treatment has the absolute advantage for the uplift over the other three types of treatment according to the results of the five uplift models. The basic binary treatments have the least uplift among the four datasets, and this may because the treatments have the conflict with each other, that the mixed treatments perform the worst.

Personnel-related treatments contain a relatively large number of adjustment acts that do not gain much uplift for all models. We think the reason may be it is challenging to standardize the treatment as it is not possible to have the same or similar human being. Even if it is the same person, we cannot say (s)he will make the same decision under the same scenario. Therefore, this kind of adjustment cannot have a good performance in uplift modeling.

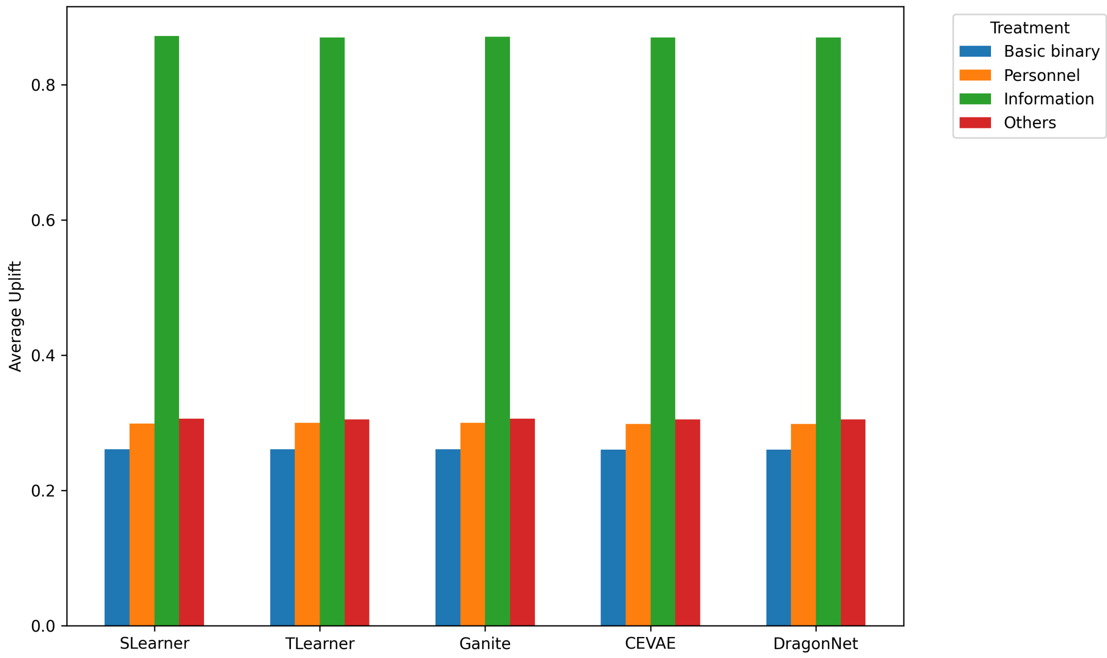


Figure 7.2: Uplift comparison for different adjustment

7.3.2 RQ2: Necessity of Considering the Sequence of Treatments

We implement the experiments on the LBR data and compare our proposed model with six other representative uplift models mentioned in the previous section. We have not been able to find any open source datasets which also involve the multiple time-dependent treatments. Therefore, we cannot evaluate our model on other datasets for comparison. In addition, there are no published uplift models that designed for handling multiple time-dependent treatments.

The original treatment contains three dimensions: value of treatment, type of treatment, and timing of treatment. We reshape the structure of the treatment in order to evaluate if it is necessary to consider all the three dimensions. For the binary treatment scenario, we ignore the type and timing of the treatment. Therefore, if $T = 1$ means that this company takes some actions. For multiple treatment scenario, we ignore the timing, and the company can have multiple adjustment without considering the sequence of adjustment.

Although MTDnet has a lower value for uplift at 30% which indicates that this is not as effective as others in the top 30% of the target group, MTDnet outperforms the other treatment scenarios for the metrics AUUC and Qini, which suggests that it is better identifies treatment effects across the entire dataset and is able to accurately predict which individuals were positively affected by the treatment over a wider range(see Table 7.3). In the future, we could consider combining the strengths of other models to build hybrid models to achieve better results both globally and locally.

Moreover, we find that the network-based models (CEVAE, DragonNet and Ganite) perform much worse than its published results, while machine learning-based models are more robust. We think this is because network-based models normally need more delicately tuned hyperparameters. As the authors [127] pointed out, there is no model that always performs the best for all the context and problems.

Table 7.3: Results of the experiments for evaluating the necessary of considering treatment sequence. The suffix of the model name represents the model is trained with different treatments. ”-bi” refers to binary treatments, ”-multi” refers to multiple treatments, and ”-original” refers to multiple time-dependent treatments.

Model	Uplift at 30%	AUUC	Qini
S-learner-bi	0.0161	0.0272	0.0478
T-learner-bi	**0.0489**	0.0422	0.0709
CEVAE-bi	0.0049	0.0094	0.0169
DragonNet-bi	0.0000	0.0008	0.0013
Ganite-bi	-0.0037	0.0206	0.0364
Ganite-multi	0.0022	-0.0160	-0.0289
MTDnet-original	0.0067	**0.0589**	**0.1880**

Chapter 8

Towards Automation: Designing a Credit Reporting System for SMEs

Contents

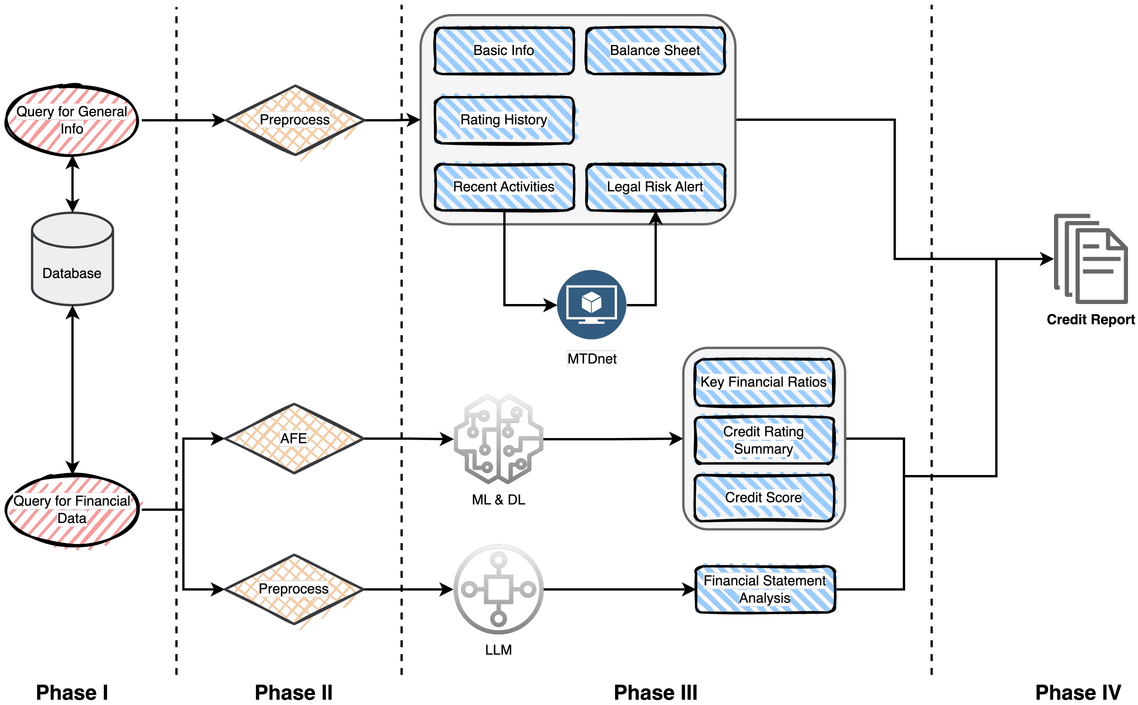
8.1 Problem statement 100
8.2 Methodology and Workflow 101
8.3 Key Features of the Automated Credit Report 102
8.4 Potential Use Cases 103
8.5 Challenges and Limitations 104

8.1 Problem statement

In today's rapidly evolving financial landscape, the ability to assess corporate creditworthiness efficiently is crucial for stakeholders, including financial institutions, investors, and corporate management. Traditionally, this process has been labor-intensive, requiring manual analysis of financial statements, market conditions, and historical data. Automating the generation of corporate credit reports addresses these challenges, offering a streamlined approach to evaluate a company's financial health and predict potential risks. The research results of previous chapters will be applied in this system. This system aims to provide a comprehensive credit risk assessment and automate the reporting process to improve efficiency and reduce manual effort. This chapter outlines the design and implementation of an automated system that produces comprehensive corporate credit reports, highlighting its methodology, key features, and potential applications. An example of the report can be found in the Appendix 10.5.

8.2 Methodology and Workflow

The automated system for corporate credit reporting is designed with modularity and scalability in mind, integrating data extraction, financial analysis, and report generation into a cohesive workflow. This structure reduces manual effort, ensures consistency, and enhances the accuracy of the generated reports.



The system comprises four main phases: data extraction, data processing, modeling, and report generation. Data extraction retrieves the company's financial data, registration details, reported files of company adjustments, and industry benchmarks from structured databases using predefined SQL queries. These queries are deployed using Python and then interface with the PostgreSQL database, ensuring seamless access to the relevant information. Once the data is extracted, it undergoes processing, during which financial ratios such as liquidity, solvency, and various statistical metrics are computed. Statistical tools and libraries, including pandas, numpy, and scipy, facilitate the analysis, while matplotlib and seaborn are employed to create visual representations of the data. Financial data prepared for ML and DL bankruptcy models is processed with the application of AFE to enhance the feature set. Financial data prepared for LLMs requires only standard preprocessing.

The final stage of the workflow involves report generation, which is executed using the ReportLab library. Results are compiled into a professional-grade PDF report that incorporates financial summaries, visualizations, and narrative interpretations. The report is dynamically generated, tailored to the specific corporate entity, and includes both quantitative data and qualitative insights.

The overall workflow of the system is characterized by a seamless transition from input data to the final report. The input comprises financial statements, corporate registration data, and raw files that are uploaded by the company to LBR. This data is processed through a series of analytical steps that clean, transform, and analyze the information to produce key financial insights. The output is a comprehensive report that offers a detailed assessment of the company's financial health, creditworthiness, and potential risks.

8.3 Key Features of the Automated Credit Report

The automated corporate credit report provides a comprehensive view of a company's financial health, delivering actionable insights across several critical dimensions.

A primary feature of the report is the financial health summary, which includes key ratios that offer a detailed understanding of the company's liquidity, solvency, and profitability. Liquidity ratios, such as the current ratio and quick ratio, evaluate the company's ability to meet short-term obligations. Solvency ratios, including the debt-to-equity ratio, assess long-term financial stability, while profitability ratios, such as return on assets and net profit margin, provide insights into the company's operational efficiency and ability to generate profits.

Another essential component of the report is the credit rating and bankruptcy prediction. The system assigns a credit score ranging from 0 to 1000, where higher scores indicate lower risk, and provides a qualitative rating (from A to E). For instance, a company with a credit score of 168 would be categorized in the lowest rating class, indicating a high likelihood of financial distress and potential bankruptcy within the next twelve months. The report not only quantifies the risk but also positions the company within its industry by comparing its financial metrics with industry benchmarks.

The report also includes comparative analysis, which highlights the company's position relative to its peers. This analysis is supported by historical performance data, showing trends in financial ratios over the past three years. These trends offer a longitudinal perspective on the company's financial trajectory, allowing stakeholders to identify patterns of improvement or deterioration.

To enhance interpretability, the report integrates visual representations, including trend charts that depict changes in financial metrics over time and risk distribution plots that compare the company's credit score to industry averages. These visual elements provide a clear and intuitive understanding of the data, facilitating quicker and more informed decision-making.

8.4 Potential Use Cases

The automated credit reporting system exhibits broad applicability across multiple industries, significantly enhancing financial analysis and risk assessment processes. In the banking and financial sector, the system is employed for credit risk assessment, streamlining the evaluation of loan applicants and providing consistent and objective risk evaluations. By reducing manual effort and improving the accuracy of risk assessments, the system expedites loan approval processes and supports better portfolio management.

Within corporate finance departments, the system serves as a valuable tool for internal financial analysis, helping CFOs and financial analysts assess their company's financial position and identify areas for improvement. It can also play a critical role in mergers and acquisitions, offering a comprehensive view of a target company's financial health during due diligence.

Investment firms can leverage the system to enhance their due diligence processes, providing detailed financial insights and risk assessments of potential investment targets. By automating these analyses, investment professionals can make more informed decisions and identify opportunities and risks with greater precision.

8.5 Challenges and Limitations

Despite its numerous advantages, the automated system faces several challenges and limitations. One significant challenge is the quality and availability of input data. The accuracy of the analysis is heavily dependent on the completeness and reliability of the financial and operational data provided. Companies with incomplete or inconsistent data records may receive less accurate reports, and smaller enterprises or startups may lack sufficient historical data for reliable trend analysis.

Another limitation concerns scalability. As the financial landscape evolves, the system must be regularly updated to accommodate new data formats, regulatory changes, and emerging financial metrics. This requires ongoing maintenance and enhancements to ensure that the system remains relevant and effective across different contexts and industries.

The automated nature of the system also presents limitations in the interpretation of results. While the system excels at providing quantitative insights, certain qualitative aspects of financial health, such as management competence, market positioning, and strategic direction, require human judgment and cannot be fully captured by automated processes.

Finally, customization poses a challenge, as different industries may have unique financial characteristics and risk factors. While the system is designed to be adaptable, further customization may be necessary to generate more industry-specific insights that align with the unique financial dynamics of various sectors.

Chapter 9

Conclusion and Future Work

This chapter provides a comprehensive summary of the key findings from the dissertation, integrating insights from the research papers and the development of the automated credit reporting system. It recaps the contributions of each chapter, highlights the critical discoveries, and discusses the broader implications for SME financial risk prediction. The chapter then discusses the limitations of this research and points out the possible directions for future research.

9.1 Summary of Key Findings

This dissertation starts from a data perspective and raises the corresponding research questions according to different datasets. Chapter 2 highlights the evolution of datasets and models used in financial health and bankruptcy prediction. It proposes a dataset taxonomy to categorize the datasets into accounting-based, market-based, macroeconomic, relational, and non-financial types, which provided a framework for understanding how different sources of data contribute to predicting financial distress. It points out the predominance of accounting-based data in training bankruptcy prediction models due to its relative ease of acquisition and high data quality. It also identifies the emerging importance of relational data in bankruptcy prediction, although this type of data is often sensitive and not easily obtainable. Furthermore, it finds that while many datasets are publicly available, most require payment or are self-integrated, with few being completely free and directly downloadable.

In Chapter 4, an automatic feature engineering approach is presented to enhance the bankruptcy prediction of companies that lack sufficient data due to incomplete financial statements for traditional risk assessment. The design is centered around generating features aimed at improving prediction based on real-world financial statements. The results of this research show that the models trained on features generated by automatic feature engineering outperform the models trained, among others, on features generated by the traditionally used financial ratios. This research thus implies that automatic feature engineering can generate effective features for model training, which is an especially useful enhancing effect for the bankruptcy prediction and risk assessment of companies lacking sufficient data in a traditional crediting setup, such as SMEs. The AFE-generated features were found to be more effective in predicting bankruptcy across various datasets, as evidenced by higher AUC scores in comparison to models trained on financial ratios, DeepFM, and deep feature synthesis. This research also demonstrated that the AFE approach is more explainable and extensible than existing methods, as the features generated are simple arithmetic expressions based on the original financial data, which can be easily understood and built upon.

Financial data is further explored in Chapter 5 by advanced models, large language models. The findings indicate that while LLMs have great potential in automating and scaling financial statement analysis with careful model selection and optimization, they are not yet at the level of precision that experienced financial analysts can provide. Also, it highlights that there are clear performance and resource trade-offs when using different LLMs for financial statement analysis. This research finds that Llama 3.1 performs with the highest accuracy, especially when using fine-tuning and RAG combined with few-shot learning. However, this model requires more GPU memory compared to Llama 3.2. Llama 3.2 offers a good balance between performance and resource efficiency, using less GPU and CPU power, making it a cost-effective choice for large-scale deployments or when resources are limited. Mistral showed mixed performance, doing well in retrieval-intensive tasks but struggling with accuracy in other areas, suggesting its architecture is more suited for efficiency-focused tasks rather than general financial applications.

In Chapter 6 and Chapter 7, company adjustments behavior is investigated. The exploration begins with the integration of company adjustments with financial data, demonstrating that hybrid datasets enhance model performance by capturing both financial ratios and the dynamics of adjustments in SMEs. The results indicate that the use of a hybrid dataset, which combines financial statements with company adjustments data, leads to an increase in the performance of bankruptcy prediction models by 4% to 13% compared to models utilizing financial data alone. This significant improvement suggests the necessity of incorporating company adjustments behavior into bankruptcy prediction models. Then the causal effect of company adjustments on financial health is further examined in Chapter 7. The uplift results of different adjustments show that information and business-related treatments had a significant advantage over other types of treatments across all models. This suggests that interventions focused on improving information and business operations are the most effective in enhancing company financial health. Personnel-related treatments also demonstrates lower uplift, possibly because of the inherent variability and unpredictability associated with human behaviors, making standardization and prediction challenging. This research also highlights the importance of the timing and sequence of company adjustments in predicting bankruptcy. The proposed MTDnet model outperforms other six uplift models in capturing these dynamics, indicating the necessity of accounting for the value, type, and timing of treatments when analyzing their effects on financial health.

Chapter 8 has provided a detailed overview of the methodology and framework for developing an automated corporate credit reporting system. By integrating data extraction, feature engineering, bankruptcy modelling and financial statment analysis, the system offers a streamlined solution for generating the credit report of a company. Its scalability, accuracy, and speed make it a valuable asset across various industries, from financial institutions to corporate finance teams.

This dissertation has contributed significantly to the field of SME financial health and bankruptcy prediction by integrating diverse data sources, advanced modeling techniques, and automation. The research findings underscore the value of hybrid datasets, the potential of automatic feature engineering, the application of large language models, and the importance of understanding the effects of corporate adjustments. Together, these insights provide a robust framework for predicting bankruptcy risk and enhancing the financial health of SMEs, paving the way for more effective, data-driven decision-making in the credit risk domain.

9.2 Limitations and Future Work

9.2.1 Limitations

While this dissertation advances the field of SME credit risk prediction, certain limitations are inherent in the research approach, data sources, and methodologies. The research relies on datasets from specific sources, such as Luxembourg Business Registers, which may limit the generalizability of findings to SMEs in other countries or industries. The findings are potentially influenced by the industry-specific characteristics of Luxembourg. Additionally, challenges related to the quality, completeness, and consistency of financial and non-financial data, such as missing values or varying reporting standards, may affect the robustness of the models. The experiments are confined to the dataset from the Luxembourg Business Register, and there is a lack of external datasets with similar characteristics to validate the model's performance.

Although this dissertation considers company adjustments behavior as the non-financial data, the scope of non-financial data is limited. Other potentially influential data, such as transactional data, tax data, relational data, are not included due to data unavailability or difficulty in quantification. This can be further studied in the future. For causal inference study, there are no published uplift models designed for handling multiple time-dependent treatments, which limits the ability to compare the proposed MTDnet model with other models that could address the same complexities. For LLMs application, LLMs are found to be useful in processing financial data, their accuracy in handling complex numerical calculations and nuanced financial analyses still falls short of expert-level performance. Furthermore, LLMs are sensitive to prompt design, which may lead to inconsistent results if not optimized carefully.

9.2.2 Future work

Expanding data sources and diversity Future research should prioritize the collection and integration of more diverse datasets across different regions, industries, and economic conditions. Incorporating alternative data sources, such as macroeconomic indicators, supply chain data, or social media sentiment, could enhance the predictive power and generalizability of the models.

Enhanced non-financial data analysis Building on the findings from corporate restructuring behaviors and uplift modeling, future work should explore additional non-financial variables. These could include environmental, social, and governance (ESG) factors, market dynamics, and industry-specific behavioral data, which are becoming increasingly relevant in financial risk assessment.

Advanced temporal and causal modeling To address the limitations of static or short-term data, future studies could develop more advanced time-series models or causal inference frameworks. For example, methods such as recurrent neural networks with attention mechanisms or temporal graph neural networks could capture long-term patterns and complex dependencies more effectively.

Improving LLM capabilities for financial statement analysis Further exploration of hybrid approaches combining LLMs with traditional financial models could mitigate the computational inefficiencies and accuracy limitations of current models. Future work could also focus on domain-specific fine-tuning of LLMs using extensive financial datasets, as well as developing lightweight and resource-efficient LLM variants tailored to SMEs.

By addressing these limitations and pursuing the outlined directions, future work can build on this dissertation to create more comprehensive, accurate, and scalable systems for SME financial risk prediction. Advancements in data integration, modeling techniques, and computational tools will not only improve predictive capabilities but also contribute to more equitable access to financial insights and resources for SMEs globally.

Chapter 10

Appendix

10.1 A

Table 10.1: Summary for related studies

Studies	Year of Publication	# of Sample	Bankruptcy rate	# of Features	Data Type	Data Source	Publicly available?	Fee applied?
[21]	2018	126	/	24	Accounting-based	Serbia	No	/
[189]	2020	343	25.36%	9	Accounting-based	CRIF-Slovak Credit Bureau	No	/
[124]	2016	2061	/	6	Accounting-based	Albertina database	Yes	Yes
[72]	2017	1115	6.66%	82	Accounting-based, Macroeconomic indicators, Corporate govenance indicators,Basic informationrmation	Standard and Poor's Capital IQ service	Yes	Yes
[42]	2016	7152	50.00%	9	Accounting-based, Basic information	Bureau Van Dijk	Yes	Yes
[110]	2019	4515	11.34%	14	Accounting-based,Tax-based	Estonian	No	
[123]	2016	318	22.96%	12	Accounting-based, Market-based, Macroeconomic indicators	Cyprus Stock Exchange	Yes	No
[144]	2018	100	22.00%	35	Accounting-based, Market-based, Macroeconomic indicators	Johannesburg Stock Exchange	Yes	Yes
[54]	2020	2860	2.17%	33	Accounting-based, Basic information	Infotel	Yes	Yes
[82]	2022	454752	0.45%	8	Market-based	Center for Research in Security Prices (CRSP) dataset	Yes	Yes
[139]	2005	33037	2.42%	20	Market-based	Compustat	Yes	Yes
[34]	2020	84	50.00%	5	Accounting-based	Bloomberg	Yes	Yes
[114]	2023	186	9.68%	5	Accounting-based	CRIF-Slovak Credit Bureau	Yes	Yes

continued on next page

Table VI continued

Studies	Year of Publication	# of Sample	Bankruptcy rate	# of Features	Data Type	Data Source	Publicly available?	Fee applied?
[182]	2024	889	/	/	Basic information,Law suit related, Knowledge graph of SMEs	SMEsD	Yes	No
[184]	2021	13489	26.44%	/	Network of board member and shareholder	HAT	Yes	No
[190]	2021	498	19.48%	/	Accounting-based	CRIF-Slovak Credit Bureau	Yes	Yes
[175],[176]	2017, 2021	2457	18.56%	55	Accounting-based, Business environment factors	Russia dataset	Yes	No
		5910	6.94%	63	Accounting-based and business environment factors	Polish dataset	Yes	No
[88]	2021	168466 167364 166527	0.64% 0.48% 0.26%	98	Basic information, Knowledge graph of SMEs, Transaction-based, Payment network–based	Business Bank of Shandong	No	/
[115]	2014	2033	22.92%	39	Accounting-based	ANS	No	/
[92]	2018	120355	0.26%	38	Accounting-based, Transactional-based variables	Korean financial company	No	/
[31]	2011	1200	50.00%	30	Accounting-based	Diane database of Bureau Van Dijk	Yes	Yes
[47]	2017	138387	1.29%	30	Basic information,Macroeconomic indicators, energy	Orbis database of Bureau van Dijk	Yes	Yes
[107]	2022	78682	0.77%	18	Accounting-based	American dataset	Yes	No

continued on next page

Table VI continued

Studies	Year of Publication	# of Sample	Bankruptcy rate	# of Features	Data Type	Data Source	Publicly available?	Fee applied?
[113]	2019	94994	0.50%	36	Accounting-based, Market-based, Text from annual reports	Compustat North America, Securities Exchange Commission, Management Discussion and Analysis section of 10-K	Yes	Yes
[99]	2016	478	50.00%	190	Accounting-based, Company goverance indicators	Taiwan Economic Journal	Yes	Yes
[97]	2009	6288	21.14%		Accounting-based, Market-based	Compustat	Yes	Yes
[65]	2013	23218	5.40%	10	Accounting-based, Market-based, Macroeconomic indicators	Datastream,Thomson One Banker and London Share Price Database	Yes	Yes
[160]	2017	2400000	/	6	Accounting-based, Relational data	Belfirst and Fame databases of Bureau Van Dijk	Yes	Yes
[185]	2013	86129	1.07%	10	Accounting-based	Compustat North America	Yes	Yes
		36637	0.16%	10	accounting	Compustat Global	Yes	Yes
[48]	2015	/	/	50	Accounting-based	Diane database of Bureau Van Dijk	Yes	Yes
[94]	2020	4358	26.78%	37	Management ability, Business feasibility, Technical ability, other	KOSME	Yes	Yes

continued on next page

Table VI continued

Studies	Year of Publication	# of Sample	Bankruptcy rate	# of Features	Data Type	Data Source	Publicly available?	Fee applied?
[132]	2013	825	15.39%	22	Financial data, Non-financial data(Firm owners' personal credit peRandom forestormance, management quality etc.)	Croatian commercial bank	No	/
[156]	2012	63107	0.41%	5	Accounting-based	SABI of Bureau van Dijk	Yes	Yes
[155]	2020	75652 75652	10.87% 12.55%	11	Accounting-based	AMADEUS of Bureau Van Dijk	Yes	Yes
[36]	2013	321	13.08%	26	Accounting-based	Taiwan Stock Exchange Corporation and Taiwan Economic Journal	Yes	Yes
[55]	2016	9000000	/	19	Accounting-based, Spatial information	Serasa Experian	No	/
[56]	2017	38036	5.00%	/	Accounting-based, Basic information, Transactional-based	UniCredit bank	No	No
[165]	2014	240	46.67%	30	Accounting-based	Polish dataset	Yes	No
		132	50.00%	24	Accounting-based	CD-ROM of Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner	Yes	Yes
[137]	2019	806	38.59%	/	Accounting-based, Basic information	Polish consultancy firm	No	/
[37]	2015	3210	50.00%	38	Accounting-based	CERVED	Yes	Yes

continued on next page

Table VI continued

Studies	Year of Publication	# of Sample	Bankruptcy rate	# of Features	Data Type	Data Source	Publicly available?	Fee applied?
		1156	50.00%	38	Accounting-based, Company goverance indicators		Yes	Yes
[85]	2018	288	50.00%	53	Accounting-based	KOSPI	No	/
		7027	3.86%					
		10173	3.93%					
[6]	2021	10503	4.71%	64	Accounting-based	Polish dataset	Yes	No
		9792	5.26%					
		5910	6.94%					
[27]	2013	1000	50.00%	6	Accounting-based	AMADEUS of Bureau Van Dijk	Yes	Yes
[71]	2016	313	33.23%	232	Accounting-based,Sentiment lexicon/variables	KOSPI and KOSDAQ	No	/
[133]	2020	101641	1.55%	660	Accounting-based	Federal Deposit Insurance Corporation	Yes	No
[152]	2019	997940	2.32%	13	Accounting-based, Basic information	NICE Information Service Co	No	/
[10]	2008	5816021	1.15%	43	Accounting-based,Basic information, Reported and compliance, Operational risk	/	No	/

10.2 B

Table 10.2: Comparison of the forecasting ability of LLMs and financial expert.

		Next Year Sales Prediction	Next Year EBITDA Prediction
Llama 3.2	zero-shot	139.6	129.6
	few-shot	137.7	146.5
	FT w/ zero-shot	132.7	142.5
	FT w/ few-shot	137.1	146.0
	RAG w/ zero-shot	134.8	134.8
	RAG w/ few-shot		
Llama 3.1	zero-shot	139.5	139.9
	few-shot	123.2	149.2
	FT w/ zero-shot	137.5	140.7
	FT w/ few-shot	138.1	152.9
	RAG w/ zero-shot	135.5	135.0
	RAG w/ few-shot		
Mistral	zero-shot		
	few-shot	136.8	152.7
	FT w/ zero-shot	124.7	131.3
	FT w/ few-shot		
	RAG w/ zero-shot	139.4	130.9
	RAG w/ few-shot		
Expert Forecasting		**25.1**	**44.9**

10.3 C

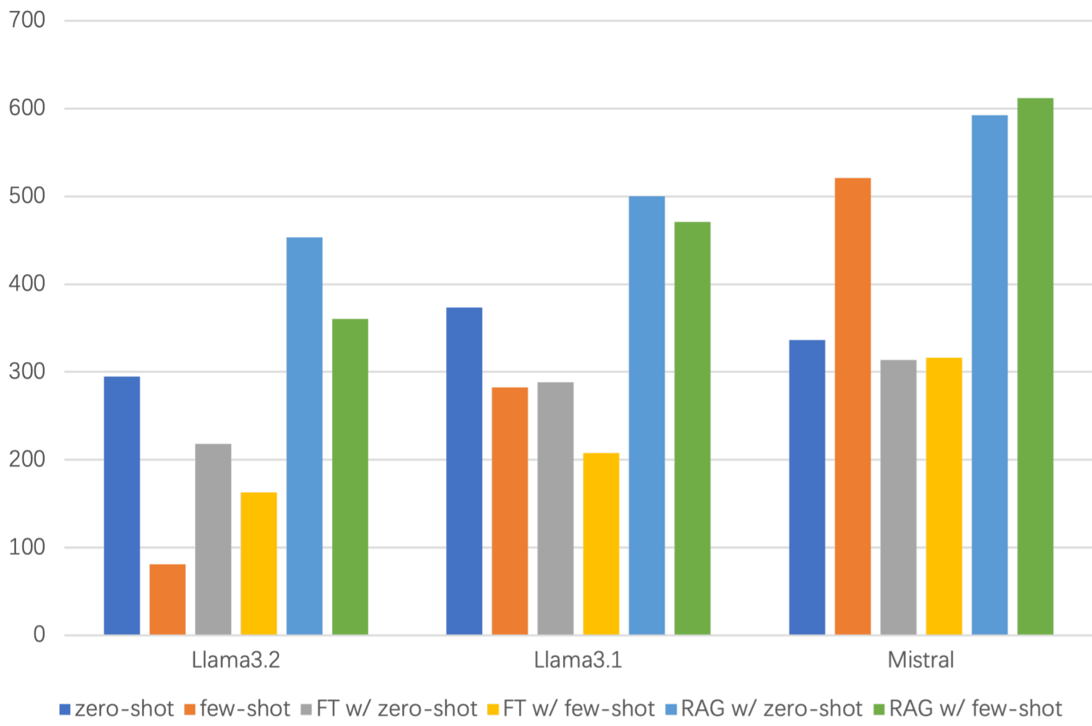


Figure 10.1: Average time consumption for each answer.

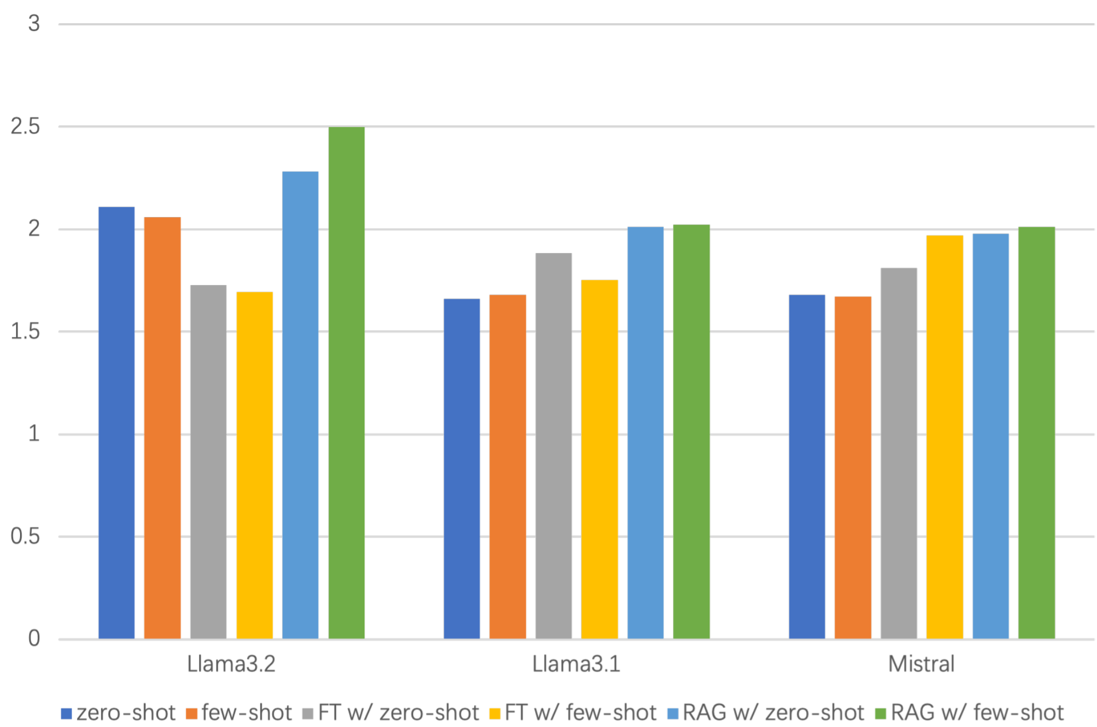


Figure 10.2: Average CPU memory consumption for inference.

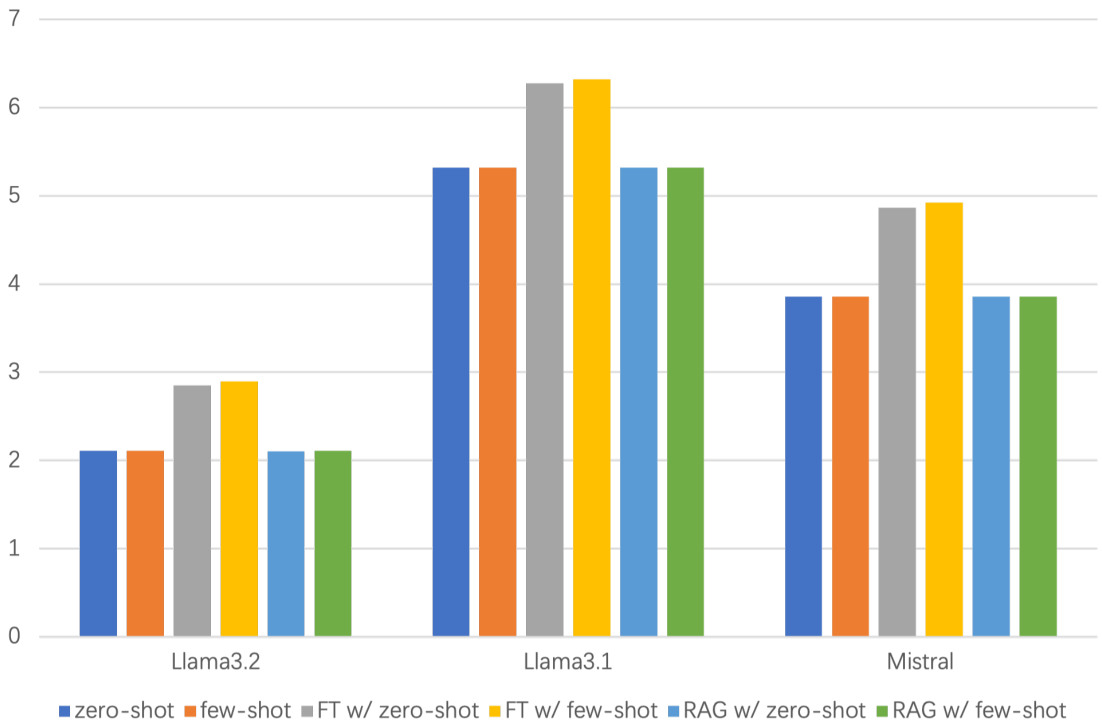
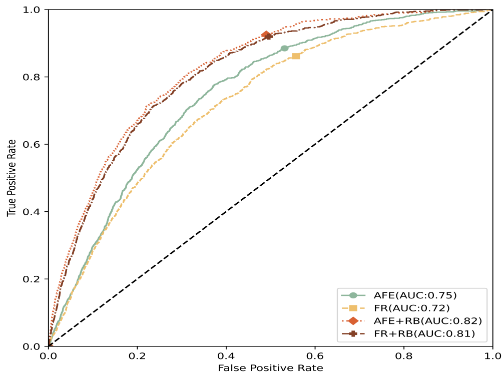
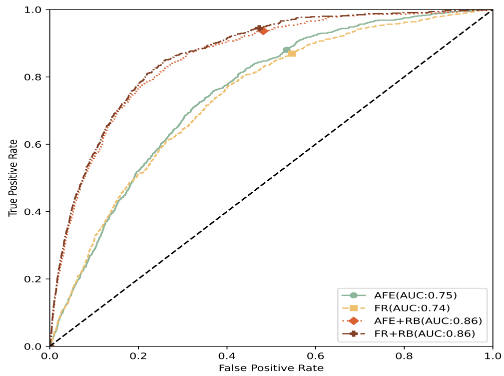


Figure 10.3: Average GPU consumption for inference.

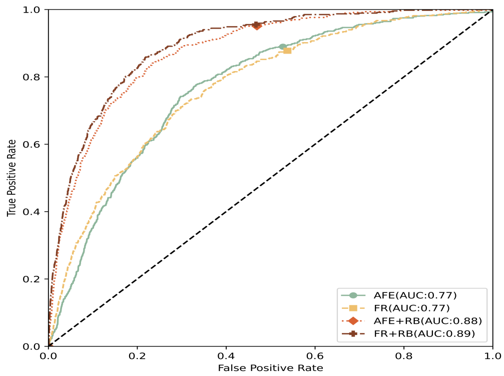
10.4 D



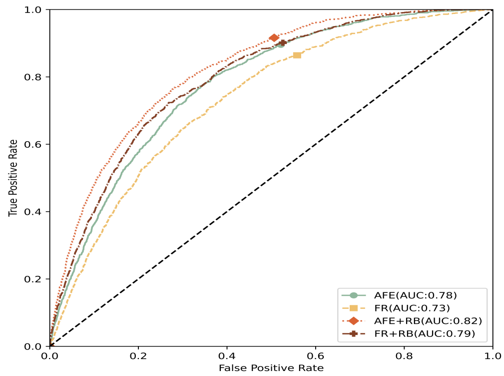
(a) LR on 1-year dataset



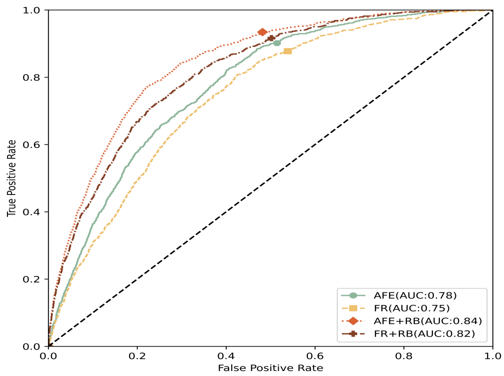
(b) LR on 2-year dataset



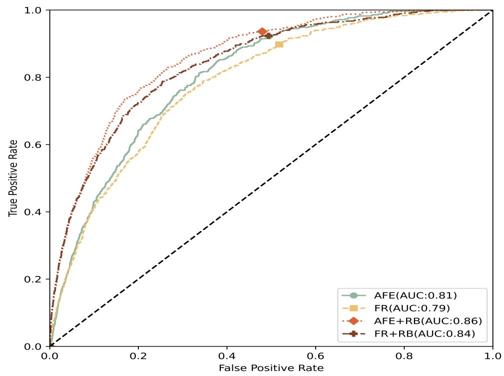
(c) LR on 3-year dataset



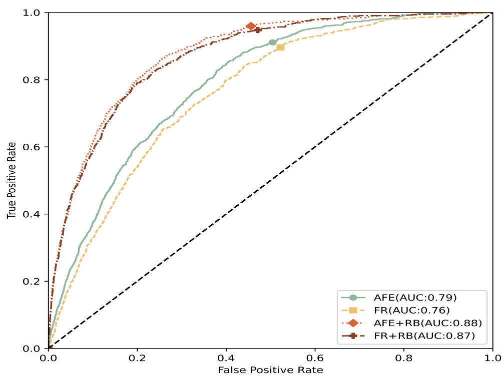
(d) RF on 1-year dataset



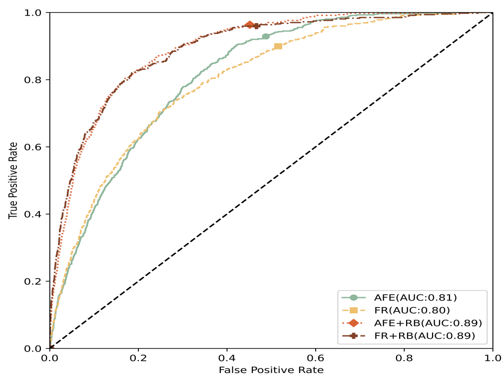
(e) RF on 2-year dataset



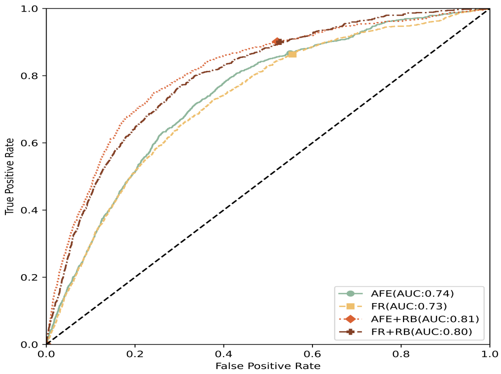
(f) RF on 3-year dataset



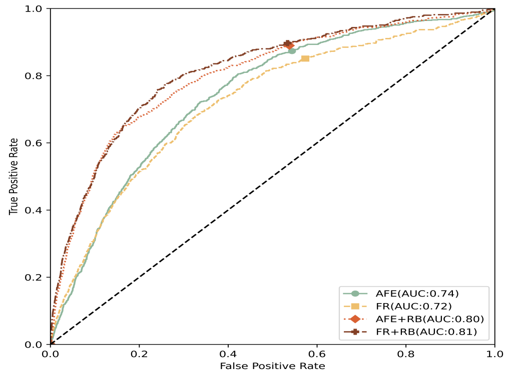
(g) LGB on 2-year dataset



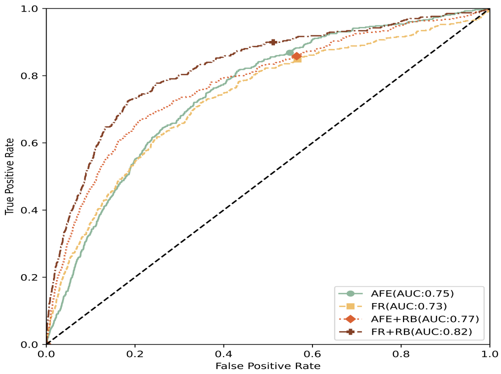
(h) LGB on 3-year dataset



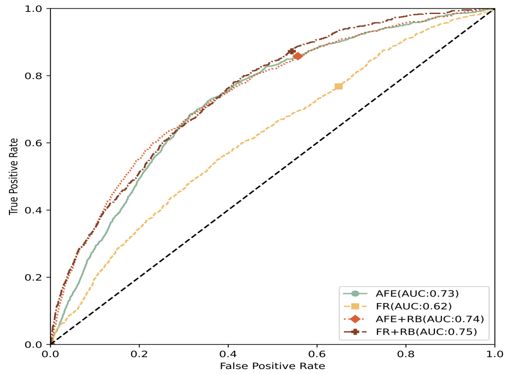
(i) MLP on 1-year dataset



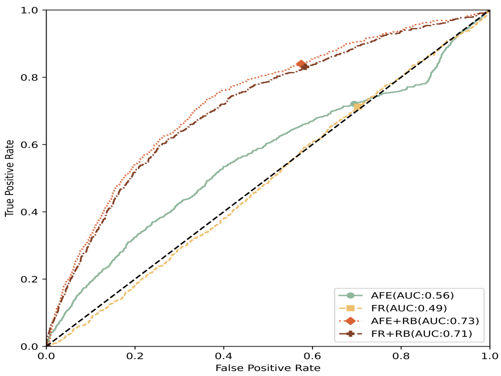
(j) MLP on 2-year dataset



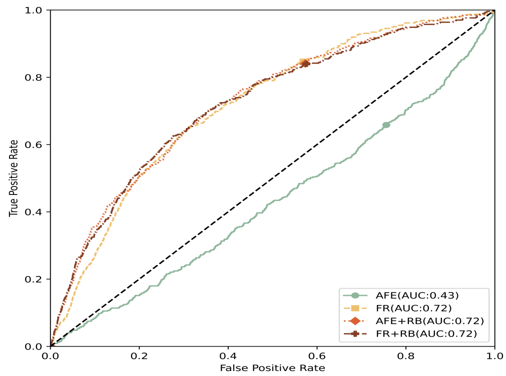
(k) MLP on 3-year dataset



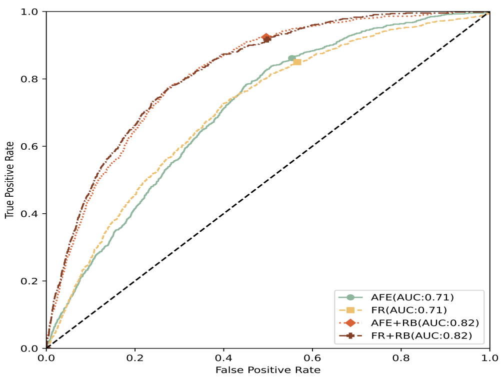
(l) CNN-1D on 1-year dataset



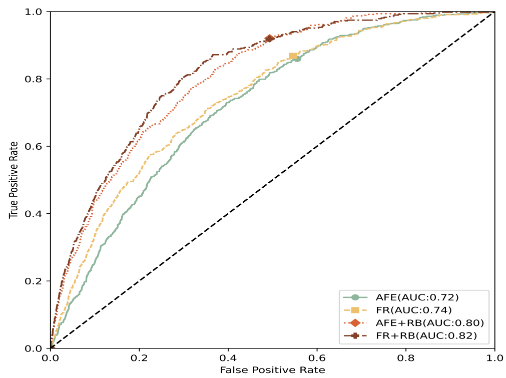
(m) CNN-1D on 2-year dataset



(n) CNN-1D on 3-year dataset



(o) LSTM on 2-year dataset



(p) LSTM on 3-year dataset

Figure 10.4: The rest results for ROC curve of 6 models on 3 datasets

10.5 E

Comprehensive Enterprise Report

TOP STAR PROMOTION-PRODUCTION, PUBLISHING AND RECORDS S.A.

Script Rating S.A

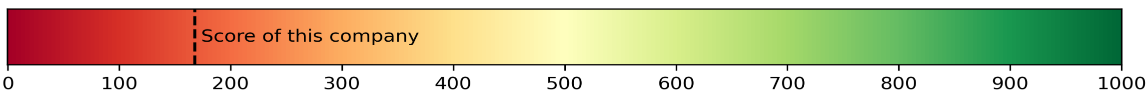
This report was generated on 2024-12-11 15:49:39 and what you are looking at is a snapshot of the data as of the cutoff point.

Basic Information

Registered name	TOP STAR PROMOTION-PRODUCTION, PUBLISHING AND RECORDS S.A.		
Business name		Registration date	1997-08-29
RCS	B60500	Legal form	Société anonyme
Nace code	82.990	Business line	Other business support service activities n.e.c.
Contact number		Email	/
Registered address	28, op der Haart, L - 9999 Wemperhardt		
Object of the company	Object from the articles of association		
Share capital	/	Number of shares	/
Shareholders	/	Number of employees	/

Credit Rating Summary

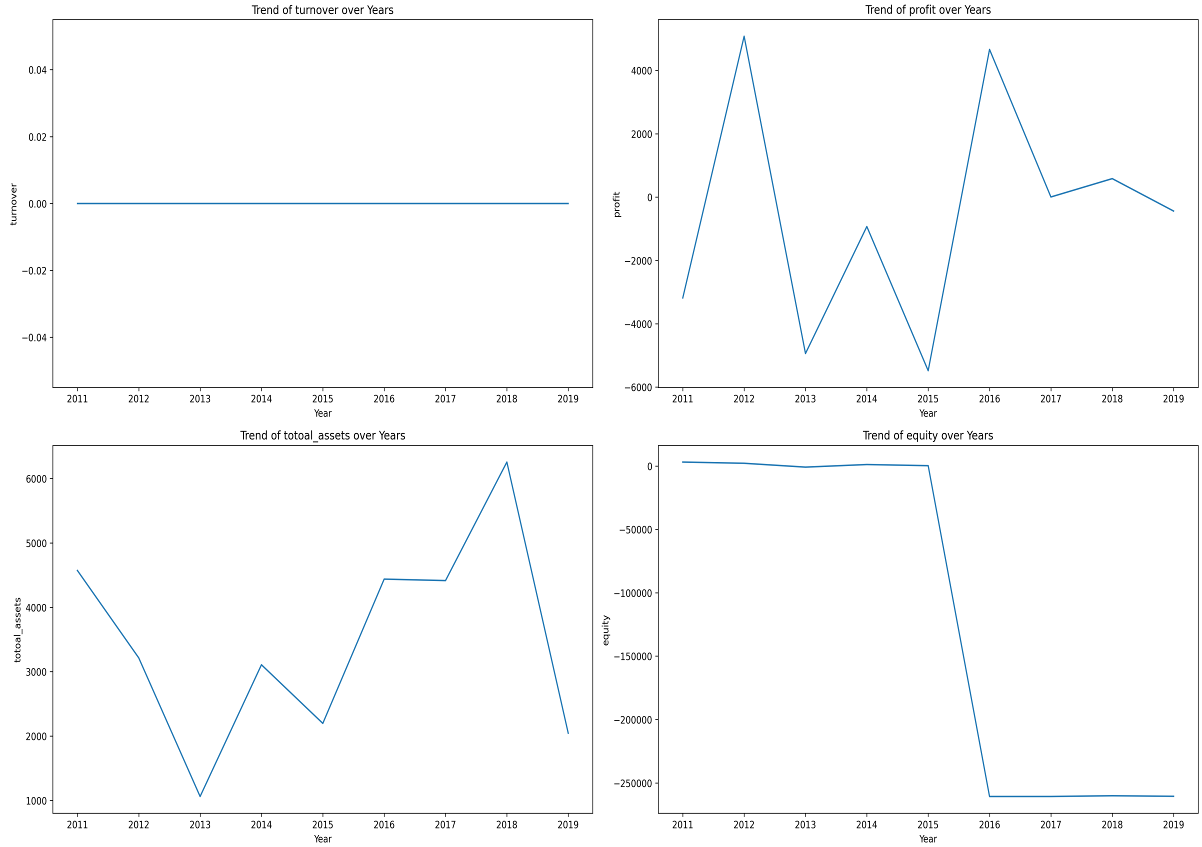
The credit score of TOP STAR PROMOTION-PRODUCTION, PUBLISHING AND RECORDS S.A. is 168 and the credit rating is E. The credit score is a numerical value that ranges from 0 to 1000, with 1000 being the best possible score. The score is calculated based on the financial ratios of the company. The credit rating is a qualitative evaluation of the creditworthiness of a company. It is based on the credit score and the probability of declaring bankruptcy in the future year.



Company position	Level	Rating	Probability of bankruptcy
	Excellent	A	<5\%
	Good	B	5\%-15\%
	Fair	C	15\%-27\%
	Passable	D	27\%-44\%
▲	Poor	E	>44\%

There is no recommended maximum credit amount for this company.

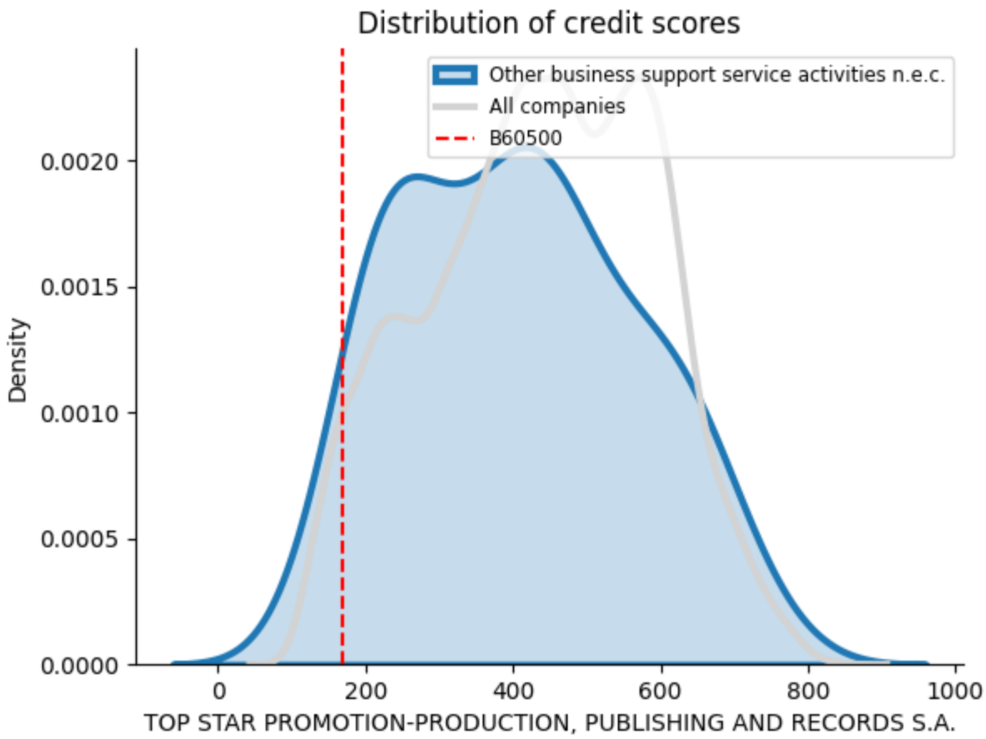
Financial Summary



Credit Score

This score is to measure if the company will go bankrupt in the future year. It is a numerical value that ranges from 0 to 1000, with 1000 being the best possible score. The score is calculated based on the financial ratios of the company.

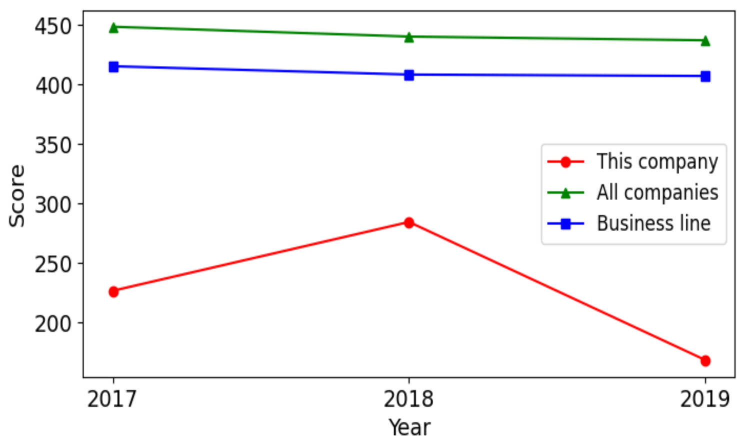
The score of TOP STAR PROMOTION-PRODUCTION, PUBLISHING AND RECORDS S.A. is 168



This company has a very low credit score, placing it in the bottom 10% of this business line. It is highly likely to go bankrupt in the future year which will be highly risky to do business with.

Rating History

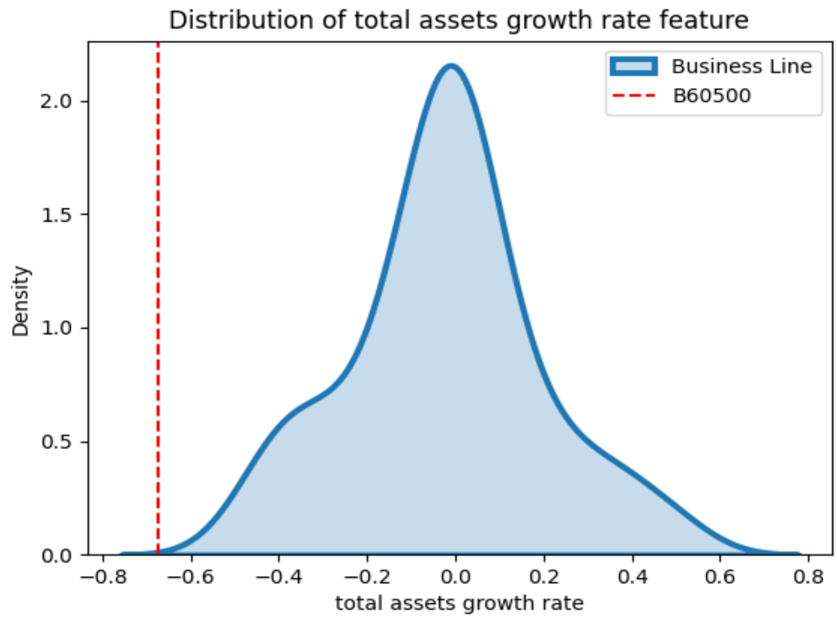
Year	Score	Rating
2019	168.0	E
2018	284.0	D
2017	226.0	D



Key Financial Ratios

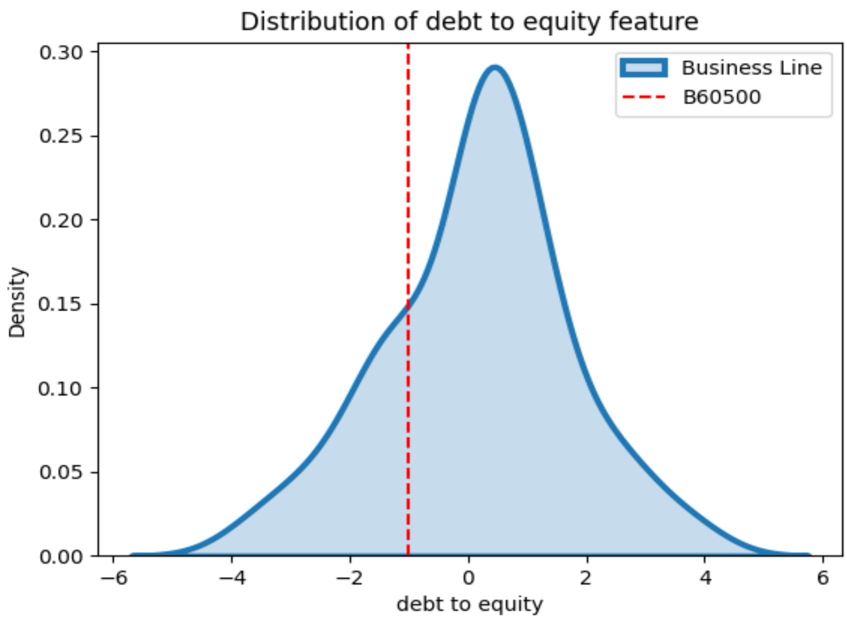
Total assets growth rate

Total assets growth rate of TOP STAR PROMOTION-PRODUCTION, PUBLISHING AND RECORDS S.A. is -0.67, and its position in this business line is displayed as below.



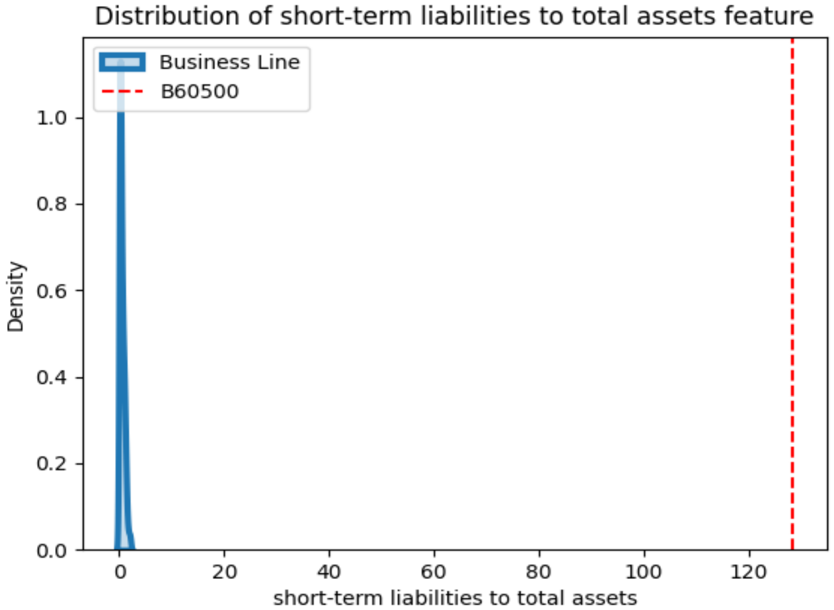
Debt to equity

Debt to equity of TOP STAR PROMOTION-PRODUCTION, PUBLISHING AND RECORDS S.A. is -1.01, and its position in this business line is displayed as below.



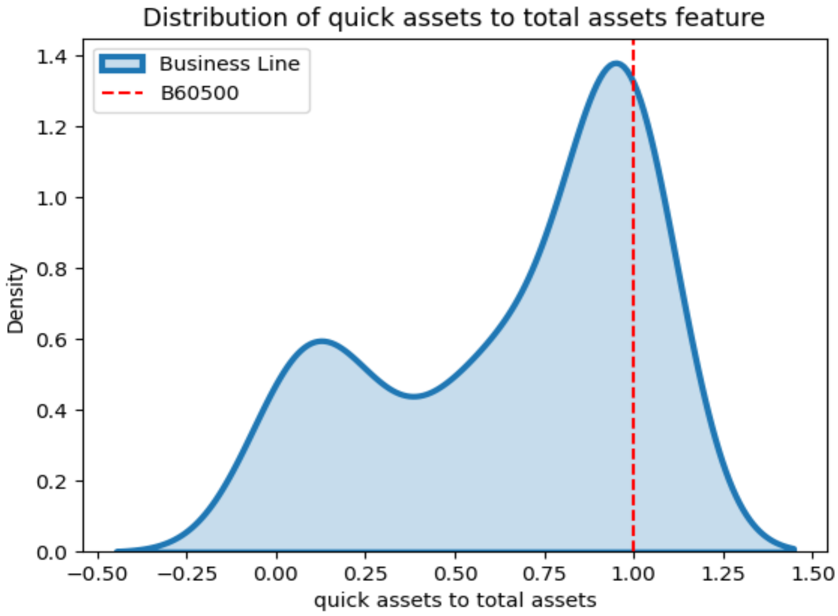
Short-term liabilities to total assets

Short-term liabilities to total assets of TOP STAR PROMOTION-PRODUCTION, PUBLISHING AND RECORDS S.A. is 128.46, and its position in this business line is displayed as below.



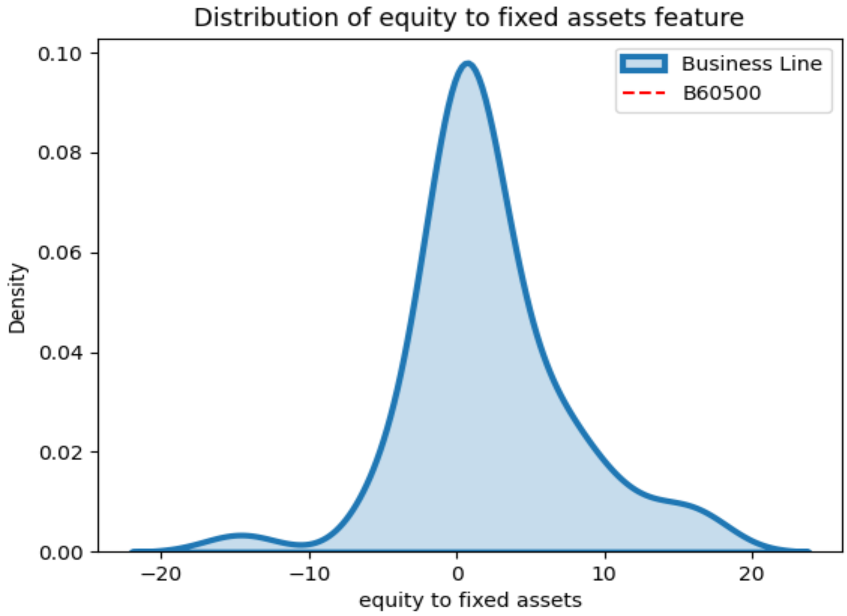
Quick assets to total assets

Quick assets to total assets of TOP STAR PROMOTION-PRODUCTION, PUBLISHING AND RECORDS S.A. is 1.0, and its position in this business line is displayed as below.



Equity to fixed assets

Equity to fixed assets of TOP STAR PROMOTION-PRODUCTION, PUBLISHING AND RECORDS S.A. is -inf, and its position in this business line is displayed as below.



Recent Activities

The company has been involved in the following activities recently.

Activity Name	Report Date	Details
Modification	20/11/2023	Voluntary liquidation
non-statutory modification of the agents	23/09/2022	Administrator(s)/Manager(s), Daily Management delegate(s), Person(s) in charge of checking the accounts
Modification	19/08/2022	Registered office
non-statutory modification of the agents	09/09/2016	Administrator(s)/Manager(s) Daily Management delegate(s) Person(s) in charge of checking the accounts
Modification	02/10/2014	Administrator(s)/Manager(s) Daily Management delegate(s)
Modification	08/03/2013	Registered office
Modification	16/09/2010	Administrator(s)/Manager(s) Daily Management delegate(s) Person(s) in charge of checking the accounts
Modification	16/11/2007	Social capital/social funds
Modification	19/10/2004	Administrator(s)/Manager(s) Daily Management delegate(s)
Modification	03/02/2004	Social capital/social funds Administrator(s)/Manager(s) Daily Management delegate(s)
Modification	17/10/2003	Administrator(s)/Manager(s) Daily Management delegate(s) Person(s) in charge of checking the accounts
Articles of association	14/05/2013	-
Articles of association	16/11/2007	-
Articles of association	24/08/2022	-

Legal Risk Alert

There is no legal risk alert for this company.

Balance Sheet

Balance Sheet Account(€)	Accounting year 2019	Accounting year 2018	Accounting year 2017
Subscribed capital unpaid	**0.0**	**0.0**	**0.0**
Subscribed capital not called	0.0	0.0	0.0
Subscribed capital called but unpaid	0.0	0.0	0.0
Formation expenses	**0.0**	**0.0**	**0.0**
Fixed assets	**0.0**	**0.0**	**0.0**
Intangible assets	0.0	0.0	0.0
Tangible assets	0.0	0.0	0.0
Financial assets	0.0	0.0	0.0
Current assets	**2044.52**	**6256.13**	**4414.85**
Stocks	0.0	0.0	0.0
Debtors	2044.52	1799.31	1620.16
becoming due and payable within one year	2044.52	1799.31	1620.16
becoming due and payable after more than one year	0.0	0.0	0.0
Investments	0.0	0.0	0.0
Cash at bank and in hand	0.0	4456.82	2794.69
Prepayments	**0.0**	**0.0**	**0.0**
TOTAL (ASSETS)	**2044.52**	**6256.13**	**4414.85**
Capital and reserves	**-260598.97**	**-260157.93**	**-260741.05**
Subscribed capital	61176.47	61176.47	61176.47
Share premium account	1121063.53	1121063.53	1121063.53
Revaluation reserve	0.0	0.0	0.0
Reserves	7461.6	7461.6	7461.6
Profit or loss brought forward	-1449859.53	-1450442.65	-1450446.51
Profit or loss for the financial year	-441.04	583.12	3.86
Interim dividends	0.0	0.0	0.0
Capital investment subsidies	0.0	0.0	0.0
Provisions	**0.0**	**0.0**	**0.0**
Creditors	**262643.49**	**266414.06**	**265155.9**
becoming due and payable within one year	262643.49	266414.06	265155.9
becoming due and payable after more than one year	0.0	0.0	0.0
Deferred income	**0.0**	**0.0**	**0.0**
TOTAL (CAPITAL, RESERVES AND LIABILITIES)	**2044.52**	**6256.13**	**4414.85**

Insights of Balance Sheet

Narrative Report

The financial data provided reveals a company with constrained liquidity, negative equity reserves, and minimal asset diversification. Significant changes are observed in current assets and creditors over the three years. There are indications of operational inefficiencies and risks due to consistent negative reserves. However, the company has slightly reduced its losses in recent years, signaling some improvement.

Key Observations:

- Liquidity Issues: The company's current assets dropped sharply from 2018 to 2019, mainly due to a reduction in cash holdings. Despite stable debtors, the significant drop in cash signals potential operational stress.
- Negative Reserves: Persistent negative reserves underscore financial instability, though the annual loss is narrowing.
- Creditors: Payables have remained steady, indicating no significant shift in short-term obligations.

Recommendations:

- Improve Liquidity: Focus on reducing unnecessary cash outflows and improving debtor collection.
- Strengthen Equity Position: Explore capital infusion or retained earnings to offset the negative reserves.
- Optimize Working Capital: Improve cash flow management by reducing reliance on creditors and enhancing debtor turnover.

Trend Analysis

1. Assets:
Current assets declined drastically in 2019 (by approximately 67%) compared to 2018. Cash reserves were entirely depleted, leaving only receivables. 2. Liabilities:
Creditors have remained steady, signaling the company's reliance on external financing. 3. Reserves:
The company's negative reserves improved slightly from 2017 to 2019. Annual losses narrowed, with the financial year loss decreasing from -583.12 in 2018 to -441.04 in 2019.

Profit Forecasting

The annual financial losses are improving, as evidenced by:
- 2017 Loss: -3.86
- 2018 Loss: -583.12
- 2019 Loss: -441.04
Using a simple trend analysis (average annual reduction in loss of approximately 142), the forecasted loss for 2020 would be approximately -299. However, without revenue or operational improvements, profitability is unlikely.

Key Financial Ratios

Liquidity Ratios

- Current Ratio = Current Assets ÷ Current Liabilities
- 2019: 2044.52 ÷ 262643.49 = 0.0078
- 2018: 6256.13 ÷ 266414.06 = 0.0235
- 2017: 4414.85 ÷ 265155.90 = 0.0167
- Quick Ratio = (Current Assets - Stocks) ÷ Current Liabilities
- The company holds no stock, so the quick ratio is the same as the current ratio.

Solvency Ratios

- Debt-to-Equity Ratio = Total Liabilities ÷ Equity
- Equity = Capital + Reserves
- 2019: 262643.49 ÷ (61176.47 - 260598.97) = -1.02
- 2018: 266414.06 ÷ (61176.47 - 260157.93) = -1.02
- 2017: 265155.90 ÷ (61176.47 - 260741.05) = -1.02

Profitability Ratios

- Return on Equity (ROE) = Net Income ÷ Shareholder's Equity
- Equity is negative, making this ratio inapplicable. Losses further reduce shareholder value.
- Return on Assets (ROA) = Net Income ÷ Total Assets
- 2019: -441.04 ÷ 2044.52 = -21.57%
- 2018: 583.12 ÷ 6256.13 = 9.32%
- 2017: 3.86 ÷ 4414.85 = 0.09%

Interpretation: ROA was positive in 2018 but turned sharply negative in 2019. This decline reflects worsening efficiency in asset utilization.

- Net Profit Margin = Net Income ÷ Total Revenue
- Revenue is not provided; thus, this ratio cannot be calculated.

Risk Alerts

- Liquidity Crisis: With a current ratio below 0.01, the company may face immediate challenges in meeting short-term obligations.
- Insolvency Risk: Negative reserves and equity make the company vulnerable to financial collapse without external support.
- Decline in Cash: A complete depletion of cash holdings in 2019 poses operational risks.

Working Capital Analysis

- Working Capital = Current Assets - Current Liabilities
- 2019: 2044.52 - 262643.49 = -260598.97
- 2018: 6256.13 - 266414.06 = -260157.93
- 2017: 4414.85 - 265155.90 = -260741.05

The negative working capital highlights a severe liquidity crunch, with current liabilities exceeding assets by a wide margin.

Recommendations for Working Capital Efficiency:

- Debtor Management: Accelerate receivables collection to improve cash inflows.
- Expense Management: Cut discretionary expenses to conserve cash.
- Explore Short-term Financing: Negotiate with creditors to extend payment terms or seek working capital loans.

Insights

The company faces significant financial challenges, including poor liquidity, negative equity, and declining cash reserves. While some improvement in annual losses is observed, the overall financial health remains precarious. Immediate actions should focus on improving liquidity, restructuring debt, and exploring revenue-enhancing opportunities to stabilize operations.

Bibliography

[1] Taiwanese Bankruptcy Prediction. UCI Machine Learning Repository, 2020.

[2] May 2021.

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[4] Vineet Agarwal and Richard Taffler. Comparing the performance of market-based and accounting-based bankruptcy prediction models. *Journal of Banking & Finance*, 32(8):1541–1551, 2008.

[5] Hafiz A Alaka, Lukumon O Oyedele, Hakeem A Owolabi, Vikas Kumar, Saheed O Ajayi, Olugbenga O Akinade, and Muhammad Bilal. Systematic review of bankruptcy prediction models: Towards a framework for tool selection. *Expert Systems with Applications*, 94:164–184, 2018.

[6] Talha Mahboob Alam, Kamran Shaukat, Mubbashar Mushtaq, Yasir Ali, Matloob Khushi, Suhuai Luo, and Abdul Wahab. Corporate bankruptcy prediction: An approach towards better corporate world. *The Computer Journal*, page 1731–1746, Nov 2021.

[7] Lucia Alessi and Roberto Savona. Machine learning for financial stability. In *Data Science for Economics and Finance: Methodologies and Applications*, pages 65–87. Springer International Publishing Cham, 2021.

[8] Edward I Altman. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, 23(4):589–609, 1968.

[9] Edward I Altman, Malgorzata Iwanicz-Drozdowska, Erkki K Laitinen, and Arto Suvas. Financial and non-financial variables as long-horizon predictors of bankruptcy. *Available at SSRN 2669668*, 2015.

[10] Edward I Altman, Gabriele Sabato, and Nick Wilson. The value of non-financial information in sme risk management. *Available at SSRN 1320612*, 2008.

[11] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010.

[12] Sanobar Anjum. Business bankruptcy prediction models: A significant study of the altman's z-score model. *Available at SSRN 2128475*, 2012.

[13] Kingsley Opoku Appiah, Amon Chizema, and Joseph Arthur. Predicting corporate failure: a systematic literature review of methodological issues. *International Journal of Law and Management*, 57(5):461–485, 2015.

[14] Amir F Atiya. Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Transactions on neural networks*, 12(4):929–935, 2001.

[15] World Bank. Improving smes' access to finance and finding innovative solutions to unlock sources of capital. https://www.worldbank.org/en/topic/smefinance, 2023. Accessed 10 April 2023.

[16] Flavio Barboza, Herbert Kimura, and Edward Altman. Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, 83:405–417, 2017.

[17] Berardino Barile, Marco Forti, Alessia Marrocco, and Angelo Castaldo. Causal impact evaluation of occupational safety policies on firms' default using machine learning uplift modelling. *Scientific Reports*, 14(1):10380, 2024.

[18] William H Beaver. Financial ratios as predictors of failure. *Journal of accounting research*, pages 71–111, 1966.

[19] Jodi L. Bellovary, Don E. Giacomino, and Michael D. Akers. A review of bankruptcy prediction studies: 1930 to present. *Journal of Financial Education*, 33:1–42, 2007.

[20] Artem Bequé, Kristof Coussement, Ross Gayler, and Stefan Lessmann. Approaches for credit scorecard calibration: An empirical analysis. *Knowledge-Based Systems*, 134:213–227, 2017.

[21] Dragana Bešlić Obradović, Dejan Jakšić, Ivana Bešlić Rupić, and Mirko Andrić. Insolvency prediction model of the company: the case of the republic of serbia. *Economic Research-Ekonomska Istraživanja*, page 139–157, Jan 2018.

[22] Lisa M Bodnar, Marie Davidian, Anna Maria Siega-Riz, and Anastasios A Tsiatis. Marginal structural models for analyzing causal effects of time-dependent treatments: an application in perinatal epidemiology. *American journal of epidemiology*, 159(10):926–934, 2004.

[23] Vytautas Boguslauskas, Ričardas Mileris, and Rūta Adlytė. The selection of financial ratios as independent variables for credit risk assessment. *Economics and management*, 16(4):1032–1040, 2011.

[24] Derek W Brown, Thomas J Greene, Michael D Swartz, Anna V Wilkinson, and Stacia M DeSantis. Propensity score stratification methods for continuous treatments. *Statistics in medicine*, 40(5):1189–1203, 2021.

[25] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[26] Brantly Callaway, Andrew Goodman-Bacon, and Pedro HC Sant'Anna. Difference-in-differences with a continuous treatment. Technical report, National Bureau of Economic Research, 2024.

[27] A.M. Callejón, A.M. Casado, M.A. Fernández, and J.I. Peláez. A system of insolvency prediction for industrial companies using a financial alternative model with neural networks. *International Journal of Computational Intelligence Systems*, page 29, May 2013.

[28] Hector F Calvo-Pardo, Tullio Mancini, and Jose Olmo. Neural network models for empirical finance. *Journal of Risk and Financial Management*, 13(11):265, 2020.

[29] D Karthik Chandra, Vadlamani Ravi, and Indranil Bose. Failure prediction of dotcom companies using hybrid intelligent techniques. *Expert Systems with applications*, 36(3):4830–4837, 2009.

[30] Huigang Chen, Totte Harinen, Jeong-Yoon Lee, Mike Yung, and Zhenyu Zhao. Causalml: Python package for causal machine learning, 2020.

[31] Ning Chen, Bernardete Ribeiro, Armando S. Vieira, João Duarte, and João C. Neves. A genetic algorithm-based approach to cost-sensitive bankruptcy prediction. *Expert Systems with Applications*, page 12939–12945, Sep 2011.

[32] Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, et al. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*, 2021.

[33] Robert Chew, John Bollenbacher, Michael Wenger, Jessica Speer, and Annice Kim. Llm-assisted content analysis: Using large language models to support deductive coding. *arXiv preprint arXiv:2306.14924*, 2023.

[34] Kwangseek Choe and Samy Garas. The graph theoretical approach to bankruptcy prediction. *The Journal of Accounting and Management,The Journal of Accounting and Management*, Jan 2021.

[35] Vineet Chouhan, Bibhas Chandra, and Shubham Goswami. Predicting financial stability of select bse companies revisiting altman z score. *International Letters of Social and Humanistic Sciences*, 15(2):92–105, 2014.

[36] Chun-Ling Chuang. Application of hybrid case-based reasoning for enhanced performance in bankruptcy prediction. *Information Sciences*, page 174–185, Jul 2013.

[37] Francesco Ciampi. Corporate governance characteristics and default prediction modeling for small enterprises. an empirical analysis of italian firms. *Journal of Business Research*, page 1012–1025, May 2015.

[38] Zeynep Cındık and Ismail H Armutlulu. A revision of altman z-score model and a comparative analysis of turkish companies' financial distress prediction. *National Accounting Review*, 3(2):237–255, 2021.

[39] Claudiu Clement et al. Machine learning in bankruptcy prediction–a review. *Journal of Public Administration, Finance and Law*, (17):178–196, 2020.

[40] Peter L Cohen. *Algorithmic Approaches to Nonparametric Causal Inference*. PhD thesis, Massachusetts Institute of Technology, 2022.

[41] CĂRUNTU Constantin and LĂPĂDUŞI MIHAELA Loredana. The analysis of the indicators which reflect the ability of companies of facing short term obligations and medium and long term maturities. *Annals-Economy Series*, 4(1):89–95, 2012.

[42] Loredana Cultrera and Xavier Brédart. Bankruptcy prediction: the case of belgian smes. *Review of Accounting and Finance*, page 101–119, Feb 2016.

[43] Catherine M. Daily and Dan R. Dalton. Corporate governance and the bankrupt firm: An empirical assessment. *Southern Medical Journal*, 15:643–654, 1994.

[44] Ali F Darrat, Stephen Gray, Jung Chul Park, and Yanhui Wu. Corporate governance and bankruptcy risk. *Journal of Accounting, Auditing & Finance*, 31(2):163–202, 2016.

[45] Floris Devriendt, Jente Van Belle, Tias Guns, and Wouter Verbeke. Learning to rank for uplift modeling. *IEEE Transactions on Knowledge and Data Engineering*, 34(10):4888–4904, 2020.

[46] Eustache Diemert, Artem Betlei, Christophe Renaudin, and Massih-Reza Amini. A large scale benchmark for uplift modeling. In *KDD*, 2018.

[47] Michael Doumpos, Kostas Andriosopoulos, EmiliosC. Galariotis, Georgia Makridou, and Constantin Zopounidis. Corporate failure prediction in the european energy sector: A multicriteria approach and the effect of country characteristics. *Research Papers in Economics,Research Papers in Economics*, Jan 2017.

[48] Philippe du Jardin. Bankruptcy prediction using terminal failure processes. *European Journal of Operational Research*, page 286–303, Apr 2015.

[49] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[50] Alexander D'Amour. On multi-cause approaches to causal inference with unobserved counfounding: Two cautionary failure cases and a promising alternative. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3478–3486. PMLR, 2019.

[51] Sophie Emerson, Ruairí Kennedy, Luke O'Shea, and John O'Brien. Trends and applications of machine learning in quantitative finance. In *8th international conference on economics and finance research (ICEFR 2019)*, 2019.

[52] Marc Eulerich, Aida Sanatizadeh, Hamid Vakilzadeh, and David A Wood. Can artificial intelligence pass accounting certification exams? chatgpt: Cpa, cma, cia, and ea. *ChatGPT: CPA, CMA, CIA, and EA*, 2023.

[53] Joseph Faello. Understanding the limitations of financial ratios. *Academy of accounting and financial studies journal*, 19(3):75, 2015.

[54] Hossam Faris, Ruba Abukhurma, Waref Almanaseer, Mohammed Saadeh, Antonio M. Mora, Pedro A. Castillo, and Ibrahim Aljarah. Improving financial bankruptcy prediction in a highly imbalanced class distribution using oversampling and ensemble learning: a case from the spanish market. *Progress in Artificial Intelligence*, page 31–53, Mar 2020.

[55] Guilherme Barreto Fernandes and Rinaldo Artes. Spatial dependence in credit risk and its improvement in credit scoring. *European Journal of Operational Research*, page 517–524, Mar 2016.

[56] Silvia Figini, Federico Bonelli, and Emanuele Giovannini. Solvency prediction for small and medium enterprises in banking. *Decision Support Systems*, page 91–97, Oct 2017.

[57] Laura Forastiere, Edoardo M Airoldi, and Fabrizia Mealli. Identification and estimation of treatment and interference effects in observational studies on networks. *Journal of the American Statistical Association*, 116(534):901–918, 2021.

[58] Mihaela Gadoiu et al. Advantages and limitations of the financial ratios used in the financial diagnosis of the enterprise. *Scientific Bulletin-Economic Sciences*, 13(2):87–95, 2014.

[59] SMAH Gillani, Suresh Ramakrishnan, Hamad Raza, and Humara Ahmad. Review of corporate governance practices and financial distress prediction. *International Journal of Engineering & Technology*, 7(4.28):30–33, 2018.

[60] Aman Gupta, Anup Shirgaonkar, Angels de Luis Balaguer, Bruno Silva, Daniel Holstein, Dawei Li, Jennifer Marsman, Leonardo O Nunes, Mahsa Rouzbahman, Morris Sharp, et al. Rag vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture. *arXiv preprint arXiv:2401.08406*, 2024.

[61] Pierre Gutierrez and Jean-Yves Gérardy. Causal inference and uplift modelling: A review of the literature. In *International conference on predictive applications and APIs*, pages 1–13. PMLR, 2017.

[62] Nicholas Harvel, Felipe Bivort Haiek, Anupriya Ankolekar, and David James Brunner. Can llms answer investment banking questions? using domain-tuned functions to improve llm performance on knowledge-intensive analytical tasks. In *Proceedings of the AAAI Symposium Series*, volume 3, pages 125–133, 2024.

[63] Dan Hauschild. Altman z-score: Not just for bankruptcy. *From Z-score to" Green Zone" survivability: AMPros Corporation*, 2013.

[64] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[65] Mario Hernandez Tinoco and Nick Wilson. Financial distress and bankruptcy prediction among listed companies using accounting, market and macroeconomic variables. *International Review of Financial Analysis*, page 394–419, Dec 2013.

[66] Tadaaki Hosaka. Bankruptcy prediction using imaged financial ratios and convolutional neural networks. *Expert systems with applications*, 117:287–299, 2019.

[67] Ronald A Howard. Information value theory. *IEEE Transactions on systems science and cybernetics*, 2(1):22–26, 1966.

[68] Bo Huang, Xiao Yao, Yinqing Luo, and Jing Li. Improving financial distress prediction using textual sentiment of annual reports. *Annals of Operations Research*, pages 1–28, 2022.

[69] Maciej Jaskowski and Szymon Jaroszewicz. Uplift modeling for clinical trial data. In *ICML workshop on clinical data analysis*, volume 46, pages 79–95, 2012.

[70] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[71] Nam-Ok Jo and Kyung-Shik Shin. Bankruptcy prediction modeling using qualitative information based on big data analytics. 2016.

[72] Stewart Jones. Corporate bankruptcy prediction: a high dimensional analysis. *Review of Accounting Studies*, page 1366–1422, Sep 2017.

[73] Stewart Jones. A literature survey of corporate failure prediction models. *Journal of Accounting Literature*, 45(2):364–405, 2023.

[74] Nathan Kallus and Angela Zhou. Policy evaluation and optimization with continuous treatments. In *International conference on artificial intelligence and statistics*, pages 1243–1251. PMLR, 2018.

[75] James Max Kanter and Kalyan Veeramachaneni. Deep feature synthesis: Towards automating data science endeavors. In *2015 IEEE international conference on data science and advanced analytics (DSAA)*, pages 1–10. IEEE, 2015.

[76] A. Katsinis, J. Lagüera-González, L. Di Bella, L. Odenthal, M. Hell, and B. Lozar. Annual report on european smes 2023/2024. https://single-market-economy.ec.europa.eu/document/download/2bef0eda-2f75-497d-982e-c0d1cea57c0e_en?filename=Annual%20Report%20on%20European%20SMEs%202024.pdf, 2023. Accessed: 2024-12-08.

[77] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[78] Edward H Kennedy, Zongming Ma, Matthew D McHugh, and Dylan S Small. Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(4):1229–1245, 2017.

[79] David M Kent, Ewout Steyerberg, and David Van Klaveren. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. *Bmj*, 363, 2018.

[80] Layla Khoja, Maxwell Chipulu, and Ranadeva Jayasekera. Analysis of financial distress cross countries: Using macroeconomic, industrial indicators and accounting data. *International Review of Financial Analysis*, 66:101379, 2019.

[81] Alex Kim, Maximilian Muhn, and Valeri V. Nikolaev. 1. financial statement analysis with large language models. *arXiv.org*, 2024.

[82] Hyeongjun Kim, Hoon Cho, and Doojin Ryu. Corporate bankruptcy prediction using machine learning methodologies with a focus on sequential data. *Computational Economics*, page 1231–1249, Mar 2022.

[83] Myoung-Jong Kim and Dae-Ki Kang. Ensemble with neural networks for bankruptcy prediction. *Expert systems with applications*, 37(4):3373–3379, 2010.

[84] Sanghyun Kim, Jungwon Choi, NamHee Kim, Jaesung Ryu, and Juho Lee. Modeling uplift from observational time-series in continual scenarios. In *AAAI Bridge Program on Continual Causality*, pages 75–84. PMLR, 2023.

[85] Sungdo Kim, Byeong Min Mun, and Suk Joo Bae. Data depth based support vector machines for predicting corporate bankruptcy. *Applied Intelligence*, 48:791–804, 2018.

[86] Dehan Kong, Shu Yang, and Linbo Wang. Multi-cause causal inference with unmeasured confounding and binary outcome. *arXiv preprint arXiv:1907.13323*, 2019.

[87] Gang Kou, Yong Xu, Yi Peng, Feng Shen, Yang Chen, Kun Chang, and Shaomin Kou. Bankruptcy prediction for smes using transactional data and two-stage multiobjective feature selection. *Decision Support Systems*, 140:113429, 2021.

[88] Gang Kou, Yong Xu, Yi Peng, Feng Shen, Yang Chen, Kun Chang, and Shaomin Kou. Bankruptcy prediction for smes using transactional data and two-stage multiobjective feature selection. *Decision Support Systems*, 140:113429, Jan 2021.

[89] Maria Kovacova, Tomas Kliestik, Katarina Valaskova, Pavol Durana, and Zuzana Juhaszova. Systematic review of variables applied in bankruptcy

prediction models of visegrad group countries. *Oeconomia Copernicana*, 10(4):743–772, 2019.

[90] Jochen Kruppa, Alexandra Schwarz, Gerhard Arminger, and Andreas Ziegler. Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications*, 40(13):5125–5131, 2013.

[91] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

[92] Tuong Le, Mi Lee, Jun Park, and Sung Baik. Oversampling techniques for bankruptcy prediction: Novel features from a transaction dataset. *Symmetry*, 10(4):79, Mar 2018.

[93] Tuong Le, Mi Young Lee, Jun Ryeol Park, and Sung Wook Baik. Oversampling techniques for bankruptcy prediction: Novel features from a transaction dataset. *Symmetry*, 10(4):79, 2018.

[94] Sanghoon Lee, Keunho Choi, and Donghee Yoo. Predicting the insolvency of smes using technological feasibility assessment information and data mining techniques. *Sustainability*, page 9790, Nov 2020.

[95] Bastien Lextrait. Scaling up smes' credit scoring scope with lightgbm. *Applied Economics*, pages 1–19, 2022.

[96] Kun Li, Jiang Tian, and Xiaojia Xiang. A new transformation approach for uplift modeling with binary outcome. *arXiv preprint arXiv:2310.05549*, 2023.

[97] Mingyuan Li and Peter Miu. A hybrid bankruptcy prediction model with dynamic loadings on accounting-ratio-based and market-based information: A binary quantile regression approach. *Journal of Empirical Finance,Journal of Empirical Finance*, Apr 2009.

[98] Deron Liang, Chia-Chi Lu, Chih-Fong Tsai, and Guan-An Shih. Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *European journal of operational research*, 252(2):561–572, 2016.

[99] Deron Liang, Chia-Chi Lu, Chih-Fong Tsai, and Guan-An Shih. Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *European Journal of Operational Research*, 252(2):561–572, Jul 2016.

[100] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[101] Fengyi Lin, Deron Liang, and Wing-Sang Chu. The role of non-financial features related to corporate governance in business crisis prediction. *Journal of Marine Science and Technology*, 18(4):4, 2010.

[102] Bin Liu, Chenxu Zhu, Guilin Li, Weinan Zhang, Jincai Lai, Ruiming Tang, Xiuqiang He, Zhenguo Li, and Yong Yu. Autofis: Automatic feature interaction selection in factorization models for click-through rate prediction. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2636–2645, 2020.

[103] Dugang Liu, Xing Tang, Han Gao, Fuyuan Lyu, and Xiuqiang He. Explicit feature interaction-aware uplift network for online marketing. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4507–4515, 2023.

[104] Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, pages 4513–4519, 2021.

[105] Claudio F Loderer and Dennis P Sheehan. Corporate bankruptcy and managers' self-serving behavior. *The Journal of Finance*, 44(4):1059–1075, 1989.

[106] Gianfranco Lombardo, Mattia Pellegrino, George Adosoglou, Stefano Cagnoni, Panos M Pardalos, and Agostino Poggi. Machine learning for bankruptcy prediction in the american stock market: Dataset and benchmarks. *Future Internet*, 14(8):244, 2022.

[107] Gianfranco Lombardo, Mattia Pellegrino, George Adosoglou, Stefano Cagnoni, Panos M. Pardalos, and Agostino Poggi. Machine learning for bankruptcy prediction in the american stock market: Dataset and benchmarks. *Future Internet*, page 244, Aug 2022.

[108] Michael J Lopez and Roee Gutman. Estimation of causal effects with multiple treatments: a review and new ideas. *Statistical Science*, pages 432–454, 2017.

[109] L. Loukas, Ilias Stogiannidis, Prodromos Malakasiotis, and S. Vassos. 2. breaking the bank with chatgpt: Few-shot text classification for finance. 2023.

[110] Oliver Lukason and Art Andresson. Tax arrears versus financial ratios in bankruptcy prediction. *Journal of Risk and Financial Management*, page 187, Dec 2019.

[111] Jing Ma, Ruocheng Guo, Aidong Zhang, and Jundong Li. Multi-cause effect estimation with disentangled confounder representation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 2021.

[112] Feng Mai, Shaonan Tian, Chihoon Lee, and Ling Ma. Deep learning models for bankruptcy prediction using textual disclosures. *European journal of operational research*, 274(2):743–758, 2019.

[113] Feng Mai, Shaonan Tian, Chihoon Lee, and Ling Ma. Deep learning models for bankruptcy prediction using textual disclosures. *European Journal of Operational Research*, 274(2):743–758, Apr 2019.

[114] José Manuel, Jarmila Horváthová, Martina Mokrišová, and Martin Bača. Bankruptcy prediction for sustainability of businesses: The application of graph theoretical modeling. 2023.

[115] A. Mendes, R. L. Cardoso, P. C. Mário, A. L. Martinez, and F. R. Ferreira. Insolvency prediction in the presence of data inconsistencies. *Intelligent Systems in Accounting, Finance and Management*, page 155–167, Jul 2014.

[116] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

[117] Felipe Moraes, Hugo Manuel Proença, Anastasiia Kornilova, Javier Albert, and Dmitri Goldenberg. Uplift modeling: From causal inference to personalization. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5212–5215, 2023.

[118] Gary P Moynihan, Vineet Jain, Robert W McLeod, and Daniel J Fonseca. An expert system for financial ratio analysis. *International Journal of Financial Services Management*, 1(2-3):141–154, 2006.

[119] Nada Mselmi, Amine Lahiani, and Taher Hamza. Financial distress prediction: The case of french small and medium-sized firms. *International Review of Financial Analysis*, 50:67–80, 2017.

[120] Setia Mulyawan. The benefits of financial ratios' as the indocators of future bankruptcy on the economic crisis. *International Journal of Nusantara Islam*, 3(1):21–30, 2015.

[121] Alhassan Mumuni and Fuseini Mumuni. Automated data processing and feature engineering for deep learning and big data applications: a survey. *Journal of Information and Intelligence*, 2024.

[122] Loris Nanni and Alessandra Lumini. An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. *Expert systems with applications*, 36(2):3028–3033, 2009.

[123] BagherAsgarnezhad Nouri and Milad Soltani. Designing a bankruptcy prediction model based on account, market and macroeconomic variables (case study: Cyprus stock exchange). *Iranian Journal of Management Studies,Iranian Journal of Management Studies*, Jan 2016.

[124] Daniel Němec and Michal Pavlík. Predicting insolvency risk of the czech companies. Jan 2016.

[125] Advocacy Office. 2023 small business profiles for the states, territories, and nation. https://advocacy.sba.gov/wp-content/uploads/2023/11/2023-Small-Business-Economic-Profile-US.pdf, 2023. Accessed: 2024-12-08.

[126] James A Ohlson. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of accounting research*, pages 109–131, 1980.

[127] Diego Olaya, Kristof Coussement, and Wouter Verbeke. A survey and benchmarking study of multitreatment uplift modeling. *Data Mining and Knowledge Discovery*, 34:273–308, 2020.

[128] Diego Olaya, Jonathan Vásquez, Sebastián Maldonado, Jaime Miranda, and Wouter Verbeke. Uplift modeling for preventing student dropout in higher education. *Decision support systems*, 134:113320, 2020.

[129] David L Olson, Dursun Delen, and Yanyan Meng. Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support Systems*, 52(2):464–473, 2012.

[130] Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. Finetuning or retrieval? comparing knowledge injection in llms. *arXiv preprint arXiv:2312.05934*, 2023.

[131] Sonali Parbhoo, Stefan Bauer, and Patrick Schwab. Ncore: Neural counterfactual representation learning for combinations of treatments. *arXiv preprint arXiv:2103.11175*, 2021.

[132] Ivica Pervan and Tamara Kuvek. The relative importance of financial ratios and nonfinancial variables in predicting of insolvency. *Croatian Operational Research Review,Croatian Operational Research Review*, Feb 2013.

[133] Anastasios Petropoulos, Vasilis Siakoulis, Evangelos Stavroulakis, and NikolaosE. Vlachogiannakis. Predicting bank insolvencies using machine learning techniques. *International Journal of Forecasting*, 36(3):1092–1113, Jul 2020.

[134] Harlan D Platt and Marjorie B Platt. Understanding differences between financial distress and bankruptcy. *Review of Applied Economics*, 2(1076-2016-87135):141–157, 2006.

[135] Paul PM Pompe and AJ Feelders. Using machine learning, neural networks, and statistics to predict corporate bankruptcy. *Computer-Aided Civil and Infrastructure Engineering*, 12(4):267–276, 1997.

[136] B Prasetiyo, MA Muslim, N Baroroh, et al. Artificial neural network model for banckrupty prediction. In *Journal of Physics: Conference Series*, volume 1567, page 032022. IOP Publishing, 2020.

[137] Aneta Ptak-Chmielewska. Predicting micro-enterprise failures using data mining techniques. *Journal of Risk and Financial Management*, page 30, Feb 2019.

[138] Zhaozhi Qian, Alicia Curth, and Mihaela van der Schaar. Estimating multicause treatment effects via single-cause perturbation. *Advances in Neural Information Processing Systems*, 34:23754–23767, 2021.

[139] Alexander Reisz and Claudia Perlich. A market-based framework for bankruptcy prediction. *SSRN Electronic Journal*, Jul 2005.

[140] Steffen Rendle. Factorization machines. In *2010 IEEE International conference on data mining*, pages 995–1000. IEEE, 2010.

[141] Donald B Rubin. Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, 81(396):961–962, 1986.

[142] Piotr Rzepakowski and Szymon Jaroszewicz. Decision trees for uplift modeling. In *2010 IEEE International Conference on Data Mining*, pages 441–450. IEEE, 2010.

[143] Piotr Rzepakowski and Szymon Jaroszewicz. Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems*, 32:303–327, 2012.

[144] SibusisoW. Sabela, LeonM. Brummer, JohnH. Hall, and HendrikP. Wolmarans. Using fundamental, market and macroeconomic variables to predict financial distress: A study of companies listed on the johannesburg stock exchange. *Journal of Economic and Financial Sciences*, 11(1), Apr 2018.

[145] Yuta Saito, Hayato Sakata, and Kazuhide Nakata. Cost-effective and stable policy optimization algorithm for uplift modeling with multiple treatments. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pages 406–414. SIAM, 2020.

[146] Christophe Schalck and Meryem Yankol-Schalck. Predicting french sme failures: new evidence from machine learning techniques. *Applied Economics*, 53(51):5948–5963, 2021.

[147] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pages 3076–3085. PMLR, 2017.

[148] Yin Shi and Xiaoni Li. An overview of bankruptcy prediction models for corporate firms: A systematic literature review. *Intangible Capital*, 15(2):114–127, 2019.

[149] Naeem Siddiqi. *Credit risk scorecards: developing and implementing intelligent credit scoring*, volume 3. John Wiley & Sons, 2012.

[150] Soekarso Soekarso. Hubungan analisis ratio keuangan dengan kesehatan perusahaan. *The Winners*, 10:156–163, 2009.

[151] Mark T Soliman. The use of dupont analysis by market participants. *The accounting review*, 83(3):823–853, 2008.

[152] Hwijae Son, C Hyun, Du Phan, and Hyung Ju Hwang. Data analytic approach for bankruptcy prediction. *Expert Systems with Applications*, 138:112816, 2019.

[153] Weiping Song, Chence Shi, Zhiping Xiao, Zhijian Duan, Yewen Xu, Ming Zhang, and Jian Tang. Autoint: Automatic feature interaction learning via self-attentive neural networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1161–1170, 2019.

[154] Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. Fine tuning vs. retrieval augmented generation for less popular knowledge. *arXiv preprint arXiv:2403.01432*, 2024.

[155] Lucia Svabova, Lucia Michalkova, Marek Durica, and Elvira Nica. Business failure prediction for slovak small and medium-sized companies. *Sustainability*, page 4572, Jun 2020.

[156] Fernando Sánchez-Lasheras, Javier de Andrés, Pedro Lorca, and Francisco Javier de Cos Juez. A hybrid device for the solution of sampling bias problems in the forecasting of firms' bankruptcy. *Expert Systems with Applications*, page 7512–7523, Jun 2012.

[157] Seçil Bal Taştan and Seyed Mehdi Mousavi Davoudi. An empirical research on the examination of the relationship between perceived workplace incivility and job involvement: the moderating role of collaborative climate. *International Journal of Work Organisation and Emotion*, 7(1):35–62, 2015.

[158] Laine E Thomas, Siyun Yang, Daniel Wojdyla, and Douglas E Schaubel. Matching with time-dependent treatments: a review and look forward. *Statistics in medicine*, 39(17):2350–2370, 2020.

[159] Mario Hernandez Tinoco and Nick Wilson. Financial distress and bankruptcy prediction among listed companies using accounting, market and macroeconomic variables. *International review of financial analysis*, 30:394–419, 2013.

[160] Ellen Tobback, Tony Bellotti, Julie Moeyersoms, Marija Stankova, and David Martens. Bankruptcy prediction for smes using relational data. *Decision Support Systems*, 102:69–81, 2017.

[161] Sebastian Tomczak. Polish companies bankruptcy data. UCI Machine Learning Repository, 2016. DOI: https://doi.org/10.24432/C5F600.

[162] Chih-Fong Tsai and Jhen-Wei Wu. Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert systems with applications*, 34(4):2639–2649, 2008.

[163] János Varga. Defining the economic role and benefits of micro small and medium-sized enterprises in the 21st century with a systematic review of the literature. *Acta Polytechnica Hungarica*, 18(11):209–228, 2021.

[164] Gang Wang, Jinxing Hao, Jian Ma, and Hongbing Jiang. A comparative assessment of ensemble learning for credit scoring. *Expert systems with applications*, 38(1):223–230, 2011.

[165] Gang Wang, Jian Ma, and Shanlin Yang. An improved boosting based on feature selection for corporate bankruptcy prediction. 2014.

[166] Xinlin Wang and Mats Brorsson. Augmenting bankruptcy prediction using reported behavior of corporate restructuring. In *BenchCouncil International Symposium on Intelligent Computers, Algorithms, and Applications*, pages 102–121. Springer, 2023.

[167] Xinlin Wang and Mats Brorsson. Which company adjustment matter? insights from uplift model- ing on financial health. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2024.

[168] Xinlin Wang and Mats Brorsson. Can large language model analyze financial statements well. ACL, 2025.

[169] Xinlin Wang, Zsófia Kräussl, and Mats Brorsson. Datasets for advanced bankruptcy prediction: A survey and taxonomy. *arXiv preprint arXiv:2411.01928*, 2024.

[170] Xinlin Wang, Zsófia Kräussl, Maciej Zurad, and Mats Brorsson. Effective automatic feature engineering on financial statements for bankruptcy prediction. In *2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, pages 1–8. IEEE, 2023.

[171] Xinlin Wang, Zsófia Kräussl, and Mats Zurad, Maciej Brorsson. Effective automatic feature engineering on financial statements for bankruptcy prediction. 2022. to be published.

[172] Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International conference on learning representations*, 2018.

[173] Umair Bin Yousaf, Khalil Jebran, and Irfan Ullah. Corporate governance and financial distress: A review of the theoretical and empirical literature. *International Journal of Finance & Economics*, 29(2):1627–1679, 2024.

[174] Qi Yu, Yoan Miche, Eric Séverin, and Amaury Lendasse. Bankruptcy prediction using extreme learning machine and financial expertise. *Neurocomputing*, 128:296–302, 2014.

[175] Yuri Zelenkov, Elena Fedorova, and Dmitry Chekrizov. Two-step classification method based on genetic algorithm for bankruptcy forecasting. *Expert Systems with Applications*, 88:393–401, 2017.

[176] Yuri Zelenkov and Nikita Volodarskiy. Bankruptcy prediction on the base of the unbalanced data using multi-objective selection of classifiers. *Expert Systems with Applications*, 185:115559, Dec 2021.

[177] Chong Zhang, Xinyi Liu, Mingyu Jin, Zhongmou Zhang, Lingyao Li, Zhengting Wang, Wenyue Hua, Dong Shu, Suiyuan Zhu, Xiaobo Jin, et al. When ai meets finance (stockagent): Large language model-based stock trading in simulated real-world environments. *arXiv preprint arXiv:2407.18957*, 2024.

[178] Liangwei Zhang, Jing Lin, and Ramin Karim. Sliding window-based fault detection from high-dimensional data streams. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 47(2):289–303, 2016.

[179] Weijia Zhang, Jiuyong Li, and Lin Liu. A unified survey of treatment effect heterogeneity modelling and uplift modelling. *ACM Computing Surveys (CSUR)*, 54(8):1–36, 2021.

[180] Yan Zhao, Xiao Fang, and David Simchi-Levi. Uplift modeling with multiple treatments and general response types. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 588–596. SIAM, 2017.

[181] Yilun Zhao, Yitao Long, Hongjun Liu, Linyong Nan, Lyuhao Chen, Ryo Kamoi, Yixin Liu, Xiangru Tang, Rui Zhang, and Arman Cohan. Docmath-eval: Evaluating numerical reasoning capabilities of llms in understanding long documents with tabular data. *arXiv preprint arXiv:2311.09805*, 2023.

[182] Yu Zhao, Shaopeng Wei, Yu Guo, Qing Yang, Xingyan Chen, Qing Li, Fuzhen Zhuang, Ji Liu, and Gang Kou. Combining intra-risk and contagion risk for enterprise bankruptcy prediction using graph neural networks. Feb 2022.

[183] Zhenyu Zhao and Totte Harinen. Uplift modeling for multiple treatments with cost optimization. In *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 422–431. IEEE, 2019.

[184] Yizhen Zheng, Vincent C. S. Lee, Zonghan Wu, and Shirui Pan. *Heterogeneous Graph Attention Network for Small and Medium-Sized Enterprises Bankruptcy Prediction*, page 140–151. Jan 2021.

[185] Ligang Zhou. Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods. *Knowledge-Based Systems*, page 16–25, Mar 2013.

[186] Xiuqin Zhu, Lian-Xin Jiang, Yixin Gao, and Youbing Yin. 1. research on financial statement analysis methods based on machine learning. 2023.

[187] You Zhu, Li Zhou, Chi Xie, Gang-Jin Wang, and Truong V Nguyen. Forecasting smes' credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach. *International Journal of Production Economics*, 211:22–33, 2019.

[188] Maciej Zięba, Sebastian K Tomczak, and Jakub M Tomczak. Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. *Expert systems with applications*, 58:93–101, 2016.

[189] Róbert Štefko, Jarmila Horváthová, and Martina Mokrišová. Bankruptcy prediction with the use of data envelopment analysis: An empirical study of slovak businesses. *Journal of Risk and Financial Management*, page 212, Sep 2020.

[190] Róbert Štefko, Jarmila Horváthová, and Martina Mokrišová. The application of graphic methods and the dea in predicting the risk of bankruptcy. *Journal of Risk and Financial Management*, page 220, May 2021.