



**HAL**  
open science

# An EXplainable Artificial Intelligence Credit Rating System

Ayoub El Qadi El Haouari

► **To cite this version:**

Ayoub El Qadi El Haouari. An EXplainable Artificial Intelligence Credit Rating System. Artificial Intelligence [cs.AI]. Sorbonne Université, 2023. English. NNT : 2023SORUS486 . tel-04472510

**HAL Id: tel-04472510**

**<https://theses.hal.science/tel-04472510v1>**

Submitted on 22 Feb 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



**SORBONNE  
UNIVERSITÉ**



École d'ingénieurs du numérique

## THÈSE

en vue de l'obtention du grade de

**Docteur de Sorbonne Université**  
délivré par Sorbonne Université

Discipline : **Informatique**

Laboratoire d'Informatique, Signal, Image, Télécommunication et Électronique  
École Doctorale ED130 Informatique, Télécommunications et Electronique

par **Ayoub El Qadi El Haouari**

---

# An EXplainable Artificial Intelligence Credit Rating System

---

Directrice de Thèse: Prof. Maria TROCAN

Devant la commission d'examen formée de:

M.	Dariusz KROL	Rapporteur
M.	Nistor GROZAVU	Rapporteur
M.	Dan ISTRATE	Examineur
M.	Christophe MARSALA	Examineur
M.	Thomas FROSSARD	Directeur Lab Innovation Tinubu
M.	Martin SHEPPARD	Directeur Risk Tinubu
Mme.	Natalia DÍAZ-RODRÍGUEZ	Co-encadrante
Mme.	Maria TROCAN	Directrice

Laboratoire d'Informatique, Signal,  
Image, Télécommunication et  
Électronique 10, rue de Vanves  
92130 Issy-les-Moulineaux

Ecole Doctorale ED130 Informatique,  
Télécommunications et Electronique  
4 Place Jussieu  
75005 Paris

# Abstract

Over the past few years, the trade finance gap has surged to an alarming \$1.5 trillion, underscoring a growing crisis in global commerce. This gap is particularly detrimental to small and medium-sized enterprises (SMEs), which often find it difficult to access trade finance. Traditional credit scoring systems, which are the backbone of trade finance, are not always tailored to assess the creditworthiness of SMEs adequately.

The term credit scoring stands for the methods and techniques used to evaluate the creditworthiness of individuals or business. The score generated is then used by financial institutions to make decisions on loan approvals, interest rates, and credit limits. Credit scoring present several characteristics that makes it a challenging task. First, the lack of explainability in complex machine learning models often results in less acceptance of credit assessments, particularly among stakeholders who require transparent decision-making process. This opacity can be an obstacle in the widespread adoption of advanced scoring techniques. Another significant challenge is the variability in data availability across countries and the often incomplete financial records of SME's which makes it difficult to develop universally applicable models.

In this thesis, we initially tackled the issue of explainability by employing state-of-the-art techniques in Explainable Artificial Intelligence (XAI). We introduced a novel strategy that involved comparing the explanations generated by machine learning models with the criteria used by credit experts. This comparative analysis revealed a divergence between the model's reasoning and the expert's judgment, underscoring the necessity of incorporating expert criteria into the training phase of the model. The findings suggest that aligning machine-generated explanations with human expertise could be a pivotal step in enhancing the model's acceptance and trustworthiness.

Subsequently, we shifted our focus to address the challenge of sparse or incomplete financial data. We incorporated textual credit assessments into the credit scoring model using cutting-edge Natural Language Processing (NLP) techniques. Our results demonstrated that models trained with both financial data and textual credit assessments outperformed those relying solely on financial data. Moreover, we showed that our approach could effectively generate credit scores using only textual risk assessments, thereby offering a viable solution for scenarios where traditional financial metrics are unavailable or insufficient.

---

## Résumé

Au cours des dernières années, le déficit de financement du commerce a atteint le chiffre alarmant de 1 500 milliards de dollars, soulignant une crise croissante dans le commerce mondial. Ce déficit est particulièrement préjudiciable aux petites et moyennes entreprises (PME), qui éprouvent souvent des difficultés à accéder au financement du commerce. Les systèmes traditionnels d'évaluation du crédit, qui constituent l'épine dorsale du financement du commerce, ne sont pas toujours adaptés pour évaluer correctement la solvabilité des PME.

Le terme "credit scoring" désigne les méthodes et techniques utilisées pour évaluer la solvabilité des individus ou des entreprises. Le score généré est ensuite utilisé par les institutions financières pour prendre des décisions sur l'approbation des prêts, les taux d'intérêt et les limites de crédit. L'évaluation du crédit présente plusieurs caractéristiques qui en font une tâche difficile. Tout d'abord, le manque d'explicabilité des modèles complexes d'apprentissage automatique entraîne souvent une moindre acceptation des évaluations de crédit, en particulier parmi les parties prenantes qui exigent un processus décisionnel transparent. Cette opacité peut constituer un obstacle à l'adoption généralisée de techniques d'évaluation avancées. Un autre défi important est la variabilité de la disponibilité des données entre les pays et les dossiers financiers souvent incomplets des PME, ce qui rend difficile le développement de modèles universellement applicables.

Dans cette thèse, nous avons d'abord abordé la question de l'explicabilité en utilisant des techniques de pointe dans le domaine de l'intelligence artificielle explicable (XAI). Nous avons introduit une nouvelle stratégie consistant à comparer les explications générées par les modèles d'apprentissage automatique avec les critères utilisés par les experts en crédit. Cette analyse comparative a révélé une divergence entre le raisonnement du modèle et le jugement de l'expert, soulignant la nécessité d'incorporer les critères de l'expert dans la phase de formation du modèle. Les résultats suggèrent que l'alignement des explications générées par la machine sur l'expertise humaine pourrait être une étape cruciale dans l'amélioration de l'acceptation et de la fiabilité du modèle.

Par la suite, nous nous sommes concentrés sur le défi que représentent les données financières éparpillées ou incomplètes. Nous avons incorporé des évaluations de crédit textuelles dans le modèle d'évaluation du crédit en utilisant des techniques de pointe de traitement du langage naturel (NLP). Nos résultats ont démontré que les modèles formés à la fois avec des données financières et des évaluations de crédit textuelles étaient plus performants que ceux qui s'appuyaient uniquement sur des données financières. En outre, nous avons montré que notre approche pouvait effectivement générer des scores de crédit en utilisant uniquement des évaluations de risque textuelles, offrant ainsi une solution viable pour les scénarios dans lesquels les mesures financières traditionnelles ne sont pas disponibles ou insuffisantes.

# Remerciements

Tout d'abord, je souhaite exprimer ma profonde gratitude envers ma directrice de thèse, Prof. Maria Trocan, ma superviseuse, Dr. Natalia Diaz-Rodriguez. Leur soutien constant, leur encouragement et leur vaste expertise ont été des éléments clés tout au long de mon aventure doctorale. Les conseils judicieux de Prof. Trocan et l'orientation de Dr. Diaz-Rodriguez m'ont été d'un grand secours durant les étapes de recherche et de composition de ma thèse. Travailler sous la direction de deux experts a été une expérience exigeante : il m'a fallu naviguer et évaluer une variété de propositions et d'approches. Néanmoins, cette collaboration m'a été extrêmement bénéfique : elle m'a offert l'opportunité d'apprendre de différentes personnalités et de me familiariser avec diverses méthodes de travail et de recherche. Je tiens également à remercier Patricia Conde-Cespedes pour son expertise dans le domaine, qui a enrichi significativement mon travail de recherche.

Je tiens à exprimer ma sincère gratitude à Tinubu pour m'avoir offert la chance de me consacrer à un sujet qui non seulement a un impact socioéconomique significatif, mais qui constitue également un challenge stimulant en matière de recherche de solutions. Plus spécifiquement, mes remerciements vont à Thomas Frossard, le directeur du laboratoire où j'ai réalisé mon doctorat. Sa constante motivation, son aptitude à suggérer des idées novatrices et son soutien inébranlable ont été des piliers essentiels de mon parcours académique. *Paper à Paper.*

Enfin, et c'est le plus important, je souhaite adresser une mention spéciale à ceux qui constituent le socle de ma vie et la source d'énergie qui m'a permis de continuer à croire en moi-même dans les moments les plus difficiles : ma famille.

À mon frère Ayman, qui a toujours été là, dans les bons comme dans les mauvais moments. À ma sœur Miriam, pour son soutien sans faille. Je suis fier de vous avoir comme frère et sœur. Pour mon père, une éternité ne suffirait pas pour exprimer toute ma gratitude pour ce qu'il a fait pour moi et ce qu'il m'a appris. Si je devais choisir une chose à retenir, ce serait les valeurs d'effort et de travail qu'il m'a inculquées. Et enfin, le plus difficile, à ma mère, qui restera toujours dans mon cœur. Son soutien inconditionnel et sa foi inébranlable en moi sont des forces sans lesquelles je ne serais pas la personne que je suis aujourd'hui. Je t'aime.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Credit Scoring . . . . .	1
1.2	Individual Consumer Credit Scoring . . . . .	4
1.3	Company Credit Scoring . . . . .	5
1.3.1	Listed Companies . . . . .	5
1.3.2	Non-listed companies . . . . .	6
1.4	Principal Company Credit Scoring Agencies . . . . .	7
1.5	Trade finance and the importance of credit scoring . . . . .	9
1.6	Machine Learning and Deep Learning in Credit Scoring: Challenges and Impediments to Implementation . . . . .	11
1.7	Problem Statement . . . . .	12
1.7.1	Contribution and Structure of the Manuscript . . . . .	13
<b>2</b>	<b>Related Work</b>	<b>17</b>
2.1	Machine Learning Terminology and Specifics of Credit Scoring . . . . .	17
2.2	Standard models and its application in credit scoring . . . . .	21
2.3	Ensemble models in credit scoring . . . . .	25
2.4	eXplainable Artificial Intelligence . . . . .	29
2.5	Natural Language Processing . . . . .	31
<b>3</b>	<b>Analyzing the Impact of the COVID-19 Outbreak on Companies Default Rates</b>	<b>33</b>
3.1	Introduction . . . . .	33
3.2	Literature Review . . . . .	34
3.3	Methodology . . . . .	35
3.3.1	Data Processing . . . . .	35
3.3.2	Sector Default Evolution . . . . .	37
3.3.3	Analyzing the Default of Companies by Sector for the Period 2008-2022 . . . . .	37
3.4	Results . . . . .	37
3.4.1	Sector Default Evolution . . . . .	37
3.4.2	Evolution of Companies by Sector for the Period 2008-2022 . . . . .	38
3.5	Discussion . . . . .	38
3.6	Conclusions . . . . .	39
<b>4</b>	<b>Aligning Feature Contributions with Expert Knowledge in Artificial Intelligence-Based Credit Scoring</b>	<b>41</b>



4.1	Introduction . . . . .	41
4.2	Related Work: EXplainable AI for Credit Scoring . . . . .	43
4.2.1	Machine Learning for Credit Risk Scoring . . . . .	43
4.2.2	EXplainable Artificial Intelligence (XAI) in Finance . . . . .	44
4.3	Methodology: Black Box Models and XAI . . . . .	44
4.3.1	Data Preprocessing Pipeline for Company Credit Risk Scoring . . . . .	45
4.3.2	PD Modeling using Machine Learning Models . . . . .	48
4.3.3	Data Oversampling using SMOTE . . . . .	49
4.3.4	SHAP for Model Explanations . . . . .	50
4.3.5	Human Expertise Alignment: Introducing Credit Risk Analysts Expertise . . . . .	50
4.4	Results . . . . .	50
4.4.1	Analyzing the Impact of Interpolation in the Credit Score . . . . .	50
4.4.2	Performance of ML Models for PD Modeling . . . . .	51
4.4.3	Mapping XGBoost Probabilities to Tinubu’s Grades . . . . .	52
4.4.4	Explaining our PD model: SHAP Value Analysis . . . . .	53
4.4.5	Feature Contribution Analysis: Assessing Explanations from risk analyst experts vs ML models . . . . .	55
4.5	Discussion . . . . .	56
4.6	Conclusion and Future work . . . . .	57
<b>5</b>	<b>Sectorial Analysis Impact on the Development of Credit Scoring Machine Learning Models</b>	<b>61</b>
5.1	Introduction . . . . .	61
5.2	Related Work . . . . .	62
5.3	Methodology . . . . .	63
5.3.1	Data Description . . . . .	64
5.3.2	Machine Learning Algorithms for Credit Scoring . . . . .	64
5.3.3	Explaining Model Behavior using SHAP Values . . . . .	66
5.3.4	Generating Ratings from Models Outputs . . . . .	66
5.4	Results . . . . .	67
5.4.1	Default Analysis by Sector . . . . .	68
5.4.2	Model Performance . . . . .	68
5.4.3	Shap Analysis . . . . .	70
5.4.4	Comparing the Risk Analysts risk scoring with the ML-based Credit Scoring System . . . . .	71
5.5	Discussion . . . . .	71
5.6	Conclusion . . . . .	73
<b>6</b>	<b>Credit Risk Scoring Forecasting using a Time Series Approach</b>	<b>75</b>
6.1	Introduction . . . . .	75
6.2	Related Work . . . . .	76
6.2.1	Credit Scoring . . . . .	77
6.2.2	Forecasting in the Financial Industry . . . . .	77
6.3	Methodology . . . . .	77
6.3.1	Data . . . . .	77
6.3.2	Forecasting Time Series . . . . .	78

6.3.3	Comparing the Forecasted Values . . . . .	80
6.4	Results . . . . .	80
6.4.1	Assessing the Performance of ARMA . . . . .	81
6.4.2	Forecasting Tinubu’s Score Using XGBoost . . . . .	81
6.4.3	Analyzing the Models Ratings . . . . .	83
6.5	Conclusions . . . . .	84
<b>7</b>	<b>Multimodal Credit Risk Scoring</b>	<b>87</b>
7.1	Introduction . . . . .	87
7.2	Related Work . . . . .	88
7.2.1	Credit Scoring . . . . .	88
7.2.2	Natural Language Processing . . . . .	88
7.3	Methodology . . . . .	89
7.3.1	Data Overview . . . . .	90
7.3.2	Text Feature Treatment . . . . .	91
7.3.3	Data Preprocessing . . . . .	92
7.3.4	Models . . . . .	92
7.3.5	Model Performance . . . . .	94
7.4	Results . . . . .	94
7.4.1	Exploratory Data Analysis . . . . .	94
7.4.2	Model Performance . . . . .	94
7.5	Conclusions . . . . .	97
<b>8</b>	<b>Predicting Corporate Solvency using Sentiment Analysis of Risk Analyst Textual Assessments</b>	<b>99</b>
8.1	Introduction . . . . .	99
8.2	Related Work . . . . .	100
8.2.1	Artificial Intelligence in Credit Scoring . . . . .	100
8.2.2	Natural Language Processing . . . . .	101
8.3	Methodology . . . . .	101
8.3.1	Data . . . . .	102
8.3.2	Model Architecture . . . . .	102
8.3.3	Experimental Context . . . . .	103
8.4	Results . . . . .	105
8.4.1	Fine-Tuned FinancialBERT . . . . .	105
8.4.2	Phrase-Level Sentiment Analysis . . . . .	106
8.4.3	End-of-Text Sentiment Analysis . . . . .	106
8.5	Conclusions . . . . .	108
	<b>Conclusion</b>	<b>109</b>
	<b>Bibliography</b>	<b>111</b>



# Introduction

The process of granting credit to lenders is undeniably one of the most essential business activities in the financial sector. This practice not only leads to the generation of significant profits for a multitude of entities, including banks, other financial institutions, and shareholders, but it also plays a huge role in contributing positively to the community.

While the advantages are numerous, this activity is also accompanied by substantial risks. One needs only to reflect on the recent financial crises that shook the global market to understand the magnitude of these risks. These crises led to devastating losses on a global scale, pushing banks and financial institutions into a state of heightened alertness. For example, the principal factor that led to the financial crisis of 2008 was the underestimation of mortgage risk of default [1].

This increased attention became particularly focused on the credit risk models that these institutions employed. It has become abundantly clear, especially in the wake of these financial downturns, that banks have to be more discerning than ever. They recognize the imperative need to incorporate rigorous credit evaluation models into their systems. Such prudence is essential whether they're contemplating granting a loan to a single individual or to a large corporate entity.

## 1.1 Credit Scoring

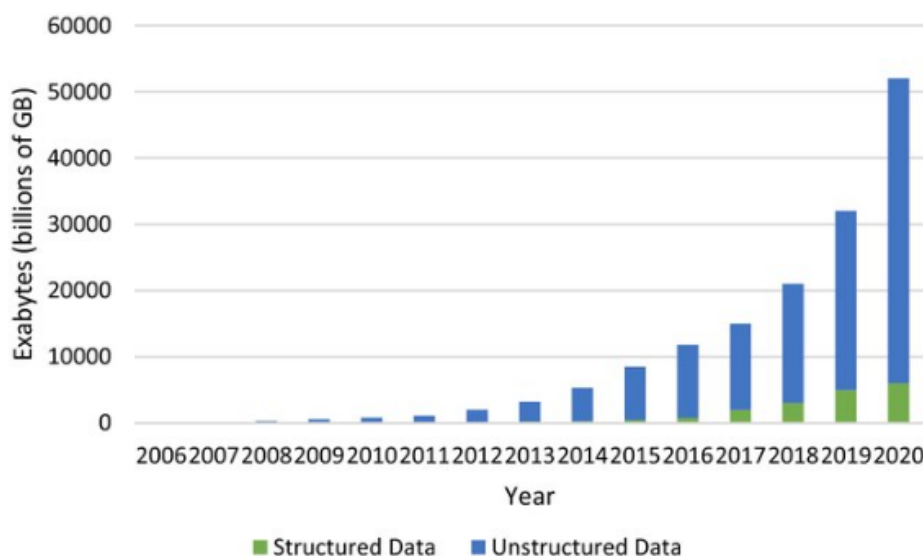
Credit scoring serves as a crucial tool in the assessment of the risk associated with lending money. It involves the use of statistical models that incorporate various factors such as the borrower's credit history, current financial standing, and other relevant socio-economic indicators. The result is a numerical score, often referred to as a credit score, which provides a quantified assessment of the likelihood of the borrower defaulting on a loan.

In an increasingly interconnected and globalized financial environment, the necessity to assess the creditworthiness of individuals and corporations has never been more paramount. Digitalization has been a driving force behind the evolution and sophistication of credit scoring, fundamentally transforming the way credit risk is assessed and managed.

Leveraging the power of big data involves utilizing cost-effective and cutting-edge processing techniques. These methods enable users to gain deeper insights, which in turn bolsters decision-making and drives process automation. With the rise of digital technologies, vast amounts of data have become accessible [2]. This spans not only traditional credit information but also extends to unconventional digital footprints like online shopping behavior and social media activities.

Structured data is akin to an organized library of information, where each piece of data adheres to a specific, predefined format. To draw a clearer picture, imagine a meticulously managed database where each entry has distinct fields like company's name, address, and net worth. This arrangement makes structured data straightforward to search, access, and manage.

Figure 1.1: Volume of structured data compared to unstructured data during the last decade ( [3])



Conversely, unstructured data is more akin to a vast ocean, teeming with a myriad of varied entities. The term "unstructured" implies that this kind of data doesn't conform to a standard format or structure. It encompasses a vast array of formats ranging from emails and images to audio recordings, videos, and even readings from various sensors. This diverse nature of unstructured data introduces layers of complexity. For example, the unstructured data utilized for credit scoring can encompass company news, consumer feedback on products, or even risk assessment documents crafted by credit risk specialists ( [4], [5]).

The landscape of credit scoring has undergone a radical transformation in recent years, primarily driven by the dual forces of the widespread availability of data and significant leaps in machine learning techniques. Traditionally, credit scoring relied on a relatively limited set of parameters, most of which were directly linked to an individual's or entity's financial history and behavior [6].

Classical models might have primarily focused on past loan payment records, current debts, income levels, and perhaps a few other factors. However, the digital revolution has brought about an era where vast quantities of data—ranging (e.g., social media activities) are easily accessible and can be harvested for insights. This data explosion, when viewed from the lens of credit scoring, offers a gold mine of nuanced information that could potentially reveal a lot more about an individual's credibility and their likelihood to repay a debt.

However, having access to a large volume of data is only one part of the equation. To transform this data into actionable insights, there's a need for sophisticated analytical tools—and this is where the advancements in machine learning come into play a huge role. Machine learning, with its ability to discern patterns from seemingly unrelated data points, can take these vast datasets and process them to extract meaningful patterns conversely to traditional credit scoring models that are able to capture linear relationships [7].

More than just processing, machine learning models can learn and adapt over time, becoming increasingly accurate as they are fed more data. This dynamic nature of machine learning means that credit scoring models can now evolve in real-time, adapting to new financial behaviors and trends as they emerge. A credit scoring system that undergoes regular retraining can adjust to evolving economic scenarios, thereby mitigating significant losses.

In the complex domain of financial lending, the methodology deployed for credit scoring is significantly influenced by the nature of the counterpart, i.e., the specific entity seeking credit.

This differentiation in counterparts mandates a diverse set of considerations, evaluation criteria, and bespoke models to accurately gauge creditworthiness. Given that each counterpart presents its own set of financial behaviors, variables, risk profiles, and borrowing histories, a comprehensive and tailored approach becomes essential for credit risk assessment. This section will systematically unpack the various credit scoring methodologies, each uniquely adapted based on the distinct characteristics of different counterparts, to elucidate the underlying complexities and specificities involved.

## 1.2 Individual Consumer Credit Scoring

Individual Consumer Credit Scoring is a system used to assess the creditworthiness of individual consumers, typically for personal financial products such as credit cards, mortgages, and loans. The important parts of individual consumer credit scoring include several factors relating to a person's management of money. The credit history is vital, as it records past payments and any failures to pay, giving an idea of past reliability. Next, the credit utilization ratio looks at existing debts compared to credit limits, showing how a person manages current debt. The time someone has had credit accounts, the various kinds of credit they have (like credit cards or loans), and any recent requests for new credit are also considered.

Figure 1.2: FICO Score description ( [8] )

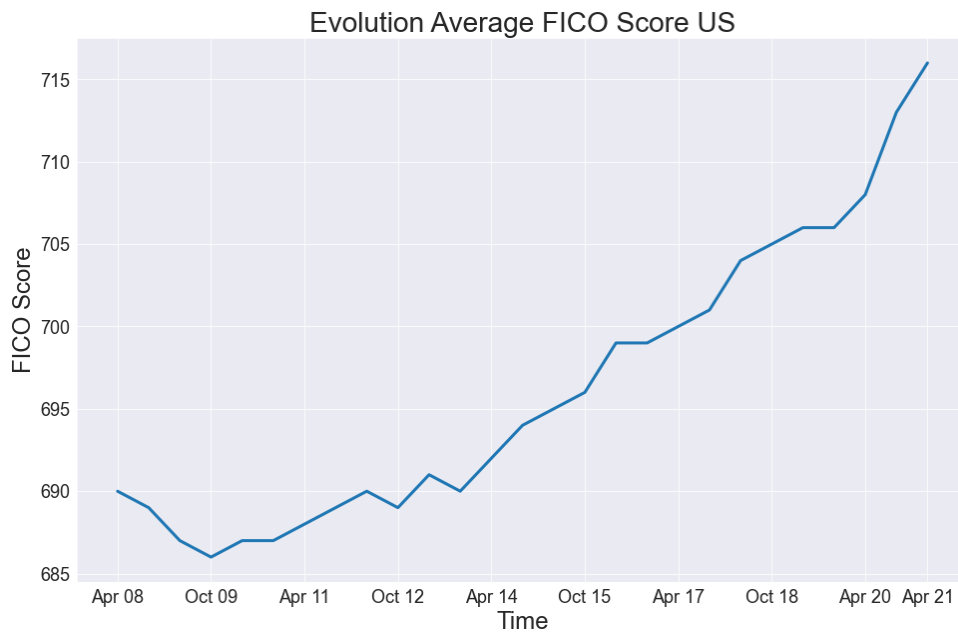
<b>FICO Score Ranges</b>	<b>Rating</b>	<b>Description</b>
<580	Poor	Your score is well below the average score of U.S. consumers and demonstrates to lenders that you are a risky borrower.
580-669	Fair	Your score is below the average score of U.S. consumers, though many lenders will approve loans with this score.
670-739	Good	Your score is near or slightly above the average of U.S. consumers and most lenders consider this a good score.
740-799	Very Good	Your score is above the average of U.S. consumers and demonstrates to lenders that you are a very dependable borrower.
800+	Exceptional	Your score is well above the average score of U.S. consumers and clearly demonstrates to lenders that you are an exceptional borrower.

For example, approximately 189 million Americans have credit scores, emphasizing the ubiquity and central role of credit scoring in the country's economic framework. Moreover, the average FICO<sup>1</sup> score, one of the most widely used credit scoring models, reached 716 in 2021, an indication of the general creditworthiness within the consumer population.

---

<sup>1</sup>A FICO Score is a three-digit number based on the information in your credit reports. It helps lenders determine how likely you are to repay a loan.

Figure 1.3: FICO Score evolution from 2008-2021 ( [9])



## 1.3 Company Credit Scoring

In the realm of corporate finance and investment, company scoring stands as a pivotal mechanism for assessing the financial health and creditworthiness of businesses. This evaluation serves as the bedrock upon which critical decisions such as loan approvals, investment allocations, and risk assessments are made ( [10], [11]). However, it's imperative to recognize that not all companies are created equal, especially when it comes to their status as either listed or non-listed entities. This distinction has far-reaching implications for how these companies are scored, the data available for analysis, and the subsequent interpretations and applications of these scores.

### 1.3.1 Listed Companies

Listed companies, commonly known as publicly traded companies, are corporations that have successfully completed an Initial Public Offering (IPO) and have their shares available for trading on stock exchanges. These companies operate under a stringent set of regulatory requirements and are obligated to disclose a wide range of financial and operational information to both the public and regulatory authorities.

One of the defining characteristics of listed companies is their high level of transparency. They are required to publish financial reports on a quarterly and annual basis, which include comprehensive financial statements such as income statements, balance sheets, and cash flow statements. This transparency is further reinforced by the regulatory oversight they are subject to, often by bodies like the Securities and Exchange



Commission (SEC) in the United States or the Autorité des marchés financiers (AMF) in France.

Another key feature is their market capitalization, which provides a readily available measure of the company's value and is calculated by multiplying the stock price by the total number of outstanding shares. The ownership of these companies is usually distributed among a diverse set of shareholders, including institutional investors, retail investors, and company insiders. Additionally, the shares of listed companies are easily bought or sold on stock exchanges, providing a high level of liquidity. Being in the public eye also means that these companies are subject to market sentiment, news coverage, and the opinions of financial analysts.

When it comes to scoring listed companies, several unique implications arise. The abundant availability of financial data allows for a more nuanced and accurate scoring process. Financial metrics such as Price-to-Earnings (P/E) ratios and Earnings Before Interest and Taxes (EBIT) are readily accessible for analysis. Stock performance and market capitalization can also serve as supplementary indicators of a company's financial health and are often incorporated into scoring models. The stringent regulatory environment and mandatory financial audits add an extra layer of credibility to the financial statements, which is usually reflected positively in the credit score.

Investor sentiment, as gauged through stock price and trading volume, can also play a significant role [12]. While positive sentiment can enhance a company's perceived creditworthiness, negative sentiment can have the opposite effect [13].

The real-time nature of stock markets means that credit scores for listed companies are dynamic and require frequent updates to reflect current market conditions. The wealth of data often allows for the use of standardized scoring models like the Altman Z-score [6] or Merton's model [14], which are widely recognized and accepted in the financial industry.

The characteristics of listed companies, such as transparency, regulatory oversight, and market-driven data, contribute to a more robust and dynamic credit scoring process. However, this also introduces additional variables like market sentiment and stock performance, requiring a more complex and nuanced approach to credit scoring.

### 1.3.2 Non-listed companies

Non-listed companies—commonly referred to as privately-held or unlisted entities constitute a distinct category of business organizations that are not traded on public stock exchanges. This absence of public trading engenders a markedly different operational landscape compared to publicly listed corporations, which are subject to rigorous regulatory oversight and disclosure mandates. The limited amount of regulatory restrictions gives non-listed companies greater freedom in their operations, allowing them to be more flexible in their strategic planning.

Ownership structures in non-listed companies are frequently characterized by a high degree of concentration, often vested in a limited cohort of stakeholders such as founding members, familial networks, or private equity consortia. This concentrated ownership paradigm expedites decision-making processes but concurrently engenders a liquidity constraint, as the shares are not readily tradable on public markets.

Unlike their publicly-listed counterparts, which can readily augment capital through the issuance of additional shares, non-listed entities predominantly rely on alternative financial instruments such as debt, venture capital, or private equity for capital infusion.

The absence of a publicly accessible trading platform further complicates the task of ascertaining an accurate market valuation, although it liberates these entities from the exigencies of quarterly financial reporting, thereby enabling a more long-term strategic focus.

The evaluation or 'scoring' of non-listed companies poses a unique set of challenges. The scarcity of publicly available financial information forces a dependence on self-reported data, which might not be as reliable or standardized as the financial statements that publicly-listed companies disclose.

This lack of transparency in financial matters, combined with minimal regulatory oversight, complicates the process of assessing risk, often requiring an in-depth due diligence approach. The shares' lack of liquidity and the generally higher perceived risk often result in an increased cost of capital, which could negatively impact the company's competitive standing and growth prospects. On the other hand, the lesser regulatory constraints provide these companies with a level of operational flexibility that can be a strategic asset, especially in fast-changing industries.

Traditional credit scoring models might not be entirely suitable for evaluating these companies, so using alternative methods like machine learning algorithms based on unconventional data sources could yield more precise and nuanced evaluations.

## 1.4 Principal Company Credit Scoring Agencies

The task of appraising a corporation's creditworthiness is an intricate and specialized endeavor, predominantly executed by agencies that are expressly devoted to the formulation of corporate credit ratings. These entities utilize a set of methodologies that are markedly divergent from those applied in the assessment of individual creditworthiness. Unlike the latter, the methodologies for corporate credit evaluation are sophisticated. These specialized agencies scrutinize an array of variables, encompassing a firm's fiscal robustness, operational efficacy, and market standing, to formulate a credit rating that cogently articulates the corporation's aptitude for fulfilling its monetary commitments.

The evaluative frameworks employed by these agencies amalgamate an extensive spec-

trum of considerations, ranging from quantitative financial metrics and business risk profiles to qualitative assessments of managerial competence and sector-specific vicissitudes. In doing so, they endeavor to furnish investors, financial institutions, and other pertinent stakeholders with a comprehensive and nuanced understanding of the corporation's credit risk. Consequently, the role of these principal agencies in the realm of corporate credit scoring is not merely instrumental but also highly specialized, necessitating a distinct arsenal of analytical methodologies and evaluative criteria vis-à-vis those employed for individual credit assessments.

Standard & Poor's (S&P), one of the preeminent agencies in the domain of corporate credit ratings, employs a multifaceted and rigorous methodology to assess a corporation's creditworthiness. At the core of this evaluative framework are four principal dimensions.

First, the Business Risk Profile is scrutinized to gauge both the inherent risks associated with the industry in which the corporation operates and its competitive positioning within that sector. This involves a nuanced analysis of market dynamics, consumer demand, and competitive advantages or disadvantages.

Second, the Financial Risk Profile is meticulously assessed, focusing on key financial ratios such as debt-to-equity and liquidity ratios, as well as the corporation's cash flow patterns and overall financial stability. This provides a quantitative foundation for the credit rating, offering insights into the firm's fiscal health and its ability to meet short-term and long-term obligations.

Third, the Country Risk dimension takes into account the macroeconomic and geopolitical landscape of the nation in which the corporation conducts its business. Factors such as economic growth rates, political stability, and regulatory frameworks are considered to understand how they might impact the company's financial standing.

Lastly, the Management and Governance aspect is evaluated to ascertain the efficacy of the company's organizational structures, leadership quality, and governance protocols. This qualitative assessment aims to shed light on the strategic direction of the company and the competence of its management team in steering the organization towards financial stability and growth.

In addition to the methodologies employed by Standard & Poor's, it's worth noting that other credit rating agencies (i.e., Moody's and Fitch) also incorporate additional factors into their evaluative frameworks. Specifically, many agencies give weight to External Influences, which encompass a broad range of macroeconomic variables, industry-specific trends, and geopolitical risks. This aspect of the methodology aims to capture the broader contextual factors that could affect a corporation's financial stability. For instance, fluctuations in commodity prices, shifts in consumer preferences, or geopolitical tensions can all have a significant impact on a company's creditworthiness.

Another critical factor considered is Legal and Regulatory Risks. This involves a thorough examination of the potential impact of changes in laws and regulations on a com-

pany's operations and, consequently, its ability to meet financial obligations. Regulatory shifts, such as changes in environmental standards or trade tariffs, can have immediate and long-term effects on a company's profitability and cash flow, thereby affecting its credit rating.

One of the contemporary challenges confronting credit rating agencies is the integration of Environmental, Social, and Governance (ESG) factors into their traditional methodologies for assessing credit risk. As societal awareness and regulatory focus on sustainability and ethical governance intensify, the imperative to incorporate ESG criteria has become increasingly salient. These factors range from a company's environmental impact and carbon footprint to its labor practices, ethical conduct, and governance structures.

The challenge lies in quantifying these often qualitative factors and seamlessly incorporating them into existing evaluative frameworks that have historically prioritized financial metrics. The objective is to offer a more holistic view of a company's risk profile, one that accounts not only for financial stability but also for long-term sustainability and ethical considerations.

This evolution in methodology is not merely a response to changing investor preferences but is also a recognition of the long-term financial risks associated with poor ESG performance. Therefore, the integration of ESG ratings is becoming a critical dimension in the comprehensive assessment of a corporation's creditworthiness, marking a significant paradigm shift in the field of corporate credit risk assessment.

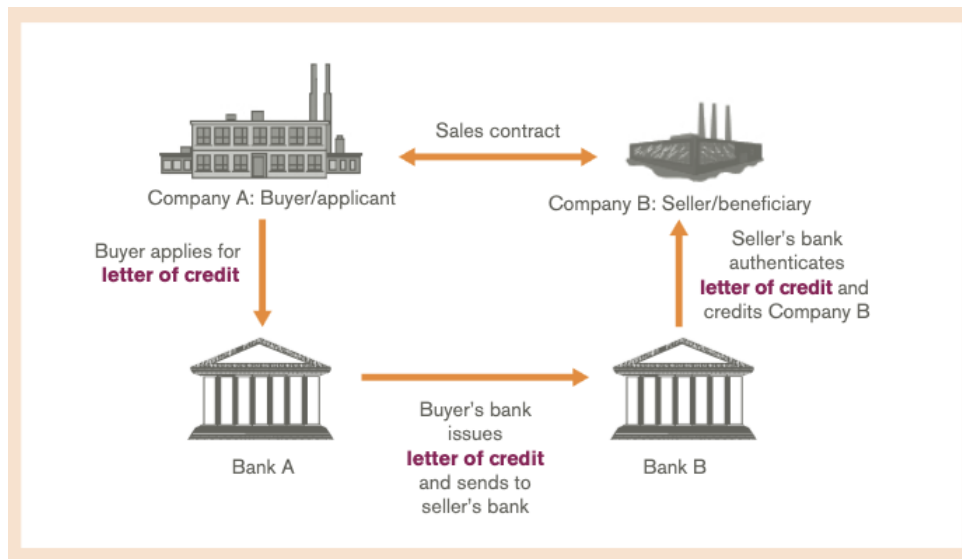
## 1.5 Trade finance and the importance of credit scoring

Trade finance frequently confronts a pronounced disparity between the demand for and supply of credit. This gap is especially acute in emerging markets and among small and medium-sized enterprises (SMEs), where access to trade finance is often constrained by various factors such as regulatory hurdles, lack of financial infrastructure, and perceived high risks from lenders perspectives [15].

Financial institutions may be reticent to extend credit due to a lack of reliable information on the creditworthiness of trading entities, particularly SMEs, which may not have an extensive credit history or may operate in jurisdictions that are perceived to be riskier.

Additionally, the complexities of cross-border transactions, including currency exchange risks and differing legal frameworks, further exacerbate the challenge of meeting the burgeoning demand for trade finance. This situation creates a self-perpetuating cycle where the absence of readily available trade finance restricts the ability of companies, especially SMEs, to engage in international trade, thereby limiting their growth prospects

Figure 1.4: Simplified structure of a credit in trade finance. Source: WTO [15]



and, in turn, making them less attractive candidates for trade finance.

As a concrete illustration of the challenges previously discussed, consider the plight of small and medium-sized enterprises (SMEs) in developing nations, where the obstacles to obtaining trade finance are particularly pronounced. In Africa, the estimated shortfall in trade finance is a staggering US\$ 120 billion, which accounts for approximately one-third of the entire trade finance market across the continent. The situation is even more acute in developing regions of Asia, where the unmet demand for trade finance soars to an astronomical US\$ 700 billion. These figures not only underscore the magnitude of the trade finance gap but also highlight the acute challenges faced by SMEs in these regions in accessing the financial resources needed to engage in international trade.

On the other hand, In certain large developed countries, as many as one-third of small and medium-sized enterprises (SMEs) grapple with difficulties in securing trade finance. To put this into perspective, SMEs are responsible for a significant portion of exports from these developed economies; they account for 20% of all U.S. exports and an even higher 40% of exports from the European Union. These statistics emphasize that the issue of limited access to trade finance is not confined to developing regions but is also a pertinent concern in developed economies, affecting a substantial number of SMEs that are key contributors to international trade.

## 1.6 Machine Learning and Deep Learning in Credit Scoring: Challenges and Impediments to Implementation

Certainly, the advent of machine learning and deep learning technologies has undeniably revolutionized the field of credit scoring. However, the industry is still grappling with several challenges that require attention for further progress.

Limited access to financial data is a significant hurdle. Data privacy concerns have made many individuals reluctant to share their personal financial information, thereby limiting the amount of data available for building robust credit scoring models. Additionally, especially for new borrowers or those without a traditional credit history, the available financial data is often sparse, making it difficult to assess creditworthiness accurately.

Financial data is also often stored in isolated databases across different institutions, complicating the task of obtaining a comprehensive view of an individual's financial health. For international borrowers, the absence of a standardized global system for financial data can further complicate credit scoring.

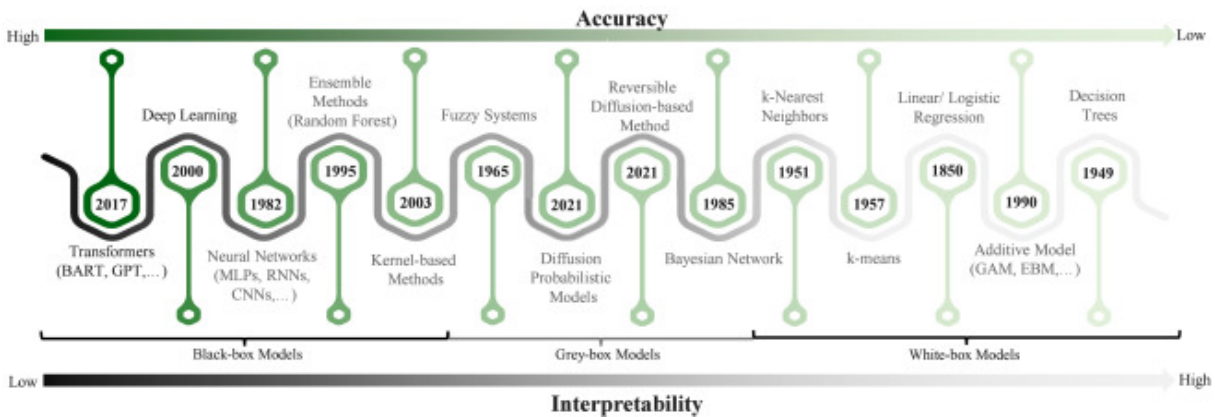
Another challenge lies in the inability of classical models to capture complex patterns in financial data. Traditional models like logistic regression or linear models are often too simplistic to capture the intricate, non-linear relationships that exist.

While machine learning models are capable of capturing these complex patterns, they are also prone to overfitting, especially when the available data is limited or imbalanced. Classical models also often rely on a limited set of features, ignoring potentially important variables that could improve the model's predictive power.

Strong regulation and the need for model transparency also pose challenges. Financial institutions are subject to stringent regulations, such as the General Data Protection Regulation (GDPR) in Europe. GDPR imposes strict regulations on the collection, storage, and processing of personal data, including financial information used in credit scoring. This has led to heightened requirements for transparency and explainability in credit scoring algorithms. Financial institutions are now obligated to provide clear justifications for their credit decisions, which poses a challenge for complex machine learning and deep learning models [16].

Advanced machine learning and deep learning models are often criticized for being "black boxes," making it difficult to interpret their decisions and thereby posing a challenge in meeting regulatory requirements for transparency. There is also an increasing focus on ethical AI, demanding that credit scoring models be free from biases related to gender, race, or socio-economic status. Complying with these ethical standards while maintaining high accuracy is a significant challenge. Moreover, different countries have different regulations concerning credit scoring, making it challenging for financial institu-

Figure 1.5: Illustration of the balance between accuracy and interpretability. Source: [16]



tions that operate internationally to maintain a consistent and compliant approach.

## 1.7 Problem Statement

In previous sections, we have delved into the principal real-world stakes associated with credit scoring, highlighting its pivotal role in various economic and financial contexts. We mentioned the biggest challenges of implementing machine-learning based credit scoring models.

This thesis aims to tackle several specific challenges that were encountered by Tinubu<sup>2</sup> when developing a credit scoring system. These challenges serve as real-world examples that highlight the complexities involved in implementing machine learning models for credit risk assessment.

One of the most volatile factors affecting credit scoring is the pace of economic changes. Economic conditions can shift rapidly, making previously reliable models obsolete or less accurate. This necessitates frequent updates and recalibrations to the credit scoring models to keep them relevant and reliable.

Another challenge is the lack of comprehensive financial features. In many cases, essential financial data may be missing or incomplete, making it difficult to assess creditworthiness accurately. This issue of missing values can severely impact the model's performance and may require sophisticated techniques for imputation or feature engineering.

In the field of credit scoring, it's common to encounter imbalanced datasets where the ratio of defaulted to non-defaulted cases is highly skewed. In Tinubu's experience, the default to non-default ratio was around approximately 1%, making it a significant challenge to train a model that can effectively identify the minority class of defaulted cases without being overwhelmed by the majority class leading the model to overfit.

<sup>2</sup>Tinubu is a company specialized in credit scoring assessments

Lastly, one of the critical issues is that the credit scoring models often yield credit ratings without providing any explanations or justifications. This lack of interpretability can be a significant drawback, especially when financial institutions or borrowers seek to understand the factors contributing to a particular credit score.

### 1.7.1 Contribution and Structure of the Manuscript

In this thesis, we present two main contributions to the field of credit scoring, particularly focusing on the challenges and opportunities arising from the application of machine learning techniques. These contributions are structured across six different chapters, each designed to delve into specific aspects of the problem, methodologies, and solutions. The chapters are organized in a way that provides a comprehensive understanding of the current landscape, the challenges witnessed by industry experts (i.e., Tinubu), and the approaches we propose to address these issues.

In the first chapter of this thesis, we lay the groundwork by introducing the field of machine learning, which serves as the cornerstone for the entire manuscript. We delve into the essential concepts, terminologies, and methodologies that are fundamental to understanding the subsequent chapters.

In the second chapter of this thesis, we examine the impact of the COVID-19 crisis on the default rates of companies, with a specific focus on those based in France and operating in various economic sectors. Through a comprehensive analysis, we examine how the pandemic has influenced the financial stability of businesses, leading to changes in their default percentages. This investigation not only provides a nuanced understanding of the crisis's economic repercussions but also serves as a foundational study for developing more resilient and adaptive credit scoring models that can better handle such unprecedented events.

#### Development of an explainable credit scoring model

In the third chapter of this thesis, we focus on a comparative analysis between Tinubu's existing automatic credit rating system and a machine learning-based model that we develop. Both systems aim to assess the risk that a borrower will default, but they do so in different ways. Tinubu's current system provides a qualitative output, offering a risk assessment in a more descriptive manner. In contrast, our machine learning-based model uses the same input variables but aims to provide both a quantitative risk assessment and explanations for its predictions. One of the key aspects of this chapter is the comparison of explanations generated by the machine learning model with the criteria used by risk analysts. By doing so, we aim to identify any divergences between the machine learning model's rationale and the human experts' judgments. This comparative approach serves to validate the effectiveness and interpretability of the machine learning model while also highlighting areas for potential improvement or alignment with human expertise.



In the fourth chapter, we take a different approach compared to the previous chapters by expanding our focus to multiple industries, each with its own unique set of challenges and characteristics. Unlike the previous chapter, which concentrated on a limited set of financial features, this chapter explores a broader feature space that encompasses variables relevant to various industries. One of the key findings in this chapter is the divergence in terms of explanations across different industries. We analyze how the machine learning model's rationale for credit risk assessment varies depending on the industry in question, providing valuable insights into the adaptability and limitations of the model in diverse settings. Additionally, we demonstrate the positive impact of this broader approach on model performance. By incorporating a more extensive set of features and considering industry-specific nuances, we show that the machine learning model's predictive accuracy improves, thereby confirming the utility of a multi-industry approach in credit scoring.

In the fifth chapter, we shift our focus from predicting the likelihood of a company going into default to predicting the future credit rating of the company. This represents a significant departure from the objectives of the previous chapters. Here, we treat each company as a time series, utilizing their historical ratings as input features to forecast future ratings. This time-series approach is particularly effective for companies that have a long track record of credit ratings, as it allows the model to capture temporal patterns and trends that may not be evident in a cross-sectional analysis. By focusing on predicting future ratings rather than default likelihood, we aim to provide a more nuanced understanding of a company's financial health over time. The results indicate that this time-series approach is highly effective for a subset of companies with an extensive history of ratings, thereby offering a new dimension to credit risk assessment that complements the more traditional methods explored in earlier chapters.

### **Multimodal credit risk scoring**

In the sixth and final chapter, we venture into the realm of incorporating textual data into the development of credit scoring models, with the ambitious goal of potentially substituting traditional financial features with textual information. Initially, we demonstrate that by adding human-generated credit reports to the existing set of financial features, the model's capability to accurately identify companies that will go into default is significantly enhanced. This suggests that textual data, often rich in qualitative insights, adds a layer of depth to the model that numerical financial features alone may not capture.

Furthermore, we make a groundbreaking revelation: it is possible to identify highly risky companies, those with poor ratings according to Tinubu's expertise, without relying on financial features at all. By solely utilizing textual data, our model is still able to flag companies that are at high risk of default or financial instability. This finding has profound implications for the field of credit scoring, especially in scenarios where financial data may be incomplete, unreliable, or unavailable. It opens up new avenues for research and application, underscoring the untapped potential of textual information in enhancing the robustness and versatility of credit scoring models.

We conclude the manuscript with a general summary that encapsulates the key findings, contributions, and implications of the research conducted. In addition to summarizing the work, we also present open questions and future perspectives that arise from our research.



## Related Work

In this particular chapter, we take a comprehensive approach to formally define the problem of credit scoring through mathematical formulations. Our aim is to provide a rigorous foundation that sets the stage for the empirical work that follows. Alongside this, we also delve into an extensive review of the existing literature and the current state of the art in several interconnected disciplines. These include not only credit scoring but also machine learning, explainable artificial intelligence, and natural language processing. This multi-disciplinary review serves to contextualize our work within the larger academic and industry landscapes.

### 2.1 Machine Learning Terminology and Specifics of Credit Scoring

When defining a credit scoring model, there are two distinct approaches based on different underlying philosophies: Point-in-Time (PIT) and Through-the-Cycle (TTC) [17]. The PIT approach focuses on assessing the credit risk of a borrower at a specific moment in time, taking into account both cyclical and structural factors. This means that the PIT model is sensitive to current economic conditions and can fluctuate based on short-term changes in the borrower's financial situation or the broader economy. On the other hand, the TTC approach aims to evaluate the credit risk of a borrower over an entire economic cycle. It smooths out short-term fluctuations and focuses on the borrower's long-term ability to meet financial obligations, irrespective of current economic conditions. While PIT models are generally more accurate for immediate risk assessment and decision-making, TTC models are often used for strategic planning and capital allocation, as they provide a more stable and long-term view of credit risk.

In light of these considerations, our work will primarily focus on Point-in-Time (PIT) approaches to credit scoring. This decision is motivated by the fact that companies

typically seek short-term credits, making the immediate and current assessment of credit risk particularly relevant. The PIT model's sensitivity to current economic conditions and short-term changes in a company's financial status makes it a more suitable framework for evaluating the types of credit commonly demanded by businesses

In this section, the focus shifts to a mathematical formalization of machine learning and deep learning tasks as they pertain to credit scoring. The aim of this section is to give a thorough but easy-to-understand foundation on how to define the credit scoring problem and to translate the business problem into a machine-learning task.

The issue of credit scoring can be rigorously defined in mathematical terms, typically falling under the category of supervised learning. In the realm of machine learning, supervised learning tasks are those where the model is trained on a dataset that includes both the input variables and the corresponding output labels. More specifically, credit scoring is most commonly approached as a classification problem, although there are instances where it can also be treated as a regression problem.

In a classification framework, the aim is to categorize borrowers into distinct classes, such as 'low risk' or 'high risk,' based on their financial attributes. On the other hand, in a regression framework, the objective might be to predict a continuous outcome, such as the probability of default. To encapsulate these concepts and provide a structured approach to the problem, a simplified mathematical model is often employed.

### Credit Scoring formalization

In the context of this thesis, it is imperative to establish a clear and consistent set of variable definitions that will be employed throughout the entire work. These variables serve as the foundational elements for constructing and evaluating machine learning models aimed at predicting credit risk.

- $\mathcal{X}$ : This is known as the feature space. It represents all the possible financial characteristics or attributes that company might have. Examples of these features include net worth, country in which the company operates, and sales. Each company will have a unique combination of these features, which will be used to assess their creditworthiness.
- $\mathcal{Y}$ : This is the label space, which is usually a binary set  $\{0, 1\}$ . Since the objective is to predict whether a company will go into a default the year following the financial statements publication, the label 1 is used to indicate that a company goes into default the year following the assessment, while the number 0 signifies that the company does not go into default. These labels serve as the outcomes that the machine learning model aims to predict based on the features.
- $(x_i, y_i)$ : This represents a single labeled example in the dataset. Here,  $x_i$  is a feature vector that belongs to the feature space  $\mathcal{X}$  and contains the financial attributes for

the  $i$ -th borrower.  $y_i$  is the corresponding label for that borrower and belongs to the label space  $\mathcal{Y}$ . The feature vector and label together provide a complete snapshot of the  $i$ -th borrower's financial situation and credit risk level.

- $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ : This is the dataset, which is a collection of  $n$  labeled examples. Each labeled example consists of a feature vector and a corresponding label, as described above. The machine learning model uses this dataset to learn the relationships between the features and labels, with the aim of making accurate credit risk predictions for new, unseen borrowers.

The formal objective of a machine learning-based credit scoring model is to construct a predictive function  $f : \mathcal{X} \rightarrow \mathcal{Y}$ .

The function  $f$  aims to map each borrower's feature vector  $\mathbf{x} \in \mathcal{X}$  to a corresponding label  $y \in \mathcal{Y}$ , such that the prediction  $f(\mathbf{x})$  approximates the true label  $y$  as closely as possible. In the context of credit scoring, the problem can be simplified to a classification task where the objective is to estimate the conditional probability  $p(y | x)$ . However, this estimation can also be interpreted as a regression problem. Specifically, the goal is to find a function  $f(x)$  that serves as the best approximation for  $F(x)$ , which is the true but unknown function governing the relationship between the features  $x$  and the labels  $y$ .

### **Bias-Variance Trade-off**

However, real-world scenarios introduce a considerable amount of noise, which can be attributed to factors such as missing financial features, incorrect labeling, and so on. We define  $\epsilon$  as the irreducible error that we cannot recover from, such that the relationship between the features  $x$  and the labels  $y$  can be expressed as  $y = F(x) + \epsilon$ .

One fundamental concept in machine learning and statistics is the bias-variance trade-off. It describes two sources of errors that affect the performance of predictive models. The bias represents the error to overly simplistic assumptions in the learning algorithm. For instance, high bias can cause the model to miss relevant relations between features and target outputs leading to what is called underfitting. On the other hand, there is the error caused by a learning algorithm with a significant level of complexity. In this case, models with high variance suffer from overfitting (i.e., the model learns the random noise from the training data which leads to poor performances in unseen data).

The expected test error  $\mathbb{E}[\text{Error}]$  can be decomposed as:

$$\mathbb{E}[\text{Error}] = \text{Bias}^2 + \text{Variance} + \epsilon$$

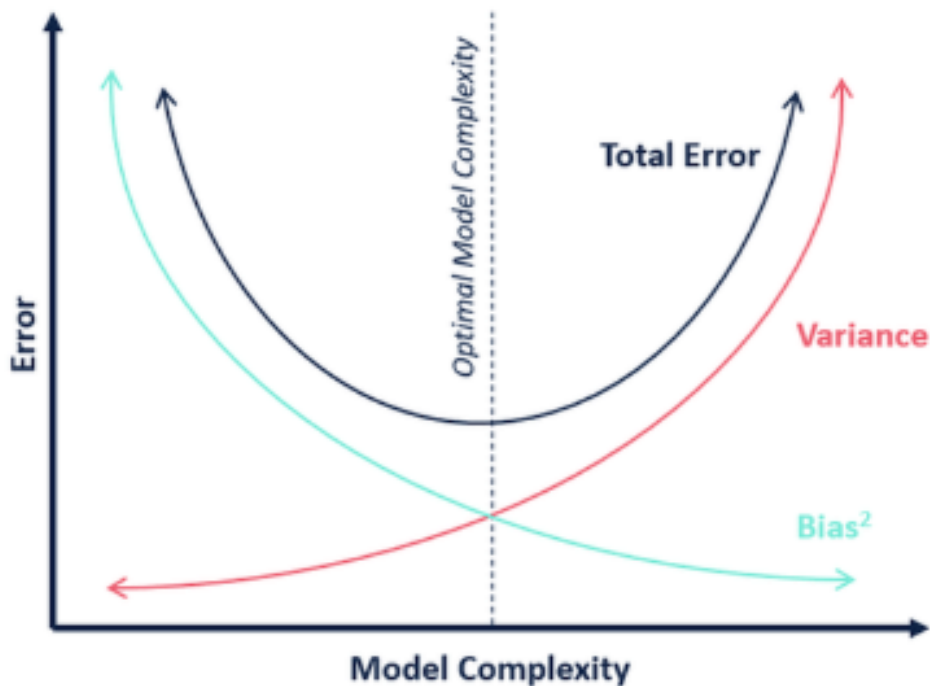
Where:

$$\text{Bias}^2(\hat{f}) = (\mathbb{E}[\hat{f}(x)] - f(x))^2$$

$$\text{Variance}(\hat{f}) = \mathbb{E}[\hat{f}(x) - \mathbb{E}[\hat{f}(x)]]^2$$

The trade-off involves minimizing  $\mathbb{E}[\text{Error}]$  by balancing Bias and Variance (see Fig.2.1).

Figure 2.1: Illustration of the common intuition for the bias-variance tradeoff. Source: [18]



A model is trained on a dataset  $\mathcal{D}$ , defined as:

$$\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$$

This dataset consists of  $n$  labeled examples. The training process involves optimizing a loss function  $\mathcal{L}$ :

$$\text{Minimize } \mathcal{L}(f(\mathbf{x}), y)$$

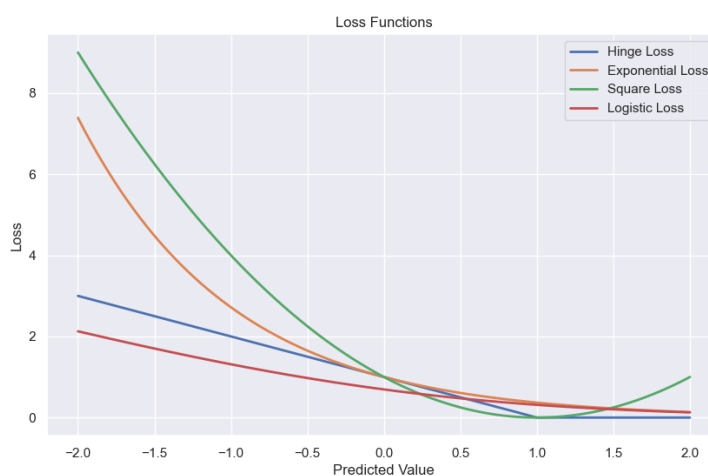
This loss function quantifies the discrepancy between the predicted labels and the true labels. The ultimate goal is to minimize  $\mathcal{L}$  while ensuring that the model generalizes well to new, unseen data.

## Loss Functions

We discuss various loss functions commonly used in machine learning:

- **Logistic Loss:**  $\log(f(x), y) = \log(1 + e^{-yf(x)})$ , often used in neural networks and logistic regression [19], measures the difference between the predicted probabilities and the actual labels. It is both differentiable and convex, making it well-suited for optimization algorithms.
- **Squared Error:**  $se(f(x), y) = (y - f(x))^2$ , the standard loss for regression tasks. In this manuscript, we use it in Gradient Boosting [20] for classification tasks.
- **Hinge Loss:**  $\text{hinge}(f(x), y) = \max(0, 1 - yf(x))$ , primarily used in Support Vector Machines (SVM) [21], [22]. This loss is zero when  $y$  and  $f(x)$  agree and linearly increases when they disagree.
- **Exponential Loss:**  $\exp(f(x), y) = e^{-yf(x)}$ , used in Adaboost [23], [24]. This loss is challenging to optimize due to its exponential nature.

Figure 2.2: Loss functions described in this manuscript.



The choice of a loss function plays a critical role in determining the accuracy of a predictive model. Different loss functions have varying sensitivities to outliers and model complexity, which can significantly impact the model's performance [25].

## 2.2 Standard models and its application in credit scoring

Linear models have long served as a foundational element in the domain of credit scoring, dating back to several decades.



The Altman Z-Score model holds the distinction of being one of the first credit scoring models to gain widespread acceptance and usage. Developed by Edward I. Altman in [6], this model was initially designed to predict the likelihood of a publicly traded manufacturing company going bankrupt within a two-year period. The model employs multiple financial ratios that are derived from a company's balance sheet and income statement, such as liquidity, profitability, and leverage ratios.

## Linear Models

During the 1980s and 1990s, statistical methods such as Logistic Regression ([26], [27], [28]) and Linear Discriminant Analysis (LDA) ([29], [30]) gained prominence in the credit scoring field. These models brought greater flexibility in accommodating various types of data and quickly became popular choices for consumer credit scoring. Logistic Regression, in particular, emerged as the industry standard for an extended period, largely due to its capability to offer probability estimates of default or non-default.

Mathematically, Logistic Regression models the log-odds of the probability  $P(y = 1|x)$  as a linear combination of the features  $x$ :

$$\log\left(\frac{P(y = 1|x)}{1 - P(y = 1|x)}\right) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_nx_n$$

The probability  $P(y = 1|x)$  can then be obtained by transforming the log-odds back:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_nx_n)}}$$

On the other hand, Linear Discriminant Analysis (LDA) aims to find a linear combination of features that best separates two or more classes. The discriminant function for LDA is given by:

$$D(x) = w^T x + w_0$$

Here,  $w$  is the weight vector, and  $w_0$  is the bias term. The class label  $y$  is then determined based on the sign of  $D(x)$ .

Regularized Linear Models such as Ridge [31] and Lasso [32] Regression have emerged as powerful alternatives to traditional methods like Logistic Regression and Linear Discriminant Analysis. Specifically, these regularized models excel in managing multicollinearity among features, a challenge that can compromise the performance of Logistic Regression and LDA [33]. By introducing regularization terms, Ridge and Lasso effectively shrink the coefficients of less important features, thereby preventing overfitting.

## Decision Trees

Decision trees are a type of supervised learning algorithm predominantly used for classification problems (e.g., [34]), but can also be employed for regression tasks (e.g., [35]).

Mathematically, a decision tree partitions the feature space into disjoint regions (see Fig.2.3). For classification tasks, each leaf node in the tree represents a class label, while for regression tasks, it represents a real value. The feature space is divided based on certain criteria that aim to maximize the information gain or minimize impurity. Two of the most commonly used criteria are Gini impurity and entropy.

The mathematical formulation for Gini impurity for a node  $t$  is given by:

$$\text{Gini}(t) = 1 - \sum_{i=1}^c p_i^2$$

Where  $p_i$  is the proportion of samples that belong to class  $c$  for the node  $t$ .

Entropy for a node  $t$  is calculated as:

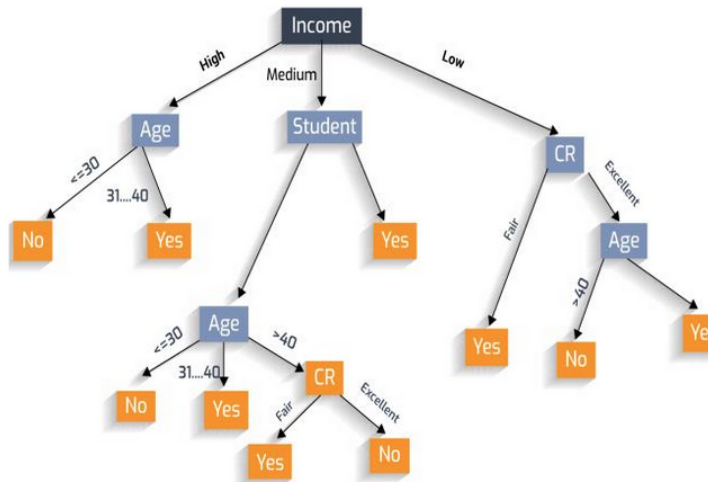
$$\text{Entropy}(t) = - \sum_{i=1}^c p_i \log_2(p_i)$$

Information Gain, which is the reduction in entropy or Gini impurity, is then used to decide which feature to split on at each step. The feature with the highest information gain is chosen for the split.

$$\text{Information Gain} = \text{Entropy}(t) - \sum_{\text{child nodes}} \left( \frac{|D_v|}{|D|} \times \text{Entropy}(v) \right)$$

Where  $D$  is the dataset,  $D_v$  is the subset of  $D$  at child node  $v$ .

Figure 2.3: Example of a decision tree in customer credit scoring [36]



In the context of credit scoring, decision trees offer several advantages. They are interpretable, which aligns well with the regulatory requirements in financial sectors. They can handle both categorical and numerical features, making them versatile for different types of financial data [37]. Moreover, decision trees can capture non-linear relationships in the data, which is often the case in credit risk assessment.

However, a single decision tree is often prone to overfitting [37], especially when the tree is deep. This is why ensemble methods like Random Forests [38], which aggregate the predictions of multiple decision trees, are often used in credit scoring to improve the model's generalization ability. In [39], it was demonstrated that ensemble models exhibit greater predictive power compared to standalone decision trees when applied to credit scoring tasks.

Regulatory constraints (e.g., [40]) often limit the use of advanced scoring methods for traditional credit products. Despite their higher predictive accuracy, complex models face adoption challenges. Simpler, more interpretable models like logistic regression or decision trees are thus more commonly used in the industry due to their regulatory compliance.

## Support Vector Machines

Support Vector Machines [41], commonly known as SVMs, serve as a category of supervised machine learning models applicable to both classification and regression problems. From a mathematical standpoint, the core aim of an SVM is to identify the optimal hyperplane that most effectively segregates data into distinct classes. While in a two-dimensional setting this separating hyperplane manifests as a simple line, in spaces with higher dimensions it takes the form of a more complex plane or even a multi-dimensional hyperplane.

The equation of the hyperplane is given by:

$$\vec{w} \cdot \vec{x} - b = 0$$

Where  $\vec{w}$  is the weight vector,  $\vec{x}$  is the input vector, and  $b$  is the bias term.

The objective is to maximize the margin, which is the distance between the closest points (support vectors) of the different classes to the hyperplane. Mathematically, the margin  $M$  is calculated as:

$$M = \frac{2}{\|\vec{w}\|}$$

The optimization problem can be formulated as:

$$\text{Minimize } c \frac{1}{2} \|\vec{w}\|^2$$

$$\text{Subject to } y_i(\vec{w} \cdot \vec{x}_i - b) \geq 1, \quad i = 1, \dots, n$$

Here,  $y_i$  are the labels, and  $\vec{x}_i$  are the data points.

SVMs come with multiple benefits. They perform well in situations where the data has many variables, a common scenario in financial data that includes diverse factors like income levels, credit history, and job status. Additionally, SVMs can handle complex, non-linear relationships between these variables by employing specialized mathematical functions known as kernels, such as the Radial Basis Function (RBF), polynomial, or

sigmoid kernels. The authors in [42] emphasize the significant impact that choosing the right kernel has on the overall performance of the model.

SVM combined with feature selection techniques have shown promising results in credit scoring tasks. This approach enhances model performance by focusing on relevant financial variables. The outcome is a more accurate and efficient credit risk assessment [43]. Furthermore in [44], the authors concluded that using logistic regression to filter dummy variables and orthogonal feature extraction improved the performance of Support Vector Machines. This approach not only simplified the model and sped up convergence but also yielded superior result

However, one of the challenges with SVMs in credit scoring is interpretability. Unlike decision trees or logistic regression models, SVMs are often considered as "black-box" models, making them less transparent. This could be a concern in financial sectors where interpretability is crucial for regulatory compliance. In recent works, the authors have addressed this concrete problem by analyzing the weights of a svm model [45].

## 2.3 Ensemble models in credit scoring

In the previous section, we highlighted that decision trees often struggle with managing the bias-variance trade-off effectively. This is a crucial aspect, especially in the context of credit scoring, where both underfitting and overfitting can lead to significant financial implications. Interestingly, in [39] the authors conducted a comparative study found that ensemble models are far more adept at handling the bias-variance trade-off in credit scoring scenarios. Given this backdrop, this section is dedicated to introducing ensemble learning models and elucidating their growing importance in modern machine learning-based credit scoring systems.

The core principle is simple yet effective: by combining the predictions of multiple models, we can often achieve greater accuracy than relying on a single model. We will specifically focus on two widely-used ensemble learning techniques: bagging and boosting.

### Bagging

In ensemble methods like bagging [46], the core idea is to amalgamate the predictions from multiple classifiers to arrive at a more accurate final prediction. What sets bagging apart is its unique approach to training each classifier on a distinct subset of the original data set, denoted as  $\mathcal{X}'$ , which is a proper subset of  $\mathcal{X}$ . This is achieved through a technique known as 'sampling with replacement,' where data points are randomly selected from  $\mathcal{X}$  and then returned, allowing for the possibility of multiple selections of the same data point.

One of the advantages of decision trees is the ease of training but it comes with an important inconvenient: overfitting (see section 2.2). Bagging addresses this issue by introducing an element of randomness through its sampling technique. This randomness foment diversity among the individual trees, making the ensemble model more resilient to overfitting, even if each constituent tree is highly variable. The final bagged model is thus more robust and less prone to overfitting, offering a balanced approach to tree-based classification (e.g., [47]) or regression tasks (e.g., [48]).

Introduced in [38], Random forest is the most widely adopted bagging method. Employing the same variable definition presented in section 2.1 and defining variables inherent to Random Forest, the algorithm can be formalized as follows:

- $T$ : Number of trees in the forest.
  - $m$ : Number of features selected at each split.
1. **Bootstrap Sampling:** For each tree  $t$ , draw a bootstrap sample  $\mathcal{D}_t$  of size  $n$  from  $\mathcal{D}$  with replacement.
  2. **Tree Building:** For each bootstrap sample  $\mathcal{D}_t$ , grow a decision tree  $f_t(x)$  as follows:
    - At each node, randomly select  $m$  features without replacement.
    - Split the node using the feature that provides the best split according to some criterion (see section 2.3).
    - Recur for the child nodes until a stopping criterion is met (e.g., maximum depth, minimum samples at leaf).
  3. **Ensemble Prediction:** The Random Forest  $F(x)$  makes a prediction by averaging the predictions of all individual trees for regression or by majority voting for classification:

$$F(x) = \frac{1}{T} \sum_{t=1}^T f_t(x) \quad (\text{Regression})$$

$$F(x) = \text{mode}\{f_1(x), f_2(x), \dots, f_T(x)\} \quad (\text{Classification})$$

Various studies have leveraged the advantages offered by the random forest algorithm. In [49], the authors have introduced a novel credit scoring model based in random forest and feature selection. This random forest based model improves the accuracy compared to different state-of-the-art machine learning models. In a comparative study of a large list of machine learning models, the authors [39] found that random forest is one of the most performant models.

## Boosting

The idea of boosting was first articulated by [50], showing that one could theoretically improve the efficacy of any learning model by amalgamating an array of weak classifiers. This theoretical underpinning has subsequently catalyzed the creation of a range of boosting algorithms.

Boosting is a sophisticated family of ensemble learning algorithms designed to enhance the predictive capabilities of weak or base learners. The fundamental principle that underpins boosting is the iterative training of a series of weak learners, where each learner in the sequence is specifically trained to correct the errors or misclassifications made by its predecessor. This iterative correction process allows the algorithm to focus on the more challenging or nuanced aspects of the data, thereby improving the model's overall performance. Once this iterative training is complete, the algorithm combines the outputs of all the weak learners to construct a single, more robust and accurate, strong learner [50].

In mathematical terms, the boosting algorithm can be formalized as follows. Let  $F(x)$  represent the final strong learner we aim to construct. This strong learner is essentially a weighted sum of  $T$  weak learners  $f_t(x)$ , each associated with a weight  $\alpha_t$ . Mathematically, this can be expressed as:

$$F(x) = \sum_{t=1}^T \alpha_t f_t(x)$$

Here,  $T$  is the total number of iterations or weak learners in the ensemble,  $f_t(x)$  represents the  $t^{\text{th}}$  weak learner, and  $\alpha_t$  is the weight assigned to this weak learner. These weights  $\alpha_t$  are computed based on the performance of each weak learner, effectively serving as a measure of its contribution to the final strong learner. The better a weak learner is at making accurate predictions, the higher its corresponding weight  $\alpha_t$  will be.

Adaptive Boosting (AdaBoost) is one of the pioneering algorithms in the field of boosting and has gained widespread popularity since its introduction in [51]. As one of the first algorithms to successfully demonstrate the power of boosting techniques, it has been extensively used in various applications, including credit scoring [52], healthcare [53], and sentiment analysis [54].

The AdaBoost algorithm starts by assigning equal weights to all the training samples. These weights are then adjusted at each iteration to give more importance to the samples that were misclassified by the previous weak learner.

Mathematically, the Adaboost algorithm is constructed as follows:

1. **Initialize Weights:** Initially, each sample  $i$  is given an equal weight  $w_i = \frac{1}{n}$ .

2. **Iterative Training:** For  $t = 1, 2, \dots, T$  (where  $T$  is the total number of iterations or weak learners):

- Train a weak learner  $f_t(x)$  using the weighted samples.
- Calculate the weighted error  $\epsilon_t$  of  $f_t(x)$  as the sum of the weights of the misclassified samples.
- Update the weight  $\alpha_t$  for the weak learner  $f_t(x)$  as  $\alpha_t = \frac{1}{2} \ln \left( \frac{1-\epsilon_t}{\epsilon_t} \right)$ .
- Update the sample weights  $w_i$  to give more importance to the misclassified samples.

3. **Final Model:** The final strong learner  $F(x)$  is a weighted sum of the weak learners:

$$F(x) = \sum_{t=1}^T \alpha_t f_t(x)$$

The sign of  $F(x)$  will give the final classification.

Another family of boosting methods is Gradient Boosting [55], which aims to construct a robust predictive model by combining multiple weak or base learners. Unlike AdaBoost, which adjusts the weights of individual samples to focus on difficult-to-classify instances, Gradient Boosting concentrates on the errors or residuals made by the preceding models in the ensemble. The algorithm iteratively fits new models to these residuals, thereby correcting the mistakes of the existing ensemble of models.

In Gradient Boosting, several key characteristics distinguish it from other ensemble methods like AdaBoost, and these can be mathematically formalized.

First, the concept of "Residual Fitting" is central to the algorithm. In this approach, each new weak learner is trained to fit the residuals  $r_i = y_i - F_{m-1}(x_i)$  of the previous ensemble model  $F_{m-1}(x)$  at each iteration  $m$ . This allows the model to focus on the errors made by the preceding learners.

Second, Gradient Boosting optimizes a differentiable loss function  $L(y, F(x))$ , offering more flexibility in model optimization compared to AdaBoost's sample re-weighting technique. The loss function is minimized through gradient descent, and the weak learners are fit to the negative gradients.

Third, the algorithm incorporates a "Learning Rate"  $\eta$  to control the contributions of each weak learner, thereby adding a regularization component to the model. Mathematically, the ensemble model at each iteration  $m$  is updated as  $F_m(x) = F_{m-1}(x) + \eta \cdot f_m(x)$ , where  $f_m(x)$  is the weak learner at iteration  $m$ .

Lastly, Gradient Boosting is highly versatile; it can be applied to both classification and regression tasks and can accommodate a wide range of loss functions. Gradient Boosting has been widely applied across various domains, demonstrating its versatility and robustness. In the field of healthcare, [56] gradient boosting has been employed in



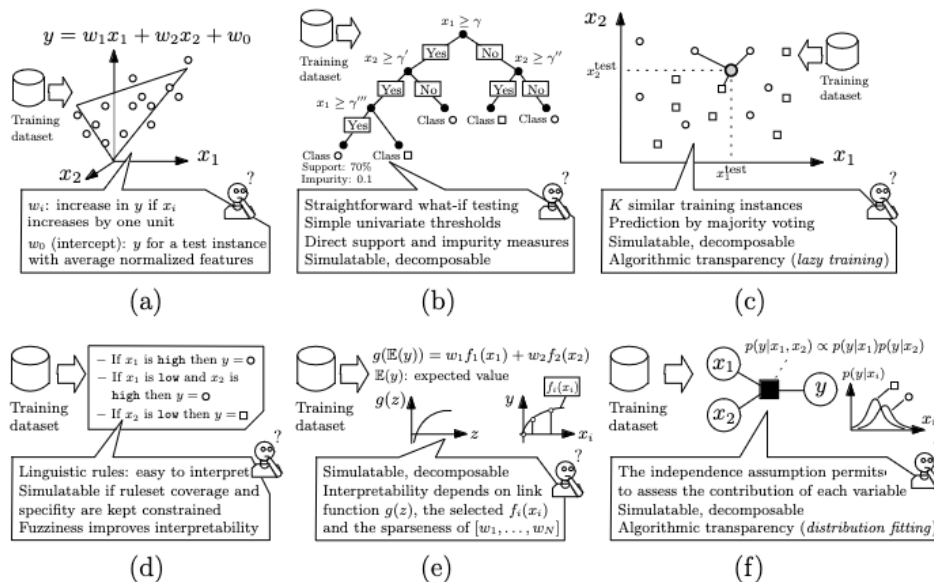
clinical. In finance, [57] utilized the algorithm for credit risk assessment, highlighting its superiority over traditional methods. In natural language processing, the technique has been used for detecting fake news [58].

## 2.4 eXplainable Artificial Intelligence

Explainable Artificial Intelligence (XAI) has emerged as a critical subfield of machine learning, particularly in contexts where understanding the decision-making process of algorithms is not just beneficial but essential. The objective is to bridge the gap between the performance of a model and the human understanding of how that model arrives at a particular decision.

The emergence of the field of Explainable Artificial Intelligence (XAI) can be traced back to the quest for increasingly accurate machine learning models. As models become more sophisticated to capture complex patterns and nuances in data, they inherently become less understandable to humans [59]. Recent research endeavors have focused on establishing a comprehensive taxonomy for explainable models within the realm of artificial intelligence [60].

Figure 2.4: Visual Representation of the Transparency Levels Across Various Machine Learning Models: (a) Linear regression; (b) Decision trees; (c) K-Nearest Neighbors; (d) Rule-based Learners; (e) Generalized Additive Models; (f) Bayesian Models. Source: [60]



First concept that needs to be clear when dealing with XAI is the definition of explainability and interpretability. Explainability refers to the extent to which the internal workings of a machine learning or artificial intelligence system can be revealed or understood. An explainable model provides insights into its decision-making process, often



through visual or textual explanations, so that a human can understand why a particular decision or prediction has been made [61].

On the other hand, interpretability is the degree to which a human can understand the cause-and-effect relationship between variables in a machine learning model. An interpretable model is one where the internal mechanics are transparent enough that they can be easily scrutinized to understand how input variables are transformed into an output decision. Interpretability is often a desired feature in models where safety or critical decision-making is involved, as it allows for easier debugging and trust-building (e.g., medicine applications [62]).

Models such as decision trees, logistic regression, and linear discriminant analysis are inherently interpretable. This means that their internal workings are transparent, allowing for a straightforward understanding of how input features relate to output predictions (see Fig. 2.4).

Due to their increasing complexity, ensemble models often require post-hoc explanations to make their predictions understandable and actionable. As explained in previous sections, ensemble methods combine multiple base learners, making the decision-making process intricate and less transparent. Post-hoc models come into play here as separate explanatory models that are applied after the original model has made its predictions. These post-hoc models aim to approximate the complex model's behavior in a more interpretable form, shedding light on the important features and decision paths. For example in [63], the authors provide a series of different types of explanations of a gradient boosting based model.

One of the most widely used post-hoc models for explaining the output of any machine learning model is SHapley Additive exPlanations (SHAP) [64]. It employs game theory to fairly allocate contributions of each feature for a particular prediction. Several. Despite its widespread use and versatility, SHAP is not without its drawbacks. One notable limitation is the interpretation of the SHAP scores themselves. While these scores offer a ranking of feature importance, the actual numerical values can be misleading, requiring the end user to focus more on the order of the features rather than the scores. Another issue is the assumption of feature independence, which may not hold true in many real-world scenarios, potentially leading to inaccurate or misleading explanations [65]. Additionally, SHAP can be computationally expensive, especially for complex models and large datasets, making it less feasible for real-time or resource-constrained applications. Recent works have focused on accelerating shap value computation for tree-based models [66].

In different research papers, the SHAP has been employed for understanding credit score. In [67], they used SHAP to find out which factors, like income or debt, are most important for a company's credit score. In [68], the authors focused on customer credit scoring. The study went beyond merely identifying important features; it also generated local explanations for individual loan decisions.

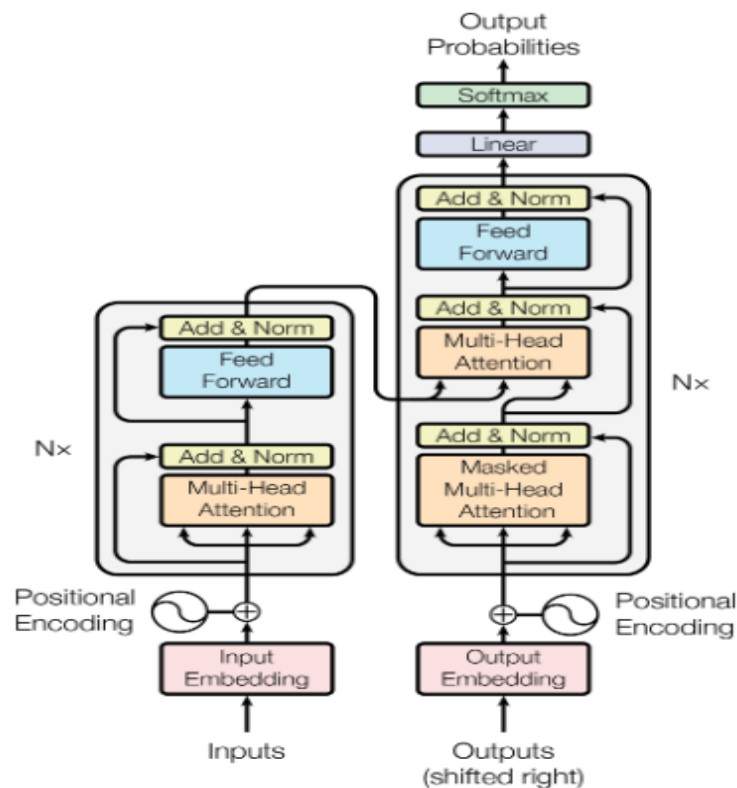
## 2.5 Natural Language Processing

The rapid advancements in the field of Natural Language Processing (NLP) have not only revolutionized various sectors but have also caught the attention of the finance industry. In this section, we will specifically focus on how these NLP methodologies have been increasingly integrated into the finance sector, including their role in credit scoring, risk assessment, and financial analytics.

One of the first NLP models to be adopted is the Bag-of-Words (BoW) [69]. The model represents text data as a 'bag' or collection of individual words, disregarding grammar and word order. It converts text into numerical vectors by counting the frequency of each word. It has been used for different application like, text classification [70], and spam filtering [71].

Term Frequency-Inverse Document Frequency (TF-IDF) which is an extension of BoW it not only counts the frequency of each word (like BoW) but also weights it based on its importance in the document relative to a collection of documents (corpus). It has been widely used for information retrieval [72]. The latter work applied directly the information retrieved from the document to perform stock trend.

Figure 2.5: Transformers architecture. Source: [73]



Before the advent of Transformer models, the most widely used family of models for tasks such as machine translation were Word Embeddings. These models, including

popular variants like Word2Vec [74] and GloVe [75], map words into a continuous vector space. They capture semantic meanings and relationships between words based on their co-occurrence in text. An interesting application of Word2Vec has been developed in [76]. The authors used this algorithm to identify stock market reactions to the COVID-19 pandemic.

In recent years, Transformer models [73] have completely revolutionized the domain of Natural Language Processing (NLP). These groundbreaking architectures (see Fig. 2.5) have set new benchmarks and reshaped the landscape of what's possible in various NLP tasks. Bidirectional Encoder Representations from Transformers (BERT) [77] a pretrained transformer has been widely used in finance. In [78], the authors employed BERT to predict stock movement. They showed that BERT is able to achieve and even outperforms state-of-the-art models.

# Analyzing the Impact of the COVID-19 Outbreak on Companies Default Rates

## 3.1 Introduction

In previous sections, we discussed how the advent of a crisis (i.e., 2008 financial crisis), has the potential to dramatically alter the landscape of the business network. In this chapter, we analyze the impact on default behavior before and after the Covid-19 outbreak from the point of Credit Insurers (CI) and Export Credit Agencies (ECA)

The business network of a country is mainly composed of Small and Medium-sized Enterprises (SMEs). They are considered the driving force of the country. In large and established economies, like France, the percentage of SMEs represents 99.8 % of the total registered enterprises and The SMEs generate 48.5 % of the total employment [79].

The definition of SMEs is not universal. However, the most common criteria used to distinguish them is the number of total employees [80] generally between 10 and 250. The simplicity of its business structure allows the company to be very flexible. Commonly, SMEs rely on the personal assets of owners to finance the company. Their lack of external financing may pose an extra risk, especially when the local economy is not stable.

On March 11, the Director-General of the World Health Organization (WHO) declared Covid-19 a pandemic, causing a severe shutdown of the global economy in an effort to contain the virus. From an economical point of view, and due to their size, SMEs were strongly affected by the temporary shutdown. Given the importance of SMEs in the global economy, it is necessary to quantify the impact of Covid-19 on the SMEs' network to understand and estimate the repercussion of the pandemic.

It is important to remark that the impact of Covid-19 on company's default depends

on the activity sectors. SMEs default occurs when the company is not able to refund the full amount of the loan. The Covid-19 impact on the SMEs' financial soundness is not homogeneous across sectors. As far as we know, no previous research has investigated the current impact of the Covid-19 outbreak on the activity sector for the French economy.

We analyze the four most represented economic sectors: Wholesale & retail trade, repair of motor vehicles; Construction; Manufacturing; Accommodation and food service activities. These sectors have been chosen regarding the number of companies that are operating in that sector.

### 3.2 Literature Review

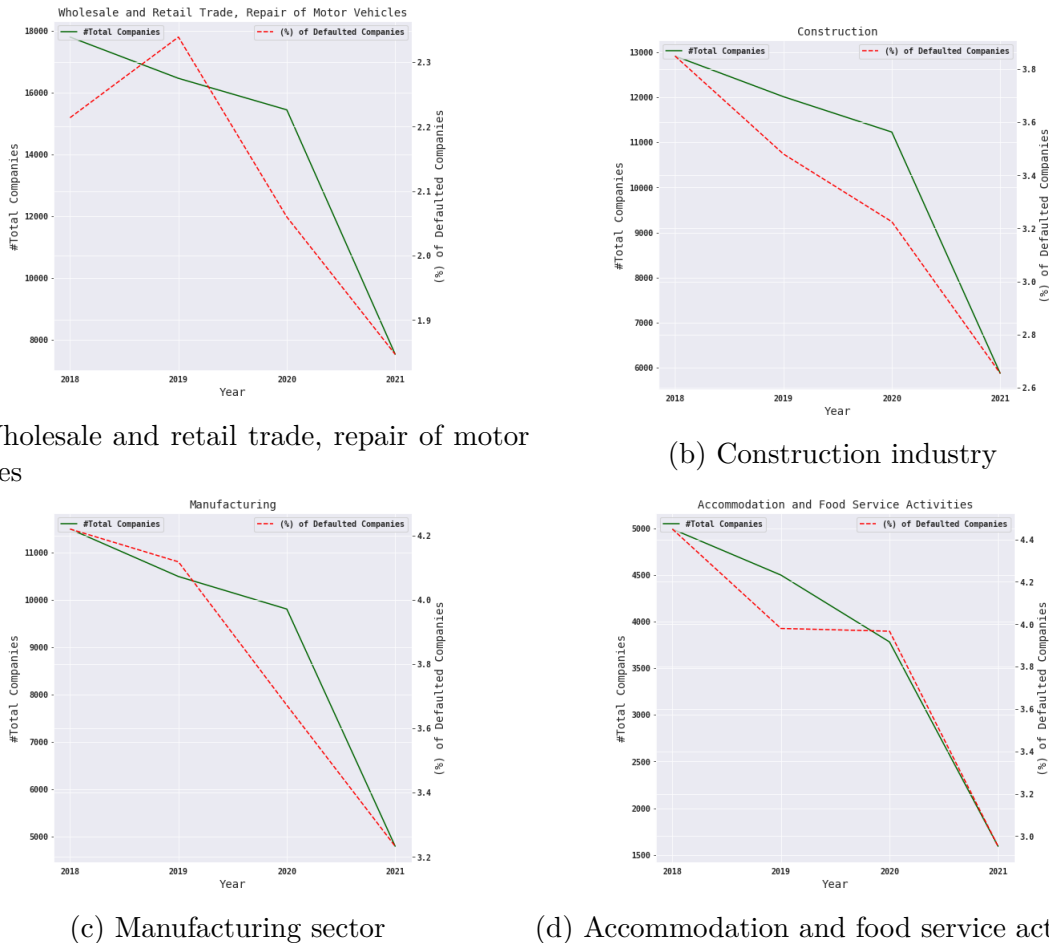


Figure 3.1: The right y-axis corresponds to the percentage of default in the sector (red line) and the left y-axis the total of companies belonging to the sector that have been assessed (green line). This analysis shows how the default by sector varies between 2018 and 2021.

A large number of researches have been focused on the impact of Covid-19 in the economy. In this section, we present literature on the economic impact of Covid-19 with

a special focus on its impact on SME's.

One of the first studies on the impact of Covid-19 on SMEs [81], the authors established, through the analysis of the answers of 4807 Chinese companies to a questionnaire, that SMEs were in a near-bankruptcy situation, mainly due to the inability to recover economic activity as well as the fact that they had to continue paying fixed costs with minimal or even no income. A similar analysis conducted by [82] shows that the main risk faced by polish SMEs during Covid-19 is a strong competition as well as the increases in energy prices and the low-profit margin.

There are studies such as [83], in which the researchers focus on the impact on the Manufacturing sector. On one hand, the authors state that in the short term the most important problem faced by SMEs is logistics management, although the impact differs between different sectors. On the other hand, they establish that in the long term the problem will depend more on the type of SME, concluding that a policy is needed that goes beyond the survival of European SMEs and that focuses on promoting the growth of companies through innovation and internationalization.

More extended analysis was conducted in [84]. In this study, the authors use cost minimization and measure each firm's liquidity shortfall during and after COVID-19. They estimate a large increase in the failure rate of SMEs absent of government support. It is found that the most affected sectors are Accommodation & Food Services, Arts, Entertainment & Recreation, Education, and Other Services.

Recent studies ([85], [86]) have been focused on the application of Artificial Intelligence methods for predicting the default of European SMEs. In [87], the authors concluded that the estimation of the probability of default can be improved by creating more granular models (i.e., a model by sector). Nevertheless, the analysis conducted to not focus on the impact of economically challenging periods (i.e., Covid-19).

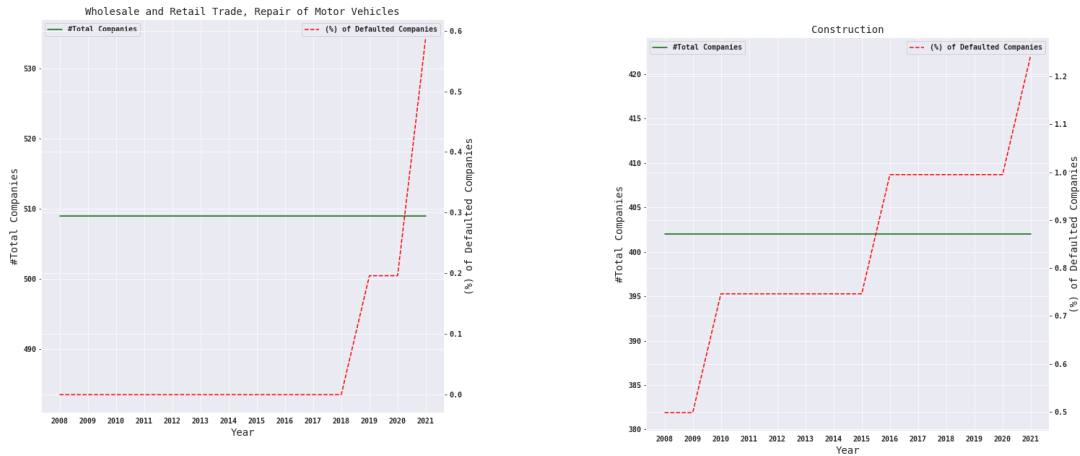
### 3.3 Methodology

The proposed experimental framework in this chapter can be divided into three main parts: first we collect and clean the data provided by Tinubu. Then we analyze two types of companies: a sector-centered analysis in which we analyze the default for the chosen sectors. To observe the impact of the Covid-19 outbreak we consider the time frame 2018-2022. The second analysis is more company-centered. We consider just the companies with the sector and default information for the complete period of 2008-2022 which will allow us to observe the complete default trend and the impact of Covid-19.

#### 3.3.1 Data Processing

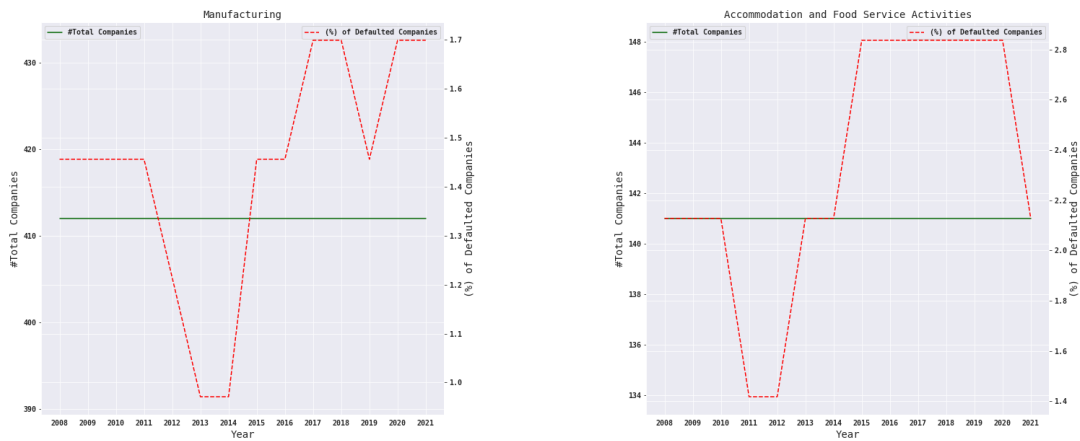
The first step is to clean the data. We remove the data points (i.e., company information) that are duplicated. We will focus on the default evolution year by year, so each company is represented by one unique row. In the case in which the company's status changed

## Chapter 3. Analyzing the Impact of the COVID-19 Outbreak on Companies Default Rates



(a) Wholesale and retail trade, repair of motor vehicles

(b) Construction industry



(c) Manufacturing sector

(d) Accommodation and food service activities

Figure 3.2: The right y-axis corresponds to the percentage of default in the sector (red line) and the left y-axis the total of companies belonging to the sector that have been assessed (green line). This analysis shows how the default by sector varies over time (2008 till 2021).

from solvent to default in a concrete year, the company is considered in payment default.

The next step is to concentrate on French companies. The original dataset contains several European countries. However, most of the companies are based in France, so for this analysis, we will keep the companies that operate in France.

The sector activity information is encoded into two features in the original dataset: the activity codification and the activity code. For french companies, the activity codification used is the NAF2<sup>1</sup>. The sectors are defined by the NAF2 which is the French nomenclature of business activities adopted since 2008. This activity codification contains 4 different levels of precision. In this chapter, we will analyze the global sector classification. We

<sup>1</sup><https://www.insee.fr/fr/information/2120875>

map the sector activity code into the real name of the sectors.

### **3.3.2 Sector Default Evolution**

The first default analysis will consider companies that belong to the same sector for a given year. In this case, we analyze the default by year as well as the volume of companies by sector (i.e., the number of companies assessed by Tinubu's Credit Risk System and Risk Analysts).

### **3.3.3 Analyzing the Default of Companies by Sector for the Period 2008-2022**

The next analysis proposed in this work consists of keeping companies with the sector and default information for the period 2008-2022. This more detailed analysis will help us obtain global conclusions about the sector's behavior.

## **3.4 Results**

In this section we will present the results yielded by the proposed framework.

### **3.4.1 Sector Default Evolution**

In Fig. 3.1 we present the default by sector and the number of companies analyzed for each sector before and after Covid-19. It is important to remark that if a certain company defaulted during the year 2020, the default information will be published the next year. In other words, the year 2021 contains what happened during the year 2020.

We focus our analysis on 4 different sectors: Wholesale & retail trade, repair of motor vehicles; Construction; Manufacturing; Accommodation and food service activities. These sectors have been chosen regarding the number of companies that are operating in that sector. They represent 77.3 % of the raw dataset.

What is interesting about the figures in Fig. 3.1 is that there is a significant decrease in terms of both default rate and size of assessed companies by Tinubu's that affects all the different sectors in a similar way. In the graph, we can see how the decrease is accentuated from 2020 onwards which can be explained by the first effects on the SME's French network by Covid-19. What should be expected, since Covid-19 impacted negatively the french economy, is an increase in the default rate which is the opposite



of what we obtained from analyzing the dataset. However, the result can be rationally explained by the policies imposed by the french government during the Covid period to avoid the default of a large number of companies.

### 3.4.2 Evolution of Companies by Sector for the Period 2008-2022

For each sector, we analyze the same companies for the period 2008-2022. We group the companies according to the sector in which they operate. In Fig. 3.2 we provide the results obtained after analyzing the companies with available data for each sector. As can be seen from Fig. 3.2, for all sectors analyzed except for the accommodation and food service activities, the default rate suffers a slight increase that starts in 2020. For the accommodation and food service activities (see Fig 3.2(d)) the default rate varies strongly due to the limited amount of data available.

## 3.5 Discussion

The analysis of the data provided by Tinubu shows the economic impact of Covid-19 on french companies from the point of Credit Insurers and Export Credit Agencies. From the first analysis, we found that there has been a decrease in the activity related to the sectors (i.e., Wholesale & retail trade, repair of motor vehicles; Construction; Manufacturing; Accommodation & food service activities) that are emphasized in the period 2020-2021. This behavior can be explained by the fact that the CI and the ECA risk aversion increased due to the uncertainties associated with the pandemic.

Moreover, the default rate found on the same dataset (see Fig. 3.1) is the opposite of what was expected. However, there is an explanation for this behavior and it is associated with the policies imposed by the french government. During the pandemic, the french government established a solidarity fund for SMEs particularly affected by the economy.

In the second analysis, we focused on the company's evolution in the period from 2008 to 2022. Regarding Fig. 3.1 and Fig. 3.2, the most surprising difference is the opposite behavior in terms of default when analyzing the companies using the different proposed frameworks. The French companies' solidarity fund varies depending on the size of the companies. Small companies are more supported financially than medium companies. The relation between the size of the company and its age explains the fact that the companies analyzed using a historic of 14 years reduced the size of the dataset biasing the dataset towards more established companies and thus less supported by the solidarity fund implemented by the french government.

## 3.6 Conclusions

This chapter has proposed a novel approach to analyze the economic impact of the Covid-19 Pandemic. The conducted analysis focus on the impact of the default of the SME from the point of view of Credit Insurers and Export Credit Agencies. The analyzed sectors in this chapter were: Wholesale & retail trade, repair of motor vehicles; Construction; Manufacturing; Accommodation & food service activities. It is particularly interesting the divergence between the two different experiments. It was found that the trends in the default rate before and after 2020 are completely different. This is due to the bias introduced to the analysis when focusing on companies with at least 14 years of available data. Small companies are, to a large extent, removed for the second analysis (see Fig. 3.2). Another important result to highlight is the decrease in the default rate during the pandemic period and how the policies implemented by the french government has helped small-sized companies has helped companies to avoid default.

This chapter represents our effort to explore empirically the effects on the SME network caused by the Covid-19 outbreak. We consider it necessary to continue to analyze the default by sector previous to and after Covid-19 from the point of view of other relevant financial institutions. This will help us understand the real impact of Covid-19 on the french economy.



# Aligning Feature Contributions with Expert Knowledge in Artificial Intelligence-Based Credit Scoring

## 4.1 Introduction

As previously introduced, credit risk assessment is a cornerstone activity for banks, financial institutions, and insurance companies. The methodology for evaluating credit risk varies based on the type of counterpart involved, which can be broadly categorized into three groups: publicly traded corporations, small and medium-sized enterprises (SMEs), and individual consumers. The risk assessment for each of these counterparts can be conducted either through the expertise of a credit risk analyst or via a mathematical approach.

The objective of credit scoring is to assess the probability that a borrower will show some undesirable behavior in the future [88]. The nature of available data to estimate the probability of default rely on the counterpart.

For publicly traded companies the literature is focused on two different approaches: the first one starts with the Z-Score [89], a model that predicts insolvency using historical accounting data. The second approach relies on securities market information [14]. In order to assess the credit worthiness of a company, financial institutions use financial indicators (i.e., financial ratios computed using financial statements) for business loans, and both personal and financial information for consumer lending.

To highlight the relevance of developing a credit score model, [90] shows that during the 2007-2009 housing crisis there was a marked rise in mortgage delinquencies and foreclosures among high credit score borrowers, suggesting that credit scoring models at

the time did not accurately reflect the probability of default for these borrowers. After the 2008 crisis, the financial institutions became more risk-averse, which provoked a substantial increase on the barriers in the process of acquiring credit [91].

As shown in [17] it is important, when developing a model that estimates the probability of default (PD) or an internal rating, to decide whether to grade borrowers using their current situation or (point-in-time, PIT) or their expected condition over a cycle (through-the-cycle, TTC). Classical Credit Rating Companies use Credit Scorecards to evaluate the risk of a counterpart. This algorithm takes as input financial information and provides a qualitative estimation of the probability of default for a company.

In the last decades, a growing number of approaches has been developed to model the credit quality of a company by exploring statistical techniques. There are three main generations of statistical techniques [92]: Discriminant Analysis [89], Binary Response Models [93], and Hazard Models [94].

Machine learning algorithms have shown an increase in the prediction power for Credit Risk Modeling [95]. Although they improve the existing credit scoring models, the AI-powered systems are regarded with suspicion because they do not provide reliable explanations for the score they provide. In this context, eXplainable Artificial Intelligence (XAI) has rapidly gained interest in the financial field.

In this chapter, we use historical financial data to predict the default of a company in one-year horizon. We focus on companies based in Europe (mainly France). The information used for modeling the default is shown in Table 4.9. As we can see in ??, The dataset provided is highly imbalanced. We apply several machine learning (ML) techniques as well as resampling techniques to address this problem. Finally, we combine the best model with the SHAP technique [96], a XAI method widely used for model interpretation based on feature attribution.

To summarize, in this chapter:

- We analyze a large dataset (around 100.000 companies) based on different European countries. Most of the literature on credit scoring algorithms predicts consumers' default, while in this work we focus on a large variety of companies (from small companies to big corporations).
- We map our machine learning model probabilities to risk score labels to compare it with an established companies credit risk scoring system.
- We perform an interpretability study, using the well-known Shapley value analysis (SHAP) , in order to understand why the algorithm made a certain decision and compare the most important features of the developed model with the expertise of several risk analysts.

## 4.2 Related Work: EXplainable AI for Credit Scoring

In the last decade the intersection between machine learning and the credit risk community has improved the performance of the credit risk models. In this section we present the main works on credit scoring focusing on machine learning models and non traded companies (i.e., companies that have their shares listed on any stock exchange). Also, we show previous studies on eXplainable Artificial Intelligence in finance.

### 4.2.1 Machine Learning for Credit Risk Scoring

During the last decade the relevance of Machine Learning (ML) has grown exponentially across all industries. The first intersection between finance, in particular credit scoring, and ML industries was in the 80s. Some ML algorithms used in credit scoring are decision trees [97], kernel-based algorithms such as Support Vector Machine (SVM) [41]. Recently, more sophisticated ML-based models have been applied to credit scoring. In [95] they compare a list of 41 different ML models for consumer credit scoring. The results show that the Random Forest Algorithm, a random version of bagged decision trees [38] outperforms the classical and widely used Logistic Regression (LR).

The scarcity of data for assessing the credit risk of non publicly traded companies has provoked research to be more focused on consumer lending. Nonetheless, some works focus on this subject. In [57], they analyze a dataset of companies based in Southern Europe for the year 2015. They use Extreme Gradient Boosting [98] for predicting whether a company will default the next year. [99] build a model to predict the default of a company in a one year horizon using a dataset composed of Italian companies over the period 2011-2017. For PD modeling they use a boosting method called LightGBM [100]. [101] they consider that a company is defaulted for the given year if the ratio of non-performing credits to total credit drawn is greater than 5%. Their best results has been obtained using Random Forest.

Companies' default, as well as consumers default, are rare events, and thus, when treating with these datasets is important to address potential data collection and reporting bias. In Table 4.1 we summarize the different datasets used in the literature. Several techniques are applied to tackle the imbalance problem . One of the common solutions is generating synthetic data of the minority class (SMOTE) [102]. In [103] they show that applying SMOTE [102] in training stage improves the performance of a large list of ML models for bankruptcy modeling.

Table 4.1: Datasets used for companies credit scoring modeling using machine learning-based models. Imbalance Ratio is the ratio between the instances of the majority class and the instances of the minority class.

Dataset Reference	Dataset Size	Features	Imbalance Ratio
Bussmann et al. 2020 [57]	15,045	Not specified	8.17
Addo, Guegan, and Hassani 2018 [104]	117,019	181	65.67
Provenzano et al. 2020 [99]	919,636	179	65
Moscatelli et al. 2020 [101]	~ 250,000	26	65
Dataset used in this chapter	138,419	15 (Table 4.7)	114.75

## 4.2.2 EXplainable Artificial Intelligence (XAI) in Finance

The pursuit of highly performant machine learning algorithms has derived in complex systems that are harder to interpret and therefore to trust [105]. The challenge for today’s ML-based credit scoring models and more generally the implementation of AI-powered systems in the financial industry is to meet strong regulations (e.g., General Data Protection Regulations (GDPR), Basel III, Solvency II). To ensure the correct, ethical and responsible development of AI in finance the implemented systems need to be explainable and interpretable. Recent works [106] discuss the requirements an AI-based system needs to meet to guarantee the fair functioning of the system. Several techniques have been developed in order to clarify opaque models interpretability problem [107]. One of these techniques is SHAP (SHapley Additive exPlanation). SHAP [64] is a framework used for interpreting predictions based on game theory. It falls into the Post-hoc explainability methods taxonomy of XAI. These family of methods target to explain the output of models that are not readily interpretable by design [108]. Recent works [63] emphasize the importance of understanding the decision-making process for ML based credit risk models. and how addressing this problem could benefit the implementations of more machine learning models in the credit risk industry.

## 4.3 Methodology: Black Box Models and XAI

In this section, we present the methods we use in our study. First, we start by presenting the State-of-the-Art ML models, and then we describe how we prepared our data for the modeling step. For the the data preprocessing we used Python frameworks Pandas<sup>1</sup> and Numpy<sup>2</sup>. We show a procedure to perform data augmentation and generate synthetic data for the minority class, based on oversampling, in order to improve the model robustness and performance metrics. Imblearn<sup>3</sup> is the framework used for data augmentation that contains the SMOTE method. We first present the different evaluation metrics used to compare the different ML models. We then briefly describe the explainability framework

---

<sup>1</sup><https://pandas.pydata.org/>

<sup>2</sup><https://numpy.org/>

<sup>3</sup><https://imbalanced-learn.org/stable/>

used in this work, SHAP values. In the last stage we detail how we conduct a survey among several credit risk analyst experts. The survey will be used to compare the results of the explainability framework with the human expert explanations.

### 4.3.1 Data Preprocessing Pipeline for Company Credit Risk Scoring

*Data Cleaning:* The data used in this study is provided by Tinubu Square<sup>4</sup>, a company that provides companies credit risk opinions as a service. Tinubu’s database is composed of financial and non-financial information about a large set of companies. The initial dataset is composed of 6,051,844 data points and 17 variables. To develop the PD model, we use the financial variables in Table 4.9. , therefore, from the original dataset we will keep those evaluations that were made using financial variables. We are interested in those companies for which we have all financial information available. At this stage, we have a total of 1,415,610 assessments (i.e., data points); the number of unique companies assessed is 418,516.

*Data Labeling:* The next step is to create the target variable of our problem. We are considering the problem of modeling the default of a company knowing its previous financial statements. Financial statements of a company non publicly traded are published yearly. Since in this study we are interested in short term PD modeling, we fix the time horizon to one year (e.g., given the financial statements of a company closed closing financial statements of a company in the year 2012, we want to estimate the probability of default for the year 2013). We select companies with financial data available for two consecutive years, knowing that the first year the company has to be a non defaulted company. This means that we check if the *Out of business* variable in to *No*), and then, if in the second year the company’s *Out of business* = *Yes*, then we set the target variable to 1; otherwise, we set the target variable to 0.

In Fig. 4.1 we observe that the years with around 100K companies assessed, the default rate is in the range [1.5%-2%]. During the period between 2015-2018, the amount of assessed companies is significantly lower than the other years, and consequently the % of defaulted companies varies heavily.

All original financial variables (see Table 4.7) are important for predicting whether a company will incur into a default. Since the number of missing values is significant (see Table 4.8), We keep those companies we have all financial variables needed to compute the ratios in Table 4.9.

For analyzing the results of the interpolation, we remove companies with missing information. We select the feature  $i$  that will be interpolated. We split the data into two datasets: a first set (95% of the data) with non-missing values and a second dataset whose feature  $i$  is full of missing values. We interpolate feature  $i$  using kNN [109]. Finally,

---

<sup>4</sup><https://www.tinubu.com/>



Figure 4.1: For a given year  $t$ , the default rate represents the percentage of defaulted companies in year  $t + 1$  over all companies rated the year  $t$  and  $t + 1$ .

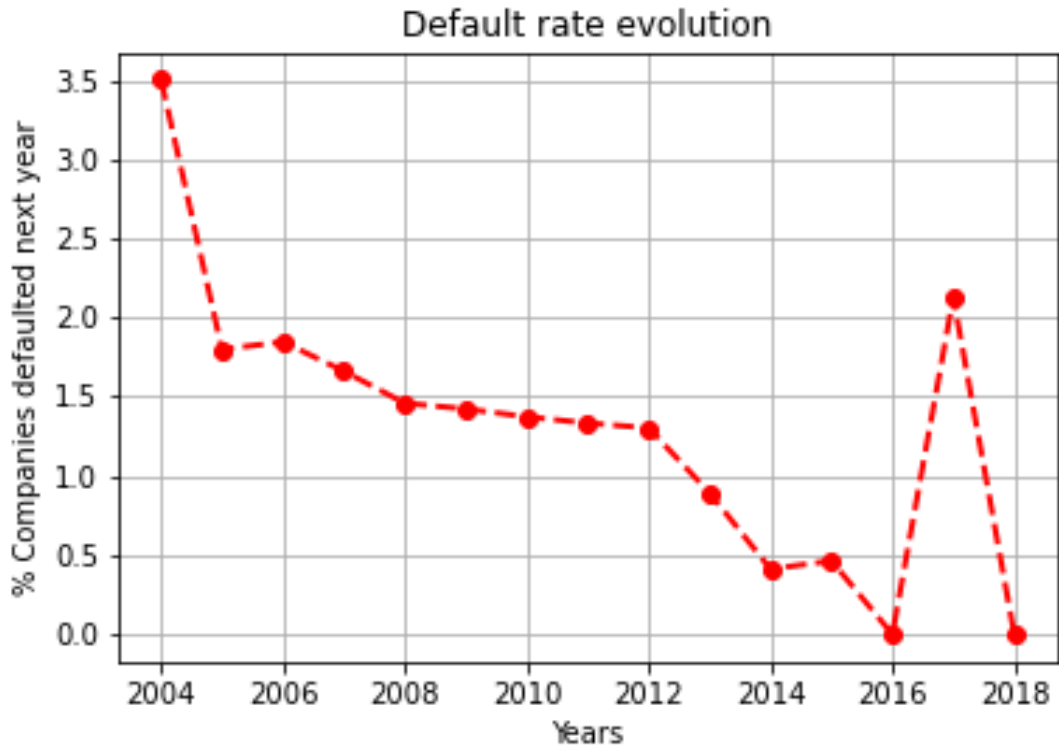


Figure 4.2: Volume of missing values for the original data provided by Tinubu and used for computing the financial ratios.



Table 4.2: Financial ratios used in the financial industry. These ratios are the inputs of Tinubu’s Scorecard Rating Algorithm and the ML models used to predict the default of a given company.

Features	Definition
Solvency $R_1$ :	$\frac{\text{Net Worth}}{\text{Total Assets}}$
Solvency $R_2$ :	$\frac{\text{Financial Debt}}{\text{Gross Income}}$
Liquidity $R_1$ :	$\frac{\text{Total Current Assets}}{\text{Total Current Liabilities}}$
Liquidity $R_2$ :	$\frac{\text{Cash Liquid Assets}}{\text{Sales}}$
Profitability $R_1$ :	$\frac{\text{Working Capital}}{\text{Sales}}$
Profitability $R_2$ :	Net Income
Profitability $R_3$ :	$\frac{\text{Gross Income}}{\text{Total Assets}}$
Time in business	Assessment year - incorporation year
Sales evolution	Current sales - previous year sales
Country code	Country codification

we score the company with Tinubu’s algorithm and compare the rating obtained for the interpolated company and the original rating of the considered company.

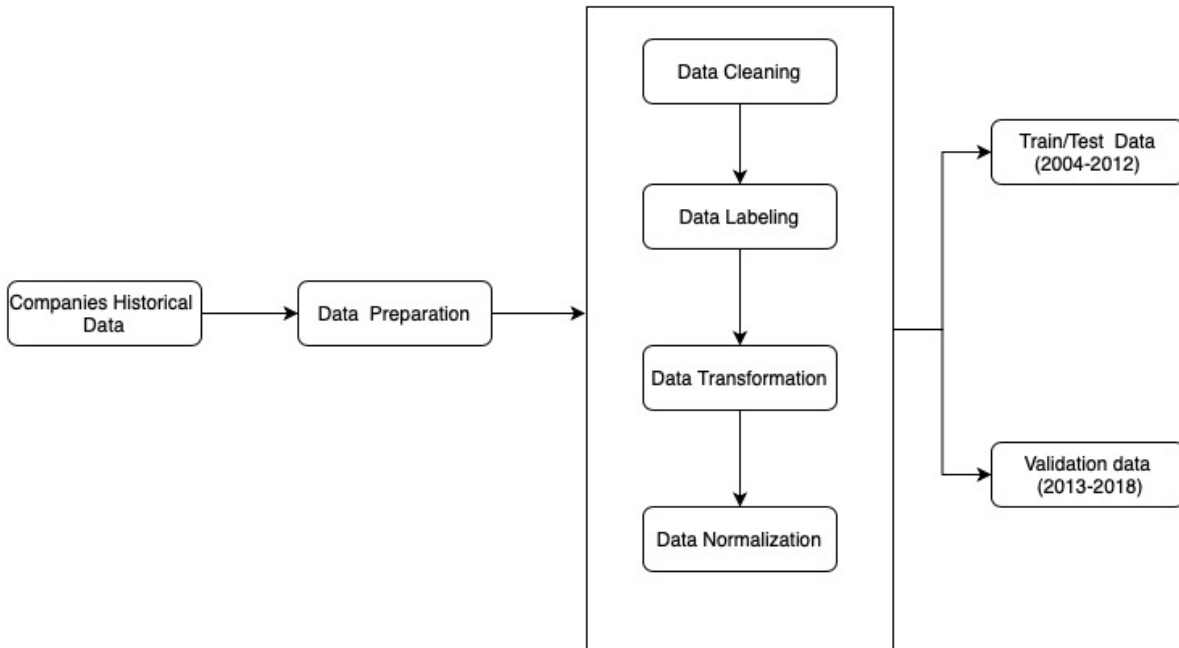
*Data Transformation:* First, we start by encoding the categorical features. The only categorical variable present in Table 4.7 is the country code. We use the one-hot encoding technique to create a new column for each country. This variable will take the value 1 if the company is located in the considered country (when a model is built per country), and 0 otherwise. We split the dataset into two main sets: the first set contains data between 2004-2012. We use this data to train and test (70% for training and 30% for testing). The data between 2013-2018 will be used to validate our model.

*Data Normalization:* The transformed data in Table 4.9 contains noise and has a different scale. We scale the data using the Standard Scaler [110]. This helps to reduce the noise by transforming the data distribution into a new one with a mean 0 and standard deviation of 1.

*Data Oversampling:* The SMOTE oversampling technique [102] consists of oversampling the minority class by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the  $k$  minority class nearest neighbors. Depending on the amount of over-sampling required, neighbors from the  $k$  nearest neighbors are randomly chosen. We resample the training set using SMOTE with the parameter  $k=10$  and the ratio between the minority class and majority class in the resampled set to 0.5<sup>5</sup>.

<sup>5</sup>After trying different combinations for both hyperparameters, these values obtain the best results for the tested ML models.

Figure 4.3: Data preparation flowchart including the years over which train and test data are split.



### 4.3.2 PD Modeling using Machine Learning Models

In this part, we present the classification models used to predict the probability of default. The output of all models is a probability of default. We reduce our problem to a binary classification model, we establish that if this probability is greater or equal to 0.5, then the model is predicting that the company will be in default the next year.

*Logistic Regression (LR)*: a linear model that makes a prediction by computing a weighted sum of the input features. The output of this model is a probability for binary classification (whether the company will fall into default or not). This probability can be mapped by setting a threshold. The threshold used for predicting whether a company will be in default is 0.5. If the probability is greater than 0.5, then the model predicts the company as defaulted, i.e., bankrupt. Logistic regression has been widely used in the credit risk prediction domain due to its simplicity and interpretability [101].

*AdaBoost (AB)*: A tree-based ensemble method whose algorithm trains a decision tree and then tries to fix the errors by training sequentially decision trees over the predecessor tree errors. For the hyperparameter optimization step, we use Grid Search 5 fold cross-validation. We obtain the best results with a *Learning rate*=0.8 and *n estimators*=100.

*Random Forest (RF)*: this algorithm generates trees from a random sample of the training data. For each generated tree, the algorithm randomly selects an attribute for splitting the tree. This randomness reduces the overfitting of the model by decreasing

the correlation between trees. The output of a Random Forest is the average of all trees predictions. Between all the parameters tested using Grid Search 5 fold cross-validation for the training set, best results (5.1) were obtained with  $n\ estimators=1500$ .

*Gradient Boosting (XGBoost)*: XGBoost consists of a sequential combination of weak learners in order to create a robust model. The difference between AdaBoost and Gradient Boost is that the drawback of having weak learners is detected using gradient descent. A variant of the Gradient Boosting, the extreme Gradient boosting has been used in this work. This variant uses a more regularized model formalization to control overfitting. In Table 5.1 results associated to the XGB model were obtained using the hyperparameters:  $learning\ rate=0.1$ ,  $n\ estimators=100$ ,  $max\ depth=10$ ,  $subsample=1$ ,  $colsample\ bytree=1$  and  $gamma=0.7$ .

All methods employed used the implementations provided by Scikit Learn [110]. XGBoost uses the one concretely from [98]. We compare the performance of the different machine learning models using the following classification metrics: precision, recall and the Area Under the Receiver Operating Characteristics (AUC).

### 4.3.3 Data Oversampling using SMOTE

The SMOTE oversampling technique [102] consists on oversampling the minority class by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the  $k$  minority class nearest neighbors. Depending on the amount of over-sampling required, neighbors from the  $k$  nearest neighbors are randomly chosen. We resample the training set using SMOTE with the parameter  $k=10$  and the ratio between the minority class and majority class in the resampled set to 0.5<sup>6</sup>.

- *Accuracy*: defined as the ratio of correct predictions
- *Precision*: the proportion of correctly predicted classes over the total of data points.
- *Recall*: the proportion of defaulted companies that are correctly predicted by the model.
- *F1-Score*: the weighted average of the precision and recall.
- *AUC*: measures model's ability to discriminate between cases. It is the area under the Receiver Operating Characteristic.

---

<sup>6</sup>After trying different combinations for both hyperparameters, these values obtain the best results for the tested ML models.

### 4.3.4 SHAP for Model Explanations

To understand the outputs of the ML model we employ SHAP [64], a framework for interpreting model predictions. SHAP uses a game-theoretic approach that explains the contribution of each feature to the final output of a given model. This method will ascertain which financial features are the most relevant to predict the default of a company.

### 4.3.5 Human Expertise Alignment: Introducing Credit Risk Analysts Expertise

We asked several Tinubu’s risk analysts (to be more precise 4 different risk analysts) to weigh what variables are the most important to rate a company. We asked them to distribute 100 points between all the variables that have been used for training our ML model. Then we compute the sum of all weights given by the different risk analysts. This sum represents the level of importance for each feature. This value will be used to rank the features by importance degree and it allows us to compare the human expert ranking with the feature importance for our model given by the SHAP value. The main point of this article is to compare the explanations given by the ML model (i.e., in this case the XGBoost) after applying the SHAP [64] framework with the credit risk analyst opinion. The way this comparison has been conducted is as follows: we asked several Tinubu’s risk analysts (to be more precise 4 different risk analysts) to weight what variables are the most important in order to rate a company. We asked them to distribute 100 points between all the variables that has been used for training our ML model. Then we compute the sum of all weights given by the different risk analysts. This sum represents the level of importance for each feature. This value will be used to rank the features by importance degree and it allow us to compare the human expert ranking with the feature importance for our model given by the SHAP value.

## 4.4 Results

In this section, we present the results of the models we described in the previous section for different settings. Then we compare the results of the best model with Tinubu’s Rating System. Finally, we discuss the differences between the decision-making process of our ML model, Tinubu’s mathematical Credit Rating System, and the credit risk experts.

### 4.4.1 Analyzing the Impact of Interpolation in the Credit Score

We analyze the impact of using an interpolation method based on kNN algorithm [109]. In this case, we consider the companies with all features available. We remove around 5%

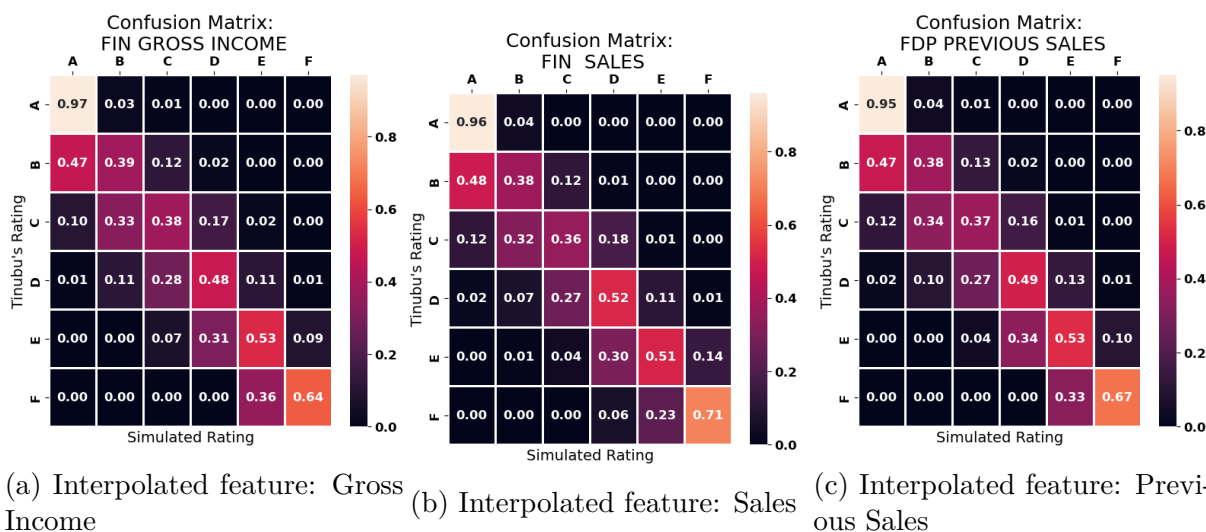


Figure 4.4: The columns of the confusion matrix represent the rating given by the Tinubu’s system after the feature interpolation. Rows correspond to the rating given by Tinubu’s algorithm with the original feature

(the number of missing values of the considered features, see Fig. 4.2) of a given variable. Then we interpolate using kNN with  $k = 2$ . The number of neighbors chosen is the one that minimizes the number of companies that after the interpolation has changed their ratings using Tinubu’s credit scoring algorithm. At this point, we run Tinubu’s algorithm and analyze the change of ratings for the companies that have been interpolated. Results in (Fig. 4.4a, Fig. 4.4b and Fig. 4.4c) mainly show that the risk score of interpolated companies tends to be slightly worse. However, it is interesting to keep in mind that we prefer a more conservative score.

#### 4.4.2 Performance of ML Models for PD Modeling

In Table 5.1, we compare several ML-models using different data strategies: without resampling (**WRS**), resampling the training data with SMOTE (**RS**) and resampling the training data and measuring the model performance over the validation dataset (**RS+VS**). Analyzing Table 5.1, we observe that the ML models do not recognize the companies that will default the next year if we train them with the original data. However, there is an improvement for models trained with generated synthetic data of the minority class (i.e., the defaulted companies). What we notice is that for the validation set, which is composed of companies between 2013-2018, the XGBoost performs better than all other tested models. The results in Table 5.1 show that the XGBoost model can to detect defaulted companies with higher precision (i.e., recall) than the rest of ML models.

Table 4.3: Models performance using the features in Table 4.9.

Model	Performance Metrics		
	Precision	Recall	AUC
LR (WRS)	0	0	0.6736
LR (RS)	<b>3.58</b>	7.44	0.6876
LR (RS+VS)	0.80	4.70	0.7292
AdaBoost (WRS)	0	0	0.7263
AdaBoost (RS)	3.18	28.18	0.7058
AdaBoost (RS+VS)	1.37	4.70	0.7324
Random Forest (WRS)	0	0	0.6551
Random Forest (RS)	2.82	12.85	0.7086
Random Forest (RS+VS)	0	0	0.6908
XGBoost (WRS)	0	0	0.6728
XGBoost (RS)	2.92	<b>30.09</b>	0.7027
XGBoost (RS+VS)	1.22	15.29	<b>0.7466</b>

### 4.4.3 Mapping XGBoost Probabilities to Tinubu’s Grades

Tinubu’s Scorecard Algorithm is the internal proprietary rating system used at Tinubu Square to evaluate the credit risk of a debtor (a company). This rating is a descriptive way to present the probability of default of the assessed entity. Tinubu’s rating scale uses letters to establish the level of credit-worthiness of a company. On one hand, *A* is given for companies that the model estimates the probability of default is close to zero. On the other hand, companies considered very likely to incur into default are rated with letter *F*. The *X* score is given to companies whose data needed to compute the PD is not available.

Table 4.4: Rating scale for tinubu’s algorithm.

Rating Scale	Probability of default level
A	Little/no default risk
B	Low default risk
C	Average default risk
D	Above average default risk
E	Increasing and high default risk
F	Extremely high default risk or in default
X	Excluded companies from the analysis due to lack of data

Our goal is to compare the two models: Tinubu’s Scorecard Algorithm and the best machine learning model (i.e., the XGBoost model). Nevertheless, the outputs of both models (Tinubu Scorecard Algorithm and ML model) are naturally different. The Tinubu Scorecard Algorithm model outputs a category, while the ML model outputs a continuous variable (probability of default). To compare both models we need to create a mapping to assign to each score letter a probability of default. As we consider the gold standard of the financial experts model of Tinubu our ground truth, this PD will need to be matched by the ML model output. Given a Tinubu’s rating class we compute the average of the probabilities of the companies being in default next year (i.e.,  $Y_t = 1$ ).

$$\mu(R) = \frac{1}{N} \sum_{i=0}^{n-1} \mathbb{P}_i(Y_{t+1} = 1 \mid Rate = R) \quad (4.1)$$

where  $R$  is the rate yielded by Tinubu's Scorecard Algorithm and rating  $R \in \{A, \dots, F\}$ .  $\mu(R)$  represents the mean of the probabilities of default given by the XGBoost model when the yielded rate by the Tinubu's Scorecard is  $R$ .

$$\hat{R} = \operatorname{argmin}_R \left\{ \mathbb{P}(Y_{t+1} \mid X_t) - \mu(R) \right\} \quad (4.2)$$

Then for each company, we estimate the rating by searching for the rate that minimizes the difference between the PD of the company given by the ML model and the average probability estimated by the ML model of all different Tinubu's Rating classes (see equation 5.2). The estimated rating  $\hat{R}$  represents the rating yielded by the XGBoost after the mapping process.

Table 4.5: Scorecard mapping: we assign a Tinubu-defined credit risk score to each probability yield by XGBoost.

Tinubu Score	$\mathbb{P}(Y_{t+1} = 1)$	
	Interval lower bound	Interval Upper Bound
A	0	0.0828
B	0.0828	0.1411
C	0.1411	0.2029
D	0.2029	0.2486
E	0.2486	0.285
F	0.285	1

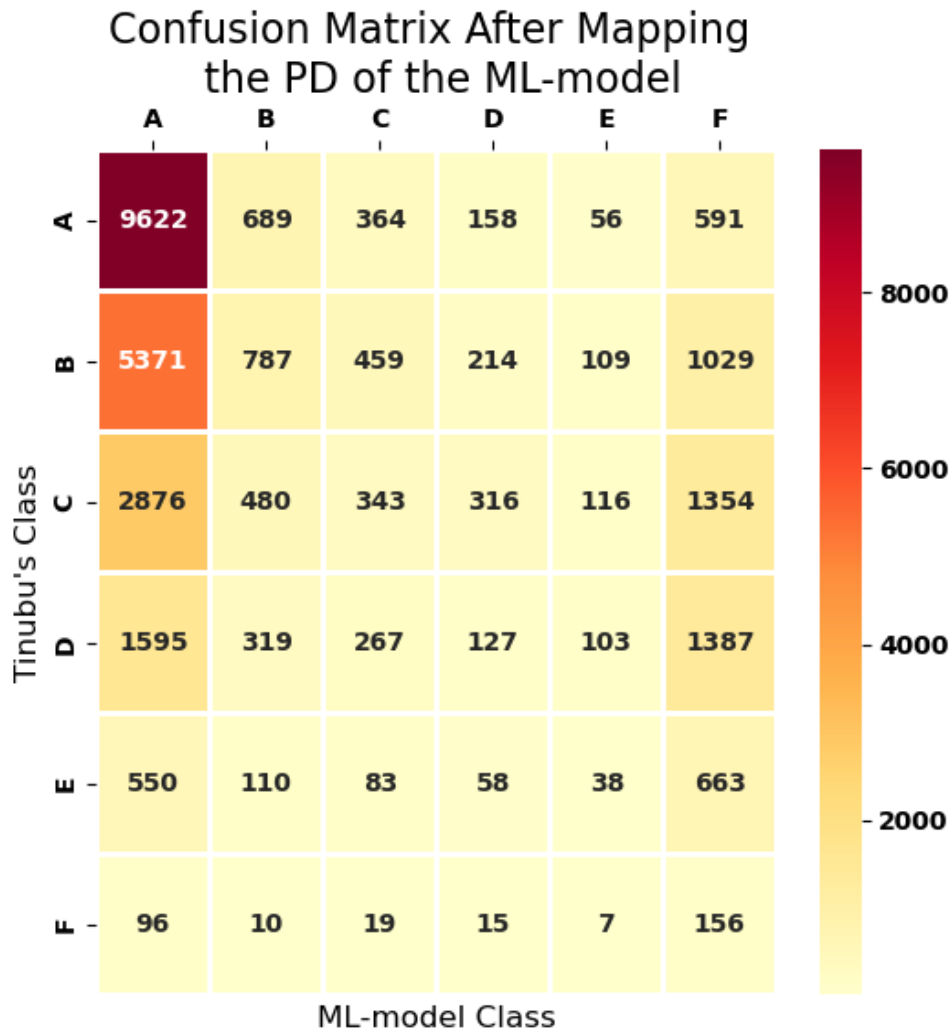
In Figure 4.5 we observe that the ML model is able to assign correctly low-risk labels to companies (A and B labels). However, the ML model overestimates the risk of a significant number of companies. This is shown in Table 4.5 in column F. In the context of Credit Scoring is important to remark the impact of underestimating the risk. The critical case occurs when the model estimates that a company has a low or near zero risk of being in default the year after the assessment, and the reality is that such company is much riskier than predicted by the model. This is why we consider that the XGBoost model works relatively well, in comparison with a well established Rating Algorithm (Tinubu's Scorecard Algorithm) that has been rating companies since early 2000s and is still being used nowadays.

#### 4.4.4 Explaining our PD model: SHAP Value Analysis

We analyze how the ML model has arrived at the results shown in Tables 5.1 and 4.5. Features in Fig. 5.4 are sorted according to their relevance (i.e., SHAP mean absolute



Figure 4.5: Confusion matrix for mapping Tinubu’s Scorecard algorithm risk labels to the ML model probabilities.

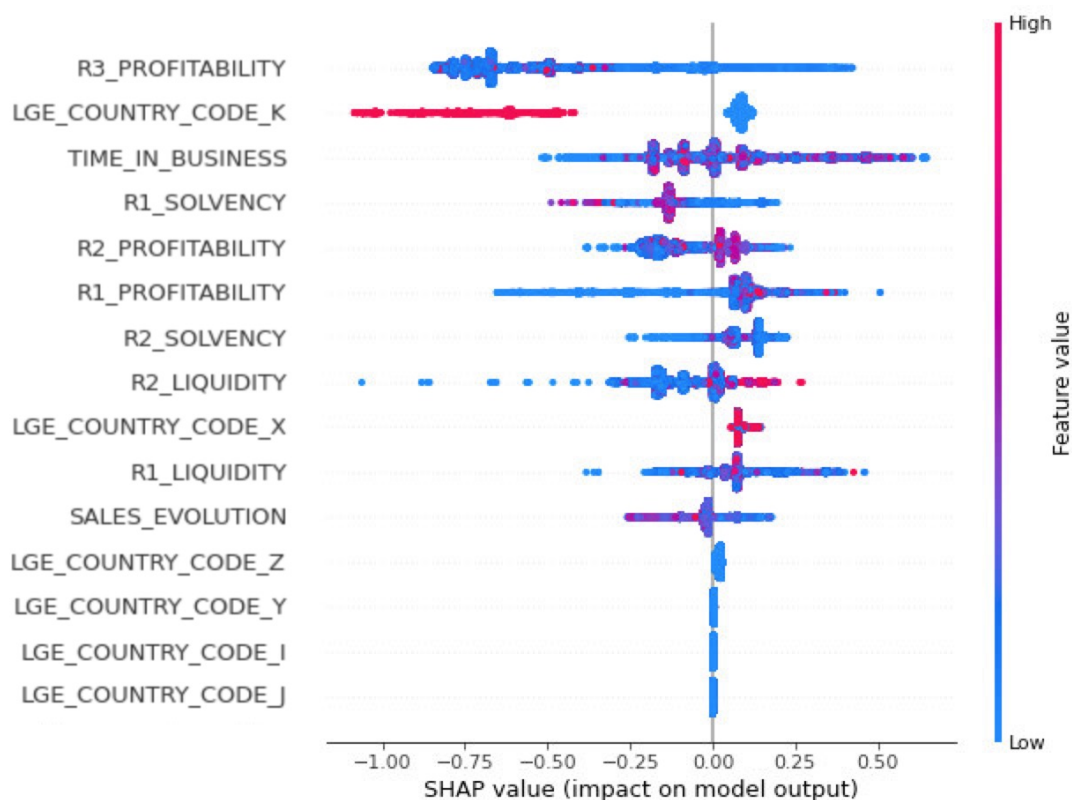


value). The most relevant (R3 profitability ratio) to the less important ratio (country code J). This analysis highlights the fact that, for the ML model (i.e., the XGBoost model), companies based in country K (i.e., the *Legal Entity (LGE) COUNTRY CODE K* binary variable is equal to 1) have a lower probability of default than companies based in other analyzed countries. This fact matches with the way Tinubu’s algorithm works. For companies with financial ratios relatively similar, the Tinubu’s Scorecard Algorithm will give a better rating to those companies based in country K. The contribution of a given feature whose data points lie next to other data points of opposite color in the same axis show that both high and low values of that feature have influenced similarly the model outcome, and thus their contribution may be contextually better determined by considering generic values of other out-weighting features that have a stronger contribution for that data point.

#### 4.4.5 Feature Contribution Analysis: Assessing Explanations from risk analyst experts vs ML models

To assess the explainability of our ML model with the respective mathematical model of Tinubu Square, we asked 4 different analysts from Tinubu Square to give a weight for all features in Table 4.9. Each analyst has 100 points to distribute between the features. All participants of this survey filled the Table independently. As all analysts work at Tinubu Square, we expect relatively uniform criteria. Comparing the weights given by the analysts in Table 4.6, and the importance of each feature for the ML model (see SHAP analysis in Fig. 5.4), we conclude that the decision-making process diverges among machine and human models. Generally, the assessment of the credit-worthiness of a given company is made by a human expert if the amount of credit demanded by the assessed company is relevant. Therefore, the dataset of companies treated by the risk analyst may be biased towards companies that can afford larger interest rates or higher insurance premiums.

Figure 4.6: Contribution of each explanatory feature to the final prediction based on Shapley analysis of contribution decomposition for the default prediction. Features are sorted according to their relevance (i.e., SHAP average absolute value)



For all analysts the  $R_3$  Profitability feature is irrelevant, while for our ML model is the most important one. Analyzing the most important features for both the human-expert and the ML model we discover that the only feature considered relevant when assessing a company credit worthiness for both human experts and the ML model is the

$R_1$  *Solvency*, which measures the ability of a company to meet short-term debts. From a credit risk analyst stand point, the most important features are those related to short-term activities (i.e.,  $R_1$  *Liquidity*,  $R_2$  *Liquidity* and  $R_2$  *Solvency*). However for our ML model, the features that contribute the most to the model output are the features that relate to activities extended in time, i.e., long term capabilities of the firm, for instance,  $R_3$  *Profitability*, *Time in Business* and *location*.

Table 4.6: Weight of each feature given by 4 different risk analyst expert (i.e, R.E) at Tinubu Square. The risk analyst expert distribute 100 points between all the features. Features are ordered by decreasing importance for the risk analysts (i.e., total points the feature has received from all experts)The sum of weights given by the analysts is the same for  $R_2$  Solvency and  $R_2$  Profitability. We ranked  $R_2$  Solvency above  $R_2$  Profitability because the weights are similar.

Features	R.E 1	R.E 2	R.E 3	R.E 4	Total Weight
$R_2$ Liquidity	20	30	30	20	90
$R_1$ Solvency	25	10	25	20	80
$R_2$ Solvency	5	10	25	15	55
$R_2$ Profitability	10	30	5	10	55
$R_1$ Liquidity	15	10	15	5	45
Sales evolution	5	3	1	10	19
Country code	10	2	2	5	19
Time in business	5	5	2	5	17
$R_1$ Profitability	5	0	5	5	15
$R_3$ Profitability	0	0	0	5	5

In Table 4.6 we rank the relevance of the features depending on the risk analysts. The latter act as a proxy of the gold standard credit risk opinions. Their expertise is captured by the Tinubu internal Scorecard Algorithm.

## 4.5 Discussion

The first issue concerns the usage of a highly imbalanced dataset. It is worth noting that the task of predicting the probability of default entails having to deal with the bias inherent to the nature of the data since the percentage of default companies is very low concerning successful companies

Therefore the ML model may have problems differentiating between both defaulted and non-defaulted companies. Since defaults are very rare events, resampling data by generating new synthetic data of the minority class (SMOTE) may not improve models' performance because each event of default has its characteristics in a particular context, and the concept of default may differ from one country to another. It is worth highlighting the comparison between the Tinubu's Scorecard algorithm and the ML model (i.e., the XGBoost model).

The comparison between Tinubu's Scorecard algorithm and the ML model shows that

the ML model does underestimate the number of highly ranked firms (i.e., ranked with top score  $A$ ,  $B$  and  $C$  by Tinubu’s Scorecard Algorithm). On the other hand, we observe that the ML (XGBoost) model rates a considerable number of companies with the  $F$  score, while the Tinubu Scorecard Algorithm maps the same companies mainly to scores in the range  $[B-D]$ . This behavior can be interpreted as the ML model is significantly more conservative than Tinubu’s Scorecard Algorithm. The property of being risk-averse is therefore desirable in our context, and thus, beneficial due to safety reasons.

The comparison showed that the criteria between human experts and the ML model are quite different. In particular:

- a. The ML model excels at being able to capture the longer term abilities of a company, for the features considered by the Tinubu risk analyst experts. The latter focuses more on shorter-term variables such as  $R_1$  *Liquidity*, or  $R_2$  *Liquidity*, while the ML model attributes higher relevance to life-long attributes of the firm.
- b. We found interesting the fact that, without explicitly implementing any constraint, XGBoost arrives at the conclusion that companies that exercise their activities in Country K are more likely to avoid financial problems, while this behavior is explicitly encoded and accounted for by Tinubu’s Scorecard Algorithm.
- c. Since credit risk scoring algorithms, especially those focused on companies rather than individuals, are in the very early stages, we acknowledge the abilities of our ML model to remain conservative when estimating risk. This is a desirable property of such complex models, since it is preferable to avoid critically large economical losses

## 4.6 Conclusion and Future work

Several state-of-the-art machine learning models have been proposed to model the probability of default of a company the year after it has been assessed. For the Tinubu dataset used, we remark the importance of dealing with a high imbalanced dataset.

In order to evaluate our ML model’s PD, casted as a regression problem, with risk analyst experts categorical scores predicting a PD in form of a score, we mapped the probabilities given by the best ML model (i.e., XGBoost, as it showed best results) to compare our ML model with to Tinubu’s Scorecard model score labels. This mapping shows that the ML model is able to tell apart companies with low risk of default (i.e., companies rated with an  $A$  and  $B$  with the Tinubu’s Scorecard Algorithm) with respect to from the rest of companies.

One of the biggest challenges for the introduction of machine learning based models in the credit scoring field, in particular for companies credit scoring, is the lack of credibility, trust, and explainability. We addressed this problem by using the explainable framework of SHAP analysis. [64]. Assessing the results of the SHAP analysis, we conclude that the difference between companies based on the country K and the rest of the companies may require a deeper analysis, since the hypothesis of different country regulations could affect the companies default analysis.

The analysis of explanations given by a SHAP analysis of feature contributions concerning explanations backed up by the expertise of a pool of credit analysts showed a certain divergence between what variables should be considered more relevant when assessing a company.

Future works should focus on studying how inductive biases can infuse expert knowledge into the ML model, for instance, by introducing credit expert opinions and preferences (at the data annotation and model design stages) to improve the model performance. Other potential avenue of research is designing a sound basis for causal explainability that can be verified and certified by human experts. Moreover, it may be interesting to further analyze the geographical locality context, i.e., contextually and historically, by assessing companies both by risk analysts and the ML model. Finally, we hope future work designs explainability metrics to programmatically assess the quality of an explanation given by a black-box model that learns, evolves and degrades over time, to continuously assess its fidelity and alignment with human expert opinions. This way we will be able to deploy AI systems that both humans and experts can mutually improve, support, and trust.

Table 4.7: Definition of original financial variables. This variables are used to compute the financial ratios in table 4.9.

Feature name	Feature Description
LGE ID	Identification number. This value is unique for each company
Statement date	Corresponds to the date in which the financial data was published
Out of business indicator	Binary variable: Yes if the company is currently defaulted
Country code	Abbreviation of the country in which the company is based
Total Employees	Number of employees
Net worth	Total amount of Equity
Total Assets	Refers to the total amount of assets owed by the entity
Gross Income	Amount of money earned before taxes
Total Liabilities	Combined debts a company owes
Current Ratio	A liquid ratio that measures the ability of a company to pay short-term
Cash and Liquid Assets	Refers to assets that can be readily convert to cash
Sales	Net sales for the period after returns, allowances, and discounts are deducted
Working Capital	Capital of the financial activity period
Net Income	Amount left over after all expenses and taxes are deducted
Incorporation Year	The year the business incorporated
Previous Sales	Financial statement date

Table 4.8: Percentage of missing values for each financial variable by year.

Financial Variable	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Total Employees	6.23	6.1	10.5	21.22	25.51	27.94	32.05	38.37	37.76	35.6
Net worth	0.77	6.94	7.94	5.57	0.97	0.54	0.71	0.23	0.1	0.07
Total Assets	0.84	6.93	7.61	5.58	1.03	0.66	1.13	0.85	0.79	0.78
Gross Income	14.74	14.35	18.84	13.31	8.99	12.4	13.13	12.03	13.78	13.75
Total Liabilities	0.87	6.97	7.49	5.27	1.18	0.86	1.37	1.06	1.04	1.01
L1 Ratio	77.59	82.23	77.82	80.47	40.12	11.43	6.47	4.43	4.61	4.04
Cash and Liquid Assets	4.44	10.74	11.12	8.95	4.48	4.11	4.53	4.47	4.28	4.03
Sales	12.5	8.7	11.79	10.02	6.72	8.82	10.92	9.1	10.12	10.47
Working Capital	7.79	13.27	30.49	52.29	58.9	60.1	63.7	69.33	70.87	71.69
Net Income	6.37	10.16	13.76	10.53	6.07	8.95	9.15	7.71	8.5	8.76
Previous Sales	18.53	14.99	18.49	17.36	12.3	15.2	14.4	11.74	12.52	12.41

Table 4.9: Financial ratios used in the financial industry. This ratios are the inputs of Tinubu's Scorecard Rating Algorithm and the ML models used to predict the default of a given company.

Features	Definition	Description
Solvency $R_1$ :	$\frac{\text{Net Worth}}{\text{Total Assets}}$	Measures enterprise ability to meet current debt obligations. High $S_1$ values is indicative of greater solvency
Solvency $R_2$ :	$\frac{\text{Financial Debt}}{\text{Gross Income}}$	Represents the percentage of the gross income that goes to debt payments
Liquidity $R_1$ :	$\frac{\text{Total Current Assets}}{\text{Total Current Liabilities}}$	The current ratio measures the ability to pay short term obligations (within one year)
Liquidity $R_2$ :	$\frac{\text{Cash Liquid Assets}}{\text{Sales}}$	Liquidity indicator that represents the percentage of liquid assets over the revenues of the company
Profitability $R_1$ :	$\frac{\text{Working Capital}}{\text{Sales}}$	Shows the relationship between the funds used to finance company's activities and the revenues a company generates as a result
Profitability $R_2$ :	Net Income	Is an indicator of company's profitability
Profitability $R_3$ :	$\frac{\text{Gross Income}}{\text{Total Assets}}$	Measures how effectively a company is using its assets to generate earning
Time in business	Assessment year - incorporation year	Years in business
Sales evolution	Current sales - previous year sales	Measures the sales evolution
Country code	Country codification	Country abbreviation in which the company is located



# Sectorial Analysis Impact on the Development of Credit Scoring Machine Learning Models

## 5.1 Introduction

As mentioned previously, in trade finance, the ability to accurately assess the creditworthiness of companies in various sectors becomes increasingly complex during economic downturns (e.g., financial crisis 2008, COVID-19 outbreak).

In [1], the authors find that the main issue that led to the financial crisis of 2008 was the underestimation of mortgage risk of default during the credit growth experienced between 2001-2007. Consequently, financial institutions became more risk-averse, which provoked a substantial increase in the barriers in the process of acquiring credits [111].

The 2008 global crisis not only affected the financial system but also international trade. Concretely, in [112], the authors found that in Europe the emerging economies were severely affected by the financial crisis. In [113], the authors conclude that financial crises also affect negatively in terms of trading costs for those countries with stable and secure trading relations. On the other hand, international trade has been associated with positive impacts on growth for Small, Medium Enterprises (SMEs). As stated in [113], there are two major challenges in international trade: the counterparty risks are relatively high for exporters and importers with limited capacities and resources. SMEs are, in a large set of cases, constrained in terms of working capital. The authors consider that digitalization can increase SMEs' access to trade finance and thus accelerate the growth of the economy.

Financial and insurance institutions can help SMEs to access trade finance by financing



their activities. It is in this context that the Credit Scoring systems play an important role. They are the systems that serve as a backbone for decision-making.

Historically, credit scoring systems have been based on statistical techniques [6, 114, 115]. Recently, machine learning (ML) algorithms have shown an improved capacity for estimating the PD [116] in different contexts. One of the main problems of ML-based approaches is the trade-off between the performance of ML-based models and the explainability given by those models [116, 117].

Previous work on credit scoring has mainly focused on consumer credit scoring. Few researchers have addressed the problem of estimating the creditworthiness of SMEs ([67, 116]). As mentioned previously, this is an important constraint for SMEs to get access to trade finance.

This chapter examines the creditworthiness of a relatively large set, compared to previous literature, of companies based in France. Furthermore, the study conducted in this chapter addresses two main questions:

1. **RQ1. Are sectorial models more appropriate than global models for predicting the company's default?** : We focus our analysis on the economic sectors and the impact on the model appropriateness.
2. **RQ2. What are the features that drive the model to consider a company highly risked for the different sectors? Are those features the same for each sector? Is there a difference between sector models and the global economic model in terms of most relevant features?:** We compare the main factors that lead the model to rate a company by sector and analyze the divergence between economic sectors.

The chapter is structured as follows: first, we present the previous research in the field of credit scoring based on ML models. Then, we explain the methodology we used to solve the problem. In the next section, we show the results obtained using the chosen approach. Finally, we present the conclusion and potential hints for future work.

## 5.2 Related Work

In this section, we present the previous work that has been done in the intersection of ML, credit scoring, and eXplainable Artificial Intelligence (XAI).

In [118], the authors analyze and compare the performance of a LOGIT model and a Support Vector Machine (SVM) model over a dataset composed of Polish companies. Both models tend to have higher accuracy in the training set. This is mainly due to the presence of a significantly low percentage of defaulted companies (i.e., 3%). Similar

work has been done in [119] for Slovakian companies. In this chapter, the authors add the featured sector to their analysis and also analyze the default over two different time frames: one year and two years. They combined two linear models. The results showed that in terms of explanatory variables both models are similar.

Besides these approaches, there has been some interesting research that focuses on the dependence between sectors and countries. Namely the work in [120]. In this chapter, the authors study the interdependence and interactions of economic sectors of several countries (e.g., USA, Russia, and China). By applying the Google Matrix Algorithm, they found that globally speaking there is a strong sensitivity of the chemicals, metals, energies, and food sectors to the price increase of the petroleum sectors. In contrast, in [121] The authors found that the economic sector features (i.e., country and sector features) are not relevant in terms of predicting the default. In [116], the analysis of the sector is introduced to the ML models using Sentence Embeddings and Autoencoders.

Furthermore in recent works [67, 116, 121], they apply a widely used post-hoc XAI method: SHapley Additive exPlanations (SHAP) Values [122]. This method is used to understand the global behavior of the model.

## 5.3 Methodology

We propose an experimental framework that can be split into three different main parts: in the first part we proceed to analyze the raw data provided by Tinubu Square, a company that specializes in providing credit risk opinions to its clients.<sup>1</sup> This study consists of three main stages. The first stage consists of obtaining, from the raw data, the data in the format necessary to create the ML-based credit scoring models.

In the second stage, we compare the performance of six different state-of-the-art ML-based models for predicting a binary class (i.e., whether a company will be in default the year after the release of its balance sheet): Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Random Forest, LightGBM, and XGBoost. The models will be compared using the following metrics: F1-Score, Accuracy and Area Under the Receiver Operating Characteristic (AUROC). To improve the appropriateness of the models, we create a model by economic sector (i.e., mainly the four most represented economic sectors in Tinubu’s dataset). Then, we compare the performance year by year. Furthermore, we compare the results of the best model with the credit rating given by human credit risk analysts by mapping the probabilities of default given by the model with the credit rating accorded by the risk analyst.

In the last stage, we compute the SHAP values for each economic sector model. We compare the model’s global explanations to find common patterns in models’ decision-making.

---

<sup>1</sup><https://www.tinubu.com/>

### 5.3.1 Data Description

In this part, we describe, step-by-step, how to obtain the data that will be used to compute the PD. *Raw Data* is composed of 9 951 981 financial and non-financial information of companies based in Europe, mainly in France. The raw dataset is composed of 81 features.

*Data Cleaning:* In this step, we remove the duplicated data points and remove the features that will not be used during the modeling stage. Since the features related to the sector are encoded using the NAF2 activity codification, we map the code to the name of the sector. We focus the analysis on the companies based in France.

*Data Labeling:* The raw data contains the financial and non-financial information about the companies for the period 2008-2020. We will analyze just those companies for which we have two consecutive financial statements. We will add to the credit record the binary variable *NextyearStatus*. This feature will take the value 1 if the company goes into default the next second year and 0 else. After this stage, the remaining dataset is composed of 171 109 credit records and 32 different financial features.

*Sector Analysis:* First, we analyze the distribution of economic sectors. We analyze graphically the default evolution over the period 2008-2020 for the sectors with more than 10 000 credit records.

*Encoding Categorical Features:* As we mentioned before, we start by creating a model over the whole dataset. For this model, we will need to encode the categorical features. In this dataset, there are two categorical features: the activity code feature and the credit rating given by Tinubu. We encode the activity feature using the One-Hot encoding technique (i.e., for the model that englobes all economic sectors) which creates a new column for each of the categories of the features. The column takes the value 1 if the company operates in the considered sector and 0 else. The rating given by Tinubu is a feature that takes 6 different values: from A (i.e., well-established company) to F (i.e., companies with a high PD). We will use Ordinal Encoding to encode these features. *Splitting the Data:* We will use the 70% for training the models and 30% for testing the models

### 5.3.2 Machine Learning Algorithms for Credit Scoring

In this section, we present briefly the models we employed in this chapter. We also define the metrics we will use to compare the results. The outcome of the models is a probability of default. Since our target variable is binary, we need to map the output of the model to binary distribution. For instance, all models will consider that a company will go into default if the PD is equal to or greater than 0.5.

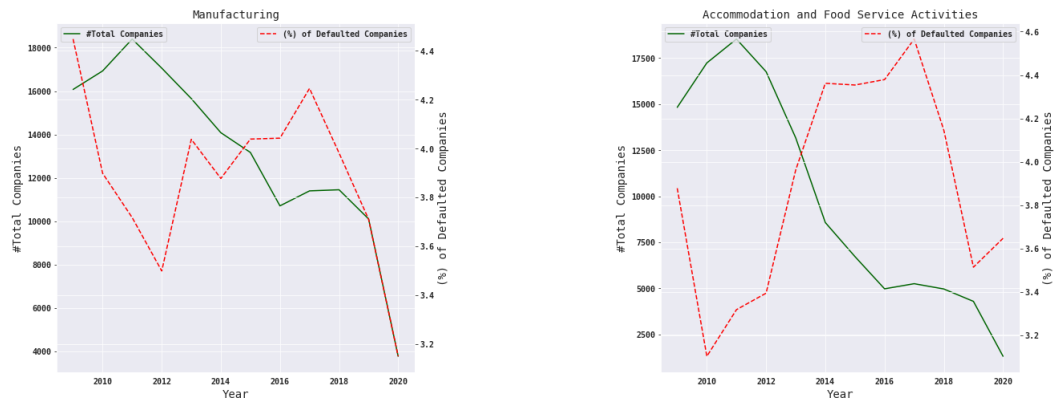
*Logistic Regression (LR):* LR is a linear model whose output is a weighted sum of independent variables. The output of the model is mapped to a probability using the sigmoid function.

Figure 5.1: Companies distribution by sector.



(a) Wholesale and retail trade, repair of motor vehicles

(b) Construction industry



(c) Manufacturing sector

(d) Accommodation and food service activities

Figure 5.2: The right y-axis corresponds to the percentage of default in the sector (red line) and the left y-axis the total of companies belonging to the sector that have been assessed (green line). This analysis shows how the default by sector varies over time.

*k-Nearest Neighbors (KNN)*: KNN is a distance-based algorithm. For predicting an example, it computes the distance between the example and the data points of the training set. Then, it predicts the example by checking the majority class of the  $k$  neighbors.

*Support Vector Machine (SVM)*: SVM is a discriminative classifier that takes training data and finds the hyperplane that best separates the elements of the training set.

*Random Forest (RF)*: Random Forest combines many simple decision trees, using the Bagging procedure, to increase their prediction power. In addition, the trees are generated more randomly, choosing arbitrarily the split variable at each node.

*Light Gradient Boosting (LightGBM)*: LightGBM is a gradient boosting framework based on decision trees to increase the efficiency of the model and reduce memory usage. It uses two different techniques: Gradient-based One Side Sampling and Exclusive Feature Bundling (EFB)

*eXtreme Gradient Boosting (XGBoost)*: The model utilizes the boosting procedure as an aggregation technique: small trees are sequentially added to the model to reduce the loss while keeping the previous trees fixed. Each tree focuses more on the individuals which have been badly predicted from the previous trees.

For all models, we performed hyperparameter tuning using the package hyperopt [123]. For the implementation of kNN, LR, SVM and RF we used the scikit-learn package [124]. XGBoost and LightGBM have been implemented using [20] and [125]. The metrics used for evaluating the performance of the different ML models are: the Area Under the Receiver Operating Characteristic (AUC ROC), F1-Score<sup>2</sup> and the Accuracy.

### 5.3.3 Explaining Model Behavior using SHAP Values

SHAP is a widely used post-hoc explainability framework used for understanding the inner functioning of any model. Developed by Lundberg et al. [122], it is a method that uses a game-theory approach to assign to each feature an importance value for a particular prediction. We will use this framework to compare the explanations given by all the different developed models.

### 5.3.4 Generating Ratings from Models Outputs

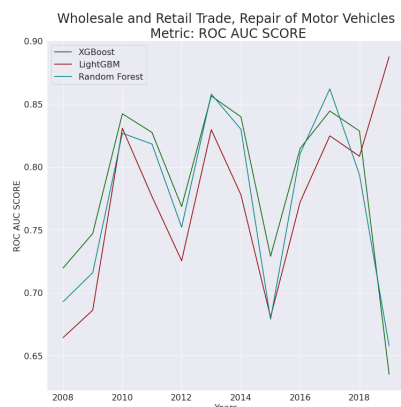
As we mentioned before, Tinubu provides credit risk ratings using labels that range from [A-F]. However, the output of the models is a PD. To compare both we need to map the PD to Tinubu's risk class. In this work, we propose the next mapping:

Given a Tinubu's risk class, we generate a PD for all companies that have been rated with this label. Then, we compute the mean of the PD associated with this rating.

$$\mu(R) = \frac{1}{N} \sum_{i=0}^{n-1} \mathbb{P}_i(Y_{t+1} = 1 \mid Rate = R) \quad (5.1)$$

---

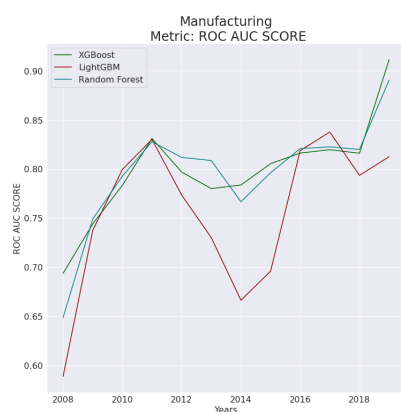
<sup>2</sup>F1 – Score =  $\frac{\text{True Positives}}{\text{True Positives} + 0.5 * (\text{False Positives} + \text{False Negatives})}$



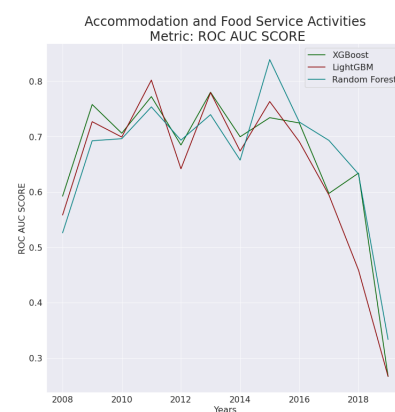
(a) Wholesale and retail trade, repair of motor vehicles



(b) Construction industry



(c) Manufacturing sector



(d) Accommodation and food service activities

Figure 5.3: Comparison of the AUC ROC over the test set for XGBoost, LightGBM and Random Forest over the period 2008-2020

where  $R$  is the rating yielded by Tinubu's Scorecard Algorithm and rating  $R \in \{A, \dots, F\}$ .  $\mu(R)$  represents the mean of the probabilities of default given by the XGBoost model when the yielded rate by the Tinubu's Scorecard is  $R$ .

$$\hat{R} = \operatorname{argmin}_R \left\{ \mathbb{P}(Y_{t+1} | X_t) - \mu(R) \right\} \quad (5.2)$$

Once calculated the mean for each risk class, we estimate the rating for every company by searching which rating class minimizes the distance between the PD for the company and the PD mean for that risk class.

## 5.4 Results

In this section, we analyze the impact of the sector on the percentage of defaulted companies to establish whether the sector variable is important when predicting the default

Figure 5.4: Contribution of each explanatory feature to the final prediction (i.e., the XGBoost model based on Shapley. analysis of contribution decomposition for the default prediction.)



of a company. We present the results obtained over the raw dataset after applying the data preprocessing pipeline shown in the previous section. The output of the most performant model (i.e., the probability of default) will be compared with the rating given by several risk analysts. Since the nature of both risk analyses is different, we proceed to map the probability of default into a credit risk rating. We simulate a scenario in which the most important features for the ML-based model (i.e., the most important features using the SHAP analysis) are missing. Finally, we compare the rating generated by both models, with the original data and the interpolated data, with the rating given by the risk analysis.

### 5.4.1 Default Analysis by Sector

Figure 5.2 shows in fact that the percentage of defaulted companies for the analyzed period (e.g., 2008-2020) differs considerably among sectors. In figure 5.1, we can observe the list of sectors in the dataset as well as the distribution. As we can observe, the percentage of default varies considerably between the analyzed sectors in figure 5.2. In this work, we will focus on the 4 principal sectors: Manufacturing, Construction, Accommodation and Food Services, and the Wholesale Retail Trade and Repair of Motor Vehicles.

### 5.4.2 Model Performance

In table 5.1, we compare the results obtained for the six state-of-the-art machine learning models over the test set. It is important to remark that all these models have been trained using the whole dataset (i.e., all sectors included in the training set). The results

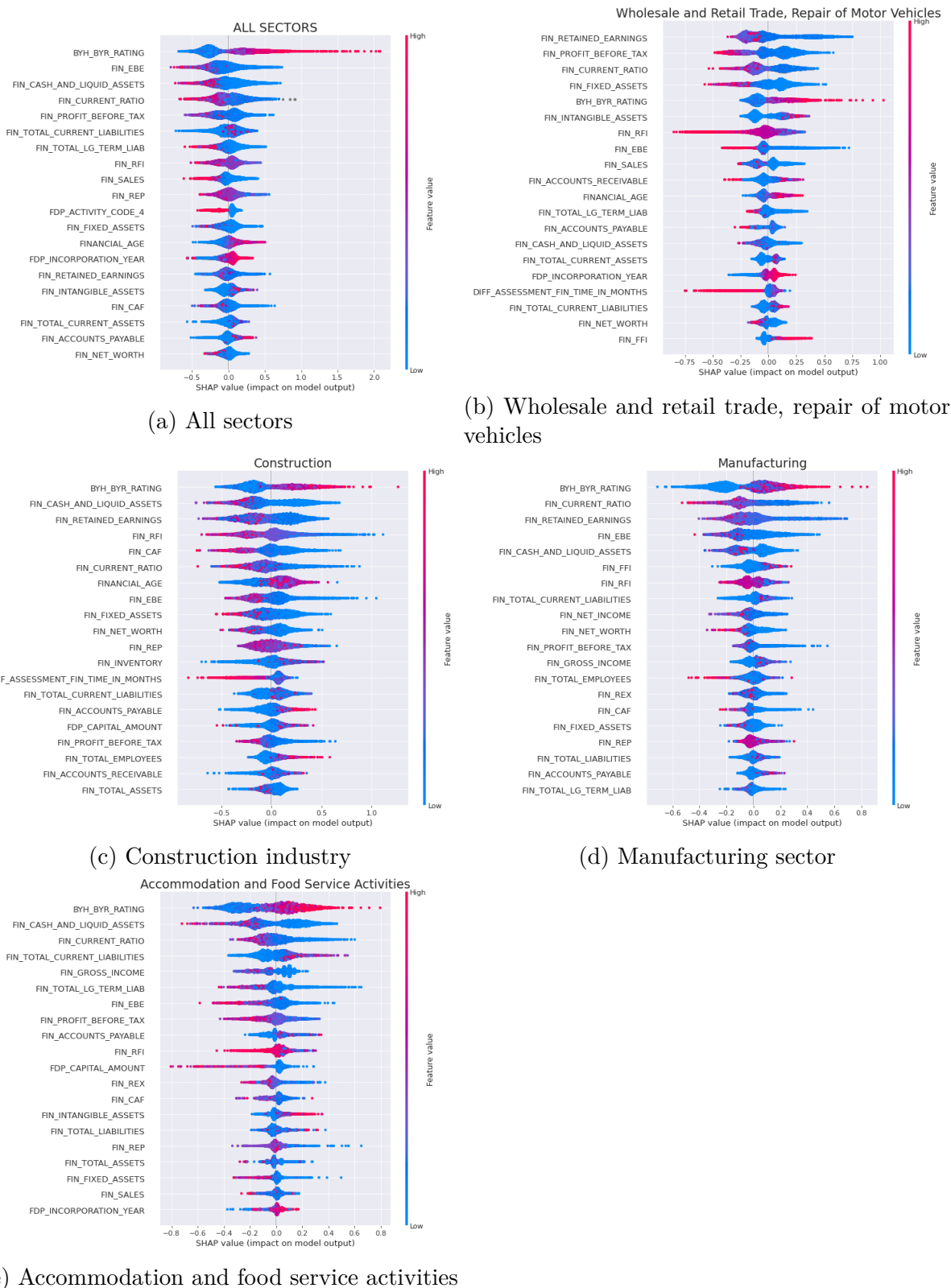


Figure 5.5: Contribution of each explanatory feature to the final prediction (i.e., the XGBoost model) based on Shapley analysis of contribution decomposition for the default prediction for every analyzed economical sector.



clearly show that the ensemble models (e.g., XGBoost, LightGBM, and Random Forest) outperform the classical ML-models. Within the trained ensemble models, the XGBoost yields the best F1-Score.

Table 5.1: Performance of the Machine Learning Models for All Economical Sectors

Model	Performance Metrics		
	Accuracy	AUC ROC	F1-Score
LR	0.695	0.722	0.14
SVM	0.653	0.698	0.07218
KNN	0.954	0.579	0.05703
Random Forest	0.792	0.755	0.18434
LightGBM	0.836	0.764	0.19804
XGBoost	0.913	0.777	0.22732

The ensemble methods outperform the more classical ML models (i.e., SVM, LR, and KNN). For the economic sector analysis, we will develop the ensemble methods we employed previously. In Fig.5.3 we compare the performance of the models for the manufacturing, wholesale, construction, accommodation, and food sectors over the period 2008-2020. We can conclude that the three models used have similar F1-Score being the XGBoost-based model, the one that generally presents the best results for all the different analyzed sectors.

### 5.4.3 Shap Analysis

In this part, we explore and understand the model’s behavior by applying the shap framework. As we can observe, the most relevant features for the model (i.e., XGBoost) for all sectors (see Fig. 5.5a) are the *Tinubu’s Rating*, *Retained Earnings* and the *EBE<sup>3</sup> Cash and Liquid Assets* and the *Current Ratio*. The analysis also highlights how the magnitude of different features

changes the output of the model. For example, the higher the risk according to Tinubu’s rating (i.e., higher values after encoding of the Rating feature) the greater the likelihood of going into default the next year. It is important to remark that, from a risk point of view, it makes sense since the most relevant financial features for the model are the ones mentioned before. These features are related to the company’s capabilities to pay short term debts, and hence, paying its debts.

On the other hand, if we diagnose the explanations given by SHAP values for the 4 different sectors in Fig.5.5, we can see that the main difference, in terms of the most relevant features, is in the wholesale and retail trade sectors (see Fig.5.5b). This sector is the only who does not consider Tinubu’s Rating feature as the most important variable

---

<sup>3</sup>EBE: *Excedent Brut d’exploitation*. It is the equivalent of the EBITDA (Earnings Before Interests Taxes, Depreciation, and Amortization)

for predicting the default. For all other sectors, the model's explanations, in terms of feature importance, vary slightly in comparison to the model that englobes all sectors.

#### 5.4.4 Comparing the Risk Analysts risk scoring with the ML-based Credit Scoring System

As mentioned in previous sections, Tinubu measures the risk using a descriptive way in which if it considers that a company will likely incur a default in the short term, it will rate the company with the letter F. On the other hand, if it considers that the likelihood of the company not repaying its debts in the short term is close to zero, they rate the company with the letter A.

The nature of the ML-based model is quantitative, in other words, it outputs a probability of default. Since both methods are different by nature, we will use the mapping described in the section methodology to compare both the ML model and the rating generated by Tinubu.

Comparing Fig.5.6a, Fig.5.6b and Fig.5.6c, the results show that the mapping is significantly better for the models that have been developed for each sector than the model that englobes all the sectors. Tinubu's extreme classes (i.e., A and F) are better mapped using the sectorial approach proposed in this work.

## 5.5 Discussion

The analysis of the data provided by Tinubu shows that the percentage of companies going into default varies significantly among the different economic sectors. This discriminant factor may play an important role in the early detection of possible company failures at a one-year horizon.

The comparison of the results between a global model and a more granular model by economic sector shows that there is a slight improvement in the ability of the models to discriminate between healthy firms and firms with a high probability of default (see Fig.5.3 vs Table 5.1)). It is important to remark when comparing the ability of the model to reproduce the rating given by Tinubu, that there is an important improvement when developing a sectorial model (see Fig.5.5)) (**RQ1**).

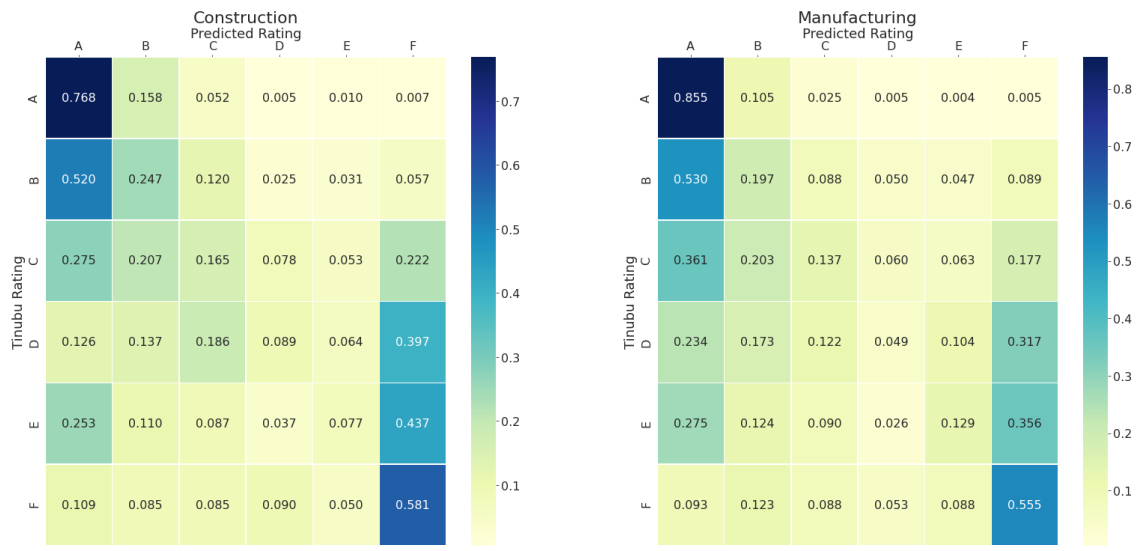
On the other hand, and thanks to the SHAP explanatory framework, it is possible to observe the divergence that exists between the sectorial models in terms of the most important variables. Mainly what we show in this study is that, except for the wholesale sector, all the models consider the Tinubu rating as the most important factor when predicting companies' default. However, for the wholesale and retail trade sector, this variable is the fifth most important variable, which leads us to think that perhaps for that sector the Tinubu rating is not sufficiently discriminating. It is important to note that, as a general rule, the models use more short-term financial variables (i.e., EBE, Retained

# Chapter 5. Sectorial Analysis Impact on the Development of Credit Scoring Machine Learning Models



(a) All Economic Sectors

(b) Wholesale and retail trade, repair of motor vehicles



(c) Construction industry

(d) Manufacturing sector

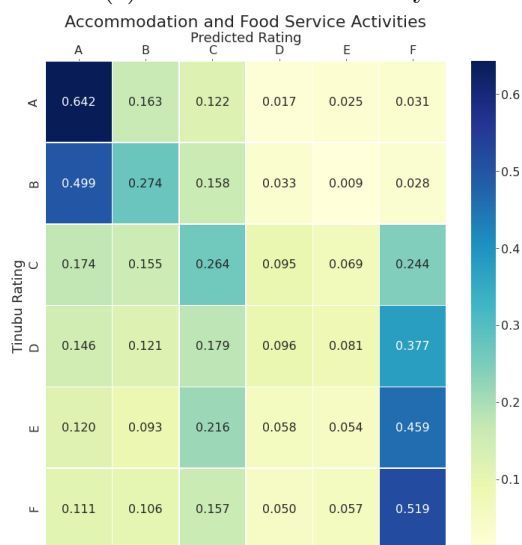


Figure 5.6: Comparison of the Rating given by Tinubu’s Risk Analysts for 4 different economic sectors with the mapped PD of the XGboost model trained using companies of the given sector. The values shown are normalized by rows, and represent the fraction of ratings given by the human risk analyst (i.e., the rows) that matches the rating given by the XGBoost model (i.e., the columns).

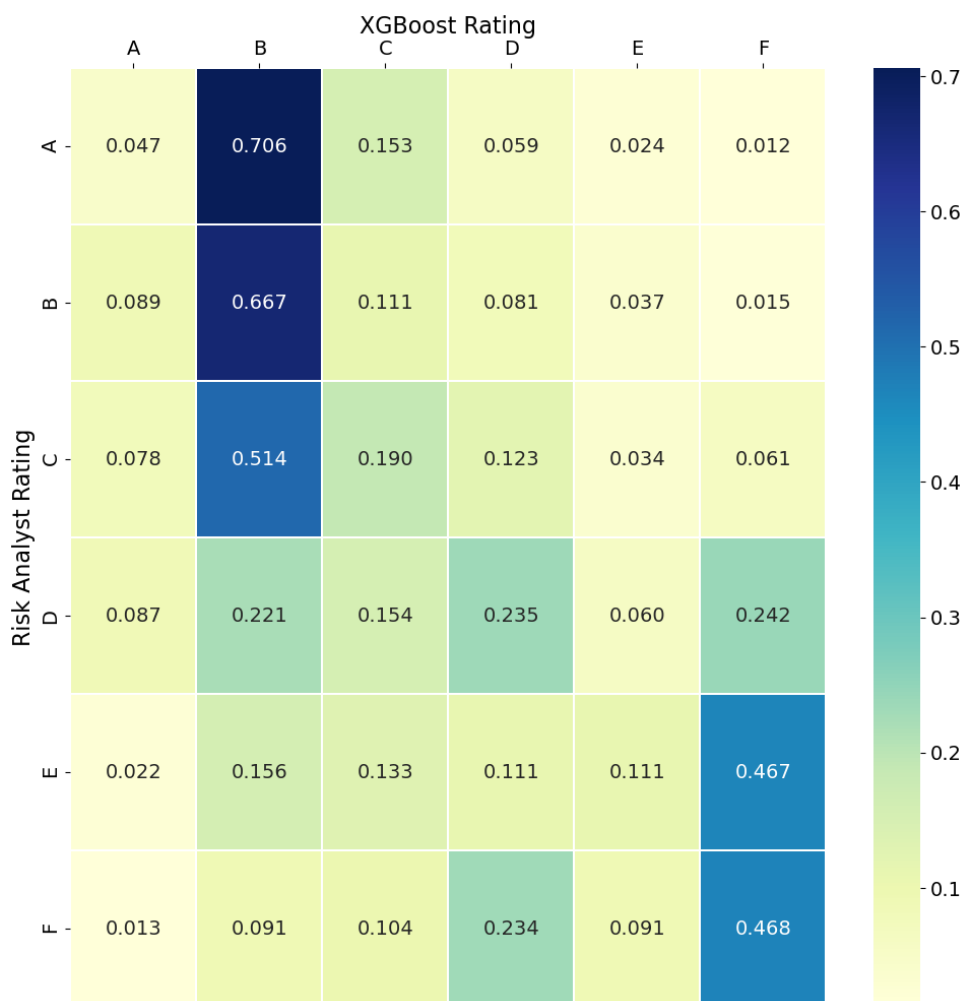


Figure 5.7: Comparison of the Rating given by Tinubu’s Risk Analysts with the mapped PD of the XGboost model. The values shown are normalized by rows. This represent the fraction of ratings given by the human risk analyst that matches the rating given by the XGBoost model.

Earnings, Current Ratio) which can be considered correct analysis from a financial point of view since the default is predicted for 12 months. **(RQ2)**

Last but not least, after applying the mapping proposed in this chapter, the comparison of the sectorial models concerning the global model shows a better alignment with Tinubu’s credit risk model.

## 5.6 Conclusion

In this chapter, we have proposed a different approach for the creation of ML-based bankruptcy models. We have focused on the study of the sectorial behavior of firms to determine the probability of bankruptcy in one year.

Furthermore, we have shown that the models show certain divergences in terms of the variables considered most important for decision-making.

In credit scoring, companies are evaluated using three different analyses: analysis of the company's accounts, analysis of the socio-economic situation of the country to which it belongs, and finally, the analysis of the sector in which it operates. In this work, we have analyzed the company's accounts as well as the sector, focusing on the creation of a model by sector. Future work should expand the framework to include the socio-economic analysis of the country to develop a model that applies to multiple countries.

The analysis in this chapter has focused on the specific analysis of the impact on the economic sector. However, it is important to note that the sectors present a certain dependence among them. Future work should also address the issue of interdependence between sectors.

# Credit Risk Scoring Forecasting using a Time Series Approach

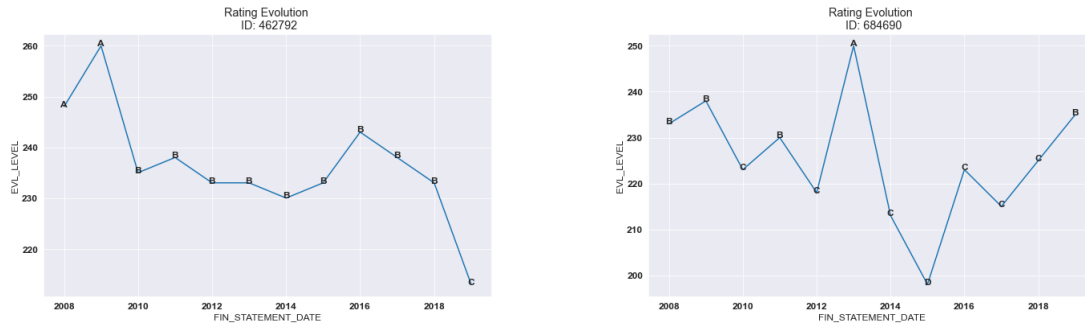
## 6.1 Introduction

In this chapter, we diverge significantly from the approach taken in previous chapters by concentrating specifically on companies with a long history of credit ratings. The primary objective here is to predict the future credit rating of these companies.

When it comes to companies, the set of financial data provided by companies differs from country to country [126]. It depends on the financial regulation adopted by the country in which the company is based. An important feature when analyzing companies is whether they are publicly traded or not. While public companies must publish audited financial statements (in most countries according to the International Financial Reporting Standards (IFRS)), the data available and analysis of their creditworthiness are completely different with respect to Small and Medium Enterprises (SME). In some countries like the United States, SMEs have no requirements [127] in terms of publishing their financial data. In other cases where the data is partially available (i.e., some financial features are missing). In [128] different data imputation techniques have been used in order to address this problem. This work has focused its efforts on a consumer's credit scoring. The latter has been widely treated in the literature ([129], [130]).

Some efforts have been made in order to create a credit scoring system for companies [131]. Nonetheless the focus on previous studies has been to build highly accurate Machine-Learning (ML) based models using a large list of financial features [85]. The problem with this approach is that they consider companies as data points and do not take into account potential trends when assessing the PD of a company.

Artificial intelligence (AI) has been widely used in the financial industry for financial



(a) Rating evolution of the company 462792 (b) Rating evolution of the company 684690

Figure 6.1: Example of two companies rated by Tinubu Square every year during the period 2008-2019. The EVL\_LEVEL represents the score given by Tinubu before being mapped to a rating class. The FIN\_STATEMENT\_DATE is the date the company has published its financial statements.

forecasting. Statistical methods, such as AutoRegressive (AR) or AutoRegressive Moving Average (ARMA), have been traditionally employed for financial forecasting. The growing capabilities of AI-based models for predicting the future of financial features based on past behavior are triggering a change in the methods used for this particular task.

In this chapter we present several contributions that are listed as follows:

- We analyze a large list of SMEs based mainly in Southern Europe.
- Our proposed framework analyzes the past behavior of companies and does not consider companies as an isolated data point.
- We forecast the rating of several companies using both a statistical traditional model (i.e., ARMA) and a ML-base model (i.e., Gradient Boosting).
- The proposed framework does not depend on the financial data, it depends on the historical behavior of companies.
- We analyze the results of the model using an out-of-time sample and compare them with the rating given by a company specialized on credit scoring.

## 6.2 Related Work

In this section we present the main works that address machine learning for credit scoring as well as the application of AI-based models for financial forecasting.

### 6.2.1 Credit Scoring

As defined in the previous section, credit scoring aims at measuring the risk for a bank or, more generally, a credit institution to grant a loan to an applicant. The most widely used algorithms for assessing the PD are logistic regression and linear discriminant ([6], [132]). The main reason why these machine learning-based models have been widely adopted in the financial industry is their simplicity, their ease of use. The latter models are very limited since they are not able to capture nonlinear relationships between features. This limitation has been addressed in the literature by applying more sophisticated machine learning-based models: Random Forests [38], Gradient Boosting ([133], [134]) and kernel-based algorithms such as Support Vector Machine (SVM) [41]. For credit scoring applications, ensemble ML-based models have shown an impressive increase in terms of accuracy when predicting whether a customer will repay the loan or not [39].

### 6.2.2 Forecasting in the Financial Industry

Financial time series forecasting has been a hot topic during the last decade. With the rise of machine learning and deep learning models, researchers have been focused on applying these models to predict the evolution of different stock markets([135], [136]). In some sense, the stock price evolution of a company can be interpreted as what the market thinks of the activities developed by the company, and thus its credit worthiness. Indeed, when there is an important event that may negatively affect the activities of a certain company, the stock price is trending downwards.

## 6.3 Methodology

In this section, we present the method used in this study. We start by presenting the raw data. Then we show the procedure employed to transform the data in order to feed the models. Finally, we introduce the models used for this work and the metrics employed to evaluate the models.

### 6.3.1 Data

The data used for this work has been provided by Tinubu Square, a company which provides credit risk assessments of potential trade partners to its customers. Tinubu Square has an internal credit risk model that has been used for 20 years to assess the creditworthiness of a company. This internal model uses a large list of financial variables to compute a score. Then, the score (i.e., EVL\_LEVEL variable) is mapped to a letter which is the final result of Tinubu's internal model. This rating represents the probability of



Table 6.1: Transformed data is the result of keeping companies rated every year during the period 2008-2019. Each row of the dataset represents an assessment of the company. Time-Series dataset is a dataset in which each row represents a company and the columns the year the company has been rated

Dataset	Size
Original Dataset	1399179
Transformed Dataset	40772
Time-Series Dataset	3395

default in a descriptive way. Highly risked companies have lower scores and are represented with the rating letter F. On the other hand, companies whose PD is close to zero are represented with higher scores and with the rating letter A.

The purpose of this work is to create a forecasting model to predict the rating evolution of companies in Tinubu's portfolio. In order to predict the future rating of a company, we need to convert the original dataset into a time series dataset. Each row of this new dataset represents a company and the columns represent the year the company has been rated.

### 6.3.2 Forecasting Time Series

In this part of the study we present the models proposed to address the problem of forecasting the creditworthiness of a company.

#### AutoRegressive Moving Average

AutoRegressive Moving Average (ARMA) is a combination of two different models: an autoregressive model (AR) and a moving average model (MA). Mathematically, ARMA processes result from the sum of both processes an AR of order  $p$  and MA of order  $q$ . An ARMA( $p,q$ ) model combines both the AR( $p$ ) and MA( $q$ ) models as follows:

The AR model assumes that we can model a time series  $x_t$  using the last  $p$  observations of the given times series plus an additional term, the white noise error  $\epsilon_t$  (see eq.6.1).

$$AR(p) : x_t = c + \sum_{i=1}^p \phi_i x_{t-i} + \epsilon_t \quad (6.1)$$

On the other hand, the MA model considers that the current value of the time series  $x_t$  is affected with the previous  $q$  white noise errors (see eq.6.2).

$$MA(q) : x_t = \mu \sum_{i=1}^q \theta_i \epsilon_{t-i} + \epsilon_t \quad (6.2)$$

$$x_t = c + \sum_{i=1}^p \phi_i x_{t-i} + \epsilon_t + \sum_{i=1}^q \theta_i \epsilon_{t-i} \quad (6.3)$$

The parameters  $p$  and  $q$  can be determined using different methods. By observing the graph of the Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF) both parameters can be estimated. In this work we will estimate the parameters using a more analytical approach by computing the Akaike Information Criterion (AIC) [137]. AIC is a statistical measure that allows the comparison between statistical models to determine which model best fits the data series. AIC considers both model goodness of fit and model complexity.

$$AIC = 2k - 2 \log(L) \quad (6.4)$$

where  $k$  is the number of the parameters of the statistical model and  $L$  is the maximum value of the likelihood function for the model. The first term represents the complexity of the model while the second term in Eq (6.4) represents how well the model fits the data.

## Gradient Boosting

The eXtreme Gradient Boosting (XGBoost) is a machine-learning based algorithm that consists of a sequential combination of weak learners that corrects the errors of the previous weak learner. XGBoost is an open-source framework proposed by [134] that has been widely used in machine learning competitions.

The outcome of a XGBoost composed of  $K$  weak learners for a data set with  $n$  instances and  $m$  features  $D = (x_i, y_i) (|D| = n, x_i \in \mathbb{R}^m, y_i \in \mathbb{R})$  is represented mathematically as follows:

$$\hat{y}_i = \Omega(X_i) = \sum_{k=1}^K f_k(X_i), f_k \in F \quad (6.5)$$

where  $F = \{f(x) = w_{q(x)}\} (q : \mathbb{R}^m \rightarrow T, w \in \mathbb{R}^T)$  is the space regression of trees.  $T$  is the number of leaves in the tree and  $q$  represents the structure of each tree that maps an example to the corresponding leaf index [134]. For each independent tree  $f_k$  there is an independent  $q$  and leaf weights  $w_q$ .

$$L^t = \sum_i (y_i, \hat{y}_i) + \sum_k \Omega(f_k) \quad (6.6)$$

where  $l$  is the loss function,  $y_i$  is the target value and  $\hat{y}_i$  is the prediction.  $\Omega$  is a term that penalizes the complexity of the function. This term is introduced to avoid overfitting.

For the hyperparameter estimation, we used scikit-learn [110] to perform a Grid Search, which consists of creating all possible hyperparameter combinations of a list of values predetermined by the user.

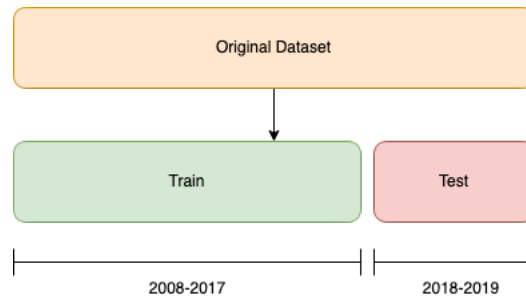


Figure 6.2: Proposed splitting strategy

### 6.3.3 Comparing the Forecasted Values

To assess the performance of both proposed models, we will split the dataset into two different parts: the train set will contain the score, and thus the rating, of companies during the period 2008-2017. The rest of the dataset, the test set, will be composed by the rating of the same companies scored by Tinubu during the period 2018-2019 (see Fig. 6.2).

We will compare graphically the results of both models. For the ARMA model, we will keep one ARMA by company with the hyperparameters  $p$  and  $q$  that minimize the AIC for the time series between the period 2008-2017. For each model we generate two predictions that correspond to the years 2018 and 2019. Then we compare the forecast of both models graphically.

The two models proposed in this chapter for forecasting the risk class of companies are different by nature. For the ARMA model we need to create several models with hyperparameters  $p$  and  $q$  and choose the one whose AIC value is the lowest. Then we will present the difference between models prediction and the Tinubu's Score, the target variable.

The prediction of the model with respect to the actual value of the test set using  $gr$ . This is a regression metric that estimates the distance between the true value and the predicted value. This metric is not used for regression problems where the target variable could take values close to zero. Since the target variable (i.e., Tinubu's Score) is in the range [100-320] the MAPE is an interesting measure to use. MAPE is defined as follows:

## 6.4 Results

In this section we present the results of the models that have been described in the previous section. First, we present the results of each model individually and then we compare the forecasted ratings for both models using several companies.

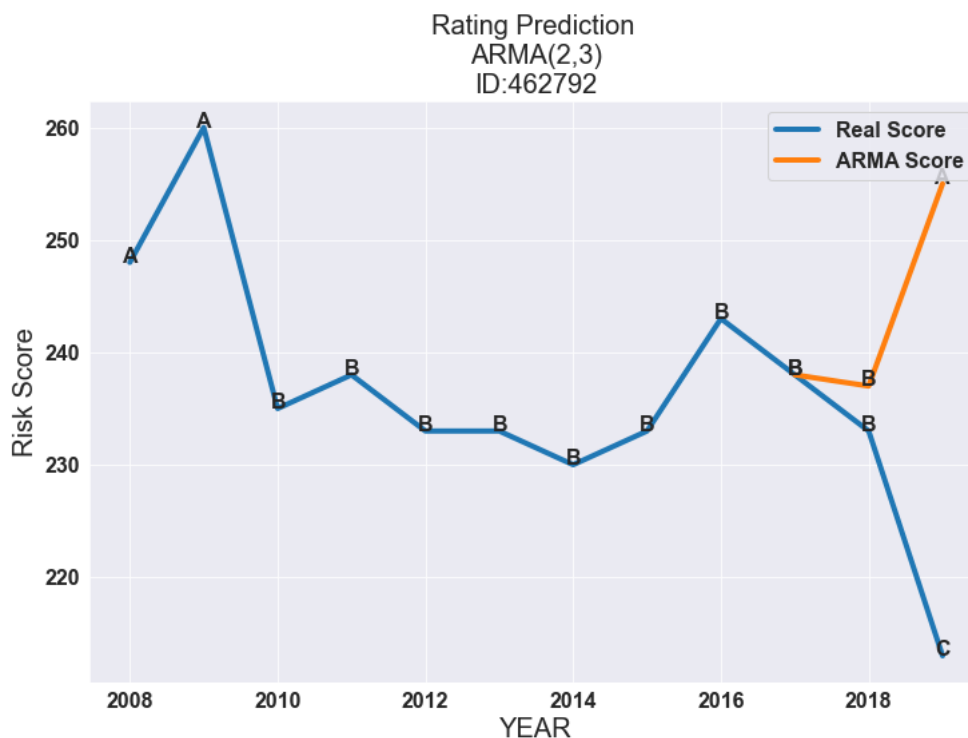


Figure 6.3: Results of the prediction of the rating of the two next years for company 462792 using an ARMA model with hyperparameters  $p=2$ ,  $q=3$  which are the hyperparameters with the lowest AIC for models trained with the data between (2008-2017).

### 6.4.1 Assessing the Performance of ARMA

As we mentioned previously, we will create several ARMA models with different hyperparameters (i.e.,  $p$  and  $q$ ). For instance, we created 9 different ARMA models and considered the best model the one with the lowest AIC. In table 6.2, we show the results obtained for the company with the id 462792. For this particular company the hyperparameters with the lowest AIC are  $p = 2$  and  $q = 3$ . When we forecast Tinubu's Score for the years 2018 and 2019 we observe that the model underestimates the risk Fig. 6.3. The ARMA captures the smooth deterioration of the company for the year 2018. However, the difference between the model in 2019 is significant.

### 6.4.2 Forecasting Tinubu's Score Using XGBoost

In this part of the study, we present the results obtained with XGBoost. In Table 6.3 we show the optimal hyperparameters found using the Grid Search optimization.

As we can see in Fig. 6.4, the XGBoost considers that the company with the ID 462792 will slightly improve during the next two years. The rating given by Tinubu and the rating generated by our XGBoost model diverges for the year 2019. This difference

Table 6.2: For each company in the dataset, we create 9 models. The best model is the one with the lowest AIC value. The AIC has been computed over the train set which consists of companies rated every year by Tinubu Square during the period 2008-2017. We show the difference between the predicted rating and the real rating.  $p$  and  $q$  are the hyperparameters of the ARMA model.

LGE_ID	2018	2019	Prediction 2018	Prediction 2019	p	q	AIC
462792	233 (B)	213 (C)	234 (B)	245 (B)	1	1	65.2517
462792	233 (B)	213 (C)	234 (B)	245 (B)	2	1	67.2505
462792	233 (B)	213 (C)	233 (B)	244 (B)	3	1	69.0822
462792	233 (B)	213 (C)	233 (B)	249 (A)	1	2	59.9773
462792	233 (B)	213 (C)	235 (B)	259 (A)	1	3	57.8751
462792	233 (B)	213 (C)	233 (B)	246 (B)	2	2	61.4822
462792	233 (B)	213 (C)	234 (B)	253 (A)	3	3	60.3457
462792	233 (B)	213 (C)	233 (B)	246 (B)	3	2	63.482
462792	233 (B)	213 (C)	236 (B)	253 (A)	<b>2</b>	<b>3</b>	<b>58.4614</b>

Table 6.3: Optimal XGBoost hyperparameters found using Grid Search.

Hyperparameter	Value
n estimators	150
learning rate	0.2
max depth	8

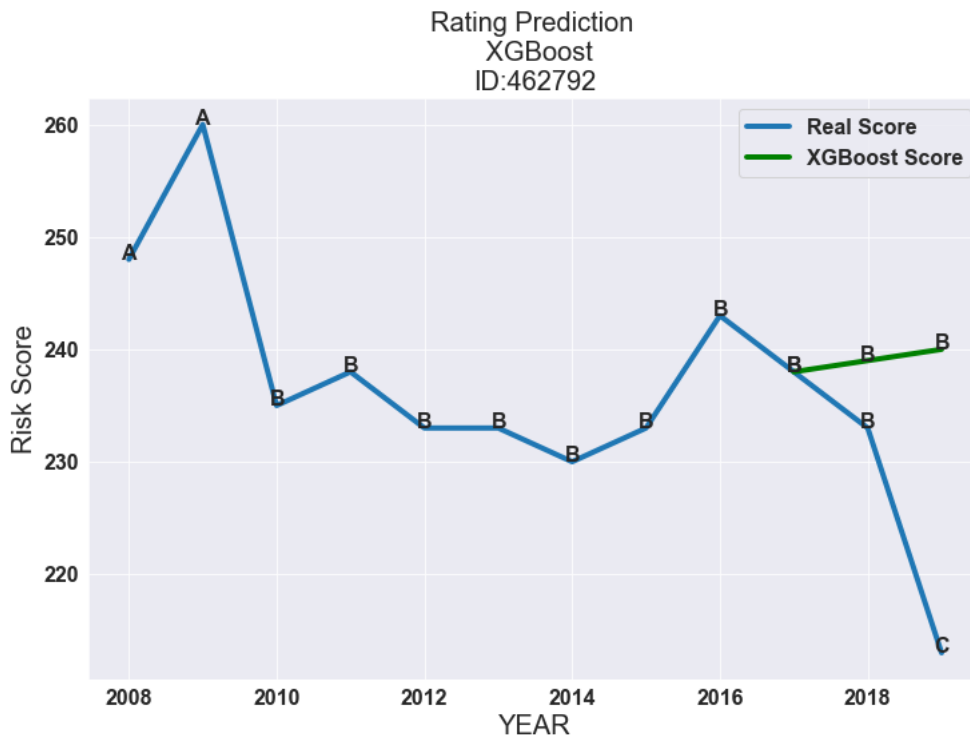


Figure 6.4: Rating prediction for the company 462792 for the years 2018 and 2019 using XGBoost with the hyperparameters presented in Table 6.3.

can be explained by the fact there has been an event that has adversely affected the company's activity.

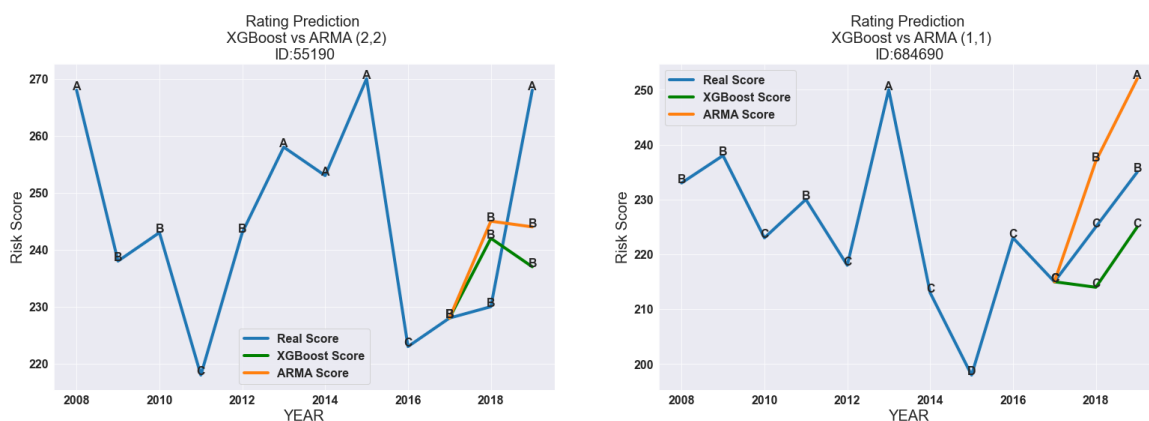
### 6.4.3 Analyzing the Models Ratings

As we mentioned in previous sections, Tinubu's rating system reflects the PD in a descriptive way. In this section we will present several examples of comparison between the behavior of both models for different companies with the rating given by Tinubu for the test period (see Fig.6.2).

In Fig.6.5(a) we compare the xgboost with an ARMA with  $p, q = 2$ . The results show that, in this particular case, both models slightly underestimate the risk. On the other side, we observe that both models overestimate the risk.

For the company 684690 (see Fig.6.5(b)), the results show a divergence between the ARMA model with  $p, q = 1$  and the XGBoost model. The latter can capture the evolution of the rating during the years 2018 and 2019 with a narrow difference between the predicted score and the Score given by Tinubu. On the other side, the ARMA model, even if it captures the trend, clearly overestimates the company's creditworthiness.

Nonetheless there are cases where both models underestimate the risk. This is the case in Fig.6.6(a). Both models consider the creditworthiness of the company for the next two years to have a downward trend.



(a) Predicted rating evolution for company 55190 (b) Predicted rating evolution for company 684690

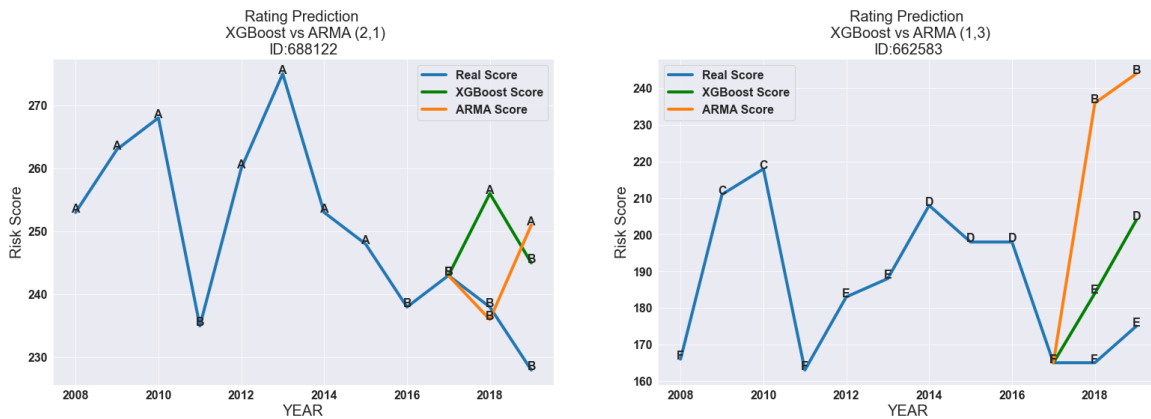
Figure 6.5: Comparison of the forecasted Tinubu's rating for the period 2018-2019 for the companies with the ID 55190 and 684690 respectively using the proposed models in this work: ARMA and XGBoost

Another relevant case is presented in Fig.6.7(b). In this one, the ARMA with  $p = 1, q = 3$  highly underestimates the risk by giving an A rate to a company for which Tinubu's rating for the year 2019 is E.



(a) Predicted rating evolution for company 162009 (b) Predicted rating evolution for company 165482

Figure 6.6: Comparison of the forecasted Tinubu’s rating for the period 2018-2019 for the companies with the ID 162009 and 165482 respectively using the proposed models in this work: ARMA and XGBoost



(a) Predicted rating evolution for company 688122 (b) Predicted rating evolution for company 662583

Figure 6.7: Comparison of the forecasted Tinubu’s rating for the period 2018-2019 for the companies with the ID 688122 and 662583 respectively using the proposed models in this work: ARMA and XGBoost

## 6.5 Conclusions

In this work we proposed a framework for the credit scoring that differs from previous work. The approach considered consists of forecasting the credit rating of a large number of companies using an historic index given by Tinubu Square for the period 2008-2019. We used the period 2018-2019 to compare both forecasting models with the target value which is Tinubu’s rating. We observed that for the set of companies analyzed, both models tend to slightly underestimate the risk. For companies specialized in credit risk, models that overestimate the risk are preferable to models that underestimate the risk.

This is mainly because underestimating the risk leads to potential economic losses. On the other hand, models that are strongly conservative (i.e., models that overestimate the risk) are not desirable for credit scoring since they impact the economic activity by reducing the volume of credits approved and thus decreasing the business between SMEs. Considering this criteria and observing the results of our work we can conclude that the machine learning-based models are closer to Tinubu standards in terms of credit scoring. It is important to remark on the fact that wealthy companies (i.e., companies with the rating classes A-B) are more stable over the period analyzed than companies with a lower rating. Future work should focus on the adequate evaluation of the performance of the models proposed in this work. It will be interesting to introduce deep learning models and compare them with results yielded by the models proposed.





# Multimodal Credit Risk Scoring

## 7.1 Introduction

As highlighted in the introduction, one of the significant challenges in credit scoring is the lack of comprehensive financial data. This scarcity often hampers the effectiveness of traditional credit scoring models, which rely heavily on financial metrics. To address this issue, researchers in the field are increasingly exploring alternative types of data that can supplement or even replace traditional financial indicators.

Traditional credit scoring models primarily rely on tabular financial data, such as credit history, income, and debt-to-income ratio [138]. However, these models may not capture the full picture of a borrower's creditworthiness, especially when dealing with complex and dynamic markets.

Textual data, such as news articles, press releases, and social media posts, can provide valuable insights into potential risks that may impact a borrower's ability to repay a loan [139]. By combining these textual risk assessments with traditional financial data, financial institutions can improve the accuracy and reliability of credit risk scoring algorithms, leading to better lending decisions and reduced credit losses.

In this context, we propose a new framework that intends to leverage the non-structured information available in the comments made by risk analysts by combining it with tabular financial data.

The chapter is structured as follows: first, we present the previous research in the field of credit scoring and in Natural Language Processing (NLP). Then, we present our framework and the results yielded. Finally, we present the conclusion and potential hints for future work.

We compare the impact of adding the text feature using different Natural Language Processing techniques on the performance of several state-of-the-art models.

## 7.2 Related Work

### 7.2.1 Credit Scoring

Credit scoring is an important tool for financial institutions and is used to assess the creditworthiness of potential borrowers. Classical credit scoring models are based on statistical techniques such as Linear Discriminant Analysis, Logistic Regression ([6], [140]).

Recently, there is a new wave of Machine Learning based algorithms that are used to estimate the probability of default. In [39], the authors have shown that high-level sub-symbolic algorithms outperform statistical-based models.

Recent works focus on the use of financial tabular data for predicting companies' future status (i.e., default/non-default). The authors in [138], center their efforts on predicting the default using a set of 81 financial features. They conclude that among all tested models, tree-based models are the best in terms of performance. In [86], the authors employed a similar approach for predicting the default of companies one year later after they published their financial sheets. An interesting different approach has been employed by the authors in [85]. They compare the output of a machine learning-based credit scoring model with the rating given by a credit risk company. The results show that the model distinguishes extremely rated companies.

A large number of works have shown that XGBoost [134], a boosting method, outperforms classical machine-learning-based methods (i.e., logistic regression, decision trees neural networks) for the task of predicting the default.

### 7.2.2 Natural Language Processing

There have been many significant advances in the field of natural language processing (NLP) in recent years. One of the most important developments has been the advancement of deep learning (DL) techniques [141], which have led to significant improvements in the accuracy and effectiveness of NLP models [142].

During the 1980s, the field of NLP experienced the first revolution with the introduction of statistical models such as Hidden Markov Models (HMM) [143] for speech recognition and N-grams models [144] for machine translation. Afterward, and as a result of the exponential growth of computational power, the deep learning models have rapidly

gained interest for solving NLP tasks like Information Retrieval (IR) [145], Named Entity Recognition (NER) [146] or Text Classification [147].

Both Machine-Learning (ML) and DL models take as input numerical features. Two popular approaches are used in NLP for text representation and feature extraction. The methods employed for representing text and words in a numerical format (i.e., high-dimensional vectors) that aim to capture the meaning and context of the words are called Word Embeddings (WE). NLP researchers have adopted several methods for creating word embeddings (i.e., Word2vec [74], GloVe [75], BERT [77]). More classical approaches like TF-IDF [148] represent the text as a sparse vector of word frequencies.

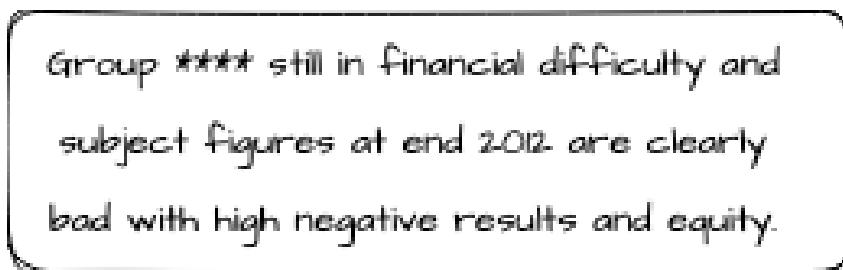
For the TF-IDF technique [148] the measurement of the importance of a word in a document is made by combining the term frequency (TF) and the inverse document frequency (IDF), which refers to how common or rare a word is across the entire document. Word2vec [74] is a WE-method that represents words in a continuous vector space in which words that have similar meanings are closer in the vector space. GloVe [75] is based on the idea of factorizing a large matrix of word-word co-occurrence counts, where the matrix is constructed from a large corpus of text. The co-occurrence information between words indicates how often they appear together in the corpus. BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art natural language processing model developed by Google in 2018. It is a pre-trained language model that uses a deep neural network architecture called the transformer, which allows it to capture the context and dependencies of words in a sentence. Sentiment analysis has been used in several industries such as Marketing [149] or Finance [150]. For example in marketing, sentiment analysis is used to identify customer needs which triggers different strategies to improve customer satisfaction and retention. In finance, multiple different approaches have been adopted. In [151], the authors analyze the impact of social media opinion on different companies for stock market predictions. In credit scoring, few works have concentrated their efforts on the impact of textual data for default prediction. In [152], the authors found that a deep learning approach, based on the BERT model for default prediction using textual data, outperforms classical ML models (i.e., Logistic Regression and Random Forest). Other approaches that have been considered for the treatment of textual data for the default prediction focus on word embedding techniques to represent in a low-dimensional vector space the economic sector of the company [116].

## 7.3 Methodology

In this section, we present our proposed framework. First, we start by presenting the data and its characteristics. Then, we describe the different strategies we propose for dealing with the text feature. We show the preprocessing pipeline and the data-splitting strategy for the training stage. Finally, we compare the performance of different state-of-the-art machine learning models for the different text feature treatment strategies.

### 7.3.1 Data Overview

The dataset used in this work is provided by Tinubu, a company specializing in credit risk assessments. The original dataset contains 4951 credit assessments. For each assessment, we have 34 standard features that represent the companies information (e.g., creation year, the country in which it operates), its financial statements (e.g., net worth, current assets), and a text feature (e.g., risk analyst comments), that typically contains a description of the company, the business context, and its activities. Each data point corresponds to the year in which the company has been assessed by credit risk analysts. Word2vec [74], [75].



Group \*\*\*\* still in financial difficulty and subject figures at end 2012 are clearly bad with high negative results and equity.

Figure 7.1: Text assessment example generated by a risk analyst.

The target variable of our models is whether the company is in financial embarrassment<sup>1</sup> or being out of business. We create the default variable by setting it to 1 if the considered company is in financial embarrassment or out of business the year after the assessment and 0 otherwise.

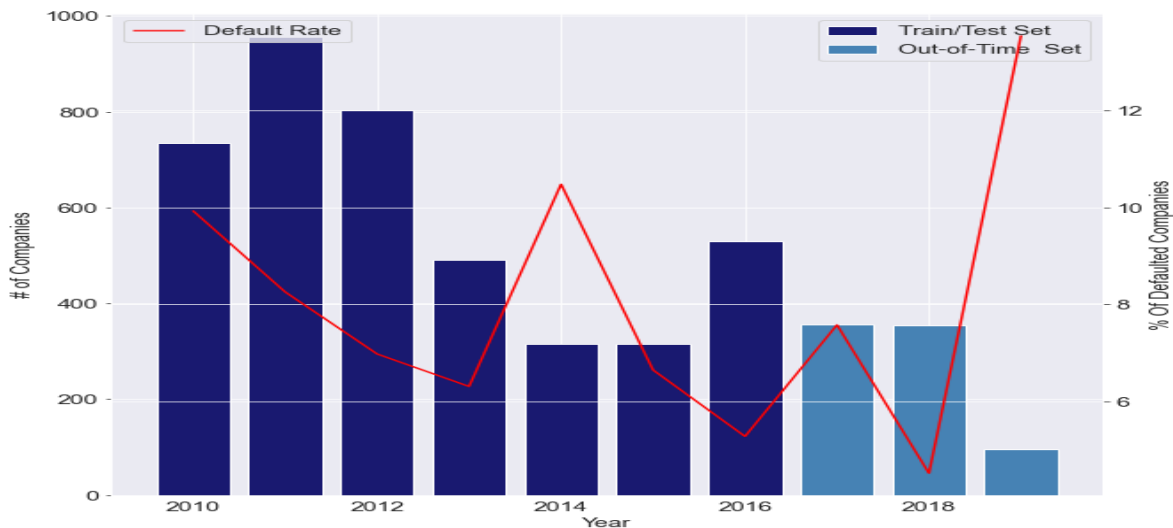


Figure 7.2: Number of companies assessed each year. The default rate represents the percentage of companies that will experience financial difficulties the year after the assessment.

<sup>1</sup>Financial embarrassment refers to a state of financial difficulty. Companies in financial embarrassment may have problems refunding their loans

### 7.3.2 Text Feature Treatment

The dataset contains a text feature in which the comments and opinions of the credit risk analysts are stored. The risk analysts' comments usually contain a summary of the industry trends and economic conditions that could impact the company as well as a description of the company and its activities.

The first step in the text feature pipeline is to remove stop words, punctuation, and other irrelevant information. Once the text has been preprocessed, we use an embedding algorithm to represent each word in the text as a vector in a lower dimensional space. For instance, in this work we employed two different word embedding techniques: Word2vec [74] and GLoVe [75].

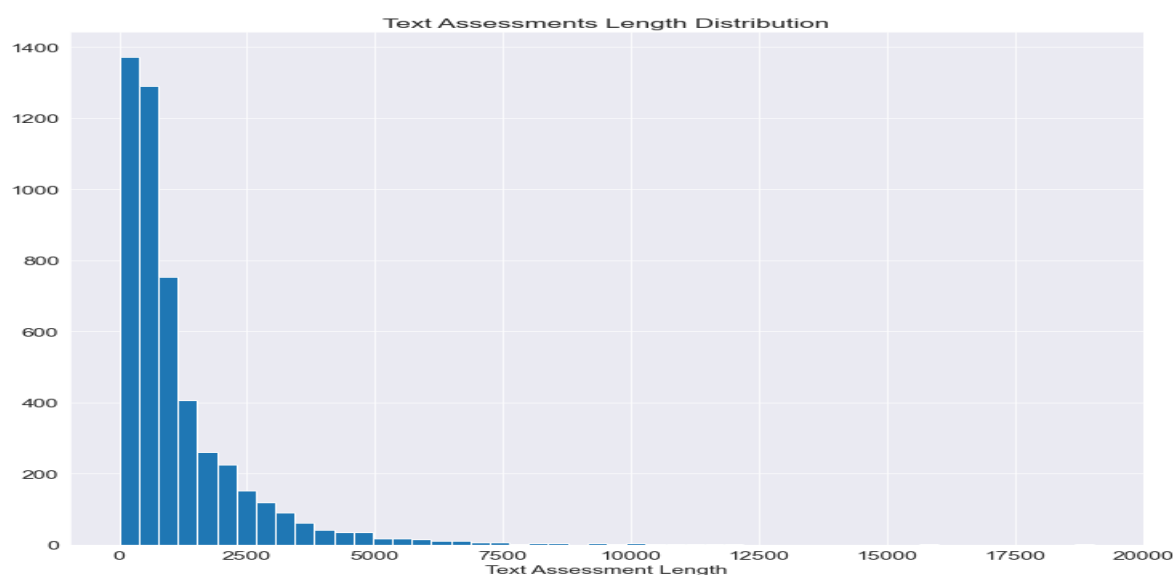


Figure 7.3: Size distribution of the text assessment generated by the risk analysts.

We used the Word2Vec algorithm trained on the Google News dataset. The model contains 300-dimensional vectors for 3 million words and phrases. Each word in the text is represented in a 300-dimensional space. The text is encoded in a matrix  $\mathbf{W} \in \mathcal{M}^{m,300}$  where  $m$  is the number of words in the text.

The GLoVe algorithm, trained on 2B tweets, transforms the text into a 200-dimensional space. In this case, the text is encoded in a matrix  $\mathbf{G} \in \mathcal{M}^{m,300}$ .

In this work, we propose a different approach when treating the text feature. We evaluate the sentiment of the text generated by the risk analyst using a sentiment analysis model (i.e., FinancialBERT [153]). FinancialBERT is a fine-tuned BERT using large corpora of financial texts. The model categorizes the text as positive, neutral, or negative.

### 7.3.3 Data Preprocessing

The preprocessing step can be split into two main parts depending on the applied text feature model. For GLoVe and Word2vec, we reduce the dimensional space of the text embedding using Principal Component Analysis ([154], [155]). For non-tree-based ML models, we scale the data using Standard Scaler [124].

The categorical features (i.e., the country in which the company is based) are encoded using the one-hot encoding technique. This method creates a new feature in the dataset for each feature label. It takes the value one if the company is based on the considered country. It is important to remark that the categorical feature derived from applying the Sentiment Analysis model to the text feature is also encoded using the same technique.

Finally, once we preprocessed our data, we split the dataset into three parts: an out-of-time dataset which includes the data for the period 2017-2019 that will be employed to valid model performance; and a train/test dataset (i.e., 70% for training) that covers the period (2010-2016).

### 7.3.4 Models

The following part presents the state-of-the-art machine learning-based models trained to predict the default the year after the assessment using the data processing pipeline explained in the previous section. The outcome of the models is the probability of being in out of business or financial embarrassments the year after the risk assessment. This probability is mapped to a binary value, 1 (i.e., default) if the probability is greater or equal to 0.5.

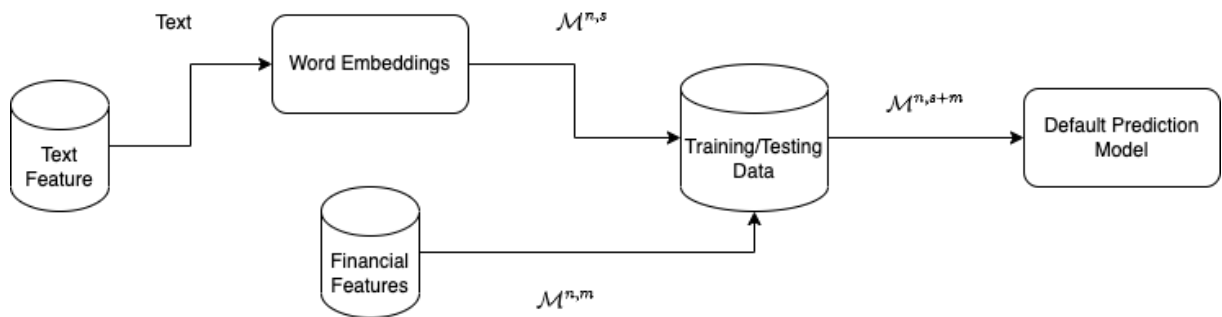
**Logistic Regression:** we implement the logistic regression using the Ridge regularisation that helps the model prevent multicollinearity. We also set the class weight parameter to *balanced* to adjust weights inversely proportional to class frequencies.

**Random Forest:** We set the number of trees to 100 and the class weight to *balanced* to deal with the imbalance in the target variable.

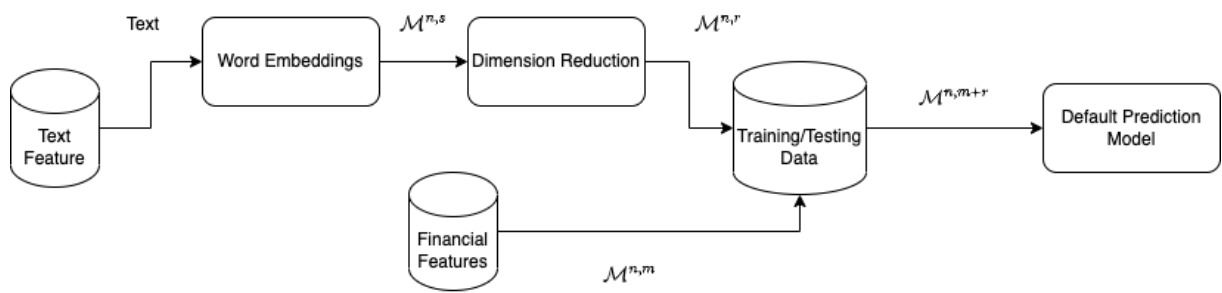
**Support Vector Machine (SVM):** In this case, we have set the hyperparameter *class\_weight = 'balanced'*. The other hyperparameters are the default ones.

**XGBoost:** This algorithm employs a sequential combination of 1000 weak learners. The hyperparameters of the xgboost are the follow: *n\_estimators = 1000*, *learning\_rate = 0.1* and *scale\_pos\_weight = 12*. The latter help the model with the imbalance dataset nad is usually set to  $\frac{\#Positive\ Instances}{\#Negative\ Instances}$ .

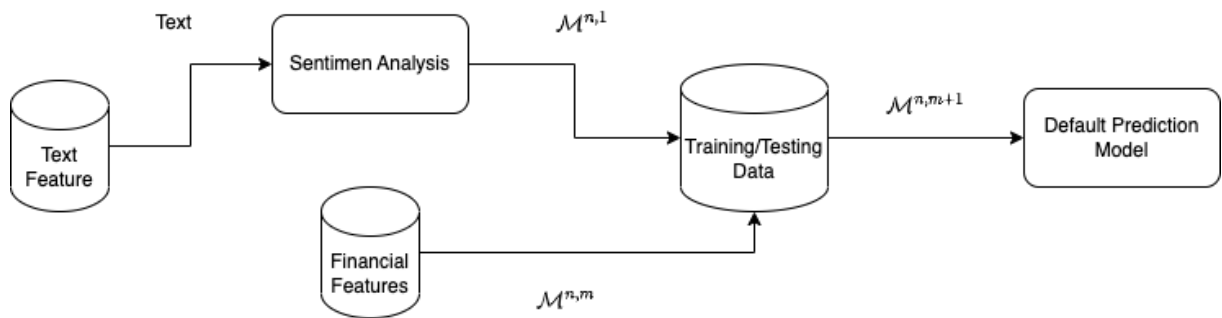
**LightGBM:** We set three different hyperparameters for this gradient boosting-based model: *n\_estimators = 100*, *learning\_rate = 0.1* and *class\_weight = 'balanced'*.



(a) Text processing using word embedding



(b) Text processing using word embedding and dimension reduction



(c) Text processing using sentiment analysis

Figure 7.4: Conceptual overview of the proposed framework.  $m$ ,  $n$  represent respectively the number of assessments and the number of financial features available in the dataset.  $s$  represents the embedding dimension (i.e.,  $s=200$  for GLoVe and  $s=300$  for Word2Vec). The variable  $r$  represents the word embedding space after the dimension reduction (i.e.,  $r=5$ ).



**Neural Network:** We trained a neural network using stochastic gradient descent. The neural network is composed by 4 hidden layers with 128, 256, 512, and 128 neurons. The activation function is the *ReLU* function.

### 7.3.5 Model Performance

Two different strategies have been used to evaluate the performance of the developed models. The first model evaluation comes from the 5 cross validation that has been employed using the train/test dataset. We also compare the performance of the models using the out-of-time sample.

For the train/test set, we employed three different classification metrics: precision, recall, and F1-Score. Precision measures the proportion of true positives among all predicted positives. Recall measures the proportion of true positives that were correctly identified. F1-Score is calculated as the harmonic mean of precision and recall.

## 7.4 Results

### 7.4.1 Exploratory Data Analysis

Our approach starts by analyzing the target variable and the text feature which is the risk analyst comments (e.g., see Fig.7.1). As we can see in Fig.7.2, the default rate experience a downward trend from 2010 to 2018. However, during 2014 and 2017 the default rate did not fit the trend.

In Fig.7.3, we show the length distribution of the analyzed risk analyst comments. The median of this distribution is 693 characters. Just 10% of the risk analysis has a length bigger than 5 pages (i.e., 2600 characters).

### 7.4.2 Model Performance

For straightforward comparison, we employ the most commonly used classification evaluation metrics that lie in  $[0,1]$ . Higher values indicate superior performance. Tab. 7.2 display the accuracy of the different models with the different text embedding techniques. The overall performance of the models improves when we raw embed the text using Word2vec and GLoVe, especially for the logistic regression and the SVM.

The imbalanced characteristics of the dataset may bias the models to predict the most represented class (i.e., the non-defaulted class). In Tab. 7.3, we compare the precision of

	Tabular	Tabular + Text (WE)		Tabular + Text (SA)	Tabular + Text (WE+PCA)	
		GLoVe	Word2vec		GLoVe	Word2vec
	Logistic Regression	0.208 ± 0.037	0.204 ± 0.023	0.196 ± 0.036	0.217 ± 0.040	0.209 ± 0.037
SVM	0.210 ± 0.038	0.225 ± 0.042	0.235 ± 0.053	0.211 ± 0.038	0.209 ± 0.038	0.210 ± 0.037
Random Forest	0.012 ± 0.014	0.005 ± 0.009	0.010 ± 0.013	0.006 ± 0.011	0.010 ± 0.013	0.000 ± 0.000
XGBoost	0.124 ± 0.073	0.101 ± 0.038	0.081 ± 0.013	0.146 ± 0.061	0.117 ± 0.047	0.132 ± 0.077
LightGBM	0.164 ± 0.075	0.083 ± 0.033	0.096 ± 0.035	0.168 ± 0.072	0.163 ± 0.062	0.133 ± 0.067
MLP	0.000 ± 0.000	0.010 ± 0.021	0.006 ± 0.036	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000

Table 7.1: F1 Score mean and standard deviation for the different strategies using 5-fold cross-validation. *WE* represent the dataset in which the text has been encoded using Word Embeddings. *SA* stands for Sentiment Analysis and is the dataset with the text categorically encoded using a sentiment analysis model. *PCA* is for the experience in which the dimensions of the numerical vector generated by the WE have been reduced using Principal Component Analysis.

the models.

Tab.7.2, we present the results using the approach described previously. In this work, we conducted 6 different experiences that related to the treatment of the textual feature. Tabular represents the model that has been trained just using the financial features. Tabular + Text (WE), we merged the financial features with the numerical representation generated by GLoVe and Word2Vec. In the WE + PCA experience, we performed a PCA to reduce the numerical representation from 200 and 300 respectively to 5 dimensions (i.e., these 5 dimensions represent 73.4 % of the variance). The SA represents the scenario in which we employ sentiment analysis to map the text to a categorical value.

In Tab.7.1, we show the F1-Score for the models and the different text treatment configurations. Since we compute 5 cross-validations, we have the mean F1-Score and the standard deviation. The table shows that both XGBoost and Logistic Regression models benefit from the addition of a text feature encoded using sentiment analysis.

The analysis of Tab.7.3 shows that the reason behind the low values of the F1-Score (see Tab.7.1) is the inability of the models to determine which companies will go into default. This is mainly due to the imbalanced characteristics of the dataset employed.

On the other hand, and regarding the F1-Score, we observe that Support Vector Machine and Neural Networks improve their performance by adding a significant number of features (i.e., the dimension of the word embeddings). This behavior is even more remarkable for Neural Networks.

In terms of the two embedding techniques proposed in this work, we see that there is no technique that improves the model performance for all the models.

	Tabular	Tabular + Text			Tabular + Text (WE+PCA)	
		(WE)			(WE+PCA)	
		GLoVe	Word2vec		GLoVe	Word2vec
Logistic Regression	0.492 ± 0.149	0.684 ± 0.069	0.717 ± 0.042	0.550 ± 0.134	0.510 ± 0.139	0.717 ± 0.042
SVM	0.499 ± 0.152	0.718 ± 0.064	0.734 ± 0.059	0.511 ± 0.169	0.512 ± 0.141	0.734 ± 0.059
Random Forest	0.919 ± 0.017	0.922 ± 0.015	0.921 ± 0.015	0.920 ± 0.017	0.921 ± 0.015	0.922 ± 0.015
XGBoost	0.905 ± 0.023	0.916 ± 0.020	0.912 ± 0.018	0.907 ± 0.022	0.903 ± 0.020	0.912 ± 0.018
LightGBM	0.897 ± 0.022	0.914 ± 0.018	0.915 ± 0.017	0.899 ± 0.024	0.901 ± 0.023	0.915 ± 0.017
MLP	0.923 ± 0.015	0.923 ± 0.015	0.920 ± 0.015	0.923 ± 0.015	0.923 ± 0.015	0.923 ± 0.015

Table 7.2: Accuracy Score mean and standard deviation for the different strategies using 5-fold cross-validation. *WE* represent the dataset in which the text has been encoded using Word Embeddings. *SA* stands for Sentiment Analysis and is the dataset with the text categorically encoded using a sentiment analysis model. *PCA* is for the experience in which the dimensions of the numerical vector generated by the WE have been reduced using Principal Component Analysis.

	Tabular	Tabular + Text			Tabular + Text (WE+PCA)	
		(WE)			(WE+PCA)	
		GLoVe	Word2vec		GLoVe	Word2vec
Logistic Regression	0.119 ± 0.024	0.128 ± 0.019	0.126 ± 0.027	0.127 ± 0.026	0.120 ± 0.024	0.126 ± 0.027
SVM	0.121 ± 0.025	0.145 ± 0.032	0.154 ± 0.041	0.123 ± 0.025	0.121 ± 0.025	0.154 ± 0.041
Random Forest	0.135 ± 0.127	0.100 ± 0.200	0.067 ± 0.133	0.000 ± 0.000	0.340 ± 0.376	0.067 ± 0.133
XGBoost	0.209 ± 0.146	0.323 ± 0.085	0.221 ± 0.041	0.262 ± 0.081	0.278 ± 0.020	0.221 ± 0.041
LightGBM	0.219 ± 0.092	0.243 ± 0.051	0.319 ± 0.111	0.227 ± 0.072	0.242 ± 0.018	0.319 ± 0.111
MLP	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000

Table 7.3: Precision Score mean and standard deviation for the different strategies using 5-fold cross-validation. *WE* represent the dataset in which the text has been encoded using Word Embeddings. *SA* stands for Sentiment Analysis and is the dataset with the text categorically encoded using a sentiment analysis model. *PCA* is for the experience in which the dimensions of the numerical vector generated by the WE have been reduced using Principal Component Analysis.

	Tabular	Tabular + Text		Tabular + Text (SA)	Tabular + Text (WE + PCA)	
		(WE)			(WE + PCA)	
		GLoVe	Word2vec		GLoVe	Word2vec
Logistic Regression	0.695	0.762	0.743	0.701	0.847	0.743
SVM	0.706	0.770	0.748	0.706	0.853	0.748
Random Forest	0.930	0.932	0.932	0.931	0.965	0.930
XGBoost	0.921	0.928	0.924	0.923	0.899	0.924
LightGBM	0.927	0.929	0.929	0.924	0.949	0.929
MLP	0.931	0.931	0.931	0.931	0.965	0.931

Table 7.4: Accuracy score of the models using the out-of-time set. *WE* represents the dataset in which the text has been encoded using Word Embeddings. *SA* stands for Sentiment Analysis and is the dataset with the text categorically encoded using a sentiment analysis model. *PCA* is for the experience in which the dimensions of the numerical vector generated by the WE have been reduced using Principal Component Analysis.

## 7.5 Conclusions

In conclusion, this chapter proposes a new approach to credit scoring that combines financial tabular data with credit risk textual assessment using word embedding techniques and sentiment analysis. Moreover, we addressed the problem of representing the text in a high dimensional space by using dimension reduction techniques (i.e., Principal Component Analysis).

Experimental results demonstrate that the addition of the textual feature slightly improves model performance. This suggests that incorporating credit risk textual assessment can provide additional information to financial institutions for more informed credit decisions. Furthermore, we experimentally demonstrate that the sentiment analysis approach tends to yield better results in comparison to the word embedding techniques. This is due to the fact that word embeddings, in essence, map words into a high-dimensional space where semantically similar words are placed closely together. However, these techniques focus primarily on individual words or at most, phrases. They do not inherently account for the larger syntactic or semantic context that extends beyond individual words and thus they might not adequately encapsulate the overall sentiment of a text, leading to a loss of crucial information.

In addition to improving model performance, a performant credit scoring system that combines financial tabular data with credit risk textual assessment can help credit risk analysts to focus on more critical cases, leaving the smaller cases to the system.

The framework proposed in this chapter offers a promising approach for credit scoring and can be applied to other text classification tasks in finance and beyond. Future research should explore the use of more advanced techniques for sentiment analysis and word embedding (i.e. LSTM [156], GRU [157]), as well as the incorporation of other qualitative factors such as news articles and social media data to further improve model performance.



# Predicting Corporate Solvency using Sentiment Analysis of Risk Analyst Textual Assessments

## 8.1 Introduction

Credit scoring has been the backbone of financial institutions' decision-making processes, allowing them to adequately evaluate the risk associated with providing credits to individual consumers and businesses. The recent progress in artificial intelligence has significantly improved the precision and efficiency of decision-making algorithms. These advances have provoked a waterfall of a wide variety of applications in several different industries.

In recent years, the financial industry has started to incorporate Artificial Intelligence (AI) powered algorithms and advanced data analytics in key tasks (e.g., credit scoring, pricing, etc...) ( [158], [159]).

In the credit scoring field, the primary role of credit risk analysts is to quantitatively estimate the risk that a borrower will experience financial difficulties afterward [39]. AI-powered algorithms allow financial institutions to estimate the probability of default of a company using typical financial statements. These models suffer from the complex and dynamic behavior of the economy. Credit risk analysts, in terms of data employed as an input for their decision-making process, may not be able to deal with more than several financial variables in comparison to the credit scoring algorithms, but they expand their knowledge of the situation of a concrete company by analyzing textual data such as news articles, social media or press releases. Subsequently, they write credit risk assessments in which they express their opinion about the company and whether credit should be granted to the company.

In this paper, we introduce a novel framework that incorporates credit risk text assessments made by credit analysts to fine-tune a pre-trained financial sentiment analysis model. Finally, we compare the sentiment analysis of the text with the real rating given by the risk analysts.

This chapter is presented as follows: first, we present recent progress in the field of credit scoring, and more specifically, the intersection between this field and the Natural Language Processing field. Then, we explain step-by-step the processes used to prepare the data and to train the model. The next stage of this paper consists of showing and analyzing the results yielded by the proposed framework. Finally, we discuss the results and potential improvements.

## 8.2 Related Work

### 8.2.1 Artificial Intelligence in Credit Scoring

Artificial intelligence (AI) has emerged as a promising tool for credit scoring due to its ability to handle large amounts of data, learn from patterns, and make accurate predictions. Several studies have explored the use of AI models in credit scoring and their potential benefits.

In [39], the authors have shown that the random forest model [38] outperforms classical machine-learning algorithms (ML). Although the authors assert that gradient boosting methods usually produce superior results, they suggest that the use of deep learning should be taken into consideration due to its greater dependence on the specific financial application [160].

In another study conducted in [161], the authors review multiple models to identify elements that influence credit scores in a microfinance setting, with the aim of constructing an ontological model to assist institutions in decision-making progress.

In the realm of credit scoring research, particularly focusing on company credit scoring, several recent studies need to be mentioned. In [86], the authors investigate using a subset of data to predict the likelihood of default within one year. This study utilizes an XGBoost model, achieving promising results in predicting default probabilities.

Another interesting approach proposed in [99] aims to reproduce rating agency ratings by employing a comprehensive list of financial variables. This approach demonstrates the potential for replicating the complex decision-making processes of rating agencies with machine learning models.

## 8.2.2 Natural Language Processing

NLP is becoming increasingly important for the finance industry due to the vast amounts of unstructured data that are generated in this field. Financial institutions generate huge volumes of data in the form of news articles, social media posts, financial statements, and other documents, making it difficult for humans to analyze and extract insights effectively [162].

Different applications of NLP techniques have been studied. In [5], the authors combine sentiment analysis and financial numeric data to predict short and long-term aerospace stock trends. Another study has shown the effectiveness of incorporating sentiment features in forecasting models [163]. An interesting approach has been introduced by [164]. In this work, the authors have developed an end-to-end model that uses formal documents (e.g., news articles) to extract and update information about customers. The model is composed of a transformer encoder [73] and the BiLSTM-CRF event recognizer [165].

One of the most widely used models in NLP is Bidirectional Encoder Representations from Transformer (BERT). It is a pre-trained model developed by Google and based on the transformer architecture [77]. In [153], the authors present FinancialBERT, a domain-specific language representation model pre-trained on a large financial corpus.

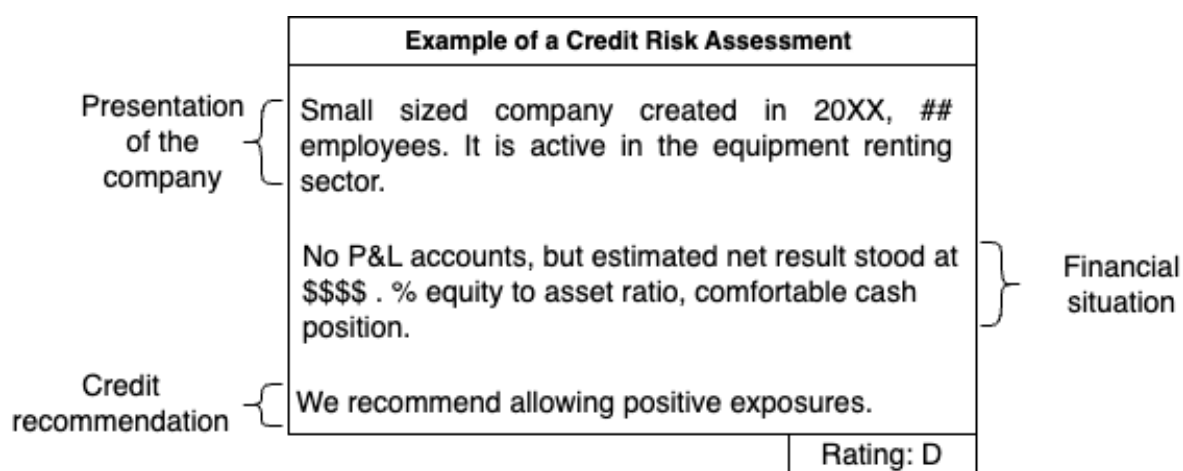


Figure 8.1: Text assessment example generated by a risk analyst

## 8.3 Methodology

In this section, we present the proposed framework. During the first stage, we prepare the text that will be used to train the model. Then, we split the data into two datasets: the train set (i.e., 70% of the data) and the test set (i.e., 30%). The next step consists on fine-tune a pre-trained sentiment analysis model. Finally, we evaluate the performance of the model using the test set.



### 8.3.1 Data

As mentioned in the introduction and in previous chapters the data used for this work has been provided by Tinubu<sup>1</sup>. One of the main activities at Tinubu is to quantitatively give credit risk recommendations. The risk analysts analyze the company using different information (i.e., financial statements, social media, news) and then they write a report in which they synthesize the company and give a recommendation in the form of a rating that range from A (i.e., very low default risk) to F (i.e., extremely high default risk).

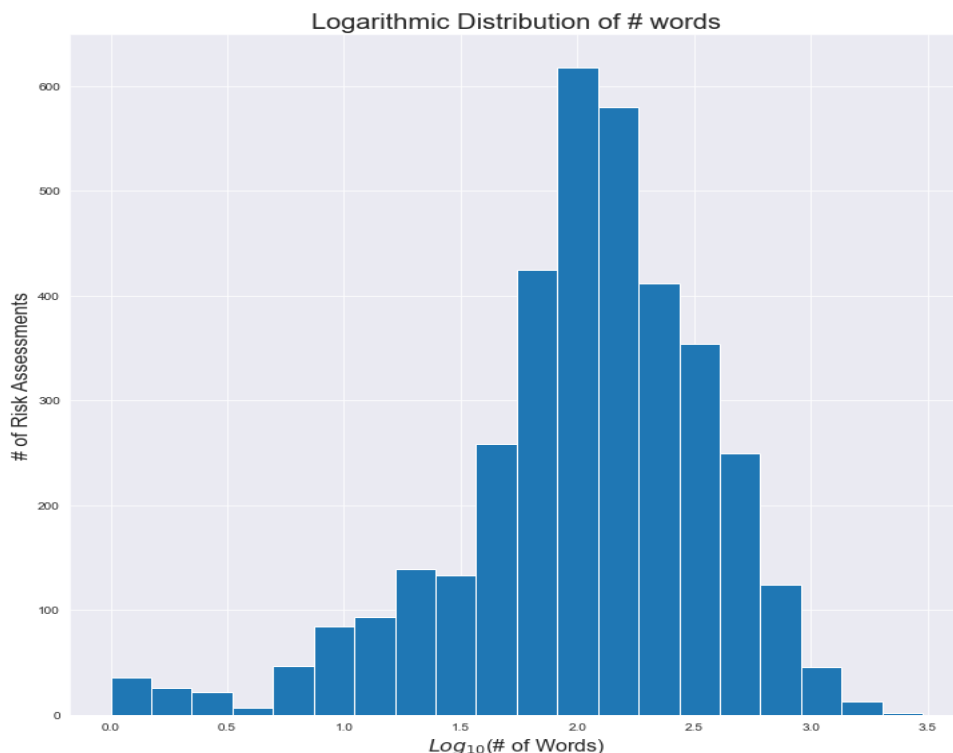


Figure 8.2: Logarithmic Distribution of Text Size (in Words)

Each text is associated with a rating. To label the text for fine-tuning the model, we label companies rated [A-B] as low-risk companies, [C-D] companies as moderate-risk companies, and finally [E-F] companies as high-risk companies.

### 8.3.2 Model Architecture

The architecture of the model used is inspired by the FinancialBERT model [153].

The model is fine-tuned on a sentiment analysis task using a supervised learning approach. The fine-tuning architecture used is the same as that used in BERT, with a dense layer added after the last hidden state of the [CLS] token.

---

<sup>1</sup>The data is not publicly available

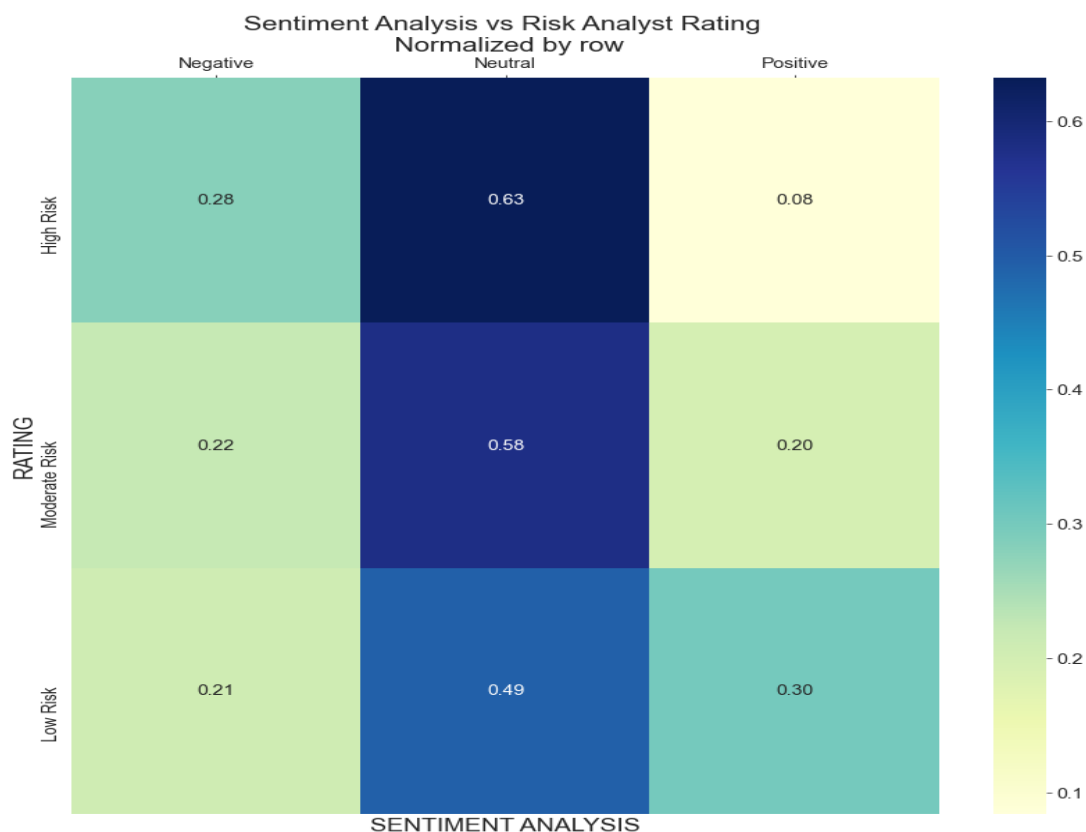


Figure 8.3: The columns represent the sentiment analysis of the Fine-Tuned FinancialBERT. The rows represent the risks associated with each assessed company

The model is trained on the labeled credit risk text assessments. As mentioned previously, each text is labeled depending on the final rating given by the expert. The loss function used is cross-entropy. In our particular case, high risk companies should be labeled by the model as negative, neutral for companies with moderate risk and low-risk companies with a positive label.

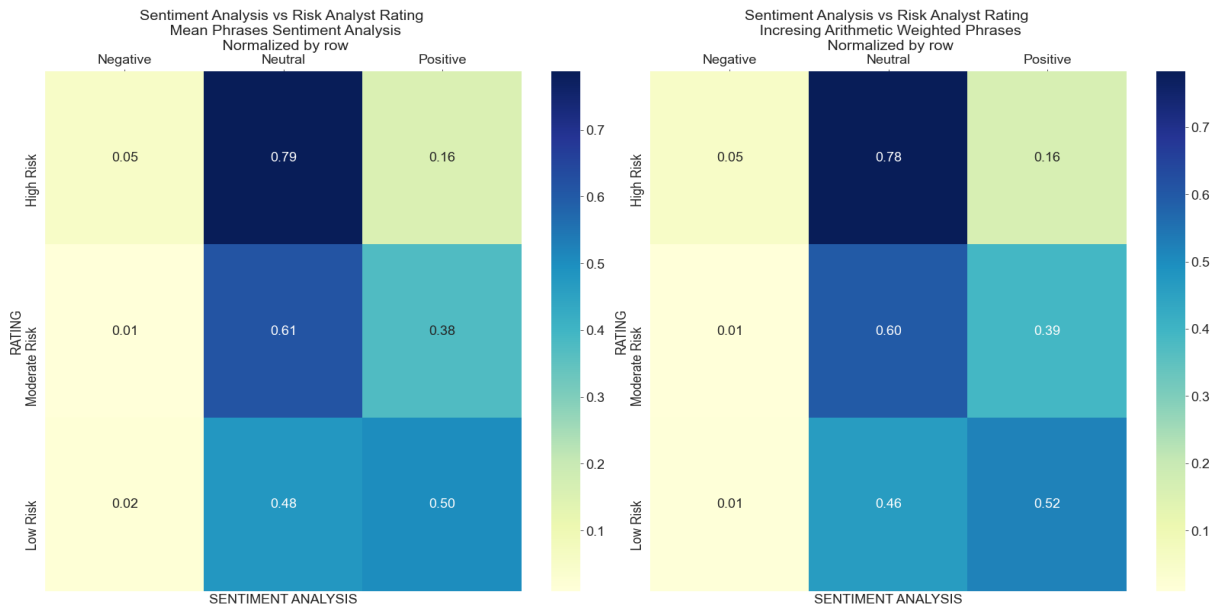
The training is performed with a batch size of 8, a maximum sequence length of is equal to the longest text assessment, and a learning rate of  $2e-5$  for 5 epochs.

### 8.3.3 Experimental Context

We define three different experiments:

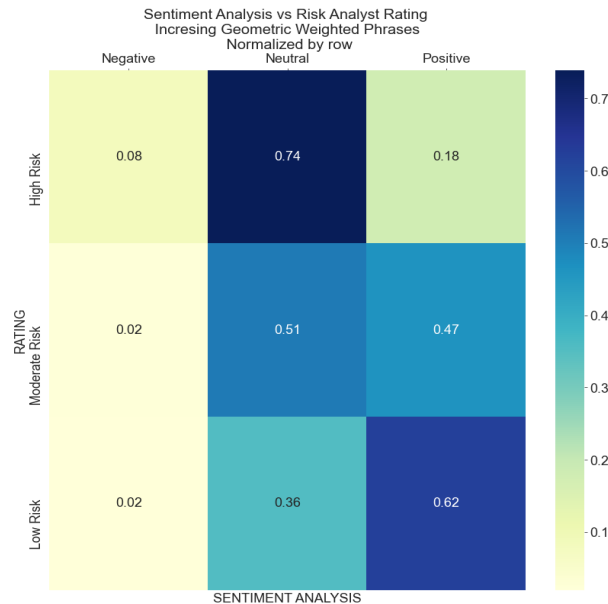
#### Fine-tuned FinancialBERT

In this experiment, we use the fine-tuned FinancialBERT model to classify financial text into three different classes: neutral, positive, and negative. We map each class into a risk



(a) Mean phrases strategy

(b) Linear Growing Importance



(c) Exponential Growing Importance

Figure 8.4: The columns represent the sentiment analysis of the model. The rows represent the risks associated with each assessed company

label (see section III. A).

### Sentiment Analysis using Phrase-Level Processing

We propose in this work a second approach in which we compute the sentiment analysis of the fine-tuned model over each phrase of the text. We define three different approaches

to label the text the split.

Let's define the target feature as follows:

$$y_i = \begin{cases} 1 & \text{Positive} \\ 0 & \text{Neutral} \\ -1 & \text{Negative} \end{cases} \quad (8.1)$$

where  $y_i$  is the label for each phrase. Then we also define  $Y$  as the label of the complete text. The output of the model is a probability of belonging to each class  $\mathbb{P}[y_i] = \mathbb{P}_i$

1. Mean Aggregation:

$$Y = \sum_{i=0}^N y_i \mathbb{P}_i \times \text{ where } N \text{ is the number of phrases present in the text.}$$

2. Linear Growing Importance:

$$Y = \sum_{i=0}^N w_i \mathbb{P}_i y_i. \quad w_i \text{ is the importance of each phrase and is calculated } w_i = \frac{i}{N(N-1)} + \frac{1}{2N}$$

3. Exponential Growing Importance:

$$Y = \sum_{i=0}^N w_i \mathbb{P}_i y_i. \quad w_i \text{ is the importance of each phrase and is calculated } w_i = \frac{r^i(1-r)}{1-r^N}$$

## Sentiment Analysis using End-of-Text Content

This approach is based on the idea that the final phrases of a text can provide valuable information about the overall sentiment expressed in the text. To test this hypothesis, we compute the sentiment analysis of the last paragraph on credit risk assessments.

## 8.4 Results

This section is structured as follows: first, we compare the labels generated by the fine-tuned FinancialBERT with the labels given associated with each risk assessment. Then, we compare the results of Fine-Tuned model using three different Phrase-Level strategies proposed in this study.

### 8.4.1 Fine-Tuned FinancialBERT

In Fig. 8.3, The results of the confusion matrix indicate that the model is overestimating the class "neutral". This means that the model is frequently predicting that the sentiment of the text is neutral, even when it is actually positive or negative. This overestimation of

the neutral class in the confusion matrix could be explained by the neutral tone often used by risk analysts in their texts. Risk text analysis tend to avoid expressions of emotions or bias, thus it is complicated for the model to find the correlation between the risk classes and the sentiment analysis of the text. It is worth noting that the extreme errors in the confusion matrix are not very high. This means that the model is not making a large number of extreme errors, such as confusing positive sentiment with high-risk companies.

### 8.4.2 Phrase-Level Sentiment Analysis

In this section, we compare the three different methods to compute the sentiment analysis using a phrase-level approach. It is worth mentioning that the model used is the fine-tuned FinancialBERT.

Comparing Fig. 8.3 and Fig. 8.4, we observe that the phrase-level sentiment analysis improves the ability to predict the extreme positive class. This indicates that the model is better able to identify positive sentiment in the text data. However, the model overestimates the risk of bankruptcy for high-risk companies and categorizes them as neutral.

Among the three phrase-level strategies used, it is worth highlighting that the strategy that produces the best results is one where the weighting of each phrase grows exponentially (see Fig. 8.4c) This can be explained by observing the typical structure of the analyses performed by experts (see Fig. 8.1). The first part of the text usually summarizes the entity. In the second paragraph, the figures are presented. Finally, the expert closes the text with his or her opinion on the solvency of the company and whether it is correct to grant it credit.

This structure of the text highlights the importance of assigning greater weight to the last phrases, as they often contain the expert's opinion and assessment of the company's solvency. By giving greater weight to these phrases, the model can better capture the sentiment expressed in the text, leading to improved performance.

### 8.4.3 End-of-Text Sentiment Analysis

As observed in the previous results (see Fig. 8.4), the best results are yielded by the exponential growing importance approach. This means that the last phrases of the text tend to better discriminate the risk associated with the text. The improvement in the results in Fig. 8.5 compared to Fig. 8.3 can be explained by the fact the end-of-text strategy provides a focused and relevant representation of the sentiment expressed in the text. By using the last paragraph, the model is able to effectively determine the sentiment of the text.

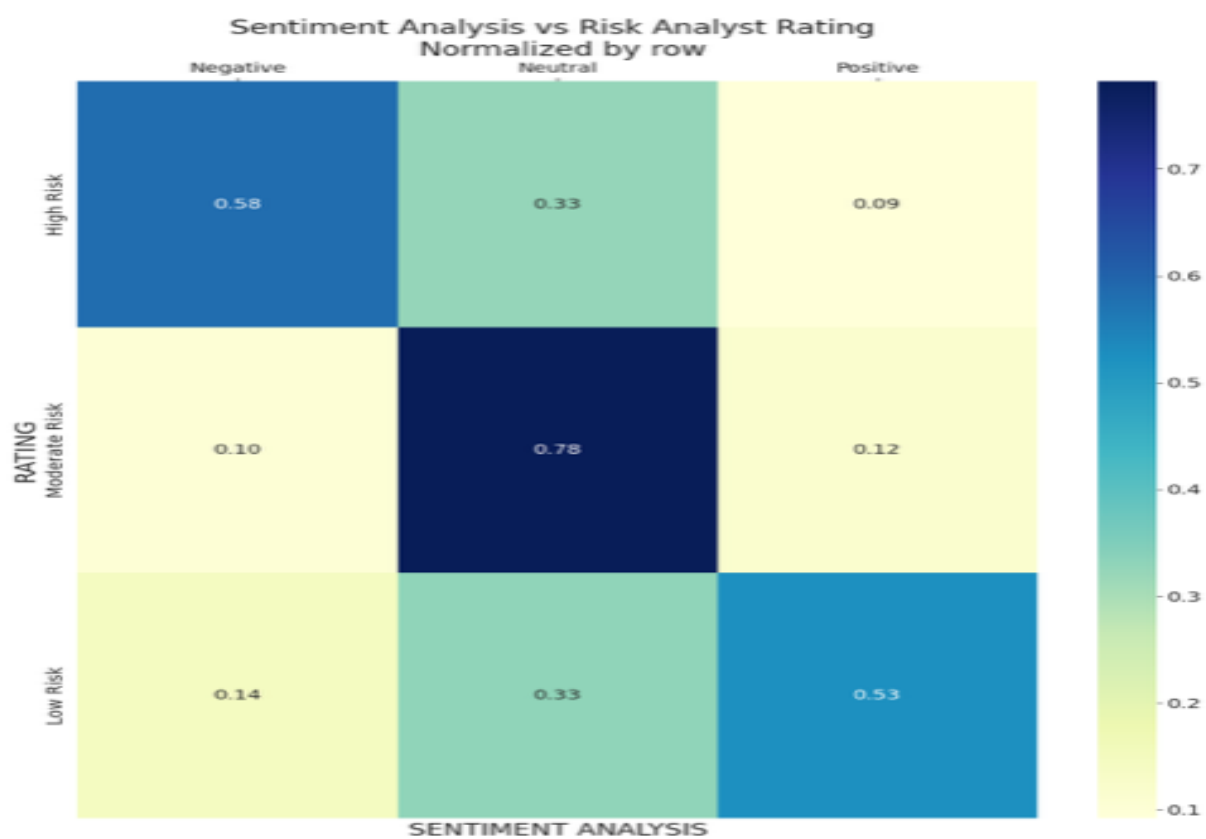


Figure 8.5: The columns represent the sentiment analysis of the End-of-Text Fine-Tuned FinancialBERT. The rows represent the risks associated with each assessed company

## 8.5 Conclusions

In this chapter, we explored the integration of credit risk text assessments made by credit analysts with a fine-tuned FinancialBERT model for sentiment analysis. Our results indicate that the model tends to overestimate the neutral class, which may be attributed to the neutral tone often employed by risk analysts in their texts. However, the extreme errors in the confusion matrix are not very high, suggesting that the model does not frequently make extreme mistakes, such as confusing positive sentiment with high-risk companies.

We further compared three different phrase-level sentiment analysis strategies and found that the one in which the weighting of each phrase grows exponentially produced the best results. Generally, the structure of a risk analysis can be split into three main parts: the presentation of the company, the financial analysis of the company, and the expert's opinion. These results are consistent with the typical structure of expert analyses, which often conclude with the expert's opinion on the solvency of a company.

The end-of-text sentiment analysis strategy, which focuses on the last paragraph, was found to be effective in determining the sentiment of the text, leading to an improvement in the results. Our findings suggest that incorporating credit analyst text assessments into a sentiment analysis model using an end-of-text approach can significantly improve the performance of credit scoring systems. Future work could further investigate the impact of the integration of this type of model with the probability of default models.

# Conclusion and Perspectives

In the introduction of this thesis, we emphasized the pivotal role that credit scoring systems play in fostering economic growth. Moreover, we showed that, from governmental (i.e., Banque de France) to intergovernmental organizations (i.e., World Trade Organization) highlight the importance of robust and ethical credit scoring.

In this thesis, we initiated our investigation by scrutinizing the impact of the COVID-19 outbreak on our dataset. The primary objective was to assess whether the pandemic had a substantial influence on the default rates of companies, which could potentially compromise the generalizability of our predictive models. We revealed that to achieve robust and generalizable default predictions, it is imperative to develop granular models. Specifically, we found that creating models tailored to individual trade sectors significantly enhances predictive accuracy. This granularity allows the models to capture the unique risk factors and economic dynamics that are specific to each sector, thereby making them more resilient to broad economic shocks, such as the COVID-19 pandemic.

In addition to the need for sector-specific granularity, one of the main contributions of this thesis is the importance of aligning the machine learning models with human risk criteria. This alignment is crucial for capturing the domain expertise that risk analysts bring to the table. This compatibility between machine-generated insights and human expertise serves to enhance the model's credibility, making it more acceptable to stakeholders and more effective in real-world applications.

In real-world scenarios, the variability in regulations concerning access to public data for small and medium-sized enterprises (SMEs) poses a significant challenge, especially at the country level. This variability often results in a lack of financial data, making traditional credit scoring methods less applicable. This leads us to the second major contribution of our framework: the development of credit scoring models that do not rely solely on financial data.

Our contribution is fundamentally rooted in the application of Natural Language Processing (NLP) to textual data generated by risk analysts. Recognizing the value of this expert-generated content, we have proposed various methods to systematically analyze



and interpret the text. These methods serve to extract meaningful insights and features that can be incorporated into our credit scoring models. Ultimately, we have developed a comprehensive framework that leverages these NLP-derived features to score companies that have not publicly disclosed their financial data. This framework not only enhances the robustness of our credit scoring system but also significantly broadens its applicability, especially for companies operating in regulatory environments where financial data disclosure is not mandatory.

## Perspectives and Future Work

Future work should concentrate on several key areas to further refine and improve the framework. One primary focus should be the incorporation of credit risk expertise directly into the training loop of the machine learning models. This would serve to align the models more closely with the domain-specific knowledge and criteria used by risk analysts.

Incorporating the unique expertise of risk analysts into machine learning models is particularly important because each company has its own distinct approach to training these professionals. This specialized training is often a reflection of the company's specific organizational philosophies, risk tolerances, and even regulatory requirements. By embedding this domain-specific knowledge into the machine learning model, one essentially captures the nuances and subtleties that come with years of specialized training and experience.

Future efforts should also prioritize the creation of dynamic strategies for monitoring and updating machine learning-based credit scoring models. The aim is to ensure that these models can adapt swiftly to shifts in economic paradigms. Economic landscapes are constantly evolving due to various factors such as technological advancements, regulatory changes, and global events like pandemics or financial crises. Traditional models may become obsolete or less accurate when faced with such rapid changes. Therefore, it's crucial to develop adaptive algorithms that can quickly incorporate new data and insights, allowing for real-time adjustments to the model's predictive capabilities.

Another avenue for future investigation that remains unexplored following the work conducted in this thesis is online data, such as social media text data and news, for credit scoring. Utilizing these additional data sources could be particularly beneficial in scenarios where financial data is sparse or unavailable. By incorporating textual data from diverse and publicly accessible platforms, we can potentially create more robust and comprehensive credit scoring models. This approach could offer a significant improvement in the accuracy and reliability of credit assessments, especially in contexts where traditional financial metrics are lacking or incomplete.

# Bibliography

- [1] S. Albanesi, G. De Giorgi, and J. Nosal, “Credit Growth and the Financial Crisis: A New Narrative,” Tech. Rep. w23740, National Bureau of Economic Research, Cambridge, MA, Aug. 2017.
- [2] M. M. Hasan, J. Popp, and J. Oláh, “Current landscape and influence of big data on finance,” *Journal of Big Data*, vol. 7, p. 21, Dec. 2020.
- [3] P. Azad, N. J. Navimipour, A. M. Rahmani, and A. Sharifi, “The role of structured and unstructured data managing mechanisms in the internet of things,” *Cluster Computing*, vol. 23, pp. 1185–1198, sep 2019.
- [4] S. Jain and E. Fallon, “Leveraging unstructured data to improve customer engagement and revenue in financial institutions: A deep reinforcement learning approach to personalized transaction recommendations,” Jul 2023.
- [5] P. Muthukumar and J. Zhong, “A stochastic time series model for predicting financial trends using nlp,” 2021.
- [6] E. I. Altman, “FINANCIAL RATIOS, DISCRIMINANT ANALYSIS AND THE PREDICTION OF CORPORATE BANKRUPTCY,” *The Journal of Finance*, vol. 23, pp. 589–609, Sept. 1968.
- [7] W. H. Beaver, “Financial ratios as predictors of failure,” *Journal of Accounting Research*, vol. 4, pp. 71–111, 1966.
- [8] “What is a fico score and why is it important?.”
- [9] E. Dornhelm, “U.s, average fico score at 716, indicating improvement in consumer credit behaviors despite pandemic.”
- [10] S. Khemir, C. Baccouche, and S. D. Ayadi, “The influence of ESG information on investment allocation decisions,” *Journal of Applied Accounting Research*, vol. 20, pp. 458–480, dec 2019.

- [11] P. C. Patel, S. Lenka, and V. Parida, “Caste-based discrimination, microfinance credit scores, and microfinance loan approvals among females in india,” *Business and Society*, vol. 61, pp. 372–388, dec 2020.
- [12] G. Zhou, “Measuring investor sentiment,” *Annual Review of Financial Economics*, vol. 10, pp. 239–259, nov 2018.
- [13] G. Serafeim, “Public sentiment and the price of corporate sustainability,” *Financial Analysts Journal*, vol. 76, pp. 26–46, mar 2020.
- [14] R. C. Merton, “On the pricing of corporate debt: the risk structure of interest rates,” *The Journal of Finance*, vol. 29, no. 2, pp. 449–470, 1974.
- [15] WTO, “Trade finance and SMEs Bridging the gaps in provision,”
- [16] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera, “Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence,” *Information Fusion*, vol. 99, p. 101805, 2023.
- [17] F. Rikkers and A. Thibeault, “The influence of rating philosophy on regulatory capital and procyclicality,” in *European Financial Management Association, Annual Meeting, June 24-28, Athens, Greece*, Nyenrode Business Universiteit, 2008.
- [18] S. Fortmann-Roe, “Bias and variance,” *Understanding the Bias-Variance Tradeoff*. Available: <http://s-cott.fortmann-roe.com/docs/Bias-Variance.html>, 2012.
- [19] A. C. Bahnsen, D. Aouada, and B. Ottersten, “Example-dependent cost-sensitive logistic regression for credit scoring,” in *2014 13th International conference on machine learning and applications*, pp. 263–269, IEEE, 2014.
- [20] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, (New York, NY, USA), pp. 785–794, ACM, 2016.
- [21] C.-P. Lee and C.-J. Lin, “A study on l2-loss (squared hinge-loss) multiclass svm,” *Neural computation*, vol. 25, no. 5, pp. 1302–1323, 2013.
- [22] G. Xu, Z. Cao, B.-G. Hu, and J. C. Principe, “Robust support vector machines based on the rescaled hinge loss function,” *Pattern Recognition*, vol. 63, pp. 139–148, 2017.
- [23] T. Hastie, S. Rosset, J. Zhu, and H. Zou, “Multi-class adaboost,” *Statistics and its Interface*, vol. 2, no. 3, pp. 349–360, 2009.
- [24] T. Chengsheng, L. Huacheng, and X. Bing, “Adaboost typical algorithm and its application research,” in *MATEC Web of Conferences*, vol. 139, p. 00222, EDP Sciences, 2017.
- [25] V. Muthukumar, A. Narang, V. Subramanian, M. Belkin, D. Hsu, and A. Sahai, “Classification vs regression in overparameterized regimes: Does the loss function matter?,” *J. Mach. Learn. Res.*, vol. 22, jan 2021.

- 
- [26] C. Bolton *et al.*, *Logistic regression and its application in credit scoring*. PhD thesis, University of Pretoria, 2010.
- [27] A. Steenackers and M. Goovaerts, “A credit scoring model for personal loans,” *Insurance: mathematics & economics*, vol. 8, no. 1, pp. 31–34, 1989.
- [28] M. Bensic, N. Sarlija, and M. Zekic-Susac, “Modelling small-business credit scoring by using logistic regression, neural networks and decision trees,” *Intelligent Systems in Accounting, Finance & Management: International Journal*, vol. 13, no. 3, pp. 133–150, 2005.
- [29] R. A. Eisenbeis, “Problems in applying discriminant analysis in credit scoring models,” *Journal of Banking & Finance*, vol. 2, no. 3, pp. 205–219, 1978.
- [30] J. Mylonakis and G. Diacogiannis, “Evaluating the likelihood of using linear discriminant analysis as a commercial bank card owners credit scoring model,” *International business research*, vol. 3, no. 2, p. 9, 2010.
- [31] A. E. Hoerl and R. W. Kennard, “Ridge regression: Biased estimation for nonorthogonal problems,” *Technometrics*, vol. 42, no. 1, pp. 80–86, 2000.
- [32] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *J. R. Stat. Soc.*, vol. 58, pp. 267–288, Jan. 1996.
- [33] F. Salehi, E. Abbasi, and B. Hassibi, “The impact of regularization on high-dimensional logistic regression,” in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019.
- [34] B. Charbuty and A. Abdulazeez, “Classification based on decision tree algorithm for machine learning,” *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, 2021.
- [35] M. Czajkowski and M. Kretowski, “The role of decision tree representation in regression problems—an evolutionary perspective,” *Applied soft computing*, vol. 48, pp. 458–475, 2016.
- [36] Y. Bengio, O. Delalleau, and C. Simard, “Decision trees do not generalize to new variations,” *Comput. Intell.*, vol. 26, pp. 449–467, Nov. 2010.
- [37] C. Kingsford and S. L. Salzberg, “What are decision trees?,” *Nature biotechnology*, vol. 26, no. 9, pp. 1011–1013, 2008.
- [38] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [39] S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas, “Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research,” *European Journal of Operational Research*, vol. 247, pp. 124–136, Nov. 2015.
- [40] S. Y. Laurent Dupont, Olivier Fliche, “Gouvernance des algorithmes d’intelligence artificielle dans le secteur financier,” 2020.

- [41] B. Baesens, T. V. Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen, “Benchmarking state-of-the-art classification algorithms for credit scoring,” *The Journal of the Operational Research Society*, vol. 54, no. 6, pp. 627–635, 2003.
- [42] I. S. Al-Mejibli, D. H. Abd, J. K. Alwan, and A. J. Rabash, “Performance evaluation of kernels in support vector machine,” in *2018 1st Annual International Conference on Information and Sciences (AiCIS)*, pp. 96–101, Nov 2018.
- [43] F.-L. Chen and F.-C. Li, “Combination of feature selection approaches with SVM in credit scoring,” *Expert Syst. Appl.*, vol. 37, pp. 4902–4909, July 2010.
- [44] L. Han, L. Han, and H. Zhao, “Orthogonal support vector machine for credit scoring,” *Eng. Appl. Artif. Intell.*, vol. 26, pp. 848–862, Feb. 2013.
- [45] S. Shirataki and S. Yamaguchi, “A study on interpretability of decision of machine learning,” in *2017 IEEE International Conference on Big Data (Big Data)*, pp. 4830–4831, 2017.
- [46] L. Breiman, “Bagging predictors,” *Machine learning*, vol. 24, pp. 123–140, 1996.
- [47] D. Zhang, X. Zhou, S. C. Leung, and J. Zheng, “Vertical bagging decision trees model for credit scoring,” *Expert Systems with Applications*, vol. 37, no. 12, pp. 7838–7843, 2010.
- [48] P. Branco, L. Torgo, and R. P. Ribeiro, “Rebagg: Resampled bagging for imbalanced regression,” in *Second International Workshop on Learning with Imbalanced Domains: Theory and Applications*, pp. 67–81, PMLR, 2018.
- [49] X. Zhang, Y. Yang, and Z. Zhou, “A novel credit scoring model based on optimized random forest,” in *2018 IEEE 8th annual computing and communication workshop and conference (CCWC)*, pp. 60–65, IEEE, 2018.
- [50] Y. Freund, R. E. Schapire, *et al.*, “Experiments with a new boosting algorithm,” in *icml*, vol. 96, pp. 148–156, Citeseer, 1996.
- [51] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.
- [52] G. Wang, J. Hao, J. Ma, and H. Jiang, “A comparative assessment of ensemble learning for credit scoring,” *Expert systems with applications*, vol. 38, no. 1, pp. 223–230, 2011.
- [53] J. Hatwell, M. M. Gaber, and R. M. Atif Azad, “Ada-whips: explaining adaboost classification with applications in the health sciences,” *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, pp. 1–25, 2020.
- [54] M. Xiao and Y. Guo, “Multi-view adaboost for multilingual subjectivity analysis,” in *Proceedings of COLING 2012*, pp. 2851–2866, 2012.

- 
- [55] J. H. Friedman, “Greedy function approximation: a gradient boosting machine,” *Annals of statistics*, pp. 1189–1232, 2001.
- [56] Z. Zhang, Y. Zhao, A. Canes, D. Steinberg, O. Lyashevskaya, *et al.*, “Predictive analytics with gradient boosting in clinical medicine,” *Annals of translational medicine*, vol. 7, no. 7, 2019.
- [57] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, “Explainable machine learning in credit risk management,” *Computational Economics*, 09 2020.
- [58] S. S. Reddy, S. Mandal, V. L. Kasyap, and R. Aswathy, “A novel approach to detect fake news using extreme gradient boosting,” in *2022 10th International Symposium on Digital Forensics and Security (ISDFS)*, pp. 1–4, IEEE, 2022.
- [59] F. K. Dosilovic, M. Brcic, and N. Hlupic, “Explainable artificial intelligence: A survey,” in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, (Opatija), pp. 0210–0215, IEEE, May 2018.
- [60] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, *et al.*, “Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai,” *Information fusion*, vol. 58, pp. 82–115, 2020.
- [61] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, “Xai—explainable artificial intelligence,” *Science robotics*, vol. 4, no. 37, p. eaay7120, 2019.
- [62] A. Shaban-Nejad, M. Michalowski, and D. L. Buckeridge, “Explainability and interpretability: keys to deep medicine,” *Explainable AI in healthcare and medicine: Building a culture of transparency and accountability*, pp. 1–10, 2021.
- [63] L. M. Demajo, V. Vella, and A. Dingli, “Explainable ai for interpretable credit scoring,” *Computer Science & Information Technology (CS & IT)*, 2020.
- [64] S. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” 2017.
- [65] A. Salih, Z. Raisi-Estabragh, I. B. Galazzo, P. Radeva, S. E. Petersen, G. Menegaz, and K. Lekadir, “Commentary on explainable artificial intelligence methods: Shap and lime,” *arXiv preprint arXiv:2305.02012*, 2023.
- [66] J. Yang, “Fast treeshap: Accelerating shap value computation for trees,” *arXiv preprint arXiv:2109.09847*, 2021.
- [67] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, “Explainable AI in Credit Risk Management,” *SSRN Electronic Journal*, 2019.
- [68] L. O. Hjelkrem and P. E. d. Lange, “Explaining deep learning models for credit scoring with shap: A case study using open banking data,” *Journal of Risk and Financial Management*, vol. 16, no. 4, p. 221, 2023.

- [69] Y. Zhang, R. Jin, and Z.-H. Zhou, “Understanding bag-of-words model: a statistical framework,” *International journal of machine learning and cybernetics*, vol. 1, pp. 43–52, 2010.
- [70] Y. HaCohen-Kerner, D. Miller, and Y. Yigal, “The influence of preprocessing on text classification using a bag-of-words representation,” *PloS one*, vol. 15, no. 5, p. e0232525, 2020.
- [71] G. V. Cormack, J. M. Gómez Hidalgo, and E. P. Sáenz, “Spam filtering for short messages,” in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pp. 313–320, 2007.
- [72] A. Thakkar and K. Chaudhari, “Predicting stock trend using an integrated term frequency–inverse document frequency-based feature weight matrix with neural networks,” *Applied Soft Computing*, vol. 96, p. 106684, 2020.
- [73] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2023.
- [74] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” Sept. 2013. arXiv:1301.3781 [cs].
- [75] J. Pennington, R. Socher, and C. Manning, “GloVe: Global vectors for word representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Doha, Qatar), pp. 1532–1543, Association for Computational Linguistics, Oct. 2014.
- [76] F. Xue, X. Li, T. Zhang, and N. Hu, “Stock market reactions to the covid-19 pandemic: The moderating role of corporate big data strategies based on word2vec,” *Pacific-Basin Finance Journal*, vol. 68, p. 101608, 2021.
- [77] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” May 2019. arXiv:1810.04805 [cs].
- [78] Q. Chen, “Stock movement prediction with financial news using contextualized embedding from bert,” *arXiv preprint arXiv:2107.08721*, 2021.
- [79] “Les PME et TPE en France : une situation financière améliorée et un accès au crédit plus facile,” Dec. 2019.
- [80] T. S. Hatten, *Coursemate with online interactive business plan and liveplan, 1 term (6 months) printed access card for hatten’s small business management: Entrepreneurship and beyond, 6th*. Taipei, Taiwan: Cengage Learning, 6 ed., Feb. 2015.
- [81] Y. Lu, J. Wu, J. Peng, and L. Lu, “The perceived impact of the Covid-19 epidemic: evidence from a sample of 4807 SMEs in Sichuan Province, China,” *Environmental Hazards*, vol. 19, pp. 323–340, Aug. 2020. Publisher: Taylor & Francis \_eprint: <https://doi.org/10.1080/17477891.2020.1763902>.

- [82] K. Grondys, O. Ślusarczyk, H. I. Hussain, and A. Androniceanu, “Risk Assessment of the SME Sector Operations during the COVID-19 Pandemic,” *International Journal of Environmental Research and Public Health*, vol. 18, p. 4183, Jan. 2021. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.
- [83] J. Juergensen, J. Guimón, and R. Narula, “European SMEs amidst the COVID-19 crisis: assessing impact and policy responses,” *Journal of Industrial and Business Economics*, vol. 47, pp. 499–510, jul 2020.
- [84] S. Kalemli-Ozcan, P.-O. Gourinchas, V. Penciakova, and N. Sander, “COVID-19 and SME Failures,” *IMF Working Papers*, vol. 20, Sept. 2020.
- [85] A. El Qadi, M. Trocan, N. Díaz-Rodríguez, and T. Frossard, “Feature contribution alignment with expert knowledge for artificial intelligence credit scoring,” *Signal, Image and Video Processing*, vol. 17, pp. 427–434, Mar. 2023.
- [86] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, “Explainable AI in Fintech Risk Management,” *Frontiers in Artificial Intelligence*, vol. 3, p. 26, Apr. 2020.
- [87] A. El-Qadi, M. Trocan, T. Frossard, and N. Díaz-Rodríguez, “Sectorial analysis impact on the development of credit scoring machine learning models,” in *Proceedings of the 14th International Conference on Management of Digital EcoSystems*, pp. 115–122, 2022.
- [88] D. J. Hand and W. E. Henley, “Statistical classification methods in consumer credit scoring: a review,” *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 160, no. 3, pp. 523–541, 1997.
- [89] E. I. Altman, “Financial ratios, discriminant analysis and the prediction of corporate bankruptcy,” *The Journal of Finance*, vol. 23, no. 4, pp. 589–609, 1968.
- [90] S. Albanesi, G. De Giorgi, and J. Nosal, “Credit growth and the financial crisis: A new narrative,” Working Paper 23740, National Bureau of Economic Research, 8 2017.
- [91] M. Cowling, W. Liu, and A. Ledger, “Small business financing in the uk before and during the current financial crisis,” *International Small Business Journal*, vol. 30, pp. 778–800, 11 2012.
- [92] H. Kim, H. Cho, and D. Ryu, “Corporate default predictions using machine learning: Literature review,” *Sustainability*, vol. 12, p. 6325, 08 2020.
- [93] J. A. Ohlson, “Financial ratios and the probabilistic prediction of bankruptcy,” *Journal of Accounting Research*, vol. 18, no. 1, pp. 109–131, 1980.
- [94] S. Chava and R. A. Jarrow, “Bankruptcy Prediction with Industry Effects\*,” *Review of Finance*, vol. 8, pp. 537–569, 01 2004.



- [95] S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas, “Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research,” *European Journal of Operational Research*, vol. 247, no. 1, pp. 124 – 136, 2015.
- [96] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *Advances in neural information processing systems*, vol. 30, 2017.
- [97] P. Makowski, “Credit scoring branches out,” *Credit World*, vol. 75, no. 1, pp. 30–37, 1985.
- [98] T. Chen and C. Guestrin, “Xgboost,” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug 2016.
- [99] A. R. Provenzano, D. Trifirò, A. Datteo, L. Giada, N. Jean, A. Riciputi, G. L. Pera, M. Spadaccino, L. Massaron, and C. Nordio, “Machine learning approach for credit scoring,” 2020.
- [100] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, “Lightgbm: A highly efficient gradient boosting decision tree,” in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.
- [101] M. Moscatelli, F. Parlapiano, S. Narizzano, and G. Viggiano, “Corporate default forecasting with machine learning,” *Expert Systems with Applications*, vol. 161, p. 113567, 2020.
- [102] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: Synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research*, vol. 16, p. 321–357, 6 2002.
- [103] S. R. Islam, W. Eberle, S. K. Ghafoor, S. C. Bundy, D. A. Talbert, and A. Siraj, “Investigating bankruptcy prediction models in the presence of extreme class imbalance and multiple stages of economy,” 2019.
- [104] P. M. Addo, D. Guegan, and B. Hassani, “Credit risk analysis using machine and deep learning models,” *Risks*, vol. 6, no. 2, 2018.
- [105] F. K. Dosilovic, M. Brcic, and N. Hlupic, “Explainable artificial intelligence: A survey,” *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 0210–0215, 2018.
- [106] O. Kuiper, M. v. d. Berg, J. v. d. Burgt, and S. Leijnen, “Exploring explainable ai in the financial sector: Perspectives of banks and supervisory authorities,” in *Benelux Conference on Artificial Intelligence*, pp. 105–119, Springer, 2021.
- [107] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, “A survey of methods for explaining black box models,” *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

- 
- [108] A. B. Arrieta, N. Díaz-Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, “Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai,” 2019.
- [109] O. Troyanskaya, M. Cantor, G. Sherlock, T. Hastie, R. Tibshirani, D. Botstein, and R. Altman, “Missing value estimation methods for dna microarrays,” 07 2001.
- [110] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [111] M. Cowling, W. Liu, and A. Ledger, “Small business financing in the UK before and during the current financial crisis,” *International Small Business Journal: Researching Entrepreneurship*, vol. 30, pp. 778–800, Nov. 2012. TO ADD: IPAD Summary and PDF.
- [112] R. C. Shelburne, “The Global Financial Crisis and Its Impact on Trade: The World and the European Emerging Economies,” no. 2010, p. 30, 2010.
- [113] “Trade finance for SMEs in the digital era,” OECD SME and Entrepreneurship Papers 24, May 2021. TO READ (100%° to understand Trade Finance). Series: OECD SME and Entrepreneurship Papers Volume: 24.
- [114] J. A. Ohlson, “Financial Ratios and the Probabilistic Prediction of Bankruptcy,” *Journal of Accounting Research*, vol. 18, no. 1, p. 109, 1980.
- [115] H. Kim, H. Cho, and D. Ryu, “Corporate Default Predictions Using Machine Learning: Literature Review,” *Sustainability*, vol. 12, p. 6325, Aug. 2020.
- [116] A. R. Provenzano, D. Trifirò, A. Datteo, L. Giada, N. Jean, A. Riciputi, G. L. Pera, M. Spadaccino, L. Massaron, and C. Nordio, “Machine Learning approach for Credit Scoring,” July 2020. Number: arXiv:2008.01687 arXiv:2008.01687 [q-fin, stat].
- [117] V. Babenko, A. Panchyshyn, L. Zomchak, M. Nehrey, Z. Artym-Drohomyretska, and T. Lahotskyi, “Classical Machine Learning Methods in Economics Research: Macro and Micro Level Examples,” *WSEAS TRANSACTIONS ON BUSINESS AND ECONOMICS*, vol. 18, pp. 209–217, Jan. 2021.
- [118] N. Nehrebecka, “PREDICTING THE DEFAULT RISK OF COMPANIES. COMPARISON OF CREDIT SCORING MODELS: LOGIT VS SUPPORT VECTOR MACHINES,” *ECONOMETRICS*, vol. 22, no. 2, pp. 54–73, 2018.
- [119] L. Svabova, L. Michalkova, M. Durica, and E. Nica, “Business Failure Prediction for Slovak Small and Medium-Sized Companies,” *Sustainability*, vol. 12, p. 4572, June 2020.

- [120] C. Coquidé, J. Lages, and D. L. Shepelyansky, “Interdependence of Sectors of Economic Activities for World Countries from the Reduced Google Matrix Analysis of WTO Data,” *Entropy*, vol. 22, p. 1407, Dec. 2020.
- [121] Y. Chen, P. Giudici, K. Liu, and E. Raffinetti, “Measuring fairness in credit scoring,” 2022.
- [122] S. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” 2017.
- [123] J. Bergstra, D. Yamins, and D. Cox, “Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures,” in *Proceedings of the 30th International Conference on Machine Learning* (S. Dasgupta and D. McAllester, eds.), vol. 28 of *Proceedings of Machine Learning Research*, (Atlanta, Georgia, USA), pp. 115–123, PMLR, 17–19 Jun 2013.
- [124] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, “API design for machine learning software: experiences from the scikit-learn project,” in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122, 2013.
- [125] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, “Lightgbm: A highly efficient gradient boosting decision tree,” *Advances in neural information processing systems*, vol. 30, pp. 3146–3154, 2017.
- [126] F. S. a. C. M. U. Financial Stability, “Review of country-by-country reporting requirements for extractive and logging industries,” Nov. 2018.
- [127] ICAEW, “SME accounting requirements: basing policy on evidence,” 2018.
- [128] R. Florez-Lopez, “Effects of missing data in credit risk scoring. A comparative analysis of methods to achieve robustness in the absence of sufficient data,” *Journal of the Operational Research Society*, vol. 61, pp. 486–501, Mar. 2010.
- [129] P. Giudici, B. Hadji-Misheva, and A. Spelta, “Network Based Scoring Models to Improve Credit Risk Management in Peer to Peer Lending Platforms,” *Frontiers in Artificial Intelligence*, vol. 2, p. 3, May 2019.
- [130] Y. Wang and X. S. Ni, “Improving Investment Suggestions for Peer-to-Peer Lending via Integrating Credit Scoring into Profit Scoring,” in *Proceedings of the 2020 ACM Southeast Conference*, (Tampa FL USA), pp. 141–148, ACM, Apr. 2020.
- [131] A. R. Provenzano, D. Trifirò, N. Jean, G. L. Pera, M. Spadaccino, L. Massaron, and C. Nordio, “An Artificial Intelligence approach to Shadow Rating,” Dec. 2019. arXiv:1912.09764 [cs, q-fin].
- [132] D. Memic, “Assessing Credit Default using Logistic Regression and Multiple Discriminant Analysis: Empirical Evidence from Bosnia and Herzegovina,” *Interdisciplinary Description of Complex Systems*, vol. 13, no. 1, pp. 128–153, 2015.

- [133] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, “Light-GBM: A Highly Efficient Gradient Boosting Decision Tree,” in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.
- [134] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, Aug. 2016. arXiv: 1603.02754.
- [135] S. T. A. Niaki and S. Hoseinzade, “Forecasting S&P 500 index using artificial neural networks and design of experiments,” *Journal of Industrial Engineering International*, vol. 9, p. 1, Dec. 2013.
- [136] H. M, G. E.A., V. K. Menon, and S. K.P., “NSE Stock Market Prediction Using Deep-Learning Models,” *Procedia Computer Science*, vol. 132, pp. 1351–1362, 2018.
- [137] H. Akaike, “A new look at the statistical model identification,” *IEEE Transactions on Automatic Control*, vol. 19, pp. 716–723, Dec. 1974.
- [138] P. Addo, D. Guegan, and B. Hassani, “Credit Risk Analysis Using Machine and Deep Learning Models,” *Risks*, vol. 6, p. 38, Apr. 2018.
- [139] B. Niu, J. Ren, and X. Li, “Credit Scoring Using Machine Learning by Combing Social Network Information: Evidence from Peer-to-Peer Lending,” *Information*, vol. 10, p. 397, Dec. 2019.
- [140] S. Y. Sohn, D. H. Kim, and J. H. Yoon, “Technology credit scoring model with fuzzy logistic regression,” *Applied Soft Computing*, vol. 43, pp. 150–158, June 2016.
- [141] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, “Review of deep learning: concepts, CNN architectures, challenges, applications, future directions,” *Journal of Big Data*, vol. 8, p. 53, Mar. 2021.
- [142] D. W. Otter, J. R. Medina, and J. K. Kalita, “A Survey of the Usages of Deep Learning in Natural Language Processing,” Dec. 2019. arXiv:1807.10854 [cs].
- [143] L. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, pp. 257–286, Feb. 1989.
- [144] P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin, “A statistical approach to machine translation,” *Computational Linguistics*, vol. 16, no. 2, pp. 79–85, 1990.
- [145] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil, “Learning semantic representations using convolutional neural networks for web search,” in *Proceedings of the 23rd International Conference on World Wide Web*, (Seoul Korea), pp. 373–374, ACM, Apr. 2014.

- [146] C. N. d. Santos and V. Guimarães, “Boosting Named Entity Recognition with Neural Character Embeddings,” May 2015. arXiv:1505.05008 [cs].
- [147] P. Conde-Cespedes, J. Chavando, and E. Deberry, “Detection of Suspicious Accounts on Twitter Using Word2Vec and Sentiment Analysis,” in *Multimedia and Network Information Systems* (K. Choroś, M. Kopel, E. Kukla, and A. Siemiński, eds.), (Cham), pp. 362–371, Springer International Publishing, 2019.
- [148] A. Berger and J. Lafferty, “Information retrieval as statistical translation,” in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, (Berkeley California USA), pp. 222–229, ACM, Aug. 1999.
- [149] M. Rambocas and B. G. Pacheco, “Online sentiment analysis in marketing research: a review,” *Journal of Research in Interactive Marketing*, vol. 12, pp. 146–163, May 2018.
- [150] A. Gupta, V. Dengre, H. A. Kheruwala, and M. Shah, “Comprehensive review of text-mining applications in finance,” *Financial Innovation*, vol. 6, p. 39, Dec. 2020.
- [151] R. Gupta and M. Chen, “Sentiment Analysis for Stock Price Prediction,” in *2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, (Shenzhen, Guangdong, China), pp. 213–218, IEEE, Aug. 2020.
- [152] M. Stevenson, C. Mues, and C. Bravo, “The value of text for small business default prediction: A Deep Learning approach,” *European Journal of Operational Research*, vol. 295, pp. 758–771, Dec. 2021.
- [153] A. R. Hazourli, “FinancialBERT - A Pretrained Language Model for Financial Text Mining,” 2022. Publisher: Unpublished.
- [154] I. T. Jolliffe and J. Cadima, “Principal component analysis: a review and recent developments,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, p. 20150202, Apr. 2016.
- [155] V. Raunak, V. Gupta, and F. Metze, “Effective dimensionality reduction for word embeddings,” in *Proceedings of the 4th Workshop on Representation Learning for NLP (Repl4NLP-2019)*, (Florence, Italy), pp. 235–243, Association for Computational Linguistics, Aug. 2019.
- [156] F. A. Gers, J. Schmidhuber, and F. Cummins, “Learning to Forget: Continual Prediction with LSTM,” *Neural Computation*, vol. 12, pp. 2451–2471, Oct. 2000.
- [157] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling,” Dec. 2014. arXiv:1412.3555 [cs].
- [158] L. Cao, “AI in finance: A review,” *SSRN Electron. J.*, 2020.

- [159] S. Bhatore, L. Mohan, and Y. R. Reddy, “Machine learning techniques for credit risk evaluation: a systematic literature review,” *Journal of Banking and Financial Technology*, vol. 4, pp. 111–138, Apr. 2020.
- [160] M. Schmitt, “Deep learning vs. gradient boosting: Benchmarking state-of-the-art machine learning algorithms for credit scoring,” *arXiv preprint arXiv:2205.10535*, 2022.
- [161] K. Ben Addi and N. Souissi, “An ontology-based model for credit scoring knowledge in microfinance: Towards a better decision making,” 08 2020.
- [162] I. Goldstein, C. S. Spatt, and M. Ye, “Big data in finance,” *SSRN Electron. J.*, 2021.
- [163] R. Gupta and M. Chen, “Sentiment analysis for stock price prediction,” in *2020 IEEE conference on multimedia information processing and retrieval (MIPR)*, pp. 213–218, IEEE, 2020.
- [164] S. Zheng, W. Cao, W. Xu, and J. Bian, “Doc2edag: An end-to-end document-level framework for chinese financial event extraction,” 2019.
- [165] Z. Huang, W. Xu, and K. Yu, “Bidirectional lstm-crf models for sequence tagging,” 2015.

