Managing Credit Risk and the Cost of Equity with Machine Learning Techniques

By

Yujing Chi

Submitted in Fulfilment of the Requirements of the Degree of Doctor of Philosophy

Adam Smith Business School
College of Social Science
University of Glasgow

July 2022

Abstract

Credit risks and the cost of equity can influence market participants' activities in many ways. Providing in-depth analysis can help participants reduce potential costs and make profitable strategies. This kind of study is usually armed with conventional statistical models built with researchers' knowledge. However, with the advancement of technology, a massive amount of financial data increasing in volume, subjectivity, and heterogeneity becomes challenging to process conventionally. Machine learning (ML) techniques have been utilised to handle this difficulty in real-life applications. This PhD thesis consists of three major empirical essays. We employ state-of-art machine learning techniques to predict peer-to-peer (P2P) lending default risk, P2P lending decisions, and Environmental, Social, Corporate Governance (ESG) effects on firms' cost of equity.

In the era of financial technology, P2P lending has gained considerable attention among academics and market participants. In the first essay (Chapter 2), we investigate the determinants of P2P lending default prediction in relation to borrowers' characteristics and credit history. Applying machine learning techniques, we document substantial predictive ability compared with the benchmark logit model. Further, we find that the LightGBM has superior predictive power and outperforms all other models in all out-of-sample predictions. Finally, we offer insights into different levels of uncertainty in P2P loan groups and the value of machine learning in credit risk mitigation of P2P loan providers.

Macroeconomic impact on funding decisions or lending standards reflects the risk-taking behaviour of market participants. It has been widely discussed by academics. But in the era of financial technology, it leaves a gap in the evidence of lending standards change in a FinTech nonbank financial organisation. The second essay (Chapter 3) aims to fill the gap by introducing loan-level and macroeconomic variables into the predictive models to estimate the P2P loan funding decision. Over 12 million empirical instances are under study while big data techniques, including text mining and five state-of-the-art approaches, are utilised. We note that macroeconomic condition affects individual risk-taking and reaching-for-yield behaviour. Finally, we offer insight into macroeconomic impact in terms of different levels of uncertainty in different P2P loan application groups.

In the third essay (Chapter 4), we use up-to-date machine learning techniques to provide new evidence for the impact of ESG on the cost of equity. Using 15,229 firm-year observations from 51 different countries over the past 18 years, we document negative causal effects on the cost of equity. In addition, we uncover non-linear effects because the level of ESG effects on the equity cost decrease with the enhancements of ESG performance. Furthermore, we note the heterogeneity in ESG effects in different regions by breaking down our sample. Finally, we find that global crises change the sensitivity of the equity cost towards ESG, and the change varies in areas.

Table of Contents

List of Tables

List of Figures

Acknowledgement

I still remember my first PhD student event when a heavy thesis was shown to us - "For my youth", said on the front page. I did not appreciate it back to that time. But now, standing at the end of my own PhD journey where I have been working at the ever-shifting boundary between the known and unknown, I would like to say the same words. While happy to achieve this milestone, I can think of so many people to thank.

First and foremost, I would like to express my deepest gratitude to my principal supervisor Professor Georgios Sermpinis. It was he who opened this door for me. Without his inspiration and patience, I would not have started a PhD in the first place. Similarly, my profound thanks go to my second supervisor Professor Serafeim Tsoukas for his guidance and support. He has shown me the attitude of a real scholar. I am sincerely grateful for my supervisors' enlightenment and support throughout the journey. They have offered me brilliant research ideas and great opportunities in academia.

My special thank goes to Jonathan Roberts, who has emotionally and physically supported and companied me in the last three and a half years. He has witnessed every step I take and has acted as my comfort zone. His words, "We learn more from failure", become my motto for life. I would also like to thank everyone from College Services of Glasgow International College, especially Donald White. They bring so many colours to my drab PhD life and the sense of belonging that makes me not a simple international student but an adult with work and friends.

Finally, I am heavily indebted to my family - my dearest parents, Mei Tang and Yaojie Chi, and my beloved boyfriend, Vladimir Revicky, for their unconditional love and support in the past, present, and future. I appreciate that my parents always let me make my own decisions and be different. I always feel lucky to have Vlad. He is my home and my guide lamp on every path. He has all the answers I am looking for and makes me a better me. My family is always my safety net and my greatest inspiration. I hope I make them happy and proud.

My PhD also belongs to the people above.

Author's Declaration

I declare that, except where explicit reference is made to the contribution of others, that this dissertation is the result of mt own work and has not been submitted for any other degree that the University of Glasgow or any other institution.

Printed Name: Yujing Chi

Signature:

Chapter 1

Chapter 1 Introduction

1.1. Motivation

Financial analysis has a significant impact on the activities of market participants. Providing powerful predictive analysis helps them mitigate risks, manage the developing strategies, and introduce productive rules. The last few decades have witnessed an exponential growth in the average person's interest in the financial market (Badolia, 2016). Along with assets worth billions of dollars traded on the financial market every day, investors acting on the market desire to obtain gain beyond their investment horizon, while firm operators seek efficient plans to minimise the cost of equity. In this study, we focus on the credit and equity market. Specifically, we examine the success of peer-to-peer (P2P) lending loans, and the relationship between social responsibility and the cost of equity.

Credit risks are responsible for most of the financial crises across the world. No matter the form of the economy, credit risk is a crucial factor in evaluating and determining the profit of the social and economic participants. "A comprehensive perspective towards credit sustainability is critical to meet the expectation of both clients and investors. Banking is one of the vital examples where the lenders and the borrower's ability to meet financial obligations are to be achieved through the use of emerging technologies for gaining sustainable competitive advantages" (Pandey et al., 2021). In addition to the traditional banks, online financial innovation services such as P2P also play an important role in serving the real economy. P2P mainly matches the investor and the borrower via their online requests. This may lead to enormous losses for the investor when the borrower defaults. P2P requires mature supporting regulatory systems, although it has gradually stabilised worldwide. Consequently, detecting, predicting, and preventing credit risk is a foremost priority for the market.

Predictive analysis for credit risk mainly estimates the default probability of a loan applicant in terms of personal and demographic information. There are two mainstreams to compute this estimation, expert experience and statistical algorithms. Traditional financial institutions used to grant loan applications based on the experience gained from previous applicants, which tends to be biased. However, the last couple of decades have seen an advancement in technology. A massive amount of credit data booming in volume, subjectivity, and heterogeneity is becoming a challenge for manual processes. Thus, machine learning approaches, including both regression and classification techniques, are employed to address this difficulty in practical applications. They utilise applicants' historical data to generate models that offer robust predictions and decisions when new inputs are given. Acting as training and predicting machines, machine learning algorithms work with data to automatically compute knowledge representation models, to automatically produce expert systems, to learn to plan and game, to construct numerical and qualitative models, to classify and analyse text, to undertake knowledge acquisition, to control dynamic process, to recognise handwriting, to identify objectives from imagines, and other real-life problems. They have been greatly helpful in credit scoring and ratings (Chen et al., 2022; Liu et al., 2022; Pandey et al., 2021). In addition to calculating the value of default probability, some machine learning approaches can report the determinants behind the default probability while making full use of the input information. This property is essential in credit risk predictive analysis. These potential determinants can be used as indicators and guidance for investors to make rational decisions, for regulators to screen, and even for the loan applicants to improve their credit as well as the chance to get the loan successfully. These motivate us to equip predictive analysis with state-of-the-art machine learning techniques in examining credit risks and other financial topics.

According to the literature, the debt-to-income ratio and inquiries in the last six months of loan applicants are positively associated with the default likelihood, while annual income is negatively correlated. Nevertheless, investors tend to emphasise different attributes when granting loans in the real world. They prefer to value loan descriptions, employment length of applicants, and even gender (Ip and Lam, 2020; Polena and Regner, 2018; Zhang et al., 2017). Additionally, economic theory suggests that credit markets are closely in relation to macroeconomic activity. Unlike traditional banks that respond to the macroeconomic changes mainly via bank lending and balance sheet channels, P2P platforms do not take collaterals and have no claims against the central bank. This leads to the indirect influence of the macroeconomic effects on P2P lending. Instead, macroeconomic factors may interact with the P2P lending market in the interest rate channel. For example, in a period of monetary policy tightening, overall leverage is reduced by disincentivising firms' interest in conventional bank credits. Thus, firms are encouraged to turn to the P2P market for alternative financing due to borrowing availability. Moreover, a climb in the monetary policy rate can transmit a decline in P2P loan lending rates. Because the borrowers' density is a driving force of P2P lending platforms' developments and the operators are likely to decrease the P2P interest rate to incentivise more loan applications (Chu and Deng, 2018; Wong and Eng, 2020). Recently, the risk-taking channel of macroeconomic factors has grabbed much attention, as the relationship between monetary policy and macroprudential management is becoming challenging for policymakers. It seems to be a trend that financial institutions take more risks during monetary policy easing by granting more risky loans (Chu and Deng, 2018; Huang, Li and Wang, 2019). Internet-based P2P lending allows investors to screen borrowers' information at reduced cost and descriptive risk, which may result in a prompt response toward the monetary policy change. However, little attention has been paid to the correlation between macroeconomic conditions and FinTech nonbank financial institutions or individual investors' risk-taking behaviour in the literature. Our study aims to fill the gap.

Inspired by the ongoing investigation on the link between equity and the credit market, we further move our attention to the cost of equity. Cost of equity is an important concept in finance research that is widely used to estimate equity risk premiums, evaluate firm valuation and capital budgeting, and manage investments. Generally, the cost of equity depends on firms' financial fundamentals, industry dynamics, and national overall economic conditions (Hasan et al., 2015). However, the last decade has witnessed a growing number of firms worldwide start to disclose non-financial. This information includes environmental preservations, employee welfare improvement, community contributions, and other corporate social responsibilities (CSR). Though the conventional view believes that CSR is a cost of resources and is better to be minimised (Frideman, 1970), the debate on the impact of CSR performance on the cost of equity has grown. Some researchers argue that better CSR performance also brings greater ethical concern, requiring higher quality financial reports (El Ghoul et al., 2011; Kim et al., 2012). Others claim CSR alleviates information asymmetry and market risks, leading to a decrease in the cost of equity (Harjoto and Jo, 2015). A survey (Lacy and Hayward, 2011) points out that the majority of global chief executive officers view sustainability as a key to future success. Besides the possibility of the new regulations and taxes that push firms to be more socially responsible, more and more investors attempt to affect the firms' operation strategies by considering their CSR performance in their investment decision. The related literature mainly focuses on the U.S. market or other developed countries, which motivates us to conduct a cross-country analysis by combining country-level and firm-level data. Armed with state-of-the-art machine learning techniques, we aim to provide a contextualised picture of the interaction between corporate social responsibility and the cost of equity.

1.2. Structure

To better organise this thesis, we divide the empirical predictive analysis into three chapters.

In chapter 2, we examine the default risk of P2P lending loans in terms of loan-specific factors and applicant-specific attributes. This internet-based marketplace dramatically reduces the cost of matching the investor and the borrower. Meanwhile, it suffers from high information asymmetry where individual investors without enough expertise fully undertake the potential credit risk. Two categories of techniques have been widely employed in the literature to estimate the default risk, traditional statistical approaches and intelligence machine learning approaches (Byanjankar et al., 2015; Emekter et al., 2014). However, the assumption of linear relationships behind the traditional statistic models results in easy implementation as well as insensitivity to complicated correlations. Consequently, we use one traditional statistical model (logistic regression) as the benchmark and four machine learning models (adaptive LASSO, LightGBM, Convolutional Neural Tensor Network and Wide and Deep Learning) to evaluate default in P2P. Our analysis is based on a cross-section of 279,512 loans for the U.S. in 2015, sourced from the Lending Club, a world-leading P2P lending platform. We contribute to the literature with the improvement of P2P lending default prediction. Firstly, we shed new light on the relatively underexplored comparison of advanced machine learning techniques in default risk and show that they play crucial roles in credit risk analysis. Second, we would like to offer insights into big imbalanced data classification.

In chapter 3, the determinants of P2P funding decisions and monetary policy effects are investigated. The funding decision's determination has gained much research attention, but most research focuses on bank loans. The literature in this field for P2P lending is still thin. A gap is left for macroeconomic effects on P2P funding decisions via the risk-taking channel. Understanding individual risk-taking behaviour in P2P lending may carry the study forward, leading us to understand the driving force behind the monetary policy risk-taking channel. Chapter 3 examines over 12 million loan requests from the Lending Club from April 2007 to December 2016. Compared to similar research in the literature (for example, Gavurova et al. use 46,916 records from 2009 to 2015, and Zhang et al. use 193,614 loan applications from January to June 2014), a wide-range sample is under study in this paper. We utilise text mining techniques to analyse the role of loan purpose in P2P funding decisions, a feature overlooked by the related literature. This feature reveals the applicants' reasoning in their fund request applications. It might disclose crucial information to credit managers and is likely to be an essential determinant of loan funding success. In order to better measure the macroeconomic effect, we introduce Taylor Residuals as an explanatory variable to access the overall stance of monetary policy. We suggest that individuals primarily formulate the risk-taking incentives based on the observed level of interest rates given the macroeconomic condition. We further document the variations of the macroeconomic influences on sub-groups concerning loan purposes.

Next, in chapter 4, we explore the link between firms' social responsibility and the cost of equity. We use an up-to-date dataset consisting of 3,055 unique firms from 51 countries over the past 18 years to document the negative and nonlinear effects of ESG behaviour on the firms' cost of equity. This study contributes to the literature in several ways. First, we use an up-to-date double machine learning (DML) approach rather than the widely employed regression analysis. Unlike regression models suffering from regularisation bias and simple linear hypotheses, the DML approach detects subtle effects while tolerating the interaction between variables. In addition to the principal ESG dimension, we include a rich selection of ESG variables. DML technique allows us to consider the three pillar scores of ESG simultaneously. Second, previous studies rely on single-country datasets that cannot offer crisp comparisons across counties. In contrast, we use a multi-country dataset that allows us to examine the heterogeneous response of ESG effects on the cost of equity. We document the curvilinearity of the ESG effects on the equity cost and further note the gaps between advanced and emerging regions. We also examine the variations of the ESG effects on the cost of equity when external shocks appear, such as the global financial crisis and the pandemic COVID-19. These reveal the heterogeneity of the ESG causal effects under different settings. To the best of our knowledge, this channel is yet to be documented. The findings in this chapter also have practical implications. The results may encourage firms to take socially responsible strategies, as they enjoy the lower equity cost. However, the power of ESG effects on the equity financing cost can get weakened in different settings. Understanding the link between them helps the firm operators to make future operational and reputational strategies.

The last chapter concludes all empirical practice by briefing the key findings. We also suggest some potential topics for future research.

Chapter 2

Chapter 2 Machine learning models and default prediction in online Peer-to-Peer (P2P) lending

2.1. Introduction

The Internet and information technology have been experiencing a fast-paced development in the last two decades. It increasingly introduces more disintermediated and democratised industries by connecting individual market participants in unprecedented ways. Such expansion has flourished peer-to-peer (P2P) economy and successful P2P platforms. For instance, Uber provides the largest peer-to-peer driving service, Airbnb offers a peer-to-peer room renting service, and the Lending Club provides an online crowdfinancing service. Online Peer-to-peer lending (P2P lending) is an innovative form of financing which directly matches lenders and borrowers without the involvement of a conventional intermediary. The borrower posts lending requests on the P2P platform whilst the investor browses listed lending requests and makes funding decisions with a click. This internet-based financial tool has gained considerable attention from investors and market participants primarily due to reduced financing costs and higher returns than traditional fixed-rate financing (Ma *et al.*, 2018). Specifically, Wei and Lin (2017) argue that the P2P lending process does not place any requirements on collateral or deposits, which leads to lower costs relative to other forms of financing. This is an appealing characteristic of the online lending market, but it comes at a cost as P2P lending is associated with the highest degree of information asymmetry. The investor only has access to the information disclosed by the borrower, which can be biased and even fabric. The upshot is that limited access to information increases the risk of default (Morse, 2015). Additionally, unlike conventional banking, online platforms cannot properly evaluate credit risk by relying on accounting standards. Individual investors without expertise are less likely to identify risky loan requests and face high default risk. For example, CrowdProperty, a P2P lending platform in the UK, reported that its default rate was 22.6 percent in 2019[1]. In the era of financial technology, where the financial services industry is changing at a rapid pace, this issue becomes of grave importance, especially for consumers and "small investors" who become adversely exposed to new financial products and risks.

[1] https://www.p2pfinancenews.co.uk/2021/08/12/p2p-default-rates-near-zero-after-pandemic-year/

The potential loss faced by P2P lending investors motivated this study. Much of the prior research has focused attention on discriminant analysis and logistic regressions. For example, equipped with discriminate analysis, Jiang *et al.* (2017) investigate the role of soft information in loan default predictions. Emekter *et al.* (2014) conduct the default risk evaluation by employing binary logistic regressions with loan applicants' credit-related features as independent variables. Serrano-Cinca *et al.* (2015) utilise binary logistic regressions to examine the determinants of Default in P2P Lending. In the same vein, Li *et al.* (2018) considered the role of economic conditions by incorporating macroeconomic factors into multivariate logistic regressions to increase the model's prediction of default and pre-payment. However, new research shows that machine learning and data-driven approaches provide good prediction accuracy in big data analysis. Byanjankar *et al.* (2015) build a credit scoring model based on a neural network and suggest it performs well in terms of accuracy. Similarly, a study undertaken by Ma *et al.* (2018) analyses a massive P2P loan dataset with the help of boosting trees. Their credit scoring model can improve the average performance rate of the historical transaction data by 1.28 percentage points. Thus, this study aims to evaluate the risk of default in P2P loans by comparing traditional statistical techniques and machine learning models. In doing so, we mark a break with prior research by offering methodological extensions and new evidence, using state-of-the-art techniques and a rich dataset.

Our analysis is based on a cross-section of 279,512 loans for the US, sourced from the Lending Club, which is a world-leading P2P lending platform. Previewing the main findings, first, we show that when applying machine learning techniques we are able to significantly improve the predictive power of our models compared to the logit model, which is the gold standard in the literature. Importantly, we note that the LightGBM model outperforms all other machine learning techniques and displays improved forecasting power. Our results are robust to tests on unemployment length, income verification, and home ownership, and a battery of stratified 5-fold cross-validations. Benefit from the feature importance generated by LightGBM, we identify the loan applicant's debt to income ratio as the most significant impact on the loan repayment, whilst monthly instalment is the second most important determinant of a successful loan. This finding is consistent with Emekter *et al.* (2014), who claim the importance of the debt-to-income ratio in measuring loan performance. By contrast, most variables associated with loan applicants' delinquency records are of low importance. Delinquency history does not dramatically affect current credit risk. Finally, we note that the predicting power of our machine learning models can supplement the traditional credit risk mitigation approaches of P2P loan providers.

Our approach is mostly related to the literature that examines the determinants of loan default, but we add to it in three important ways. First, we contribute methodologically by employing some of the most promising machine learning models (adaptive LASSO, LightGBM), including the latest deep learning models (Convolutional Neural Tensor Network and Wide and Deep Learning) to the task of forecasting the default of P2P loans in the US. Unlike most studies (Ince and Aktan, 2009; Blanco et al., 2013; Bekhet et al., 2014) that utilise a single machine learning technique against the statistical approach in default risk evaluation, we conduct a horsing racing of several representative machine learning techniques against the statistical approach. The techniques we employ are well suited to model large and complex datasets such as ours and have provided promising empirical evidence in other fields. Second, we allow for three dimensions in the data that are critically important in determining defaults. Specifically, we split our sample according to the length of unemployment, income verification, and home ownership. All aspects offer significant heterogeneity at the borrower level, and we are able to tease out differences between the groups of borrowers. Third, we demonstrate the value of machine learning and deep learning in credit risk management of P2P loan providers.

The remainder of the chapter is structured as follows. Section 2.2 provides the literature review, and Section 2.3 presents the data used in the empirical analysis and provides summary statistics. Section 2.4 describes our methodology. Section 2.5 reports the main empirical results. Section 2.6 concludes.

2.2. Literature Review

2.2.1. P2P Lending and Risks

A growing number of business models that employ digital services without intermediations has been seen in the past decade. Online Peer-to-Peer lending is one of the most popular innovations attracting wide attention from market participants, financial supervision agencies, and researchers. Peer-to-peer (P2P) lending provides services based on the Internet. In this marketplace, borrowers submit loan applications as listings where they are required to offer information on themselves and the loan details, including their age, occupation, loan amount, and loan purpose. It can be regarded as a kind of crowdfunding that enables the lender to make a partial or total investment on loan according to the loan information provided by the borrower (Cai et al., 2016).

P2P lending offers benefits to market participants. Trades can be settled effectively via P2P. The Internet-based process requires no collaterals or capture deposits, lowering financing costs (Wei and Lin, 2017). The financial exclusion that misses some conventional credit assessment parameters and is refused by traditional institutions may get credit from P2P lending (Komarova Loureiro and Gonzalez, 2015). Meanwhile, lenders can construct their investment portfolios on P2P loans to minimise risks. They are able to optimise the P2P portfolio to hedge economic situations, employment conditions, or other exposures. P2P is believed to help with the equalisation of economic opportunities. This is termed the "flat-world" hypothesis by Singh et al. (2018). They investigate 660 thousand loans in 220 regions between 2005 to 2013. Not in accord with the hypothesis, P2P lending seems to move away from flatness. A spatial preference in lending for geographically proximal borrowers is presented in the study via the regression analysis. After analysing the opportunities and risks in terms of the *regulatory framework*, Lenz (2016) draws attention to whether the credit assessment should be fully automated algorithm-based or more human-based. They point out that though P2P lending shows advantages, it suffers from several drawbacks. It is short of risk-preventing mechanisms as banks do. P2P platforms spread risks to lenders lacking expertise or experience, while P2P lenders tend only to access the information that loan applicants want to provide.

This information asymmetry received the attention of researchers. Zhang et al. (2017) examine the Determinants of successful loans in online P2P lending. They are inspired by the underdeveloped Chinese marketplace, where the high investment returns of risky borrowers attract a large number of investors and lead to various practical problems. Approximately 200 thousand loan requests are collected in 2014 from the largest Chinese P2P platform, but only a quarter of these requests are successfully funded. Logistic regression reports the results of the analysis. Annual interest rate, repayment period, description, credit grade, successful loan number, failed loan number, gender, and borrowed credit score are significant contributors to successful loans. Equipped with the data from the same platform and a similar methodology, Cai et al. (2016) suggest it matters in the funding decision if it is a first-time request. Lenders tend to adapt their investment decisions when more verifiable information accumulates. Besides the credit-related attributes, Gonzalez and Loureiro (2014) believe that personal characteristics such as perceived attractiveness, age and gender affect online peer-to-peer lending decisions. An online experiment based on two heterogeneous consumer samples is performed to collect data and test the relationships of interest. They conclude that loan success is sensitive to age which signals competence and the attractiveness of lenders and borrowers.

Emekter et al. (2015) explore P2P loans and assess the credit risk using the empirical data from the Lending Club, one of the largest P2P platforms in the world. According to the results from logistic regression (LR) and the Cox Proportional Hazard test, they reveal the contributors to credit risk. Credit grade, debt-to-income ratio, FICO score, and revolving line utilisation play vital roles in loan defaults. Additionally, a concern is presented in the study. The higher interest rate from the risker borrowers may not compensate for the default risk. Guo et al. (2016) suggest that traditional rating-based assessment models may not meet the needs of individual investors in P2P lending and propose a data-driven investment decision-making framework. This framework combines an instance-based credit risk assessment model and a portfolio selection technique. Data from two P2P lending firms prove its efficiency in improving investment performance. Recent literature discloses an increasing scope for the employment of big data and machine learning in P2P credit evaluation. Malekipirbazari and Aksakalli (2015) employ random forests (RF) to predict loan status instead of FICO credit scores. Using 350 thousand samples from the Lending Club for the period between 2012 and 2014, the results indicate that the RF-based model is superior in detecting good borrowers to the credit grades computed by the Lending Club. Similarly, Ma et al. (2018) examine the potential roles of two RF-based methods, LightGBM and XGbbost, in P2P loan assessment experimented on the Lending Club. They select 0.57 million observations with 24 characteristic features after data cleaning. Both of the methods demonstrate superiority, but LightGBM performs slightly better. It seems to raise the average performance rate of the Lending Club historical transaction data by 1.28 percentage points, which may reduce loan defaults by approximately $117 million. Li et al. (2017) point out a sample bias in the standard credit scoring where the scoring is built on the accepted applicants but used on new applicants. To help with the bias, they attach a technique that solves reject inference to a semi-supervised support vector machine (SVM). The integrated model beats the benchmark LR in terms of prediction accuracy. Alternatively, Byanjankar et al. (2015) introduce a neural-network-based credit assessment model to screen applicants effectively. The proposed neural network (NN) model also outperforms LR with the sample from Bondora, a leading European P2P lending marketplace. Bastani et al. (2019) generate a two-stage framework to analyse the P2P default loans, as most credit scoring methods are not free from the imbalance problem where most of the past loans are non-default. Stage 1 identifies non-default loans, while the imbalanced nature of loan status is considered in the probability of default prediction. The loans identified as non-default are then moved to stage 2 to predict profitability, measured by the internal rate of return. Their model relies on an up-to-date technique termed Wide and Deep Learning (WDL), which is widely used in recommendation systems. The numerical results claim a robust performance of WDL. It benefits the credit evaluation.

2.2.2. Credit Risk and Machine Learning Techniques

Credit risk is a primary concern of the financial and banking industry regarding loss, defraud, and even financial crisis. They seriously damage the global economic system. The US subprime mortgage crisis and the European sovereign debt crisis are the most notable cases which can be referred to. Accordingly, reliable credit scoring could effectively decrease the chance of credit fraud and financial crises.

Credit scoring is essentially a classifier for identifying risky loan requests from all requests with the help of some specific characteristics. The original idea of discriminating between groups in a population is generated by Fisher (1936). Their study on the iris proposes a statistical way to differentiate two varieties by their physical measurements. This technique is later introduced by Sowers et al. (1942) to identify bad loans. It pushes the birth of credit scorecards. Credit analysts write down the discriminating characteristics to help non-experts make credit decisions. Thomas (2000) summarises these characteristics into 5Cs: The character of the loan applicant, the capital, the collateral, the capacity of repayment, and the condition of the market. The answers to the 5Cs are used to input statistical or operational research methods for credit scoring. After going through the literature, Thomas (2000) notes the statistical tools are mainly discriminant analysis, logistic regression, and classification trees, whilst the operational research techniques include variants of linear programming. Rivai (2006) extends the 5Cs in their book by adding one more item: constraint. It includes all limitations and barriers that may affect repayments, such as regulation and resource scarcity. Rivai (2006) further explains that the credit assessment should be done by the banking account officer and the assessment committee, and all 6Cs factors should be collected from the loan requests. According to a Federal Reported on American credit development (2003), credit information is revealed to be incomplete and contains duplications and ambiguities at times. Information users are supposed to make assumptions about omissions and limitations of the information when applying certain reported items to developing a credit profile for a consumer. In the previous literature, studies widely use traditional statistical approaches to estimate the risk and determine a client's good or bad status. For instance, Li et al. (2018) employ multivariate logistic regression to analyse the repayment and default risks of online lending. They collect unsecured consumer lending data and directly estimate the probabilities of payment and default. But the average accuracy of the proposed model is only above 75%. However, along with the pace of technology development, machine learning techniques have been introduced into credit risk analysis in the last decades.

A study by Kruppa et al. (2013) suggests the superiority of the machine learning techniques in credit scoring. Using LR as the benchmark, they employ RF built on the decision tree (DT), k-nearest neighbours (KNN), and bagged k-nearest neighbours (BNN) to investigate over 64 thousand customers' payments histories of short-termed instalment credits. The attributes of samples are selected according to expert knowledge, comprising demographic information and the instalment details. Among the four models, RF provides the best accuracy. It outperforms a standard LR and even a well-turned LR. The advantages of DT-based methods are also represented by Khandani et al. (2010). They test the DT performance against LR in identifying risky clients with a high-dimensional dataset sourced from a bank's customer base from January 2005 to April 2009. The results suggest that DT credit scoring can potentially reduce total losses by 6% to 23%.

Many other machine learning algorithms are provably consistent for probability estimation, such as NN and SVM. They can be promising alternatives to the conventional-statistical credit scoring system. Zhao et al. (2015) examine a feed-forward NN with three layers on a widely used German credit dataset. It is an imbalanced dataset with 1000 instances, of which 700 approved cases and 300 rejected ones. They utilise a method named Average Random Choosing to alleviate the imbalance and enhance the prediction accuracy to 87%, which is 5% higher than the previous literature. Based on the same dataset, Khashman (2010) compares the various structures to find the most efficient NN in credit risk analysis. Three-layer NNs are tested under three different training-to-validation data ratio settings. The proposed models show an average accuracy rate of 80%. Besides the fundamental NNs, researchers also explore some sophisticated architectures. Khashman (2011) first introduces emotional NN, an approach for pattern recognition, into credit risk evaluation. 12 NN models in total are applied to 690

Australian credit cases with 14 numerical attributes. Both emotional and conventional NNs provide effective credit risk evaluations. However, the emotional NNs consume less computation time with better accuracy, making them desirable for implementing fast processing of credit applications. Correspondingly, Kvamme et al. (2018) utilise convolutional neural networks (CNN), another pattern recognition technique, to investigate mortgage data. This dataset is from banking services, and it consists of 20,989 records. In addition to the single CNNs, they also combine CNN with RF to enhance the performance. However, the added benefit is smaller than the backwards of the added complexity of the model. The first study that applies SVM to credit scoring is undertaken by Li et al. (2006). They collect 600 loan applications during a period of 2001 to 2002 from a local bank in Taiwan, and the attributes variables are selected based on the literature review. Employing linear discriminant analysis (LDA) and LR as benchmarks, they validate the effectiveness of SVM against NN via the paired T-test based on a five-fold-cross validation. Their study presents that SVM overcomes NN-based models, dramatically strengthening the credit estimation. But they also point out that the black-box nature of SVM lacks financial explanatory ability. Hens and Tiwari (2012) believe on top of the high predicting accuracy, SVM advantages in less computational cost. They propose an SVM model incorporated with the concept of dimension reduction utilising F score, and they generate it on an Australian credit dataset and a German credit dataset. The SVM model is compared with several other approaches, including NN, generic programming (GA), DT. Thanks for the SVM model is built on a quadratic form. Its running time is faster than the rest models while achieving a robust credit status estimation. Harris (2013) investigate the effect of the default definition in credit risk assessment with the help of SVM. SVM models are built with Board (less than 90 days delay) and Narrow (greater than 90 days delay) default definitions. Over 250,000 credit card records dating from 1997 to 2012 are employed. The results show that models based on a Broad definition of default can outperform models developed with a Narrow default definition. To further enhance the befits of SVM, Yu et al. (2010) propose a four-stage SVM structure. The first stage is the in-sample and the out-of-sample data splitting. The second stage is SVM learning paradigms with many dissimilarities utilised for credit risk evaluation, and the next stage is multiple individual SVM agents are generated. In the last stage, all the individual results are constructed together for the outcome. This new approach, along with other machine learning techniques including single SVM, LR, DA, and NN, is tested on 800 empirical samples. All results reported in the experiments clearly show that the proposed SVM ensemble learning approach can consistently outperform the other comparable single models.

Having said that, researchers have extended their interests to ensemble and hybrid approaches as the statistical and machine learning techniques for credit risk have been extensively studied. Models combing various techniques are proposed. Tsai and Chen (2010) explore the architecture combining classifier and cluster. Their fundamental methods include DT, NN, Naive Bayes (NB), K-means, and Expectation Maximisation. With 12,929 observations from a business bank in Taiwan, they illustrate that 'classifier + classifier' serves the best performance. A study undertaken by Finlay (2011) compares different methods in credit risk assessment. LR and LDA represent the traditional approaches, while DT, NN, and KNN present the machine learning approach. They build multi-stage classifiers with the five approaches and evaluate them based on the data from several lending institutions in the UK from April to June 2002. The results find that bagging with decision trees performs the best. It beats the traditional methods in terms of detecting the potential bad loans and profit loans. In the same vein, Xiao et al. (2016) ensemble DT, LR, and SVM and test the models with 6000 mortgage loans and more than 100 attributes; Abellán and Castellano (2017) employ LR, SVM, and DT to compose the models and test them on a very large database; Xia et al. (2018) use SVM, RF, XGBoost, and Global Product Classification (GPC) to build hybrid models and use empirical data from various source to evaluate the models. Their results also suggest that DT-based techniques offer outstanding performance in credit risk assessment.

2.3. Data and Summary Statistics

2.3.1. Data Description

We apply our models to the task of forecasting the default of 279,512 P2P loans from the Lending Club, the largest online P2P lending platform. It has settled down loans of more than 50 billion dollars and served more than 3 million customers, accounting for the half industry until 2019[2]. Unlike other P2P lending platforms such as Prosper, which only offer data access to verified clients, all the Lending Club loan records except personal data are open to the public and are updated quarterly. It makes the Lending Club the most widely used in the literature for P2P lending analysis (see among others, Jiang et al., 2017; Polena & Regner, 2018). Our dataset consists of loans issued in 2015 with a maturity of 36 months. When we collected the data in 2019, we wanted to ensure all the loans had reached maturity and were latest enough to embed

[2] See https://www.lendingclub.com for more information on the Lending Club and its database

new information. Thus, 2015 is the year. Additionally, microeconomic and macroeconomic conditions change over time, and it makes far past loans carry less information value and result in biased forecasting. 2015 is an ideal year also because there was a massive amount of new loans without any significant changes in politics. A large fraction of the loans (84.90%) is considered "non-default", and the remaining (15.10%) are labelled as "default" loans.

Our sample presents two characteristics that make it especially appealing for our analysis. First, it is of big scale and high dimension as we cover approximately four times more loans relative to previous studies. For example, Emekter et al. (2014) employ 61,451 loan applications with 14 explanatory variables, and Xia et al. (2017) employ 49,795 loans with 17 explanatory variables. Hence, our dataset is more than four times larger, with approximately ten more explicators compared to prior research. Imbalanced credit data of small size tends to miss information and may cause specification errors (Florez-Lopez & Ramon-Jeronimo, 2014). This allows us to capture a more representative sample of loans and the underlying contributors to the credit risk, offering more accurate default predictions. Second, compared with previous studies, which mainly restrict their attention to a "narrow default definition", we offer a broader definition. For example, Emekter *et al.* (2014) only include loans that are already charged off and fully paid to estimate the credit risk. By contrast, we also incorporate default loans and loans delayed repayment over 30 days. This is because 75% of delayed-over-30-day loans later become charged off and default, according to the Lending Club statistic (Polena and Regner, 2018).

The choice of explanatory variables is motivated by previous empirical work on the determinants of default probability in P2P lending (see, among others, Emekter *et al.*, 2014 and Li *et al.*, 2018) and can be divided into the following broad categories: (1) loan characteristics (e.g. loan amount, loan title), (2) borrower characteristics (e.g. annual income, home ownership), (3) borrower indebtedness (e.g. debt to income ratio), and (4) credit history (e.g. delinquencies number, utilization rate). Data cleaning is conducted considering the sheer volume of data collected from the Lending Club. There are 68 features in the original dataset, some containing duplicate information or missing values that may degrade the model performance later. Hence, repeated contents are dropped, and empty variables or variables with more than 50% missing values are removed from the dataset. Furthermore, samples with any missing variables are excluded since we have a massive number of samples, most of which are full size. In sum, we employ twenty-four explanatory variables. Three variables (home ownership, loan title, verification status) are categorical, while the rest are numerical variables.

Table 2-1: Variables definitions

Variable	Definition
Annual income	The self-reported annual income provided by the borrower during registration
Average balance	Average current balance of all accounts
Bankruptcies	Number of public record bankruptcies
Charge-offs number	Number of charge-offs within 12 months
Collection amounts	Collection amounts ever owed
Collections number	Number of non-medical collections in 12 months
Credit line length	Length of the earliest reported credit line (calculated by days)
Credit lines	Number of open credit lines
Current delinquency amount	Currently delinquency amount
Current delinquency number	Number of currently delinquent accounts
Debt to income ratio	Borrower monthly debt payments divided by monthly income
Delinquencies number	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
Employment length	Employment length in years. The values range between 0 and 10 where 0 means less than one year and 10 means ten or more years
Home ownership	The home ownership status provided by the borrower during registration or obtained from the credit report. Values are: RENT, OWN, MORTGAGE, OTHER
Inquiries	Number of credit inquiries in the last 6 months
Loan amount	The listed amount of the loan applied for by the borrower

Loan title	Purpose of loan
Monthly instalment	The monthly payment owed by the borrower if the loan originates
Open accounts	Accounts opened in the past 12 months
Total balance	Total current balance of all accounts
Total credit balance 1	Total credit revolving balance
Total credit balance 2	Total credit balance excluding mortgages
Utilization rate	Revolving line utilization rate
Verification status	Indicates if income was verified by Lending Club, not verified, or if the income source was verified

Note: The table defines our explanatory variables. These variables act as inputs to the five models under study.

2.3.2. Summary Statistics

We report summary statistics of the numerical variables used in the empirical models in Table 2-2. In addition, we present p-values for the tests of equality of means across the default and non-default loans in column 4 of Table 2-2. We observe, as expected, differences in several predictors between default and non-default loans. For example, non-default loan recipients are equipped with higher and stabler annual income on average, while risky loan recipients tend to have lower annual income. This finding is consistent with the research conducted by Emekter et al. (2014), where "bad loans" present lower monthly incomes in general when they summarise the dataset from the Lending Club. Similarly, Serrano-Cinca et al. (2015) notice approximately a ten-thousand gap between the average annual income of the two groups. They also note a small difference in loan recipients' credit history, and comparably greater heterogeneity in current delinquency number among default loan recipients is noted by the standard derivation in Table 2-2. For other variables, we note no significant differences. For instance, the default and non-default groups have similar situations in terms of charge-offs number. The descriptive statistics provide a first glimpse regarding the importance of some variables in predicting P2P loan default. In order to deal with the different scales and to assist our algorithms in their forecasting task, we normalise all variables before incorporating them into the models.

Table 2-2: Summary statistics

		All instances (1)	Default (2)	Non-default (3)	P-value (4)
Annual income	mean	74.49	66.48	75.92	0
	std	74.39	82.19	72.82	
Average balance	mean	12.33	9.43	12.84	0
	std	15.74	12.35	16.22	
Bankruptcies	mean	0.14	0.17	0.14	0.98
	std	0.4	0.44	0.39	
Charge-offs number	mean	0.01	0.01	0.01	1
	std	0.12	0.12	0.12	
Collection amounts	mean	0.28	0.28	0.27	0
	std	2.27	1.98	2.32	
Collections number	mean	0.02	0.03	0.02	0.99
	std	0.16	0.18	0.16	
Credit line length	mean	37125.27	37693.55	37024.22	0
	std	5051.93	5000.04	5054.43	
Credit lines	mean	11.61	11.81	11.57	0.81
	std	5.54	5.63	5.53	
Current delinquency number	mean	0.01	0.01	0.01	1
	std	0.09	0.1	0.09	
Currently delinquency amount	mean	0.01	0.02	0.01	0
	std	0.79	0.94	0.76	
Debt to income ratio	mean	18.46	20.23	18.14	0.04
	std	8.6	8.8	8.52	
Delinquencies number	mean	0.35	0.39	0.35	0.97
	std	0.94	1.01	0.93	
Inquiries	mean	0.58	0.76	0.55	0.83
	std	0.87	0.99	0.85	
Loan amount	mean	12.8	12.46	12.86	0
	std	8.06	7.94	8.08	
Monthly instalment	mean	0.42	0.42	0.42	0.08
	std	0.27	0.27	0.26	
Open accounts	mean	2.17	2.63	2.08	0.58
	std	1.86	2.06	1.81	
Total balance	mean	128.97	101.42	133.87	0

	std	153.83	124.47	157.98	
Total credit balance 1	mean	16.21	14.04	16.59	0
	std	23.73	18.92	24.46	
Total credit balance 2	mean	48.24	45.55	48.72	0
	std	47.78	42.8	48.6	
Utilization rate	mean	0.53	0.57	0.52	0.96
	std	0.5	0.5	0.5	

Note: The table reports summary statistics of the explanatory variables used in the empirical models. All $ values are in thousands. Column 4 reports the p-value for the test of equality of means between the default and non-default groups.

Table 2-3 illustrates an important dimension in P2P lending, namely the employment spell. Specifically, the table presents the percentage of the default and the non-default loans according to the employment length of the loans' recipients within our dataset. We breakdown the information for unemployed recipients as well as those employed for more or less than ten years. We note considerable differences in the default/non-default ratio based on the employment status of the loan recipients. The reason may be that unemployment is associated with an unstable income, which naturally leads to a higher default loan probability. However, Serrano-Cinca et al. (2015) report no divergence in average employment length between the default and non-default instances. In our case, it is interesting to note that there is no notable difference across employed recipients when we use ten years as a cut-off point.

Table 2-3: Default rates according to employment length

	All instances	Default	Non-default	P-value
Unemployed	6.67%	21.89%	78.11%	0.00
Employed <10 years	61.66%	15.17%	84.83%	0.71
Employed ≥10 years	31.66%	13.53%	86.47%	0.00

Note: The table reports the percentage of default and non-default loans according to the applicants' employment length within our dataset. The p-values refer to the test of equality of frequencies between the default and non-default loans of the related group.

Next, we explore the extent to which verification of the recipients' income is a significant factor in P2P loans default. Verifying the income and the income source of the loan applicants is a way to mitigate credit risk. However, Lending Club does not always conduct such checks given the scale of loan applications. The decision to verify the income information of the loan applicants is based on their risk profile. The verified loans are associated with loan applicants of higher risk profiles, which is typically translated to a higher percentage of default. If an applicant fails to provide the requested information, the loan application is rejected. Table 2-4 presents the percentage of loan recipients that have their income and sources verified by Lending Club.

Table 2-4: Default rates according to loans' income verification

	All instances	Default	Non-default	P-value
Not verified	29.47%	11.41%	88.59%	0.00
Income source verified	42.15%	14.92%	85.08%	0.22
Income size and source verified	23.38%	19.19%	80.01%	0.00

Note: The table reports the percentage of default and non-default loans according to the recipients' income source verification within our dataset. The p-values refer to the test of equality of frequencies between the default and non-default loans of the related group.

We observe that for almost 30% of the loans provided, the decision to award the loan is based on the information provided by the recipients. For these loans, the Lending Club has not verified whether the information by the recipients is accurate or not. However, these loans are not associated with a higher probability of default. On the contrary, for loans whose income and the source of income of the recipients have been verified, the percentage of default loans is higher.

We also explore the role of home ownership in our study. Home ownership is associated with higher income and can act as an asset that the loan applicant can defer in case of financial distress. On the other hand, a mortgage or renting a house implies a monthly income outflow that adds to the P2P loan repayments of the loan recipient. Emekter et al. (2014) find that "good loans" tend to have more control over their home ownership. Table 2-5 presents the related summary statistics of the three groups on our dataset. Renting a house is associated with a higher probability of default, while having a mortgage has the opposite effect. Owning a house seems to have no obvious effect on the default of P2P loans.

Table 2-5: Default rates according to home ownership

	All instances	Default	Non-default	P-value
Mortgage	45.85%	37.15%	47.40%	0.00
Own	11.21%	11.96%	11.08%	0.00

Rent	42.94%	50.89%	41.52%	0.00

Note: The table presents the percentage of loan recipients that have mortgages, own or rent a house. The p-values refer to the test of equality of frequencies between the default and non-default loans of the related group.

In summary, these preliminary statistics suggest that borrowers' default rates may be related to their characteristics and credit history. In the sections that follow, we provide a formal econometric analysis of the links between these variables.

2.4. Empirical Implementation and Methodology

The extant literature on loan default prediction considers various models and techniques to improve prediction accuracy. In this study, we propose four representative methods against the traditional benchmark, aiming to select the most important predictors and at providing accurate P2P default loan forecasts using big data. However, different machine learning techniques require varied implementations of categorical variables. We usually use integer encoding to transfer categorical variables into numerical variables. Take the categorical variable "home ownership" as an example: assigning 1, 2, and 3 to "mortgage", "own, and "rent", respectively. This method works with algorithms like LASSO or LightGBM that cannot understand ordered relationships in a variable. By contrast, NN relevant models give weights to independent variables according to the order. "mortgage" is more preferred than "rent" as the value 3 is greater than 1 by their default. To avoid this discrimination, we introduce a technique termed one-hot encoding to NN-based models. The single variable "home ownership" is extended into three variables "mortgage", "own, and "rent", and values 0 and 1 stand for "no" and "yes" in the three variables.

A common feature of all data-driven models is their sensitivity to under-fitting, over-fitting, and data snooping. Machine learning models with complex topologies and parametrisation are particularly prone to these biases. In order to shield our results and the generalisation of our findings, we employ a stratified 5-fold validation in all estimations. We train all models on 80% of the loans and evaluate their performance in the remaining 20%. We repeat the process five times, and each loan acts only once as out-of-sample. We present the average of these five estimations.

2.4.1. Logistic Regression (LR)

LR is the most popular statistical model in credit scoring (Nikolic *et al.*, 2013). Let us define D ={$x_{i,n}, y_n$} as our dataset, where $x_{i,n}$ is a matrix with $279{,}512 \times 24$ (279,512) samples of our *i*=1…24 predictors. y is the dependent binary variable that takes the value of 1 if the loan defaults, and 0 otherwise. LR takes the form:

$$prob(x) = \frac{e^{b_0 + \Sigma b_i x_{i,n}}}{1+e^{b_0+\Sigma b_i x_{i,n}}} \tag{1}$$

where b_0 is a bias term (intercept) and b_i are the regression coefficients.

2.4.2. Adaptive LASSO (AdLASSO)

AdLASSO is a shrinkage regression method that performs regression and variable selection simultaneously. It applies weighting parameters, and penalises the different coefficients in the l_1 penalty and achieves model selection consistency. Suppose the linear regression:

$$prob(x) = \beta_i x_{i,n} + b \tag{2}$$

where β_i *and* b are coefficients and bias, respectively. The AdLASSO estimation is:

$$\widehat{\beta_{AdLASSO}} = arg\left|\left| prob(x) - \sum_{i=1}^{24} \beta_i x_{i,n}\right|\right|^2 + \lambda \sum_{i=1}^{24} w_i |\beta_i| \tag{3}$$

where λ is the penalty and w_i stands for the adaptive weight. In the practical modelling process, the initialised weights for variables are required in adaptive LASSO to be optimised in the training process. A common option for the initial weights is the corresponding LASSO estimates (Bühlmann & Van De Geer, 2011). For λ, we randomly initialise 100 different numbers as candidates and then generate models with them on in-sample data. The λ, offering the best accuracy according to the Bayesian information criterion (BIC) is chosen in order to build the final AdLASSO and test the out-of-sample prediction.

2.4.3. LightGBM

LightGBM is a gradient boosting decision tree method that is used for large dataset classification. It produces more complex trees than any other boosting algorithm, as the trees grow “vertically” rather than “horizontally”, an element that leads to superior accuracy. LightGBM uses a histogram-based algorithm that leads to fast training speed, low memory usage, and high efficiency. It also supports parallel processing. These elements have led LightGBM to be increasingly popular in large datasets classification tasks and an appropriate candidate for our study.

A single decision tree can be expressed as:

$$f_k = arg \sum \quad L\left(y_n, \hat{y}_n^{(k)}\right) + \ \Upsilon J + \frac{1}{2} \lambda \sum_{j=1}^{J} \quad \omega_j^2 \tag{4}$$

where $L\left(y_n, \hat{y}_n^{(k)}\right)$ is a loss function. Υ and λ denote penalty parameters, J is the maximum number of leaves, and ω denotes the weights of the leaves. The LightGBM forecasts take the form:

$$prob(x) = \sum \quad g_n f_k\left(x_{i,n}\right) + \frac{1}{2} h_n {f_k}^2\left(x_{i,n}\right), f_k \in F \tag{5}$$

where g_n and h_n are first and second-order gradient statistics of the loss function. If N_j is a space that contains leaf j, the benefit-maximised leaf weight scores of nodes in LightGBM is $-\frac{\sum_{n \in N_j} \quad g_n}{\sum_{n \in N_j} \quad h_n + \lambda}$. Then equation (5) is transformed to:

$$prob(x) = \frac{1}{2}\left(\frac{\left(\sum_{n \in N_L} \quad g_n\right)^2}{\sum_{n \in N_L} \quad h_n + \lambda} + \frac{\left(\sum_{n \in N_R} \quad g_n\right)^2}{\sum_{n \in N_R} \quad h_n + \lambda} - \frac{\left(\sum_{n \in N_j} \quad g_n\right)^2}{\sum_{n \in N_j} \quad h_n + \lambda}\right) \tag{6}$$

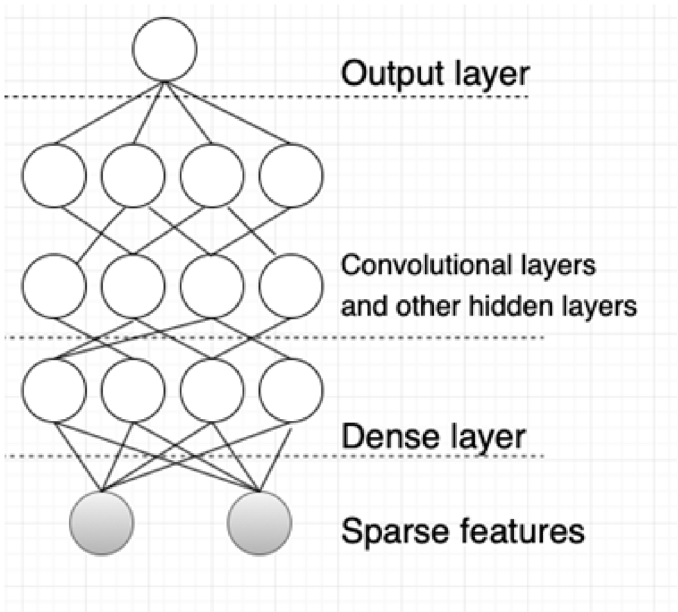
where I_L and I_R represent the left and right branches of the sample, respectively.

In practical modelling, LightGBM mainly relies on three parameters, the learning rate, the number of leaves, and the number of estimators. They significantly impact its configuration. The learning rate is a parameter that controls the speed of iterations. A small learning rate may require numerous updates before reaching the loss minimum point, whereas a greater learning rate means drastic updates, resulting in divergent behaviours. Thus, we perform cross-validation on all in-sample data with five widely used learning rates, 0.0050, 0.01, 0.015, 0.1, 0.15. We note that 0.01 is a better choice. The number of leaves measures included feature variables on a single decision tree, which controls the complexity of the model. Here, we would like to know the contribution of each feature to the final loan status, so it is reasonable to keep them all in the model. In our case, the number of leaves is 24 (see **table 2-1** for feature variables understudy), the same number as feature variables. The number of estimators determines the number of single decision trees in LightGBM. We set it as 500 since there is no significant decrease in loss function when it increases from 500 in our experiments. The optimised number of iterations (steps) can be found appropriately during the model training process and automatically employed in the prediction part.

2.4.4. Convolutional Neural Tensor Networks (CNTN)

Convolutional Neural Networks (CNN) is a widely employed method in large-scale image and pattern recognition. Its success in these fields has led recently to a series of applications in finance (Kvamme *et al.*, 2018). Figure 1 shows the main structure of a CNN.

Figure 2-1: Main structure of a CNN



Note: This figure illustrates the general structure of a Convolutional Neural Network.

CNN consists of several hidden layers, typically convolutional layers, pooling layers and fully connected layers. The convolutional layer consists of features extraction from input samples. It detects the relationship between features of each sample by scanning input data with small squares. The process is realised via a mathematical operation on the sample-data matrix and filter (trainable convolutional kernel) matrix. The convolution of a sample-data matrix is multiplied with a filter matrix to generate an output feature map. The filter is a matrix that contains weights, and its sum with the sample-data matrix measures the correlation between the filter and the relevant part of the input. To exhibit the desired behaviour, CNN may recruit hundreds of dimensions, which can be easily stored in a tensor. Convolutional Neural Tensor Networks (CNTN) extends CNN by transferring hidden layers into tensor layers, resulting in higher efficiency.

Firstly, we apply one-hot encoding to categorical variables in dataset D. Let * denote 1D convolution and d denote the filter. If l is the layer number, the convolutional layer is described:

$$a_{(m)}^{l} = \Omega(\textstyle\sum_{o} \quad a_{(o)}^{l-1} * d_{(0,m)}^{l-1} + c_{(m)}^{l}) \tag{7}$$

where $a_{(m)}^{l}$ represents m-th output map in layer l and$*$ $d_{(i,m)}^{l-1}$ represents the kernel connecting m-th and o-th output map in layer l and l-1. $c_{(m)}^{l}$ is a bias term and Ω is the activation function.

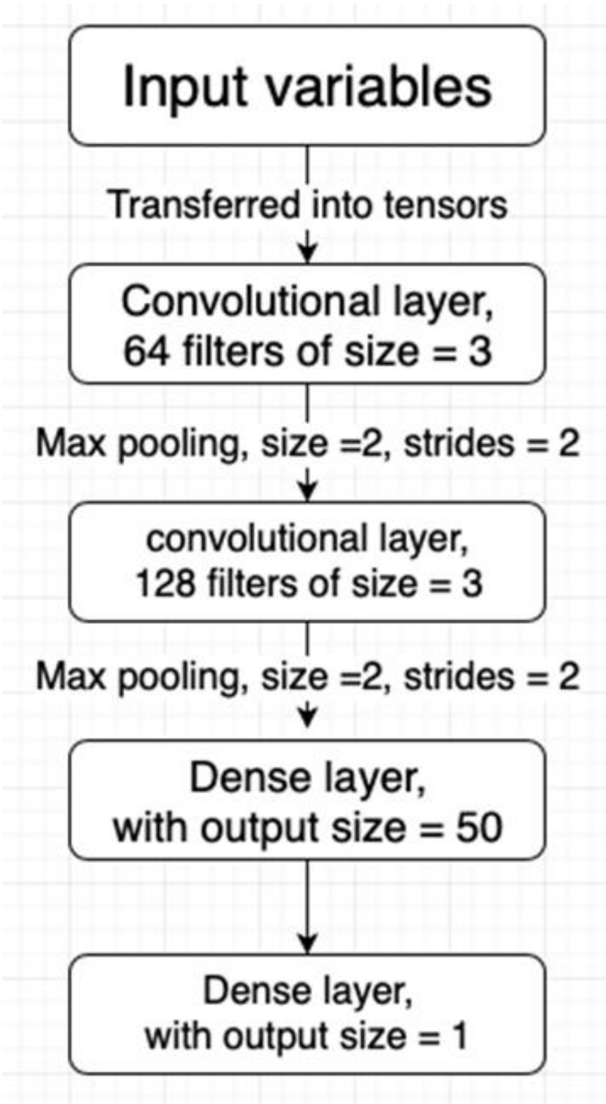
In CNTN the data is transferred into tensor form without employing any new parameters:

$$a_{(m)}^{l} = \Omega\left(\sum_{i} \quad v_{(o)}^{l} * d_{(o,m)}^{l-1} + c_{(m)}^{l}\right) = f(z_{o}^{l}(v^{l})) \tag{8}$$

where v^{l} is the tensor form of a^{l-1}. In a convolutional layer, we basically apply filters over the dataset to extract different features. Then a max-pooling layer reduces the parameters by down-sampling.

The architecture of the proposed CNTN in this chapter is shown in Figure 2-2. Convolutional layers and max-pooling procure are sequentially involved twice, followed by two fully connected dense layers.

Figure 2-2: Architecture of the proposed CNTN



Note: This figure illustrates the architecture of the proposed Convolutional Neural Tensor Network in the study

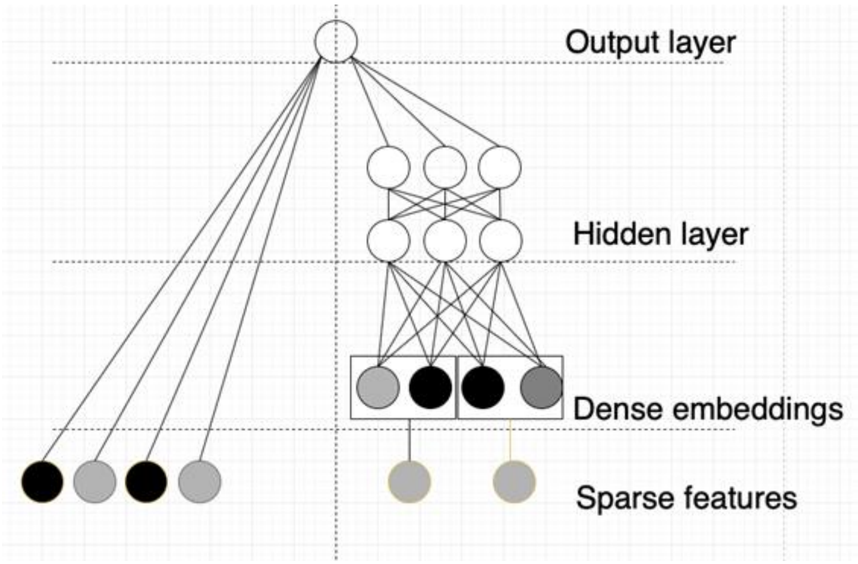
Figure 2-2 also exhibits the number and the size of filters for convolutional layers, the size and strides of max pooling kernels, and the number of output nodes for the dense layers. Suggested by Ju et al. (2019), we use a filter of size 3 for higher accuracy. Since CNN is initially developed for images and pixels, it is common to set multiplies of 32 as the number of filters. 64 is widely employed as the filter number for high-volume datasets (Chen et al., 2015; Krizhevsky, Sutskever and Hinton, 2017; Kvamme et al., 2018; Mane and Kulkarni, 2018). Specifically, we have 41 features after one-hot encoding, which is the size of the dataset. The filter identifies features (different from feature variables), and then the first convolutional layer convolves them with 64 kernels of size 3. In this way, we have the first output with the size of 41. Similarly, 64 filters with size 3 are employed and compute a 41 output in the second convolutional layer. In the next step, each convolutional layer is followed by a pooling layer using overlapping max pooling with the window of size 2. It means two values of each column of the convolved features are pooled into one. Stride is the number of data elements that shifts over the whole input matrix. We set 2 as the value of stride in the chapter. It means we move filters to 2 elements to extract the feature simultaneously. In the last two dense layers, the first one has 50 neurons, whereas the last one contains only 1 neuron because our loan status is a single channel. The process is based on 10 epochs. We replace gradient descent with the Adam optimisation algorithm to learn and correct the model parameters better.

2.4.5. Wide and Deep Learning (WDL)

Credit scoring can be viewed as a credit ranking system. The inputs are features that present loan information. The output is a list of loans ranked by their default probabilities. Thus, a recommendation system is potentially a solution for the credit prediction problem. Wide and Deep Learning (Choi *et al*., 2016) is a recommendation system approach that takes advantage of memorisation and generalisation. Memorisation enables us to learn frequently-occur features and correlations in the historical data (e.g. "Mortgage" for *home ownership* always happens with "Home buying" for *loan title*), and it can be realised by a linear model with cross-product transformations. In comparison, generalisation means exploring new features that rarely or never appear in the past (e.g. "Mortgage" for *Home ownership* and "10" for *Annual income* never occur together in our dataset).

WDL is composed of 2 parts, the wide learning and the deep learning. Its structure is shown in figure 2-3.

Figure 2-3: The structure of WDL



Note: The figure illustrates the general structure of Wide and Deep Learning

In wide learning, a full connected single hidden layer Neural Network (NN) is fed by our dataset D after one-hot encoding. The first step is to compute cross-product transformation, which represents the interaction between features:

$$\varphi_p(x) = \prod_{q=1} \quad x_q^{c_{(q,p)}} \quad c_{(q,p)} \epsilon \{0,1\} \tag{9}$$

where $\varphi_p(x)$ denotes p-th transformation and $c_{(q,p)}$is defined s a boolean variable takes the value of 1 when q-th feature belongs to φ_p or otherwise is 0. Wide learning is then expressed as a general linear model:

$$prob(x) = \varrho_i X + \varrho_0 \tag{10}$$

where $\varrho_o$is a bias and ϱ_i are regression coefficients. Note that X contains both x and $\varphi_p(x)$.

On the other hand, the deep learning part has hidden layers of the form:

$$u^{(l+1)} = \Omega(\theta^l u^l + \mu^l) \tag{11}$$

where $\theta \; and \; \mu$ are weights and bias, respectively. We also employ a world embedding layer that captures attributes of the textual variable *Loan title*.

2.5. Empirical findings

2.5.1. Predictive Ability

In order to evaluate the accuracy of our models in predicting the default of P2P loans, we estimate the area under the curve (AUC), the Youden’s index (Y index), the F-score and the misclassification cost (miscost). For all metrics except miscost, the higher the value the more accurate the model under study is.

AUC is a non-parametric measure generated from the receiver operating characteristic curve, which is frequently applied to evaluate the ability of a model to discriminate between binary events. We compute this statistic by plotting the true positive rate (the ratio of correctly classified default loans to all real default loans) against the false positive rate (the ratio of incorrectly classified default loans to all real default loans) at various threshold settings. If the value of AUC is above 0.8, the predictive ability may be considered accurate. The Youden's index is used to capture the model's diagnostic ability in imbalanced data. We compute this statistic by taking the sum of the true positive rate and true negative rate (the ratio of correctly classified non-default loans to all real non-default loans) minus one. We employ the F-score to measure classification performance in the imbalanced dataset. It is the harmonic mean of positive predictive value (the ratio of correctly classified default loans to all labelled default loans) and the true positive rate.

The costs of misclassifying default and non-default loans can be high. This is because potential profit is degraded if profitable loans are mislabelled as default, but misclassifying risky loans will result in unpredictable losses. Thus, we introduce the misclassification cost to estimate the model's performance using the following equation:

$$miscost = Cost_{12}\pi_2 * (false\ positive\ rate) + Cost_{21}\pi_1 * (false\ negative\ rate) \tag{12}$$

where $Cost_{12}$ is the cost of granting credit to a high-risk applicant while $Cost_{21}$ stands for the cost of rejecting a low-risk applicant. This concept is firstly implemented by West (2000) in their study on Neural network credit scoring models. Following them, the former is assigned the value 5 and the latter is given the value 1 in this study. Further, π_1 and π_2 are the rates of non-default and default loans, respectively. The higher the misclassification cost, the worse the model performance is.

In addition, we employ the Delong statistic to test the equality of AUC between benchmark LR and reference methods (Delong *et al.*, 1988). The null hypothesis of the Delong test for the two models is that the AUCs of the two models are not statistically different. We apply this procedure in pairs to examine the gain of the machine learning models over LR.

We present the performance of our models in our whole dataset based on the employment status and income verification of the loan applications. Additionally, we present the performance of our models based on home ownership as a robustness check.

2.5.2. Main Results

Table 2-6 presents the out-of-sample performance of our models.[3]

Table 2-6: Out-of-sample predictions

	AUC	Y index	F score	Miscost
LR	61.32%	22.64%	31.93%	60.68%
AdLASSO	59.01%*	18.02%	30.53%	69.60%
LightGBM	**82.22%***	**64.43%**	**78.37%**	**30.20%**
CNTN	82.12%*	64.24%	78.08%	30.35%
WDL	79.69%*	59.37%	74.51%	34.49%

Note: The table reports the accuracy ratios for the estimated models in the out-of-sample. LR represents the logit model; AdLASSO refers to the Adaptive Lasso model; CNTN is the Convolutional Neural Tensor Networks; WDL represents the Wide and Deep Learning model. "AUC" refers to the area under receiver operating characteristic curve, "Y index" stands for Youden's index, and "Miscost" is the misclassification cost for all models under study. The * in the AUC denotes that the difference against the benchmark LR model is significant at 99% confidence interval in the Delong test. The best values for each metric are in bold.

According to the AUC test, the LightGBM outperforms all models with 82% correct predictions. This is a marked improvement compared to the benchmark model (LR), which attains 61% correct predictions. In addition, the difference between the benchmark and the top-performing model is statistically significant, as shown by the DeLong test. Our ranking is consistent when we consider the Y-index, the F-scores, and the miscost. For example, the Y index shows that the percentage of correct predictions increases from 22% in the LR model to 64% in the LightGBM model, and the cost of misclassifying labels is two times higher for LR than LightGBM. Zhou et al. (2019) point out that because of the high dimension and class imbalance of P2P lending credit data, most models cannot effectively and accurately predict default probability. But they show that decision-tree-based models solve this challenge and beat LR and NN when analysing the default risk of a well-known Chinese P2P lending company. As a tree-based classifier, LightGBM considers the order of variable values rather

[3] We present the related in-sample performance in the Appendix.

than variable values themselves, allowing extreme outcomes in our imbalanced credit data. Its high discrimination ability presented by high AUC value comes from Gradient-Based One-side Sampling (GOSS) where small-gradient unrepresentative instances are removed. In the same vein, Ma et al. (2018) demonstrate that the average performance rate of the historical transaction data of the Lending Club platform rose by 1.28 percentage points if the credit assessment was armed with LightGBM, which reduced loan defaults by approximately $117 million.

CTCN performs slightly worse than LightGBM, and WDL follows it. Their good discrimination abilities are presented by high AUC scores (82.12% and 79.69%, respectively) and high F scores (78.08% and 74.51%, respectively). However, West (2000) investigates 5 NN-based credit scoring models with two real-world datasets and concludes that they may not be the most accurate model. Additionally, NN-based algorithms are criticised for their black-box process (Balabin and Lomakina, 2011; Verbraken, T., 2014). We are unable to interpret the modelling process or figure out the determinants of default. Besides, expertise and time cost are required when generating CTCN and WDL. The complex algorithms behind them make it necessary to repeat the training process for better parameters (Balabin and Lomakina, 2011). Thus, they are not the best choices for credit classification in this case.

AdLASSO shows poor performance in the study with an AUC value of 0.59, only slightly beating a completely random classifier whose AUC value should be 0.5. This may be caused by our imbalanced dataset, where there are far more non-default loans than default ones. Linear models may fail to tell subtle non-linear relationships in big data (Khandani et al. 2010).

In addition to the high accuracy, LightGBM is admirable for further credit analysis because it imports feature importance in the model (Ma et al., 2018). Histogram subtraction buckets attribute and Leaf-wise learning allows LightGBM to find the essential features fast and generate a model with less time and storage cost (Li.F et al., 2018; Zhang et al., 2019). Only the most profitable features are split whilst other less valuable information is discarded. It potentially guides both loan applicants to improve credit reliability and loan investors to examine the credit risk. Table 2-7 outputs the variable importance generated by LightGBM.

Table 2-7: Variable importance computed by LightGBM

Variable	Importance	Variable	Importance
Debt to income ratio	**10.92%**	Utilization rate	3.17%
Monthly instalment	9.43%	Verification status	3.11%
Average balance	8.95%	Total credit balance 2	2.62%
Open accounts	8.62%	Credit lines	2.60%
Annual income	8.49%	Collection amounts	1.48%
Credit line length	7.47%	Delinquencies number	1.26%
Total credit balance 1	7.19%	Loan title	1.10%
Employment length	5.88%	Bankruptcies	0.81%
Inquiries	5.02%	Current delinquency amount	0.80%
Home ownership	3.69%	Collections number	0.13%
Total balance	3.62%	Charge-offs number	0.09%
Loan amount	3.47%	Current delinquency number	0.07%

Note: The table reports the variable importance generated by LightGBM in descending sort. The most important variable is in bold.

According to LightGBM, the loan applicant's debt to income ratio (DTI) has the most significant impact on the loan repayment. A low DTI ratio indicates sufficient income relative to debt servicing, making a borrower less risky. Monthly instalment is also an important determinant of a successful loan. A high monthly instalment requires high repayability, which may imply an increased risk. This result is consistent with prior studies in the literature. In a study on the Lending Club default risk conducted by Emekter et al. (2014), the debt-to-income ratio is believed to play an important role. Having said that, Serrano-Cinca et al. (2015) use survival analysis to study the default loans from the same P2P lending company and reach similar results. Determinants of credit risk lay in loan recipients' indebtedness and annual income. The right panel of the table reports the variables with lower values in the LightGBM default prediction. Most of these variables are associated with loan recipients' delinquency records of low importance, such as the number of delinquencies in the last 12 months. Different from Serrano-Cinca et al. (2015)'s idea that delinquency history is crucial for credit assessment, it seems not to be a concern for P2P loan investors in our study.

2.5.3. The Role of Employment Length

The length of employment may be considered a key contributor to loan repayment. In research on mortgage repayment, Quercia *et al*. (2012) present that employment affects annual income and further affects loan recipients' repayability. From Table 2-3, unemployed borrowers are more likely to default, while most borrowers who have been employed for 10 years or more successfully repay loans. Thus, the whole sample is split into 3 groups: unemployed borrowers, borrowers have been employed for less than 10 years, and borrowers who have been employed for 10 years or more.

The reason may be that the longer a loan applicant has been employed, the more promising their income tends to be. Stable income is believed to represent less uncertainty in repayability. Conversely, the unemployed group may carry more risks. Unemployment may suggest insufficient income to cover the debt. However, there are some exceptional cases. For example, a student who just graduated from a high-ranked institution may be unemployed temporarily but tends to get a bright job quickly to pay back debt. Self-employed recipients are unemployed, but it is likely to lead to solid repayment ability. Varied situations behind the same label "unemployed" add difficulty to the model training process and degrade prediction accuracy. Nevertheless, a robust model should outperform other models even in the unemployed group. Thus, we re-run our five models using three employment length subgroups and expect LightGBM to outperform the rest models. Table 2-8 presents these results.

Table 2-8: Predictions based on employment length

	Unemployed				
	LR	AdLASSO	LightGBM	CNTN	WDL
AUC	58.52%	57.04%	**77.48%***	76.79%*	58.02%*
F score	39.10%	24.68%	**70.93%**	69.75%	27.64%
Y index	17.04%	14.08%	**54.96%**	53.58%	16.03%
Miscost	65.30%	72.95%	**38.24%**	39.41%	71.29%
	Employed <10 years				
	LR	AdLASSO	LightGBM	CNTN	WDL
AUC	60.12%	59.00%*	**82.63%***	82.62%*	77.59%*
F score	31.19%	30.50%	**78.97%**	78.96%	71.11%
Y index	20.25%	18.00%	**65.25%**	65.23%	55.17%
Miscost	62.73%	69.62%	**29.50%**	29.52%	38.06%
	Employed ≥10 years				

	LR	AdLASSO	LightGBM	CNTN	WDL
AUC	58.36%	59.15%	**83.27%***	82.81%*	60.46%*
F score	27.41%	30.93%	**79.89%**	79.03%	34.61%
Y index	16.72%	18.30%	**66.53%**	65.61%	20.93%
Miscost	65.51%	69.37%	**28.41%**	29.18%	67.13%

Note: The table reports the accuracy ratios for the estimated models based on the length of employment. LR represents the logit model; AdLASSO refers to the Adaptive Lasso model; CNTN is the Convolutional Neural Tensor Networks; WDL represents the Wide and Deep Learning model. “AUC” refers to the area under receiver operating characteristic curve, “Y index” stands for the Youden’s index and “Miscost” is the misclassification cost for all models under study. The * in the AUC denotes that the difference against the benchmark LR model is significant at 99% confidence interval in the Delong test. The best values for each metric are in bold.

Table 2-8 corroborates our assumption that “unemployment” adds uncertainty to prediction since models perform weakly for unemployed instances in general. LightGBM still performs best, ranked the first for all evaluations, showing good discrimination ability. The CNTN continues to follow with the second-best performance. It is interesting to note the difference in the performance of WDL across the different groups, which can be interpreted as an indication on the sensitivity of the model to different datasets.

2.5.4. The Role of Income Verification

The Lending Club, like all other P2P loan providers, try to mitigate credit risk by verifying the income and the income source of a percentage of loan applications that are labelled “risky”. Our summary statistics in Table 2-4 reveal that the verification process of the Lending Club does not mitigate the risk of default. In Table 2-9, we present the performance of our models on the different income verification subgroups.

Table 2-9: Predictions based on information verification subgroups

	Not verified				
	LR	AdLASSO	LightGBM	CNTN	WDL
AUC	58.47%	61.91%*	**79.81%***	79.28%*	64.20%*
F score	24.38%	38.49%	**74.68%**	73.82%	44.16%
Y index	16.94%	23.83%	**59.62%**	58.55%	28.40%
Miscost	66.84%	64.67%	**34.28%**	35.19%	60.78%
	Income source verified				
	LR	AdLASSO	LightGBM	CNTN	WDL
AUC	56.94%	58.52%*	**82.86%***	82.51%*	67.37%*
F score	28.62%	29.12%	**79.31%**	78.59%	51.55%

	LR	AdLASSO	LightGBM	CNTN	WDL
Y index	13.89%	17.04%	**65.72%**	65.02%	34.73%
Miscost	67.54%	70.43%	**29.11%**	29.69%	55.41%
	Income size and source verified				
	LR	AdLASSO	LightGBM	CNTN	WDL
AUC	59.71%	58.09%*	**82.83%***	82.46%*	62.38%*
F score	36.72%	27.86%	**79.27%**	78.01%	39.69%
Y index	19.41%	16.19%	**65.66%**	64.92%	24.76%
Miscost	63.05%	71.16%	**29.15%**	29.73%	63.88%

Note: The table reports the accuracy ratios for the estimated models based on the income verification. LR represents the logit model; AdLASSO refers to the Adaptive Lasso model; CNTN is the Convolutional Neural Tensor Networks; WDL represents the Wide and Deep Learning model. “AUC” refers to the area under receiver operating characteristic curve, “Y index” stands for the Youden’s index and “Miscost” is the misclassification cost for all models under study. The * in the AUC denotes that the difference against the benchmark LR model is significant at 99% confidence interval in the Delong test. The best values for each metric are in bold.

We note that the ranking of our models is retained, with LightGBM performing best with CNTN following closely. We also note that the accuracy of all models worsens for the non-verified group. This finding can be associated with the level of uncertainty of these loans’ applications. Loan applicants can fabricate their information to be more attractive without verification. Models built on inaccurate information fail to extract the actual attributes of default loans, leading to biased predictions. Nevertheless, the LightGBM and the CNTN provide staggering accuracy metrics. These results highlight the value of these two machine learning models in P2P loans’ default prediction. These two models can supplement the existing credit risk mitigation processes of P2P loan providers that, based on our dataset, fail to mitigate credit risk.

2.5.5. The Role of Home Ownership

We learn from Table 2-5, that home ownership potentially affects the default risk. Additionally, Serrano-Cinca et al. (2015) study empirical default P2P loan data from the Lending Club from 2008 to 2014 and note that loan recipients’ current housing situation explains potential risk. Thus, in this section, we present the performance of our models on the three sub-groups based on home ownership as the robust check.

Table 2-10: Predictions based on home ownership subgroups

	Own				
	LR	AdLASSO	LightGBM	CNTN	WDL
AUC	54.50%	57.97%*	**82.82%***	81.65%*	50.40%*
F score	28.89%	27.50%	**79.26%**	74.99%	1.57%
Y index	9.00%	15.94%	**65.64%**	63.29%	0.79%
Miscost	71.13%	71.37%	**29.17%**	31.02%	84.23%
	Mortgage				
	LR	AdLASSO	LightGBM	CNTN	WDL
AUC	60.04%	58.64%*	**82.09%***	81.88%*	70.20%*
F score	26.55%	29.48%	**78.17%**	77.87%	57.28%
Y index	20.08%	17.29%	**64.17%**	63.76%	40.40%
Miscost	63.29%	70.22%	**30.42%**	30.76%	50.59%
	Rent				
	LR	AdLASSO	LightGBM	CNTN	WDL
AUC	58.79%	59.59%	**82.27%***	82.09%*	69.92%*
F score	34.04%	32.19%	**78.45%**	78.15%	56.98%
Y index	17.58%	19.19%	**64.54%**	64.19%	39.84%
Miscost	65.30%	68.61%	**30.11%**	30.40%	51.08%

Note: The table reports the accuracy ratios for the estimated models based on the home ownership status. LR represents the logit model; AdLASSO refers to the Adaptive Lasso model; CNTN is the Convolutional Neural Tensor Networks; WDL represents the Wide and Deep Learning model. “AUC” refers to the area under receiver operating characteristic curve, “Y index” stands for Youden’s index and “Miscost” is the misclassification cost for all models under study. The * in the AUC denotes that the difference against the benchmark LR model is significant at 99% confidence interval in the Delong test. The best values for each metric are in bold.

We note that LightGBM continues to outperform for all metrics, with the CNTN having a close second performance. Most models present a similar performance across the three different sub-groups, with the notable exception of WDL, which confirms its sensitivity to the dataset.

2.6. Conclusion

Peer-to-peer lending (P2P) is an innovative form of financing. It benefits both the lender and the borrower by directly matching them according to their needs. However, P2P lending also suffers from drawbacks. Unlike traditional banks that screen loans through collaterals, deposits, and extensive datasets, P2P lending platforms transfer risk to the investor as intermediaries. There is a gap in information between the investor and the borrower. The investor can hardly make rational and informed decisions without an appropriate level of expertise, especially in the presence of information asymmetry between the two parties. Hence, P2P is inherently risky, which motivates an in-depth assessment of credit risk.

This chapter examines the determinants of P2P lending default chances using a rich dataset from the Lending Club. We study 279,512 loans and 24 explanatory variables for each loan with Logistic Regression (LR), Adaptive LASSO (AdLASSO), LightGBM, Convolutional Natural Tensor Network (CNTN), and Wide and Deep Learning (WDL). LR is the benchmark as it has been the key approach to identifying high-risk loans since the 1980s (Chen, Härdle and Moro, 2011). It binarily classifies samples without assumptions, distinguishing itself from other reduced-form models. Later in the literature, AdLASSO is put forward when more variables are considered. It is equipped with a penalty term to shrink variable coefficients for automatic feature selection. This is achieved by the process where coefficients of less valuable features tend to be pushed to zero while only curial ones are left. On the contrary, such linear models have been criticised for poor performance in massive and high-dimensional data (Khandani et al. 2010). They fail to determine the subtle non-linear relationship. Therefore, we are interested in non-linear approaches. We utilise LightGBM, CNTN, and WDL in our study. LightGBM is a decision-tree-based ensemble model specialised in big data classification with high classifying accuracy and low storage cost; CNTN is a convolutional neural network-integrated with tensors for better efficiency; WDL is a combination of general linear and deep learning models, serving as a credit ranking system that offers both generalisation and memorisation. In this chapter, these four machine learning models above are recruited along with LR to determine a better prediction model for P2P loans on the basis of a stratified 5-fold validation.

Our results show that machine learning classifiers offer significant gains in predictive ability. When comparing all candidate models, the LightGBM model performs better in out-of-sample prediction than logit models, mostly adopted in previous studies. CNTN and WDL present second-best and third-best performance, respectively. For the robustness test, we split the original dataset according to employment length, income verifications, and home ownership of the loan recipients. We find that the differing results hold for different subgroups. Because these three attributes are likely to be associated with loan repayment. LightGBM shows superiority in every subgroup and is followed by CNTN closely. On the other hand, our WDL approach seems sensitive to our dataset, reporting varied accuracy in different subgroups. Notably, in addition to the high accuracy, LightGBM generates feature importance atomically, making it a desirable credit scoring approach. Consistent with the literature, the debt-to-income ratio of the loan recipients is revealed by LightGBM to be the most valuable attribute in default prediction. Monthly instalment is also highly correlated with repay-ability. By contrast, loan recipients' delinquency history variables, such as the number of delinquencies in the last 12 months, seem not to be concerned for P2P loan investors.

Our findings highlight the value of machine learning, and more especially of LightGBM and CNTN, in handling large datasets and predicting the P2P loans' default rate accurately. This performance is robust to a set of different sub-groups characterized by heterogeneous levels of uncertainty. Our models are capable of supplementing the credit risk mitigation procedures of P2P loan providers. Traditional procedures such as income verification seem unable to reduce the default rate of P2P loans. On the other hand, we note the superiority of machine learning in predicting loan default risk, which can strengthen financial organisations' current credit scoring systems by enhancing the accuracy of abnormal loan request detection. Our study also offers insights into the attributes of risky loan requests that can guide individual investors to optimise their P2P investment portfolios and regulators to implement rules on loan applicants.

Appendix

2.A.1 In-sample Prediction

Table 2-11 presents the in-sample predictions of our models.

Table 2-11: In-sample predictions

	AUC	Y index	F score	Miscost
LR	61.36%	22.71%	31.96%	60.62%
AdLASSO	59.00%*	17.91%	30.38%	69.69%
LightGBM	**85.74%***	**71.48%**	**83.37%**	**24.22%**
CNTN	85.63%*	71.26%	83.22%	24.40%
WDL	80.46%*	60.92%	75.71%	33.18%

Note: The table reports the accuracy ratios for the estimated models in the in-sample. LR represents the logit model; AdLASSO refers to the Adaptive Lasso model; CNTN is the Convolutional Neural Tensor Networks; WDL represents the Wide and Deep Learning model. "AUC" refers to the area under receiver operating characteristic curve, "Y index" stands for Youden's index and "Miscost" is the misclassification cost for all models under study. The * in the AUC denotes that the difference against the benchmark LR model is significant at 99% confidence interval in the Delong test. The best values for each metric are in bold.

Chapter 3

Chapter 3 P2P loan funding decision: Determinants and macroeconomic effects

3.1. Introduction

Online Peer-to-Peer lending (P2P lending) is an innovative form of financing which directly matches lenders and borrowers without the involvement of a conventional intermediary. Since the first P2P lending platform, Zopa, was initiated in the UK in 2005, numerous P2P lending firms have launched worldwide. The Prosper is the first P2P lending platform in the US, established in 2006, followed by the Lending Club, one of the largest P2P lending firms in the world. In the same year, Renrendai and CreditEase were founded in China. Europe created its platforms such as isePankur and Auxoney. This internet-based financial tool has gained considerable attention from investors and market participants primarily due to reduced financing costs and higher returns than traditional fixed-rate financing (Ma *et al*., 2018). Tao and Lin (2016) note that in the P2P lending market, loan borrowers disclose their personal and credit information, for instance, demographic features, current and historic financial situation, and the reason for the loan. The lender uses this information to determine if their investment criteria are met and make the investment decision. However, the market can break down because of information asymmetry. Information asymmetry is reduced by banks in the traditional finance industry with extensive experience and expertise to assess the credit risk and allocate capital accordingly. On the contrary, the low cost of P2P lending is achieved by its lending process does not place any requirements on collateral or deposits like banks (Wei & Lin, 2017). This is an appealing characteristic of online lending, but it comes at a cost, as information asymmetry is associated with investors' risky reaching-for-yield behaviour. In addition to the loan-level information, macroeconomic situation, such as monetary policy, may also affect the risk-taking behaviour of the P2P lending investors. Current literature has mainly concentrated on bank loan funding decisions that banks are believed to take more risk by granting loans to risker borrowers during monetary policy easing. Evidence of lending standards in a FinTech nonbank financial organisation may carry the study a step forward.

|Understudying macroeconomic effects on P2P lending may also lead us to understand better the driving force behind the momentary policy risk-taking channel.

The aim of this study is to evaluate the decision to accept or reject P2P loans by accessing both loan-level information and macroeconomic information with the help of machine learning techniques. Our analysis is based on a cross-section of 12.4 million loan requests for the US, sourced from the Lending Club, which is a world-leading P2P lending platform. The data covers ten years, from April 2007 to December 2016, consisting of 12,401,233 instances. We constructed four common loan features for both successful and rejected loan requests (the amount requested, the loan purpose, the debt-to-income ratio and the employment length of the loan applicant). Among the four loan-level variables, loan purpose seems to be overlooked in the literature though it may carry important information for risk managers. It is a descriptive feature that is free typed by loan applicants. The context of this variable varies from one to the other and is unreadable for most statistic models. Thus, we utilise a text mining technique termed word2vec to convert the words into vectors and then use a clustering approach to classify them. In this way, we transfer the textual variable loan purpose into a categorical variable which our predictive models can process. Meanwhile, we consider the month a loan is funded or the loan application is generated to represent the periodicity. A series of macroeconomic variables are also included as features. We employ the quarterly growth of the gross domestic product (GDP), the quarterly growth of the consumer price index (CPI), the monthly growth of the Federal Funds Rate (FFR), and the monthly growth of the 90-day treasury bill rate (TBR). The Taylor residuals (Taylor) are used to assess the overall stance on monetary policy. To analyse the funding decision, we employ some of the most promising machine learning models, including SVM, Xgboost, and random forest, along with the latest variable importance analytical technique, Shapley value.

Previewing the main findings, first, we show that when applying machine learning techniques we are able to significantly improve the predictive power of our models compared to the logit model, which is the gold standard in the literature. Importantly, we note that the SVM model outperforms all other machine learning techniques and displays improved forecasting power. Our results are robust to tests on a battery of stratified 5-fold cross-validations. Second, we note that macroeconomic conditions affect P2P loan funding decisions as well as loan-specific attributes. Among them, a borrower's employment length is the main factor in the preference of lenders making decisions. The Federal Funds Rate (FFR) is the most powerful macroeconomic factor affecting individual risk-taking behaviour. Finally, we further document the variations of the macroeconomic effects on sub-groups in relation to debt-to-income ratio and loan purpose. Individual risk-taking behaviour is encouraged more towards the risker loan requests when monetary policy eases.

Our approach is mostly related to the literature that examines the determinants of P2P loan funding decisions. Recent studies by Tao & Lin (2016), Gavurova et al. (2018), Ip & Lam (2020) suggest that a borrower's income and employment length are the main factors in the preference of lenders making funding decisions. We complement and extend this finding in several ways. First, our focus is on granting loans, paying attention to both borrower-specific and macroeconomic indicators. We document the role of macroeconomic conditions in individual risk-taking behaviour. Second, we contribute methodologically by employing some of the most promising machine learning models (SVM, Xgboost, and random forest) along with the latest variable importance analytical technique (Shapley value) to the task of forecasting the granting decision of P2P loans in the US. The techniques we employ are well suited to model complex datasets such as ours and have provided promising empirical evidence in other fields. Third, we allow for two dimensions in the data that are critically important in determining the funding decision. Specifically, we split our sample according to the debt-to-income ratio and loan purpose. These aspects offer significant heterogeneity at the borrower and loan level, and we are able to tease out differences between the groups of borrowers. Finally, we demonstrate the value of machine learning and text mining in loan funding decisions because the way of processing our rich sample offers a low-computation-cost solution for big data.

3.2. Literature Review

With the pace of information and technology development, the popularity of bank loans has been transferred to electronic markets. Peer-to-peer (P2P) lending is one of the electronic financing markets where borrowers and lenders are matched directly. Most research focus on predicting default risk, and only a handful of studies are on funding success and macroeconomic factors. Thus, we also extend our topic into associated topics such as "monetary policy and lending standards in bank loans" and "macroeconomic factors in non-performing loans".

3.2.1 Successful P2P Loans and Monetary Policy

Based on the conceptual characteristics, P2P lending is supposed to be advantageous for the market participants in terms of cost. However, current literature shows some differences in determinants of the decision-making process. Tao and Lin (2016) explore the effect of borrowers' financial and demographic information, loan features, and lending models on the P2P funding outcomes. Their data sample is collected from Renrendai, a leading P2P lending firm in China, consisting of 220 thousand loans from 2013 to 2015. The independent variables in the study come from the information that the investors see. They fall into six categories: variables measuring the outcome and the performance of the loan request, credit and financial records of the borrower, specific characteristics of the loan request, demographic attributes, listing type, and other control variables such as geographic region dummies. The study applies regression analysis for the research task and notes that borrowers with higher income or cars are more likely to secure a loan while paying lower interest rates. Similarly, Zhang et al. (2017) examine the determinants of successfully funded P2P loans focusing on Chinese data. Approximately 200 thousand loan requests on Paipaidai are extracted from January to June 2014. Among them, only a quarter is granted requests. A binary logistic regression is utilised, along with eight loan-level independent variables: annual interest rate, repayment period, loan description, credit grade, previous successful-loan number of the borrower, previous unsuccessful-loan number of the borrower, gender, and the credit score. They consider all the variables understudy determinants of funded loans but with different effects. The annual interest rate, credit grade, previous successful-loan number of the borrower, and the credit score positively affect the funding decision, while the rest have a negative impact on the funding outcome. Besides the regression analysis, analysis of variance and the test of residuals normality are valuable tools. To investigate the successful loan applications, Gavurova et al. (2018) gather around 44 thousand loan application records from Bondora, a well-known European personal loan online provider, from 2009 to 2015. First, they hypothesise the impact of the variables on the investors' decisions. A sensitivity analysis follows this in the form of logistic regression. Finally, testing of the residuals is performed to verify the desired outcomes. The results indicate that the history of the borrower has no significant influence on the lending decision, while the debt-to-income ratio and liabilities are very important. Notably, they prove the gender discrimination in funding decision-making where women have lower failure rates on the platform Bondora. In addition to the loan-level features, some studies argue that market environment or macroeconomic conditions also play important roles. Ip and Lam (2020) undertake a study on funding decisions in online P2P lending based on the Lending Club. They retrieve 28 million loan listings from June 2007 to December 2018 by merging the rejected loan records and funded loan records. By doing so, four loan-level features are constructed from the intersection of data columns, requested amount, borrowing purpose, debt-to-income ratio, and employment length. Periodicity variables are created using the month and quarter of the loan application. Based on them, Ip and Lam introduce the monetary policy regime dummy variable. Unlike the main steam of the literature where conventional methodologies are widely used, this study employs a machine learning approach termed the decision tree. The outcome shows that investors tend to have a high preference for borrowers' employment length, as it may be the proxy for borrowers' income stability. The requested amount and the debt-to-income ratio are also contributors to a successful funding decision. However, monetary policy does not matter much.

It is challenging to examine macroeconomic effects and lending behaviour for traditional financing intermediaries such as banks. Because the monetary policy also changes loan demand via the interest rate channel or the balance sheet channel. However, it is possibly tackled in the P2P lending market by full access to the information given to the lender. Having said that, Chu and Deng (2018) define low federal funds rate and quantitative easing to represent eased monetary policy and investigate how it affects Prosper.com from 2006 to 2013. Based on the results from linear regressions, they point out that momentary policy softening induces individual investors to fund riskier loans. They note that other macroeconomic variables, including inflation rate and GDP growth, do not eliminate monetary policy effects. Wong and Eng (2020) argue that "P2P lending could potentially defy the two notable monetary policy transmission mechanisms". Due to no claims being requested for the P2P lending platforms against the central bank, it causes no leverage on the interest rate investors would like to offer in the market. It possibly enables the interest rate channel. On the other hand, collaterals are not required for a P2P loan, degrading the relevance of the balance sheet channel. A defined policy rate and a New Keynesian model reveal that a tightening policy may induce P2P borrowing to leverage a stronger business investment. Having the same research interests, Huang, Li, and Wang (2019) claim that evaluations of monetary policy that only affect banks may no longer be sufficient to present the whole finance market or to support the policymakers. This is because nonbanking institutions are of increasing importance. 73,264 loan applications during 2017-2018 are employed in their study to estimate loan granted probability by probit models. A similar conclusion is reached that monetary policy easing is associated with a higher probability of granting loans to risky borrowers and greater riskiness of credit allocation. "But these changes do not necessarily relate to a larger loan amount on average", they added.

3.2.2 Monetary Policy and Bank Lending

Limited by the number of studies on P2P loans, we also draw a lesson from the bank lending standards. Den Haan et al. (2010) examine the behaviour of loan components and firstly document a robust response of bank loans to monetary policy. They measure the federal funds rate (FFR) and consumer price index (CPI) with a vector autoregression model (VAR) to estimate the monetary policy shocks on bank loans. Driscoll et al. (2012) enrich the research using the same methodology but considering year-head gross domestic products (GDP), unemployment rate, treasury bill rate, and treasury yield. They confirm that a sharp widening of credit spreads always occurs with easing the monetary policy. A common path is suggested by Ibrahim & Shah (2012) that links bank lending, macroeconomic conditions, and financial market uncertainty. Specifically, they employ total bank credit to the private sector deflated by the GDP deflator to present the real bank credit. They use GDP, lending rate, stock price and volatility to deliver macroeconomic conditions. A vector autoregression framework is adopted to examine the interactions between them. They note long-run positive relations between real output and both real bank credits and real stock prices.

Reduced-form models are widely used in estimating monetary policy shocks. To estimate the monetary policy effects on the risk-taking behaviour in U.S. bank loans, Delis et al. (2017) collect 1987 to 2012, 22,592 unique loan requests to bring together the change in the real federal funds rate with information on loans, firms, and characteristics. Taylor residuals are employed in the study to measure the monetary policy, and simple logit models are employed to compute the default probability. They document that risk-taking by banks (as measured by corporate loan spreads) is negatively associated with expansionary monetary conditions. Furthermore, Popov (2016) suggests that risky applicants may benefit from it because monetary policy loosening leads to higher lending and higher credit risk-taking of banks. Their study is conducted on eight economies that use the euro from 2004 to 2007. It reveals that banks, especially banks with lower capital ratios, increase credit in general and especially the credit to ex-ante risky firms. Based on 18,907 loans from 235 banks, Paligorova & Santos (2017) reach a similar conclusion. Their study is based on loan pricing policies of U.S. banks in the last two decades analysed by loan spread regression. Loan spreads for riskier firms are relatively lower during periods of monetary policy easing compared it tightening. A potential explanation is that a lower overnight interest rate that is consistent with monetary policy easing induces lowly capitalised banks to grant more loan applications to the ex-ante risky firms and to commit larger loan volumes with fewer collateral requirements to these firms. This is claimed by Jimenez et al. (2014) when they undertake a study on over 241 thousand bank loan applications in Spain, a country with an economic system dominated by banks. They utilise a two-stage model whose first stage is investigating the granting of loan requests, and the second stage is examining the given loan applications. Meanwhile, the model controls for time-varying, firm, and bank heterogeneity. They note that a lower interest rate may lead to higher risk-taking in bank lending. In contrast, higher rejection of loan applications, reduced volume of new loans, and higher loan rates result from tightening credit conditions. By introducing a 7-day interbank rate, real GDP growth, inflation, bank liquidity, and bank capital, monetary policy tightening shock is believed to be stronger for banks with less capital and greater exposure to sovereign debt, based on a study on 26,363 loan applications from 15 Ugandan banks (Abuka et al., 2015). Furthermore, a tightening monetary policy reduces the supply of bank credit to loan applicants and dampens economic activity. This is argued by Abuka et al. (2019), who use 2010 to 2014 Uganda's supervisory credit register records for loan data while using interbank rate, policy rate, T-bill rate, GDP change, and CPI change for macroeconomic data. The linear probability model is utilised for statistical analysis and shows a significant impact of monetary policy on the quantity and price of credit. Gete (2018, p.21) proposes a trackable way to integrate bank lending standards into DSGE models of financial frictions. They define "tighter lending standards as a non-price mechanism that may generate misallocation between safe and risky borrowers." Associated with higher borrowing costs, banks reject the riskier borrowers because no lending rate compensates banks in the case where the borrowers' default. However, it seems not the case for P2P lending platforms as they take no responsibility for default risk.

3.2.3. Macroeconomic Factors and Non-performing Loans

Macroeconomic factors are believed to affect no-performing loans (NPLs), which may also change the lending standards in return. Most loans fail to get funds due to the possibility of default. Thus, it is worth exploring the relationships between macroeconomic factors and

NPLs. 5 categories of macroeconomic factors are commonly considered in the top: (1) economic general state (e.g. GDP, unemployment rate); (2) price stability conditions (e.g. inflation rate, the annual growth rate of the M3 monetary aggregate); (3) cost of servicing debt (e.g. 3-month Euribor rate); (4) debt burden (e.g. the ratio of loans to disposable income); (5)financial and real wealth (e.g. the growth rate of the Italian stock prices index) (Bofondi & Ropele, 2011). However, Bofondi & Ropele (2011) reveal that the quality of lending is mainly determined by a small number of macroeconomic variables in relation to the general state of the economy with a lag. Specifically, a study on the Romanian banking system during the financial crisis conducted by Vogiazas and Nikolaidou (2011) points out the key influencers. This study is motivated by the hypothesis that non-performing loans are affected by macroeconomic-cyclical indicators, monetary aggregates, interest rates, financial markets, and bank-level variables. Based on the monthly series from December 2001 to November 2010 of Greek banks in Romania, they perform time series modelling and note that macroeconomic determinants of non-performing loans are the construction and investment expenditure, the inflation and the unemployment rate, the country’s external debt to GDP, and M2 money.

GDP, interest rate, and inflation rate are widely mentioned in NPLs research. To examine the potential influences of both macroeconomic and bank-level variables on the quality of loans, Abid et al. (2014) employ dynamic panel data methods over loans from 16 Tunisian banks for a period from 2003 to 2012. Their choice of variables is inspired by the literature. Macroeconomic variables are composed of GDP growth, the inflation rate, and the real lending rate. They reveal the coefficients of the lagged macroeconomic variables are statistically significant. The result is compatible with the hypothesis that a decline in the inflation rate boosts financial conditions, and the slow growth in economics affects the NPL ratio negatively. Furthermore, Koju et al. (2018) argue that low economic growth is regarded as the primary cause of high NPLs in Nepal. In addition to dynamic panel estimation, they also utilise static approaches on 30 Nepalese commercial banks from 2003 to 2015. The dataset includes 5 macroeconomic variables and 7 loan-specific variables subject to data availability. The macroeconomic data consists of GDP growth, per capita outstanding debt, inflation rate, remittance, and export to import ratio. It is sourced from the annual monetary policy report and economic survey report. In the same vein, Grace et al. (2016) documents that the inflation rate along with the 90-day treasury bill rate is important to NPLs in Ghana's banking system. They collect quarterly data from commercial banks in Ghana from 2008 to 2015 and consider inflation, exchange rate, money supply (M2), GDP, and treasury bill for 90 days to present the macroeconomic state. Armed with the Auto-Regressive Distributed Lag bounds test of co-integration, they show that macroeconomic instabilities, especially changes in inflation and T-bills, have a dramatic impact on loan performance. On the other hand, the impact of the inflation rate on NPLs is criticised as it can be statistically insignificant. Tanasković and Jandrić (2015) collect loan data combined with country-specific macroeconomic and financial indicators from different countries over the period 2006-2013. They use a static panel model approach to approve that non-performing loans are linked to various macroeconomic factors, but a significant negative relationship between GDP growth and NPL ratio is noted. The research also claims that the inflation rate impact is statistically insignificant for sample countries. Zainol et al. (2018) also hold the same conclusion by applying auto-regressive distributed lag to Malaysian bank loans. They collect a set of time series data from Q1 2006 to Q4 2015, and they consider macroeconomic effects by including GDP, base lending rate, inflation, and household income distribution as variables. In line with the expectation, GDP significantly affects the non-performing loans in a negative direction, while base lending rate and household income distribution are significant and positively associated with non-performing loans. But contrary to our expectation, the impact of inflation on loan performance is insignificant. This result is robust based on several diagnostic tests.

3.3. Methodology

In this section, we present our machine learning approaches for forecasting the loans' outcomes.

3.3.1. Predictive Models

Logistic Regression (LR):

Even though the new toolbox for researchers keeps updating, linear regression models remain the cornerstone of empirical work in finance and other scientific fields (Romano & Wolf, 2016). We add instance-weights to this most fundamental binary linear approach and execute it as the benchmark in this study. Suppose we have N loan application instances and M features for each of them. Let us define it as $X_{n,m}$ (n =1, 2, .., N ; m = 1, 2, .., M; $X_{n,m} \in X$) the vector of features for our loan applications and $\mathcal{W} = \{w_1, \dots, w_{\mathrm{N}}\}$ as the corresponding instance weights. If $Y \in \{0,1\}$ is a binary variable that takes the value of 1 if the loan application has been rejected and 0 if the loan is granted, then LR conditional probability can be expressed as:

$$Probability(Y = 1) = \frac{e^{\beta_0 + \sum \beta X_{n,m} * W}}{1 + e^{-\beta_0 - \sum \beta X_{n,m} * W}} \quad (1)$$

where β is a vector of covariate effect parameters and β_0 is a scalar parameter.

Naïve Bayes (NB):

NB is a machine learning statistical classification algorithm based on Bayes Theorem. It simply builds the connection between the possibility of the event $X_{n,m}$ given outcome Y_n and the possibility of Y_n given $X_{n,m}$. The NB with the maximum posterior probability is defined as:

$$Probability(Y_n | X_{n,m}) = \text{arg}max P\text{robability}(Y_n) \prod_m^M Probability(X_{n,m} | Y_n) \quad (2)$$

where $Probability(Y_n) = \frac{\sum_n^{\mathrm{N}} w_n \vartheta(Y_n, Y) + \frac{1}{2}}{\sum_{n=1}^{\mathrm{N}} w_n + 1}$, $Probability(X_{n,m} | Y_n) = \frac{\sum_n^{\mathrm{N}} w_n \vartheta(X_{n,m}, X_m) \vartheta(Y_n, Y) + \frac{1}{\mathrm{N}_m}}{\sum_n^{\mathrm{N}} w_n \vartheta(Y_n, Y) + 1}$, and $\vartheta(\varphi_1, \varphi_2)$ is a binary function that takes the value of 1 when $\varphi_1 = \varphi_2$ and 0 otherwise.

Random Forest (RF):

RF is an ensemble machine learning method which is based on sets of Classification and Regression Trees (CARTs). It is a computationally efficient method and has been widely used in many areas, especially in credit assessment (Chang, Chang & Wu, 2018; Pan and Zhou, 2019; Pal, Kapali & Trivedi, 2020).

If f (x) presents the CART, K denotes the number of the CART. We split the sample (X, Y) into K subsets for each tree. Thus, a Random Forest (RF) with K trees is generated as the average prediction value:

$$\hat{Y} = \frac{1}{K} \sum_{k=1}^{K} f_k(X^k), \quad f_k \in \mathcal{F} \quad (3)$$

Here $\mathcal{F}$ is the set of all possible CARTS. The following function is minimised to compute the tree in the RF:

$$\Phi = \sum_n \varphi \, w_n(\hat{Y}_n, Y_n) + \sum_k \Omega(f_k) \text{ and } \Omega(f) = \Upsilon\eta + \frac{1}{2}\lambda \parallel weights \parallel^2 \quad (4)$$

Where φ is a differentiable convex loss function that measures the error between the predicted $\hat{Y}_p$ and the actual Y_p. Ω term penalises the classification tree function from being complex. It requires RF to select and use a small subset of "strong explanatory variables" among all variable candidates. η is the number of leaves in each tree, and Υ and λ are constants. $weights$ is the leaf weight.

Xgboost:

Xgboost is a gradient tree-boost machine learning algorithm with a similar function to RF. It also incorporates CARTs. More specifically, it sharpens weak CART classifiers' performance by continuously superimposing. The algorithm is based on gradient boosting. In RF each tree is built independently, while gradient boosting builds one tree after the other. Xgboost uses CARTs as base classifiers and applies a loss function to control their complexity. Assume we have $\mathcal{K}$ CARTs in the Xgboost. The integrated classifier is computed as follows:

$$\hat{Y} = \sum_{\mathscr{k}=1}^{\mathcal{K}} f_b(X), \quad f_b \in \mathcal{F} \tag{5}$$

Here b is used to summarise the iteration in Xgboost. Given each sample, the final $\hat{Y}$ is calculated by retaining (b-1) rounds of model prediction in each step and intruding a new $f_b(X)$ at the end. Similar to RF, our target is minimising the equation (6). But an additional term f_b is added, and the second-order approximation is introduced to obtain the optimised objective:

$$\Phi' = \sum_{\mathscr{j}}^{\eta'} [(\sum g) weights_{\mathscr{j}} + \left(\frac{1}{2}\sum h + \lambda' weights_{\mathscr{j}}{}^{2}\right) + \Upsilon'\eta' \tag{6}$$

We denote $\mathscr{j}$ is the leaf node and $weights_{\mathscr{j}}$ is the weight of the newly-grew leaf. g and h are first and second-order gradient statistics on the loss function. η' is the number of leaves in each tree, and Υ' and λ' are constants. The optimal $weights_{\mathscr{j}}$ is defined by:

$$weights_{\mathscr{j}}{}^{*} = -\frac{\sum g}{\sum h + \Upsilon'} \tag{7}$$

Support Vector Machine (SVM):

The aim of SVM is to separate the instances into two classes via an optimum hyper-plane that maximises their margin. This is achieved with the assistance of a kernel function which projects the samples into a higher dimensional plane where they can be separated. The SVM classifier is expressed as:

$$Sign\ (\sum_{n=1}^{\mathrm{N}} \sigma_n Y_n \mathbb{K}(\mathrm{X}, X_n) + \mathfrak{b} \tag{8}$$

$$\min_{X} \sum_{n=1}^{\mathrm{N}} \mathrm{X}, X_n Y Y_n \mathbb{K}(\mathrm{X}, X_n) - \sum_{n}^{\mathrm{N}} X_n \qquad s.t. \sum_{\mathrm{n}}^{\mathrm{N}} X_n Y_n = 0,\ 0 \leq X_n \leq \mathcal{C} w_n \tag{9}$$

σ_n stands for the Lagrange multiplier in terms of sample X_n. $\mathbb{K}(\mathrm{X}, X_n) = \exp(-r \parallel X_n - \mathrm{X} \parallel^2)$ is the radial basis function kernel. r $(r > 0)$ refers to a kernel parameter. $\mathcal{C}$ is a constant that balances the maximisation of classification margin and the minimisation error. We optimise ƅ and $\mathbb{K}$ 5-fold by 5-fold cross-validation in the in-sample.

3.3.2 Feature Importance

Interpretability of a predictive model is crucial. It shields user trust, supports the understanding of the modelling process, and adds benefits to further analysis and model improvement. A metric that can reveal the relationship between our machine learning forecasts and features is the Shapley value. It calculates the importance of a feature by comparing what our model predicts with and without that feature. In other words, it matches the goal of a cooperative game – distributing the worth of the grand coalition among players in a fair way:

$$\psi_m(X) = \sum_{\mathcal{M} \subseteq S_M \backslash \{m\}} \frac{|\mathcal{M}|!(|S_M| - |\mathcal{M}| - 1)!}{|S_M|!} \left(\Delta_{\mathcal{M} \cup \{m\}}(X) - \Delta_{\mathcal{M}}(X)\right) \tag{10}$$

where $\mathcal{M} \subseteq S_M$ is a subset of the features. $\Delta_{\mathcal{M}}$ is the importance of the feature subset $\mathcal{M}$. However, it requires an exponential time complexity to compute, which makes it inappropriate for practice use. Thus, we equip a structured random sampling to generate the approximate Shapley value with reduced computational cost. We start by introducing permutations to the equation (8):

$$\psi_m(X) = \frac{1}{M!} \sum_{\mathcal{O} \in \pi(M)} \left(\Delta_{pre^m(\mathcal{O}) \cup \{m\}} - \Delta_{pre^m(\mathcal{O})}\right) \tag{11}$$

We denote $\pi(M)$ as the set of all permutation orderings of the feature indexes {1, 2,…, M}, and $pre^m(\mathcal{O})$ as the set of all indexes that precede m in premutation $\mathcal{O} \in \pi(M)$. To reduce the computational complexity, we follow Štrumbelj and Kononenko (2013) and limit ourselves to such distributions of instances n that individual features are distributed independently. In this manner, the contribution of a subset feature is defined:

$$\Delta_{\mathcal{M}}(X) = \not{f}_{\mathcal{M}}(X) - \not{f}_{\{\ \}}(X) = \sum_{\ell \in X} n(\ell) \left(\not{f}\left(\ell_{[\ell_m = X_{m,} m \in S_m]}\right) - \not{f}(\ell)\right) \tag{12}$$

where $\not{f}()$is a predictive model, $\ell_{[\ell_m = X_{m,} m \in S_m]}$ presents instances ℓ with the value of feature m replaced with the feature's value in instances X, for each $m \in S_m$. For instance, with $\ell = \{3,5,7\}$ and $X = \{4,6,8\}$, $\ell_{[\ell_m = X_{m,} m \in \{1,2\}]} = \{4,6,7\}$. We can substitute the time-consuming Δ terms in equation (9) with equation (10). We have approximate Shapley value:

$$\widehat{\psi_m} = \frac{1}{\mathrm{N}} \sum_{\mathcal{s}=1}^{\mathrm{N}} (\mathcal{f} \left(\ell_{[\ell_{\mathcal{s}}=X_{\mathcal{s}}, \mathcal{s} \in pre^m(\mathcal{O}) \cup \{m\}]} - \ell_{[\ell_{\mathcal{s}}=X_{\mathcal{s}}, \mathcal{s} \in pre^m(\mathcal{O})]} \right) \tag{13}$$

3.4. Data Description and Compact

3.4.1. Data and Variable Source

We study loan applications from the Lending Club, the biggest peer-to-peer lending platform worldwide. The Lending Club data is ideal for exploring the funding decision for a number of reasons. First, while the historic loan requests are hard to access on most platforms like Prosper, the Lending Club data is publicly available for investment advice or research[4]. We examine the same information as potential investors do. Second, it is the largest P2P lending platform in the world which has issued over USD 60 billion in loans and has worked with over 3 million customers since 2007, offering a wide range of samples (Polena & Regner, 2018). Our sample employs loan requests from April 2007 to December 2016, consisting of 12,401,233 instances. Compared to other studies in the same vein, for example, a study on individual risk-taking behaviour and monetary policy (Chu & Deng, 2018) that employs loan requests from 2006 to 2013, our data is of huge volume over a wide period. We merged the information provided for funded and rejected loans and we constructed four loan features that are common for all loan requests (the amount requested, the loan purpose, the debt-to-income ratio and the employment length of the loan applicant).

The loan purpose or the reason for which the applicant is requesting the funds contains descriptive facts that reveal important information for credit risk managers. This textual feature is free-typed and can be only incorporated into predictive models through text mining and natural language processing (NPL). In our study, we utilize word2vec with continuous bag-of-words (CBOW) in order to extract previously unseen information from the loan applications under study. CBOW aims at computing the possibility of a word presenting a context. It transfers the context of words from the textual parts into vector representations and predicts the central word given these vector representations. For example, if the context words are "guy, attempt, over, puddle, fall", CBOW can output the central word "jump" (Othman, Faiz and

[4] LendingClub Dataset retried on 31 December, 2020. But they are working with a new bank aiming at offering more in the future. Please refer https://www.lendingclub.com/investing/peer-to-peer

Smaïli, 2019). Assume that vocabulary list V contains all unrepeated words in the loan applications under study. Once word embedding dimension d is confirmed, word2vec firstly maps the vector representation of context words into a |V| dimensional vector. The output matrix $O \in \mathbb{R}^{|V| \times d}$ which represents the central word w_c can be determined by maximising the following conditional probability:

$$\text{Probability}\left(w_c \middle| w_{[-\alpha,\alpha]-[c]}\right) = \frac{\exp\left(V_c^T O_c\right)}{\sum \exp\left(V^T O_c\right)} \tag{14}$$

where α defines the window of context words, and T is the sequence of the text. Following the central words, we can further classify the loan textual feature loan purpose into categories via a clustering technique.

The month a loan is funded or the loan application is generated (for the cases a loan is rejected) is included as a variable to represent the periodicity[5]. Additionally, we also include a series of macroeconomic variables as features. This choice is motivated by previous research on macroeconomic effects and lending standards (Popov, 2016; Chun and Deng, 2018; Abuka *et al.*, 2019). We consider the quarterly growth of the gross domestic product (GDP), the quarterly growth of the consumer price index (CPI), the monthly change of the Federal Funds Rate (FFR), and the monthly growth of the 90-day treasury bill rate (TBR)[6]. We also consider the overall stance of monetary an essential macroeconomic variable by including the Taylor residuals (Taylor) in the study. This is because macroeconomic variables and credit risk are affected by each other simultaneously - exogenous monetary policy may contribute to P2P loan funding decisions. Taylor rule is the basis for federal funds rate decisions, and the residual of its regression implies the direction of the monetary policy. That is, a negative residual suggests softening and vice versa. Following the literature, Taylor residuals are obtained by running rolling regressions of the Federal Funds Rate on the deviation of CPI from the 2% target rate and the difference between the actual and potential GDP with the data from 1980 to 2020. These macroeconomic indicators mentioned above are widely studied in bank loan lending standards. But their roles in P2P lending and its investors' lending preference are rarely revealed in the literature. This gap motivates us to study them in our context further.

[5] By constructing this feature, we assume that the loans' application and funding are happening in the same month. This assumption is consistent with the literature (Ip & Lam, 2020) and the information provided by the Lending Club which states that the whole process (application, approval and funding) takes on average 7 days. It is also necessary as the date that a loan is rejected or the date that the application of a successful loan is made is not available by the Lending Club.
[6] All macroeconomic variables are collected from https://fred.stlouisfed.org/categories/32991.

3.4.2. Summary Statistics

In this section, we present the summary statistics of our features. In Table 1, the summary statistics of the amount requested and the debt-to-income ratio of applicants are presented. It should be noted that the Lending Club assigns "-1" to the debt-to-income ratio of applicants with no income but debt.

Table 3-1: Statistic summary

		All	Accepted	Rejected	p-values
Sum		12,401,233	1,321,847	11,079,386	-
Amount Requested	mean	13.54	14.75	13.39	0.00***
	std	15.57	8.62	16.20	
	min	0.00	0.50	0.00	
	Q3	20.00	20.00	20.00	
	max	1400.00	40.00	1400.00	
Debt-To-Income Ratio	mean	174.01	18.87	192.51	0.00***
	std	15611.44	70.99	16516.37	
	min	-1.00	-1.00	-1.00	
	Q3	31.81	24.24	33.60	
	max	50000031.49	9999.00	50000031.49	

Note: The table presents the summary statistics of the Amount Requested (presented in a thousand of $) and debt-income ratio (presented in%) for the whole dataset, for the accepted loans and the rejected applications. Q3 stands for the third quartile, which is the middle value between the median and the highest value of the data set. In the last column, the p-values present the test of the equality of mean in accepted loans and rejected loans. *** indicates that the null hypothesis of equal means is rejected at the 99% level.

We note that our sample is heavily imbalanced. 90% of loan requests are rejected. The p-values reveal that the two features do not have equal means. This is also evident from the substantial differences in the summary statistics of the two variables. In Table 2, we present the employment length of the loan applicants. We used integer-encoding to represent the employment length in a time series.

Table 3-2: Variable employment length

Categories	Integer codes	All	Accepted	Rejected	P-values
Unemployed	0	3.45%	5.53%	3.20%	0.00***
< 1 year	1	**72.27%**	7.76%	**79.97%**	0.00***

1 year	2	1.27%	6.53%	0.64%	0.00***
2 years	3	1.62%	8.96%	0.74%	0.00***
3 years	4	1.41%	7.93%	0.63%	0.00***
4 years	5	1.05%	5.93%	0.47%	0.00***
5 years	6	10.59%	6.22%	11.11%	0.00***
6 years	7	0.81%	4.66%	0.35%	0.00***
7 years	8	0.74%	4.45%	0.29%	0.00***
8 years	9	0.77%	4.75%	0.29%	0.00***
9 years	10	0.61%	3.90%	0.22%	0.00***
>=10 years	11	5.43%	**33.41%**	2.09%	0.00***

Note: The table presents the employment length of the loan' applicants in terms of %s of the total applications. The second, the third, and the fourth column present the related %s for the whole dataset, for the accepted and the rejected loan applications, respectively. In the last column, the p-values refer to the test of equality of means in accepted and rejected applications. *** indicates that the null hypothesis of equal means is rejected at the 99% level. The values in bold represent the highest figures in each category.

From the table above, we note a positive (negative) relationship between employment length and loan acceptance (rejection). The highest percentage of employment length of the successful applicants is more than 10 years, and for the unsuccessful applicants, less than 1 year. Table 3-3 below presents the loan purpose percentages for the loan applications under study. We extracted this series with word2vec and CBOW. The loan purpose reveals the reasoning the applicants have stated in their applications for their fund requests. This feature is overlooked by the related literature, although it might disclose crucial information to credit managers and is likely to be an important determinant of loan funding success.

Table 3-3: Loan purpose

Categories	Integer encodes	All	Accepted	Rejected	P-values
Medical purpose	0	2.69%	1.06%	2.88%	1.00
Other	1	16.75%	6.65%	17.96%	0.00***
Credit card	2	12.79%	22.53%	11.62%	0.00***
Debt consolidation	3	**47.87%**	**58.49%**	**46.60%**	0.00***
Major purchase	4	3.56%	2.16%	3.73%	0.00***
Car	5	5.90%	2.18%	6.34%	0.00***
Moving	6	2.77%	0.65%	3.02%	0.00***

Home improvement/buying	7	5.73%	6.28%	5.67%	0.00***
Business	8	1.94%	0.00%	2.17%	0.00***

Note: The table presents the loan purpose of the loan applications in terms of %s of the total applications. The second, the third, and the fourth column present the related %s for the whole dataset, for the accepted and the rejected loan applications, respectively. In the last column, the p-values refer to the test of equality of means in accepted and rejected applications. *** indicates that the null hypothesis of equal means is rejected at the 99% level. The values in bold represent the highest figures in each category.

Debt consolidation is the main reason for funding requests in both the rejected and accepted groups. The p-values reveal a significant difference in the means of the accepted and rejected loans with a notable difference in medical purpose. It is noteworthy that no loan was taken when "business" is stated in the loan purpose. In Appendix 3.A.1, we present the fluctuations of the macroeconomic level features in relation to the number of loan requests and the amount of money requested.

3.4.3. Data Clustering

Our dataset contains more than 12.4 million loans. Examining a dataset of this size with machine learning is computationally expensive (even with parallel computing), and the inherent noise and outliers will cripple the performance of our forecasting models. Clustering can reduce the computational cost and the noise of the dataset while retaining the information of the original data.

BIRCH (balanced iterative reducing and clustering using hierarchies) is a data mining algorithm that performs hierarchical clustering. BIRCH uses tree structures (clustering feature trees) where each node is composed of clustering features. Each clustering feature is defined by the number of data points in the cluster, the summation of these datapoints and their squared summation. The fraction of the summation of the datapoints to the number of the datapoints is termed centroid (x_0).

Each clustering feature tree has two parameters, the branching factor ($\mathcal{B}$) and the threshold ($\mathcal{T}$). $\mathcal{B}$ is the maximum number of subclusters in each node, and $\mathcal{T}$ is the maximum radius of a subcluster obtained by merging a new sample and the closest subcluster. These two parameters try to balance computational cost and informational value. We set $\mathcal{B}$ as 150 and adjust $\mathcal{T}$ from 0.5 until the representation is capable of the memory and no duplicate medoids are reached for different subclusters (Yu et al., 2003). Harrington et al. (2018) point out that the optimality of the representation is only slightly affected by the parameters after experimenting with potential parameter values in BIRCH.

Given $\mathcal{N}$ loan instances with associated features, we have $S = \{x_1, \dots, x_{\mathcal{N}}\}$ to be grouped. The subcluster radius R is:

$$R = \sqrt{\frac{\sum_{n=1}^{\mathcal{N}} (x_{\mathrm{n}} - x_{0})^{2}}{\frac{1}{2}}} \tag{15}$$

Using average values for the numerical features is acceptable. But for categorical features (e.g. employment length and loan purpose) whose integer values stand for categories, simply taking the average values damages the underlying information. Real instances from the subclusters are better positioned to act as representations in this case. Consequently, we choose the closest real instance to the centroid (medoid) alternatively. Medoid is captured by calculating and comparing the Euclidean distance between each instance in the subcluster and its centroid. BIRCH transforms our dataset to two matrices. One with a set of presentative vectors and one with the corresponding weights for these presentative vectors. We extracted 41,170 weighted medoids to represent the original sample. Among them, 32,945 are rejected, and 8,225 are accepted medoids.

3.5. Empirical Findings

3.5.1 Predictive Performance

We measure the predictive performance of all our models in terms of area under the curve (AUC), precision, recall and F1 score. We estimate the statistically significant difference between our machine learning models AUCs and those obtained by the LR benchmark with the Delong (Delong et al., 1988) test. The relevant formulas of these metrics and the test statistic of the Delong test are presented in Appendix 3.A.2.

We split our dataset into five equal subsamples. We use four subsamples as in-sample and the last one as out-of-sample. We repeat this process five times, so each of the five subsamples acts once as an out-of-sample, and we estimate the average performance of our models.

The extant literature on loan lending standards mainly considers conventional models such as time-series analysis and Ordinary Least Square. In this study, we propose a horse-racing where four representative machine-learning methods are against a traditional approach, aiming at exploring the determinants of P2P loan lending decisions using big data.

Machine-learning algorithms with complex topologies and parametrisation are particularly prone to sense subtle non-linear connections between variables that transitional models may miss. To guard the robustness of our results and the generalisation of our findings, we normalise data into a range of [0,1] before training the models. This alleviates the misleading influence of various units of feature measurement where features with large-scale value take the dominance from small-value features. Representations with corresponding weights are stratified-sampled into five subsets with equal sizes. Each time four subsets integrate to train the model (for Logistic Regression and Naïve Bayes) or perform an exhaustive grid-search for optimal parameters (for machine learning models). At the same time, the remaining one acts as the test set to evaluate the model and obtain the feature importance of each variable. The process above repeats five times. Each subset has a chance to test the models. Finally, model properties and feature importance are gathered by averaging the five results of the test subsets.

The out-of-sample performance of our models is presented in Table 3-4[7] , and the in-sample performance is shown in Appendix 3.A.3.

Table 3-4: Out-of-sample predictions

	AUC	Precision	Recall	F Score
LR	54.41%	53.29%	66.34%	35.88%
NB	47.37%*	45.68%	1.28%	1.80%
SVM	**72.62%***	**76.16%**	**70.40%**	**62.41%**
RF	55.46%*	58.85%	42.85%	42.96%
Xgboost	61.78%*	69.46%	60.11%	55.35%

Note: The table reports the accuracy ratios for the estimated models in the out-of-sample. LR represents Logistic Regression; NB refers to Naïve Bayes; SVM is Support Vector Machine; RF represents Random Forest. * in the AUC denotes that the difference against the benchmark LR model is significant at 99% confidence interval in the Delong test. The best values for each metric are in bold.

[7] Models are built using python scikit-learn library and DeLong Test is built with the help of https://github.com/yandexdataschool/roc_comparison

SVM beats the rest four models in terms of all metrics. Starting from AUC, SVM states the best discrimination ability regardless of the threshold, increasing the overall accuracy of the benchmark LR by almost 18%. Though the high precision and the relatively low recall obtained by SVM suggest it tends to be slightly biased towards the bigger-sized class (rejected applications), the best F score among the candidate models claims that SVM performs classification well. Concerning other models, Xgboost is ranked second after SVM in all evaluations. RF slightly outperforms the benchmark with respect to AUC, precision, and F score, but it does not show a significant superiority. We note NB is unable to reach a robust prediction in this study. Both recall and F score are merely over 1%, which reveals it fails to identify accepted instances. This result is consistent with the literature where SVM is proved to have better classification performance in imbalanced credit data (Zhang et al., 2020).

To further assess the robustness of the proposed models, we split the original 12.4 million samples into subgroups in terms of Debt-to-income Ratio (DIR) and Loan Purpose. The same process is repeated. Each subset is re-clustered by BIRCH and re-analysed based on 5-fold cross-validation.

DIR is used to measure the risk level in this study because it reflects repayability by comparing the borrower's debt to their income. We consider 50 percent as the maximum value for healthy DIR and the gap between subgroups[8]. Starting with the lowest value, "-1" in DIR refers to applicants with no income. Most instances in this subgroup are rejected; only 1 out of 793,992 is a successful application granted in 2016. This suggests that income is crucial for the funding decision. We classifier the rest instances into four subgroups shown in table 5. A statistical summary comparing subgroups is provided in Table A.2.

Table 3-5: Debt-to-Income Subgroups

Subgroup name	Condition	Description
DIR 0	$DIR = 0$	Instances have no debt
DIR 50	$0 < DIR \leq 50$	Instances with relatively healthy DIR
DIR 100	$50 < DIR \leq 100$	Instances with high DIR but debt can be covered by income
DIR101	$100 < DIR$	Instances with high DIR where debt cannot be covered by income

Note: The table shows how subgroups are set up according to the variable Debt-to-Income Ratio.

[8] The Lending Club advises healthy DIR should be lower than 43 percent.

We assume that the risk of the loan climbs with the increase of DIR. Among the four subgroups, DIR 0 and DIR 50 are composed of instances with healthier DIR. Instances with heavier debts are assigned to DIR 100 and DIR 101. Table 3-6 presents the corresponding model performance in each subgroup.

Table 3-6: Out-of-sample predictions for Debt-to-Income ratio

		AUC	Precision	Recall	F1
DIR 0	LR	47.93%	98.57%	90.52%	93.26%
	NB	**60.15%***	97.86%	80.63%	85.77%
	SVM	52.79%*	98.57%	**98.26%**	**98.36%**
	RF	54.42%*	96.84%	42.33%	50.35%
	XGBoost	54.84%*	**99.23%**	97.68%	97.70%
DIR 50	LR	68.90%	67.76%	87.86%	60.76%
	NB	61.39%*	55.04%	73.51%	56.98%
	SVM	**73.94%***	**73.00%**	**93.92%**	**77.75%**
	RF	49.44%*	53.17%	45.87%	46.95%
	XGBoost	62.18%*	70.77%	67.44%	61.22%
DIR 100	LR	55.13%	**98.12%**	98.37%	98.23%
	NB	55.04%	97.49%	81.93%	87.77%
	SVM	52.20%*	98.02%	**98.86%**	98.41%
	RF	59.08%*	97.45%	89.49%	92.63%
	XGBoost	**70.29%***	98.05%	99.50%	**98.75%**
DIR 101	LR	50.27%	**98.20%**	80.55%	80.15%
	NB	39.24%*	96.79%	33.91%	51.27%
	SVM	50.00%	**98.20%**	**100.00%**	**99.09%**
	RF	**54.47%***	80.43%	77.94%	78.41%
	XGBoost	49.75%*	80.20%	80.01%	79.34%

Note: The table shows the performance of the five estimated models in the four Debt-to-Income Ratio subgroups. LR represents Logistic Regression; NB refers to Naïve Bayes; SVM is Support Vector Machine; RF represents Random Forest. * in the AUC denotes that the difference against the benchmark LR model is significant at 99% confidence interval in the Delong test. The best values for each metric are in bold.

We note that the performance of all models is robust in this subsample, as SVM still shows superiority over the rest four models in general. However, SVM is occasionally and slightly outperformed. It achieves over 95% accuracy for precision, recall, and F score in most subgroups but shows fewer advantages in AUC, suggesting its performance relies on thresholds. Only specific thresholds lead to high true positive rates and low false positive rates of SVM. It is interesting to note the difference in the performance of NB across the subgroups and the whole sample, which can be interpreted as an indication of the sensitivity of the model to different datasets.

Though the textual feature Loan Purpose may offer extra information on the loan and the applicant, its role in funding decision-making is barely investigated. We aim to fill the gap in this section. The original sample is grouped according to their loan purpose and the statistics summary of each subgroup are compared in Table 3-13. Noted from Table 3-3, investors are least interested in loans with a business purpose since there is no accepted business loan in our sample. We then exclude the business purpose and examine the remaining subgroups by applying the proposed models. The corresponding performance evaluation is provided in Table 3-7.

Table 3-7: Out-of-sample predictions for Loan Purpose

		LR	NB	SVM	RF	Xgboost
Medical purpose	AUC	62.21%	62.63%	**71.81%***	58.12%*	64.72%*
	Precision	87.25%	84.65%	**91.67%**	84.57%	87.33%
	Recall	63.47%	61.48%	**95.68%**	39.61%	67.18%
	F Score	70.67%	66.19%	**93.49%**	51.62%	71.48%
Other	AUC	57.15%	49.30%*	**64.76%***	63.43%*	63.28%*
	Precision	63.05%	48.63%	66.78%	**74.06%**	69.55%
	Recall	**78.27%**	24.98%	66.10%	46.38%	52.57%
	F Score	**58.57%**	26.65%	51.64%	42.66%	46.07%
Credit card	AUC	50.00%	67.50%*	**74.14%***	65.43%*	66.87%*
	Precision	51.23%	87.65%	**89.95%**	80.90%	71.83%
	Recall	**100.00%**	41.09%	54.70%	41.03%	57.37%
	F Score	67.75%	55.59%	**68.03%**	54.44%	63.79%
	AUC	62.92%	56.39%*	**65.12%***	56.68%*	52.08%*

Debt consolidation	Precision	61.50%	59.15%	48.39%	**66.41%**	58.75%
	Recall	51.75%	20.73%	**57.54%**	35.72%	29.81%
	F Score	51.28%	23.80%	**52.51%**	38.41%	33.16%
Major purchase	AUC	65.80%	**70.57%***	64.41%*	53.25%*	55.75%*
	Precision	84.03%	**88.43%**	84.77%	79.47%	80.55%
	Recall	47.08%	48.80%	**93.13%**	30.41%	37.65%
	F Score	57.16%	57.36%	**87.52%**	36.80%	41.69%
Car	AUC	**67.67%**	58.61%*	57.71%*	49.97%*	50.11%*
	Precision	78.32%	87.26%	**88.38%**	82.07%	83.56%
	Recall	48.05%	28.48%	**95.80%**	14.90%	21.05%
	F Score	53.85%	36.27%	**91.25%**	23.49%	30.23%
Moving	AUC	50.16%	62.57%*	**71.80%***	48.60%*	59.88%*
	Precision	84.05%	83.37%	84.55%	84.22%	**85.49%**
	Recall	70.44%	44.02%	**99.96%**	17.11%	47.92%
	F Score	76.65%	49.22%	**91.61%**	22.47%	57.15%
Home improvement /buying	AUC	54.21%	55.38%*	**78.61%***	52.05%	58.05%*
	Precision	92.22%	**94.20%**	91.31%	83.14%	93.12%
	Recall	11.58%	13.64%	**84.85%**	13.37%	23.54%
	F Score	20.58%	23.83%	**87.96%**	23.03%	37.58%
Business				-		

Note: The table shows the performance of the five estimated models in the eight loan purpose subgroups. LR represents Logistic Regression; NB refers to Naïve Bayes; SVM is Support Vector Machine; RF represents Random Forest. * in the AUC denotes that the difference against the benchmark LR model is significant at 99% confidence interval in the Delong test. The best values for each metric are in bold.

We note that the ranking of our models is retained with SVM performing best, obtaining robust accuracy in terms of all metrics in general. Although in some cases, SVM is outperformed, it closely follows the best. We also note that the accuracy of five models, especially SVM worsens for the "Other" group across all subgroups. This is a subgroup that contains all the remaining loan purposes, including blank and broken context that we are unable to analyse. Consequently, the model performance may be damaged by the uncertainties.

3.5.2. Feature Importance

Financial crises and the following monetary policies have triggered renewed interest in their effect on bank risk-taking behaviour. However, not much attention is paid to the individual investor risk-taking channel. We aim to shed light on this by generating variable importance from our best predictive model - SVM. Approximate Shapley value is computed for each feature in each fold, and Table 3-8 shows the average importance in percentage[9]. The sum of the loan-level variables' importance and the sum of the macroeconomic-level variables' importance are also reported.

Table 3-8: Variable Importance in SVM

Amount Requested	Debt-To-Income Ratio	Employment Length	Loan Purpose	Loan-level variables
12.76%	14.85%	**19.59%**	15.95%	63.15%

CPI	GDP	FFR	TBR	Taylor	Macroeconomic level variables
8.97%	8.83%	**11.74%**	8.99%	9.33%	47.85%

Note: The table reports the average variable importance (in percentage) for the estimated Support Vector Machine (SVM) in the out-of-sample. The top part summarises loan-level variables and shows the sum of the importance of loan-level variables. The bottom part summarises macroeconomic level variables and shows the sum of the importance of the macroeconomic level variables. The best values for variables on loan and macroeconomic levels are in bold, respectively.

Variable importance stated in the table credits our variable selection. They contribute to the P2P loan lending decision without any variable showing significant low importance. Notably, Employment Length is recognised as the main determinant on average, which is consistent with the prior research. For example, Ip and Lam (2020) undertake a study on funding decisions in the online marketplace with the Lending Club data. They note that though the requested amount, borrowing purpose, debt-to-income ratio, and employment length are all crucial to the funding decision, a borrower's employment length is the main factor in the preference of lenders making decisions. In the study, they also examine the macroeconomic effects by adding a dummy variable of the monetary policy regime. However, we consider a wider range of macroeconomic variables and investigate their roles.

[9] Approximate Shapley values are computed using python SHAP library, available at https://github.com/slundberg/shap

Based on the variable importance produced by Shapley values, it seems P2P lending investors are sensitive to the macroeconomic situation, such as inflation (CPI), when they make funding decisions. Among the five macroeconomic variables understudy, FFR is indicated to be the most powerful macroeconomic variable, showing over 10% more importance in P2P funding decisions. It suggests that FFR can influence investors' behaviour when managing P2P investment portfolios. This result is consistent with the literature where Chu & Deng (2019) focus on P2P loan applications during 2006-2013 to study the risk-taking channel of monetary policy. They compute an intersection term of FFR and loan credit score in the regression of default and suggest the positive relationship between loose monetary policy and individual risk-taking behaviour. Their hypothesis is robust based on a test that replaces FRR with Taylor in the regressions. Similarly, results also note that Taylor representing the monetary policy fluctuations is associated with P2P loan funding decisions, shown in Table 3-8. It presents a slightly greater impact than CPI or GDP does. This figure fits the assumption that individual (also the Lending Club) mostly formulates the risk-taking incentives based on a sense of the observed level of interest rates given the macroeconomic conditions.

Macroeconomic impacts are believed to change the probability of granting a bank loan according to its risk level. Riskier loans get more influence. P2P lending applicants with no income are already revealed to be least likely to get granted in 3.5.1. We would like to examine if the rest subgroups draw a similar pattern. Therefore, the feature importance for Debt-to-Income Ratio subgroups is presented additionally.

Table 3-9: Variable Importance in Debt-to-Income Subgroups

	DIR 0	DIR 50	DIR 100	DIR101
Amount Requested	14.48	12.72	2.68	9.76
Debt-To-Income Ratio	0.00	0.26	12.50	1.35
Employment Length	**21.65**	**23.17**	**25.60**	12.06
Loan Purpose	15.37	14.54	3.87	**17.90**
Loan-level variables	51.50	50.69	44.64	41.07
CPI	10.89	10.43	9.23	10.76
GDP	11.92	9.03	6.55	9.45
FFR	**12.23**	**13.36**	**21.43**	**18.47**
TBR	7.46	8.78	12.50	13.94
Taylor	6.00	7.71	5.65	6.31
Macroeconomic-level variables	48.50	49.31	55.36	58.93

Note: The table compares variable importance (in percentage) in Loan Purpose subgroups. Loan-level variables importance is the sum of importance for all loan-level variables, and

Same as our expectation, the macroeconomic impacts are likely to be correlated with DIR. Notable differences between the subgroups at different risk levels are revealed in Table 3-9. We note that macroeconomic-level variables account for greater values in subgroups of unhealthier DIRs. Their change possibly encourages individual risk-taking behaviour in P2P lending. Specifically, individual risk-taking behaviour towards riskier loan requests is encouraged more, as the macroeconomic variables weigh heavier with the risk level increase. This tendency is also noted by Lian et al. (2019). They analyse over 400 surveys and demonstrate that individuals have a greater "appetite" for taking risks when interest rates are low. In addition, we note that variable importance varies in different subgroups. If the applicant's DIR is still in a healthy position, investors seem to care less about DIR while valuing other loan-level features over it (subgroup DIR 50). In the country, in the case where the applicant's DIR is unhealthy, but their income is still able to cover the debt, DIR becomes more important, ranked the second top loan-level estimator for the lending decision (subgroup DIR 100). The investor evaluates their loan purpose first for applicants whose debt is far more than income. Employment length comes second (subgroup DIR 101).

Feature importance in loan purpose subgroups offers extra insights on its role in funding decision-making. To this end, we output the average feature importance for Loan Purpose in Table 3-10. Loan purpose " Business" is excluded since there are no successful instances with a "Business" purpose. This purpose is believed to be at the bottom of the lending preference.

Table 3-10: Variable importance for Loan Purpose subgroups

	Medical purpose	Other	Credit card	Debt consolidation	Major purchase	Car	Moving	Home improvement/ buying
Amount Requested	10.77	6.55	8.87	19.00	22.28	4.29	16.15	23.11
Debt-To-Income Ratio	0.37	2.90	24.36	7.57	1.90	5.02	0.82	2.10
Employment Length	26.89	19.23	10.55	10.28	14.96	45.71	16.44	18.25

Loan-level variables	38.04	28.68	43.78	36.86	39.15	55.0 2	52.83	43.46
CPI	10.27	15.19	6.87	7.42	6.24	15.2 2	8.54	10.71
GDP	8.51	3.00	6.74	3.52	7.69	5.88	12.75	11.38
FFR	14.34	15.22	14.72	37.06	22.42	9.97	9.75	16.88
TBR	14.24	24.45	27.67	5.73	9.08	6.54	4.65	6.95
Taylor	14.60	13.45	0.21	9.41	15.43	10.8 3	11.47	10.62
Macroecono mic level variables	61.96	71.32	56.22	63.14	60.85	48.4 4	47.17	56.54

Note: The table compares variable importance (in percentage) in Loan Purpose subgroups. Loan-level variables importance is the sum of importance for all loan-level variables, and macroeconomic level variables importance is the sum of importance for macroeconomic level variables.

We note that the requested amount contributes more to the lending decision for applicants whose loan purposes mention purchasing (e.g., major purchase and home improvement/ buying), compared to other loan purposes. Similarly, DIR accounts for a greater share of importance in funding decisions if the applicant requests a loan for credit cards or debt consolidation. The potential reason may be these loan purposes are closely associated with the applicant's debt. "Other" is a subgroup that we are interested in, as applicants in this subgroup provide blank or random words as their loan purpose. This results in insufficient loan-level information and potentially raises their risk level. We notice that macroeconomic variables almost weigh double that of loan-level variables in the funding decision in this subgroup, ranked at the top among all subgroups. Investors fail to understand the objective of the investment and seem to rely on macroeconomic conditions. It suggests that macroeconomic condition changes P2P loan funding decisions and influences more on individual risk-taking behaviour towards loans without a proper purpose.

3.6. Conclusion

Funding decision determinants and macroeconomic effects have gained much attention as it is linked with financial stability. Though previous research mostly relies on survey analysis or simple linear regression, it is believed that a financial institution tends to take more risks when monetary policy eases by turning to risker borrowers. But a gap is left for research on

P2P lending. This research aims to fill the gap by employing a rich dataset from the Lending Club with state-of-the-art machine learning approaches. Over 12.4 million instances extracted from the Lending Club with a time span of 10 years are under study. We construct the dataset by merging the rejected loan requests and successfully funded loans. We obtained four common loan-specific variables, amount requested, debt-to-income ratio, employment length, and loan purpose. Among them, loan purpose is a variable that is overlooked by the literature, as it is textual and incomparable to most statistic models. However, we believe it reveals important information for funding decisions and risk management. Armed with text mining techniques, we enrich our loan-level variables by analysing this textual variable. We convert free-typed words into vectors and then transfer them into a categorical variable by a clustering technique. To test the P2P investors' sensitivity toward macroeconomic conditions, we include four widely studied macroeconomic variables, CPI, GDP, FFR, and TBR. In addition, Taylor residuals are introduced into the study on individual funding behaviour as an indicator of monetary policy. Finally, we collapse the original database into subclusters and compute dense representations with subcluster size as the weights. Based on these, four machine learning prediction models (NB, SVM, RF, and Xgboost) are trained against the widely employed regression model – LR. We use AUC along with the Delong test, Precision, Recall, and F Score to evaluate the proposed models. Among them, SVM provides superiority in terms of all the metrics. Xgboost performs the second-best with significant superiority over the benchmark and is followed by RF which only slightly outperforms LR. By contrast, NB fails to provide a reliable analysis with given variables. This result is robust based on a robustness test with subgroups of DIR and Loan Purpose.

Considering the approximate Shapley value from SVM as the variable importance, we note that both loan-specific features and macroeconomic condition understudy affect individual lending decisions and risk-taking behaviour. Generally, a borrower's employment length is the main factor in the preference of lenders making decisions with the greatest feature importance in the study. But loan applicants with no income but debt and applicants who claim business-associated loan purposes are highly likely to be rejected. In terms of the five macroeconomic variables, the Federal Funds Rate has the most powerful impact. Additionally, the feature importance of Taylor residuals insists on the effects of monetary policy on P2P loan funding decisions. The results are in line with two hypotheses in the literature. First, monetary policy affects individual risk-taking behaviour. Second, individual mostly formulates the risk-taking incentives based on a sense of the observed level of interest rates given the macroeconomic conditions. To further examine heterogeneity at the loan application level towards macroeconomic factors, we also break the original database into sub-groups according to debt-to-income ratio and loan purpose, respectively. We note apparent macroeconomic effect variations between loan requests from applicants of healthy DIR and applicants of high DIR. We also note that loan purpose plays a vital role in loan application. Applications with purposes that do not state loan purposes properly are sensitive to macroeconomic variables.

Appendix

3.A.1 Macroeconomic-level Variables and the Loan Requests

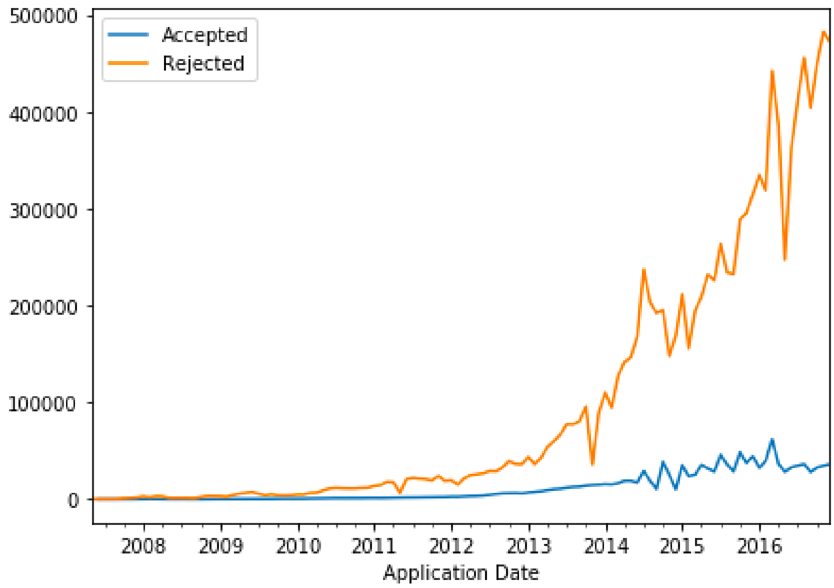
Figure 3-1 compares the quarterly average growth of the five macroeconomic variables from April 2007 to December 2016. Figures 3-2 and 3-3 compare the quarterly-average number of loan requests and the quarterly average amount requested, respectively. In general, the number of loan requests and the amount requested have experienced an increase since the Lending Club was established. This trend reveals the growth of the Lending Club and the P2P lending industry.

Five macroeconomic features fluctuated dramatically during the 2008 financial crisis, while a similar drop occurred in the loan requested amount in Figure 3-3. Besides, during 2014 and 2015, the growth of CPI became negative along with sharp decreases in the number of loan requests and amount requested. This may imply that the pressure of the macroeconomic condition also affects the lending standards.

Figure 3-1: Quarterly average growth of CPI, GDP, FFR, TBR, and Taylor (by %)

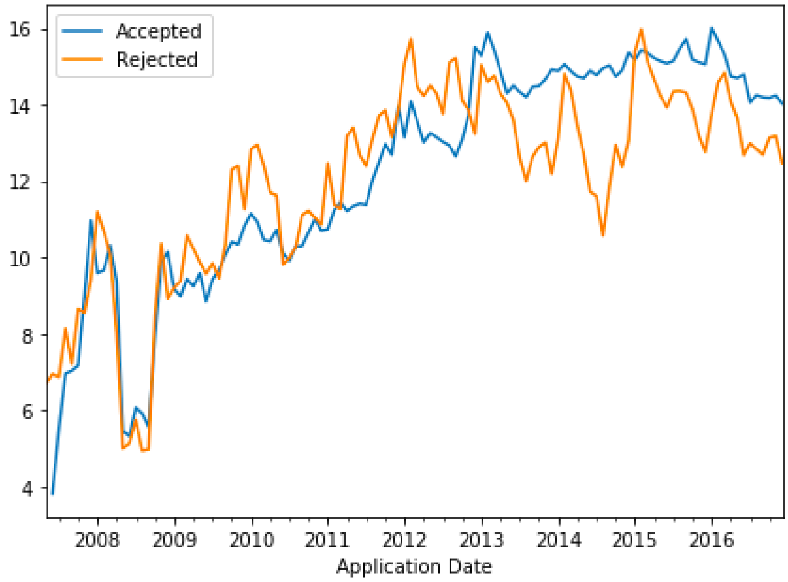
Note: The figure compares quarterly average growth by the percentage of CPI, GDP, FFR, TBR, and Taylor

Figure 3-2: The quarterly average number of loan requests



Note: The figure compares the changes in the number of loan requests from the Lending Club. The changes are presented on a quarterly average.

Figure 3-3: Quarterly average amount requested of loan requests (by million$)



θ

Note: The figure compares the changes in the amount of rejected and accepted loan requests from the Lending Club. The changes are presented on a quarterly average.

3.A.2. Performance Evaluations

In this study, we use AUC, precision, recall, and F score to evaluate the predictive performance of the proposed models. They are defined by the following formulas.

$$\widehat{\theta^{\iota}} = \frac{1}{\mathrm{PQ}} \sum_{p=1}^{P} \sum_{q=1}^{Q} \Psi(Y_p^{\iota}, Y_q^{\iota}) \tag{8}$$

$\widehat{\theta^{\iota}}$ is the empirical AUC of ιth classifier where $\iota \in (1,2,3,4,5)$ in our case. P and Q are the number of rejected instances and accepted instances. Y_p^{ι} and Y_q^{ι} denote the estimated funding results from classifier ι are rejected and accepted, respectively. Ψ is a kernel function. The value of AUC ranges from 0 (no discrimination ability) to 1 (perfect discrimination power), regardless of the threshold. It plots the true positive rate against the false positive rate with varied discrimination thresholds in a binary classifier. DeLong test is a nonparametric test for comparing the AUC of two or more classifiers. Let $\hat{\theta} = (\widehat{\theta^1}, \dots, \widehat{\theta^{\iota}})$ be the vector of ι empirical AUCs, $\theta = (\theta^1, \dots, \theta^{\iota})$ be the vector for true AUCs, $\lambda \in \mathbb{R}^{\iota}$ be a fixed vector of coefficients, S_{10} and S_{01} are $\iota \times \iota$ matrix defined byY_p^{ι} and Y_q^{ι}. To test the equality of two AUCs, we set $\lambda = (1-1)^T$. The null hypothesis of the Delong test is

$$H_0: \theta^1 = \theta^2 \qquad i.c.\ \lambda^T\theta = 0$$

Under the hypothesis, the asymptotic distribution of the quantity is the standard normal distribution.

$$\frac{\lambda^T\hat{\theta} - \lambda^T\theta}{\sqrt{\lambda^T(\frac{1}{P}S_{10} + \frac{1}{Q}S_{01})\lambda}} = \frac{\widehat{\theta^1} - \widehat{\theta^2}}{\sqrt{\lambda^T(\frac{1}{P}S_{10} + \frac{1}{Q}S_{01})\lambda}} \tag{9}$$

However, AUC may be biased in imbalanced datasets. Thus, we add precision, recall, and F1 score to screen the accuracy in each class.

$$Precision = \frac{True\ rejected * w_{TP}}{True\ rejected * w_{TP} + False\ rejected * w_{FP}} \tag{10}$$

$$Recall = \frac{True\ rejected * w_{TP}}{True\ rejected * w_{TP} + False\ accepte * w_{FN}} \tag{11}$$

$$F1\ score = 2 \times \frac{precision \times recall}{precision + recall} \tag{12}$$

Precision and recall measure Type I and Type II error of the prediction, and F score is employed to balance them.

3.A.3. In-sample Predictions

Table 3-11: In-sample predictions

	AUC	Precision	Recall	F Score
LR	73.71%	94.50%	88.13%	90.28%
NB	88.37%*	98.12%	90.27%	92.81%
SVM	95.22%*	97.03%	**96.87%**	**96.92%**
RF	**95.43%***	**99.58%**	90.93%	94.05%
Xgboost	**95.43%***	**99.58%**	90.93%	94.05%

Note: The table reports the accuracy ratios for the estimated models in the out-of-sample. LR represents Logistic Regression; NB refers to Naïve Bayes; SVM is Support Vector Machine; RF represents Random Forest. * in the AUC denotes that the difference against the benchmark LR model is significant at 99% confidence interval in the Delong test. The best values for each metric are in bold.

Table 3- 11 summarises the in-sample prediction performance of the five proposed models in terms of AUC, precision, recall, and F score. All five models perform well, with over 88% accuracy in most cases. SVM is slightly over-performed by RF and Xgboost regarding AUC and precision. But it is ranked best in recall and F score, followed by RF and Xgboost. We note RF and Xgboost output the same results in in-sample prediction, which may be explained by their same fundamental structure - CART. NB beats the benchmark LR in in-sample prediction but performs poorly out of sample. This suggests the low generalisation ability of NB.

3.A.4. Statistic Summary for Subgroups

Table 3-12: Statistic summary for Debt-to-Income subgroups

		DIR = -1	DIR 0	DIR 50	DIR 100	DIR 101
Percentage		6.40%	3.65%	80.53%	6.75%	2.66%
Amount Requested	mean	25.24	7.59	12.75	15.26	13.09
	std	45.2	9.25	10.29	11.28	11.55
	min	0.00	0.50	0.00	0.50	0.50
	Q3	25.00	10.00	20.00	24.00	20.00
	max	600.00	1200.00	1400.00	128.00	100.00
Debt-To-Income Ratio	mean	-1	0.00	19.70	65.41	5781.93
	std	0	0.00	12.62	13.88	95554.53
	min	-1	0.00	0.01	50.01	100.01
	Q3	-1	0.00	28.60	72.75	742.17
	max	-1	0.00	50.00	100.00	50000030.00
Employment Length	Most frequent category	< 1 year	< 1 year	< 1 year	< 1 year	< 1 year

	Percentage of the most frequent category	83.61%	82.34%	69.00%	87.88%	93.15%
Loan Purpose	Most frequent category	Debt consolidation	Other	Debt consolidation	Debt consolidation	Debt consolidation
	Percentage of the most frequent category	25.23%	36.04%	50.00%	56.09%	46.13%

Note: The table presents statistics for variables amount requested (shown in a thousand$), debt-to-income ratio (presented in%), employment length, and loan purpose in subgroups based on the debt-to-income ratio. "DIR =-1" refers to applicants with no income but debt. Q3 stands for the third quartile, the middle value between the median and the highest value of the data set.

The table reports the statistics summary of the five debt-to-income ratio subgroups. We note that the subgroup for applicants who have debts without income shows the greatest mean in terms of the amount requested, indicating that applicants in this group request more money on average. By contrast, applicants with no debt but income (subgroup DIR 0) request the least amount of funds on average. This is the only subgroup where most loan purposes are "other".

Table 3-13: Statistic summary for Loan Purpose Subgroups

		Medical purpose	Other	Credit card	Debt consolidation	Major purchase	Car	Moving	Home improvement /buying	Business
Amount Requested	mean	6.41	8.51	14.89	14.29	10.54	11.91	5.17	14.16	59.76
	std	7.28	9.50	9.96	10.32	10.18	10.73	6.27	11.44	69.29
	min	0.50	0.00	0.50	0.50	1.00	0.50	0.50	0.50	1.00
	Q3	8.00	10.00	20.00	20.00	15.00	18.00	5.00	20.00	75.00
	max	40.00	500.00	150.00	1200.00	50.00	1400.00	50.00	80.00	600.00
Debt-To-Income Ratio	mean	136.51	363.18	164.68	121.93	154.98	205.37	197.84	118.75	7.63
	std	5936.53	35749.37	7620.25	5819.42	4246.51	7374.61	5163.29	2656.63	647.75
	min	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00
	Q3	26.56	26.21	36.68	34.97	22.38	22.33	21.75	27.71	-1.00
	max	1924800.00	50000031.49	4408800.00	7374826.00	1690800.00	3967200.00	900000.00	813112.00	199998.00
Employment Length	mean	1.91	2.32	3.05	2.35	2.16	2.09	1.83	2.54	1.58
	std	2.34	2.66	3.28	2.92	2.52	2.38	2.07	3.09	2.08
	min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Q3	1.00	2.00	6.00	1.00	1.00	1.00	1.00	2.00	1.00
	max	11.00	11.00	11.00	11.00	11.00	11.00	11.00	11.00	11.00

Note: The table presents statistics for variables amount requested (presented in a thousand$), debt-income ratio (presented in%), and employment length in subgroups based on loan purpose in 9 columns. Q3 stands for the third quartile, the middle value between the median and the highest value of the data set.

Table 3-13 reports statistics across the loan purpose subgroups. Subgroup “Business” is the group where no accepted instances are included. Compared to other subgroups, its statistics do not suggest any significant differences that lead to loan application failure. However, we note that most instances with “business” purposes show no income but debt, indicated by Q3 for the debt-to-income ratio.

Chapter 4

Chapter 4: ESG Effects on the Cost of Equity

4.1. Introduction

It has become a new challenge to explore a sustainable development path that meets the need of the current generation without compromising the benefits of the future generation. Negotiators from various countries closed deals setting rules for sustainable development at the United Nations COP26 Climate Summit in Glasgow in November 2021. Along with the increased attention to the transition to a sustainable development model, regulators have started implementing corporate social responsibility requirements. For instance, the United Nations 2030 Agenda for Sustainable Development launched in 2015 claimed that well-being must include certain factors like fairness within and across generations. The guidelines of the European Central Bank also suggest companies may wish to develop value and adhere to pertinent ideals like environmental conservation. On top of it, the last decade has seen the development of burgeoning literature on financial-economic sustainability disclosure. This interest creates opportunities and challenges for firms and fuels their participation in socially responsible projects. Corporate actions towards sustainability and social welfare are often termed Environmental, Social, and Governance (ESG) or Corporate Social Responsibility (CRS). The Governance & Accountability Institute indicates that over 90% of S&P500 index companies published their sustainability report in 2019, compared to only around 20% in 2011[10]. Investments also soar in ESG, which is illustrated by the fact that global assets in ESG Funds hit a new record, nearing 2 trillion dollars, boosted by record inflows in 2021[11].

To cope with the high attention given to ESG, researchers conduct studies on the impact of ESG on firms' operation and performance. After reviewing a significant amount of research in this vein, Gillan et al. (2021) summarise the five most vital and debatable questions: Firms' ESG attributes and the corresponding market attributes; the interactive relationship between ESG practice and firms' executives; ESG and firms' ownership; ESG effects on firms' value; ESG effects on firms' risks. The cost of equity, a proxy of investors' evaluation of firms' risks, lies in the middle of the stream. However, this line remains developing. For instance, Chen & Zhang (2021) and Wang et al. (2021) investigate samples from Asian countries, where the former research includes only Chinese samples, and the latter uses broad East-Asian data. Based on the same regression techniques, they find that Chinese samples exhibit a negative

[10] https://www.morningstar.com/articles/961765/sustainable-fund-flows-in-2019-smash-previous-records

[11] https://www.fundssociety.com/en/news/markets/global-assets-in-esg-funds-neared-2-trillion-dollars-boosted-by-record-inflows

correlation between the two parties, while East-Asian observations exhibit a positive relationship. Dhaliwal et al. (2014) and Breuer et al. (2018) believe that the correlation between CSR disclosure and the cost of equity varies among countries, and a threshold may exist. Compared to a vast literature on the link between financial information and the cost of equity, the empirical work on non-financial information ESG and the cost of equity is scarce. Our interest is also motivated by the following considerations: First, the cost of equity represents the required return rate based on risk evaluation from the market participants. If ESG negatively affects the cost of equity, as we hypothesise, the firms should benefit from operating socially responsibly. They can distinguish themselves from firms with lower ESG scores and enjoy cheaper financing. Second, information asymmetry is prevalent in the finance industry. Studying how ESG performance and disclosure affect the cost of equity also contributes to the broad literature on information asymmetry. Third, the cost of equity implies the investors' expectation of the firm, which is crucial for firms' long-term plans. Understanding the link between ESG and the cost of equity helps firm operators make future operational and reputational strategies.

This chapter proposes that the relationship between ESG and equity financing cost is not necessarily linear; instead, it is curvilinear. We perform our investigation in a worldwide setting by collecting data from various databases for 3,055 unique firms from 51 different countries over 18 years. The state-of-the-art machine learning approach XGBoost is integrated with an up-to-date machine learning framework, double machine learning (DML), to examine the causal effects of ESG performance on the cost of equity. We document the negative effects of ESG dimensions on the cost of equity and confirm the nonlinearity. This result is consistent with the literature. Our results are robust to a battery of sensitivity tests, including using alternative ESG components and subgroups. For example, Chava (2014) and Ng and Rezaee (2015) examine firms’ social responsibility and note that investors expect higher returns on firms with environmental issues than on firms without such concerns. However, we extend the literature by breaking down this analysis by regions where the firms operate. We note significant gaps between emerging and advanced countries - the relationship between ESG and the equity cost is stronger but less heterogeneous in developed countries. Good ESG performance is preferable in advanced regions in terms of the financing cost. In addition, we test the sensitivity of ESG effects towards external shocks such as the global financial crisis and the COVID-19 pandemic in two subgroups and note that the magic power of ESG factors on the equity cost is weaker during these external shocks.

Our study contributes to the literature in several ways. First, we use an up-to-date DML approach rather than the widely-employed regression analysis (El Goul et al., 2011; Dhaliwal et al., 2014; Breuer et al., 2018). Unlike regression models suffering from regularisation bias and simple linear hypotheses, the DML approach detects subtle effects while tolerating the interaction between variables. In addition to the principal ESG dimension, ESG Combined Score, we include a rich selection of ESG variables. DML technique allows us to consider the three pillar scores of ESG simultaneously, namely Environmental Score, Social Score, and Government Score. Second, previous studies rely on single-country datasets that cannot offer crisp comparisons across counties (Breuer et al., 2018; Gupta et al., 2018a; Ke, 2021). In contrast, we use a rich, multi-country dataset that allows us to examine the heterogeneous response of ESG to the cost of equity. In doing so, we document the curvilinearity of the ESG effects on the equity financing cost and further note the gaps between advanced and emerging regions. We also examine the variations of the ESG effects on the cost of equity when external shocks appear, such as the most significant financial crisis and the most up-to-date pandemic COVID-19. These reveal the heterogeneity of the ESG causal effects under different settings. To the best of our knowledge, this channel is yet to be documented.

The rest of the study proceeds as follows: Section 4.2 provides a background history of ESG and the cost of equity. Section 4.3 illustrate the methodology. Section 4.4 specifies the variables and describes the data set. In section 4.5, we present the empirical results. We conclude in Section 4.6.

4.2. Literature Review

4.2.1. ESG/CSR Attributes and Market Characteristics

ESG (and CSR) is believed to contribute to many financial characteristics, such as firm value and risks. ESG disclose increases the firm value but can also decrease the value, according to the study by Fatemi et al. (2018). With 1640 US firm-year observations from 2006 to 2011, they utilise a two-stage least squares model and note that ESG strength improves firm value while ESG weakness damages firm value. Duque-Grisales and Aguilera-Caracuel (2019) widen the research by examining firms worldwide. Their sample is collected from 104 developing nations. It leads to an opposite conclusion where the correlation between ESG scores and financial performance is significantly and statistically negative. In the same vein, Albuquerque et al. (2019) propose an industry equilibrium model that allows firms to choose CRS or non-CRS activities and embeds the choice within a standard asset-pricing framework. They test the model on 4,670 distinct US firms from 2003 to 2015 and suggest CRS increase product differentiation as an investment. This differentiation leads to higher profit margins. Thus, CRS affects firm value positively, and the effects are more substantial for firms with high product differentiation. Besides, systematic risk is statistically and economically significantly lower for firms with higher CRS scores.

Many other researchers also document a negative relationship between ESG and systematic. For example, Oikonomou et al. (2012) employ around 7,000 firm-year observations from 1991 to 2008 and utilise reduced-form models to zoom in on the correlation between CSP variables and financial risk. A positive and strong CRS impact is noted. On the contrary, they find that corporate social irresponsibility negatively but weakly affects systematic risk overall. Specifically, most of the individual CRS strengths' impacts are negative but insignificant, while most individual CRS concerns' effects are positive and significant. It implies asymmetry. Farah et al. (2021) point out the nonlinearity. With the help of 26,621 firm-year observations from 43 countries and a quadratic model for panel regression, their results show that the initial systematic risk gets greater with an increase in CSR but reduces after CSR reaches a threshold. In addition to systematic risk, ESG/CSR and firms' idiosyncratic risks are also examined. Humphrey et al. (2012) believe that CSR may not impact unsystematic risk, as there is no significant difference in the risk-adjusted performance of portfolios with or without high CSR. 256 UK firms are collected from 2002 to 2010 and are tested by regression models. No evidence to show the role of CRS in this study. Becchetti et al. (2015) introduce a bigger-scaled dataset to investigate the same topic. With regressions, they insist that idiosyncratic volatility is positively associated with aggregate CSR but is negatively associated with an individual CSR component, stakeholder risk factor.

Besides systematic risk, studies on credit risk indicate that ESG/CSR also plays an important role. Dumitrescu et al. (2020) employ a rich US dataset from 1991 to 2015 and use reduced-form models to test the effects. Their evidence shows that ESG/CSR is valuable for the stock and debt markets because it can predict financial distress. Goss & Roberts (2011) study 4,586 North American firms from 1991 to 2006 using similar approaches. By analysing CSR and credit risk measured by bond yields, they note that firms with the worst CSR performance pay up to 20 basis points more than firms with the highest CSR scores. This CSR impact is not economically significant, they add. Having said so, Mahmoud et al. (2017) use monthly data collected from 2005 to 2013 but find no apparent correlation between CSR and bond yields. However, they find socially responsible firms raised debt with lower spreads and longer maturities during the financial crisis in 2008. These firms tend to have better credit ratings during the financial crisis. Stellner et al. (2015) study 872 bonds issued by non-financial European firms. Zero-volatility spreads and credit ratings measure credit risk. Their results support the argument that superior CSR performance leads to lower credit risks. It is also argued by Devalle et al. (2017) that ESG/CSR should be taken into consideration for creditworthiness evaluation. Because examing 56 Italian and Spanish firms, they find that ESG performance is positively correlated with credit ratings, especially two individual ESG components termed Community Score and Shareholder Score. Correspondingly, Seltzer et al. (2020) claim that firms performing poorly in ESG/CSR environmental components such as carbon footprints are likely to have lower credit ratings and higher bond yield spreads. However, the country's ESG performance determines the relationship between credit risk and firms' ESG/CSR performance. Meanwhile, Jiraporn et al. (2014) argue that CSR policies in different regions should be a control viable in the study. They use an identification strategy according to the geographic similarity in CSR policies and examine over 2,000 observations. The same result shows that socially responsible firms gain better credit ratings. In addition to credit risks, Lin & Dong (2018) document the effects of ESG/CSR on bankruptcy probabilities during financial distress. They analyse 4,163 US firm-level observations from 2000 to 2014 and consider firms' previous (two years ago) CSR performance. It is noted that firms that do well in CSR tend to recover from financial distress instead of facing bankruptcy.

4.2.2. ESG/CRS and Stock Performance

Reflecting upon the importance and popularity of ESG factors, publicly traded firms are increasingly assessed on corporate social responsibility. A great volume of research has emerged to examine the implications of such ratings for investing and stock performance. Kerkemeier and Kruse-Becher (2022) study the convergence behaviour of global ESG indices. They employ 18 ESG stock market indices with convergence tests and a clustering procedure that is based on a time-varying nonlinear panel factor model. The results imply that currently, there may be less diversification between ESG indices than in May 2019, when this index category emerged. Shanaev and Ghimire (2022) investigate the correlation between ESG rating changes and stock returns. Rather than focusing on the ESG rating levels, they extract ESG rating changes for a representative sample of all US-traded firms available on MSCI from 2016 to 2021. Armed with a calendar-time portfolio methodology, they tease out the unsymmetric effects on the stock performance. Though ESG rating upgrades do not lead to significant positive abnormal returns, downgrades are significantly related to negative abnormal returns. A potential explanation is that ESG is a component of corporate reputation, and investors tend to be more sensitive toward a negative reputation. To study the integrated effects of ESG and corporate reputation on stock prices, Wong and Zhang (2022) collect stock prices and ESG ratings for the US publicly traded companies from 2007 to 2018 and analyse them with regressions. They prove that ESG disclosure significantly impacts firm valuation via media channels. But when they investigate firms according to the industry classification, it is revealed that the stock performance of companies in the ‘sin’ triumvirate (i.e., alcohol, tobacco, and gaming that tend to conflict with the nature of social responsibilities) is not significantly influenced by negative ESG media coverage. On the other hand, Hong and Kacperczyk (2009) only focus on the sin of the US. They find that sin stocks are less invested by norm-constrained institutions such as pension plans. More significant litigation risks heightened by social norms resulted in higher expected returns than other comparable stocks. “Sin” is mainly associated with the social factor under the ESG factors, and environment and governance are the rest two factors. However, using both survey and trading data from 9,286 retail investors for the 2005–2011 period, D’Hondt et al. (2022) claim that the three ESG factors are heterogeneous and should be considered separately. Similarly, Luo (2022) undertake a study on ESG, liquidity, and stock returns based on UK securities from 2003 to 2020. In addition to the ESG effects in general, they unpack it and demonstrate that environmental and social factors have greater power on stock performance while governance premium is insignificant.

4.2.3. Cost of Equity and ESG/CSR

Cost of equity refers to the rate of return required by the shareholder for their equity investment. It is a direct measure of external equity financing costs, and as such, it is closely related to financial risks and affects investment and financing decisions. Implied cost of equity is believed to perform better in predicting future market returns than traditional estimators, including book-to-market ratio and payout yield (Rjiba et al., 2021).

In addition to financial ratios, non-financial factors can contribute to the cost of equity. This argument has been studied in the literature. External influence is one factor. Boubakri et al.

(2012) study the role of government and politicians in the cost of equity based on 1,248 observations from 35 countries. With regressions, they find that firms gain advantages in the cost of equity by being politically connected. In the same vein, Li et al. (2018) note that the cost of equity increases with political uncertainty, especially if the firm's CEO is politically connected or during a bear market. Gupta^{a} et al. (2018) use 7,380 observations from 22 developed countries to test governance attributes' effects on the cost of equity. Using the same method, they show these effects are more potent in Common Law countries with high levels of financial development. In the meanwhile, pandemic shock may act as an external influence. Based on update-to-date US firm monthly data, Ke (2021) regress the relationship between the COVID-19 pandemic and the cost of equity. They note a climb in the cost of equity during the pandemic. Information availability can be another main factor. Having said so, Saci and Jasimuddin (2021) undertake a study on institutional investor research and the cost of equity. 4,928 samples are collected from Shenzhen Stock Exchange from 2013 to 2017, and multivariate regression analysis is utilised. Their evidence shows that the cost of equity is lowered when there is a more significant proportion of field research in the total investment activity. Rjiba et al. (2021) investigate the correlation between firm annual report readability and the cost of equity. They study a large dataset consisting of around 40 thousand US firm observations with regression analysis. If the annual report is textually hard to interpret, the information risk tends to be higher as the investor undertakes uncertainty. This results in a higher risk premium and cost of equity. Information on social media such as Twitter may also reduce the cost of equity. Al Guindy (2021) point this out by analysing all the firms listed on NYSE, AMEX, and NASDAQ since 2006, when Twitter first appeared. Using a difference-in-difference analysis, they further note that some firms profit more from posting financial information on Twitter. These firms include firms that suffer from the most significant information asymmetries, firms with minor institutional holdings or analyst followings, and small firms that benefit most. He et al. (2013) reach the same conclusion by testing the Australian Securities Exchange firms. The study documents that the cost of equity increases information risk, including uncertainty and asymmetry.

The last two decades witnessed an increasing interest in socially responsible investment. Investors consider the possibility of new regulations towards firms' social responsibility performance and long-term picture of the firm, and screen stocks based on firms' social behaviour such as the amount of CO2 emission and employee rights. A sufficient number of social-responsible-sensitive shareholders or lenders may affect the expected return on firms with socially responsible concerns. Thus, firm leaders tend to face litigation and reputation risks if they increase firms' socially responsible concerns. Information on ESG/CSR serves as an extra contributor to investors' expected risk premium. Motivated by these theoretical arguments, Dhaliwal et al. (2014) undertake a study with worldwide samples. The dataset consists of over 5 thousand observations sourced from 31 counties. This allows them to compare ESG/CSR effects on the cost of equity in different regions. Same as the hypothesis, they find a negative correlation between CSR disclosure and the cost of equity. Notably, the correlation is more potent if the country or firm has a higher level of financial opaqueness. Breuer et al. (2018) manage a dataset sourced from 39 counties worldwide. They investigate the ESG/CSR effects on the cost of equity and point out that a threshold may exist in the level of investor protections. This threshold determines the effects. For instance, they show that the cost of equity rises with increased CSR activities in a country with low-level investor protection but falls with more CSR activities in a country with strong investor protection. A similar pattern is noticed by comparison. Chen & Zhang (2021) and Wang et al. (2021) only study Asian countries. The former research includes over 7 thousand Chinese firm samples, and the latter uses data from East Asia. Based on the same regression techniques, they reach different conclusions. Chinese samples exhibit a negative correlation between the two parties, while observations from East Asia exhibit a positive relationship. CSR activities add the cost of equity in East Asia.

However, different components under ESG/CSR are likely to have various levels of impact on the cost equity. Chava (2014) examines the impact of firms' social responsibility, specifically the environmental profile, on the cost of equity and debt capital. 13, 114 firm-year observations with financial characteristics abstracted from Compustat, CRSP, and I/B/E/S database. Their environmental profiles are sourced from KLD, where the concerns and strengths of the firms are listed. Using simple regressions, they noted that investors expect higher returns on firms with environmental issues such as climate change and substantial emissions than firms without such concerns. Ng and Rezaee (2015) also focus on the environmental profile. They examine a rich dataset and employ market beta and firms' financial ratios as control variables. Their study leads to the argument that the economic sustainability disclosure (ECON) and ESG determine the cost of equity interactively. El Goul et al. (2011) employ a sizeable US-firm sample where the ESG/CSR data is sourced from the KLD database. Multivariate analysis and several other approaches modify the relationship between ESG/CSR and cost equity. Besides the fact that socially responsible firms enjoy cheaper equity financing, they further point out that responsible employee relations, environmental policies, and product strategies are the most effective ESG/CSR components.

4.3. Methodology

4.3.1. Double/Debiased Machine Learning

In this section, we illustrate the Double/debiased machine learning (DML) approach. It is proposed by Belloni et al. (2014) and Chernozhukov et al. (2018) to estimate the role of a low-dimensional parameter of interest in the presence of a high-dimensional nuisance function with many covariates.

Given that N is the number of observations, Y is the cost of equity, D is the vector of ESG variables, X is the firm-year observation containing p features: $X^1,\ X^2, \dots,\ X^p$. We assume the data $(Y_i, D_i, X_i')', i = 1, 2, \dots, N$, are independent and identically distributed. Suppose a partially linear regression (1), the regularisation-basis-free approach directly starts patriating out the effects of D on X in (2) or (3).

$$Y = D\theta_0 + f_0(X) + U \ \ and \ \ D = \ m_0(X) + V \tag{1}$$

$$\ell_0(X) = \mathbb{E}(Y|X) \ = m_0(X)\theta_0 + \ f_0(X) \tag{2}$$

$$Y - \ell_0(X) = \ V\theta_0 \ + U \tag{3}$$

where f_0 and m_0 are unknown nuisance functions, U and V are the errors. θ_0 is the leading regression coefficient for the variable of interest. Following it, we split the sample into k parts. For simplicity, we suppose k = 2 and generate two subsets $\mathcal{I}$ and $\mathcal{I}^c$ of equal size n=N/2. A machine learning approach is used to estimate ℓ_0 and m_0 based on subset $\mathcal{I}^c$. $\tilde{\ell}_0$, the estimator for ℓ_0, is computed from Y_i and X_i, while $\widetilde{m}_0$, the estimator for m_0, is obtained from D_i and X_i. We then formulate the parameter of interest θ_0 with $\mathcal{I}$ as:

$$\tilde{\theta}_0 = \left(\sum_{i\in\mathcal{I}} \tilde{V}_i \tilde{V}_i\right)^{-1} \sum_{i\in\mathcal{I}} \tilde{V}_i\, [\, Y_i - \tilde{\ell}_i(X_i)]), \quad i \in \mathcal{I} \tag{4}$$

where $\tilde{V}_i = D_i - \widetilde{m}_0(\,X_i)$. Suppose that $\widetilde{M} = \ \mathbb{E}\big(\tilde{V}_i^2\big)$, we compute the scaled estimation error between $\tilde{\theta}_0$ and θ_0 to investigate the regularisation basis:

$$\sqrt{n}(\tilde{\theta}_0 - \theta_0) \ = a^* + b^* + c^* + \ o_p(1) \tag{5}$$

where

$$a^* = \ \widetilde{M}^{-1} \frac{1}{\sqrt{n}} (\textstyle\sum_{i\in\mathcal{I}} V_i U_i\,) \tag{6}$$

$$b^* = \ \widetilde{M}^{-1} (\frac{1}{\sqrt{n}} \textstyle\sum_{i\in\mathcal{I}} [m_0(X_i) - \widetilde{m}_0(X_i)][\ell_0(X_i) - \tilde{\ell}_i(X_i)]) \tag{7}$$

$$c^* = \tilde{M}^{-1}(\frac{1}{\sqrt{n}}\sum_{i\in\mathcal{J}}\{V_i[\ell_0(X_i) - \tilde{\ell}_o(X_i)] + U_i[m_0(X_i) - \tilde{m}_0(X_i)]\} \tag{8}$$

$$\tilde{M} = \mathbb{E}(\tilde{V}_i^2) \tag{9}$$

We note a is asymptotically normally distributed, so we focus on b and c. Assume $||\ell_0(X_i) - \tilde{\ell}_o(X_i)||_2 = O\ (n^{-\varphi\ell})$, and $||m_0(X_i) - \tilde{m}_0(X_i)||_2 = O\ (n^{-\varphi m})$. They lead us to

$$b^* = o_p(n^{1/2-(\varphi\ell+\varphi m)}) \tag{10}$$

One can tell from equation (10) that b shrinks to zero if $\varphi\ell + \varphi m > \frac{1}{2}$. It is an achievable requirement to get machine-learning-based estimators $\tilde{\ell}_o$ and $\tilde{m}_0$ with $\frac{1}{4} < \varphi\ell = \varphi m < \frac{1}{2}$. As a nonparametric estimator, c is correlated with U_i and V_i when the same data compute the parameter of interest and the nuisance functions, DML avoids this correlation by using independent subsets. Recall that subset $\mathcal{J}^c$ is employed to learn $\tilde{\ell}_0$ and $\tilde{m}_0$ while $\mathcal{J}$ obtains U_i and V_i. c can shrink under certain regularity conditions. In this manner, the scaled estimation error between $\tilde{\theta}_0$ and θ_0 is also asymptotically normally distributed. We set k =5 in the following study because Chernozhukov et al. (2018) suggest it provides better sample performance.

4.3.2. XGBoost

Within the DML framework, we utilise a machine learning technique named XGBoost to estimate the effects. XGBoost is a gradient tree boosting machine learning algorithm. It incorporates Classification and Regression Trees (CART) by building them one after the other. XGBoost sharpens the performance of weak CART classifiers by continuously superimposing. Suppose we have a dataset $\mathbb{S} = \{Y_i,\ X_i'\}$. X_i is a matrix with N observations and p predictors. Y_i is the outcome variable. A loss function is applied to control the complexity of XGBoost, which aims at minimising the equation:

$$L_S = \sum_{N=1} l(Y_i, \widehat{Y}_i) + \mathcal{K}(\mathscr{k}_q) \tag{11}$$

where $l(Y_i, \widehat{Y}_i)$ measures the difference between the actual and the forecasted value, $\mathcal{K}(\mathscr{k}_q)$ is a regularisation term, and $\mathscr{k}_q$ represents the tree structures. The number of trees and parameters in the regulation term are selected through 5-fold cross-validation in the in-sample.

4.4. Empirical Data

4.4.1. Dataset Construction

To examine the treatment effects of ESG on the cost of equity, we consider various variables by merging several databases. We collect ESG data from ASSET4. Firm-level financial indicators, including consensus forecasts on earnings and dividends, are sourced from Thomson Reuters I/B/E/S and Worldscope. We also employ country-specific data from World Bank Open Data Project. Our choice of variable closely follows the literature.

Our choice of variable closely follows previous literature. The cost of equity is typically measured by both the implied approach and the realised approach. The implied approach calculates the discount rate applied to a firm's expected future cash flows to form its current stock price. The realised approach estimates the cost of equity in terms of ex-post stock returns. However, proxies choosing is a contentious topic. Dhaliwal et al. (2014) suggest "there is no consensus on the 'best' proxy, or even on how to evaluate the merits of the various measures proposed in the literature." However, the ex-post approach has been criticised in the literature. Because realised return tends to be unobservable and noisy (Ding et al., 2015; Gupta^{b} et al., 2018), it suffers from measurement basis such as risk loading (Hasan et al., 2015). Hence, studies are increasingly utilising implied approaches.

Following the most recent literature, we derive the cost of equity as the internal rate of return in four different estimation models. The four models are Gebhardt et al. (2001), Claus and Thomas (2001), Easton (2004) and, Ohlson and Juettner-Nauroth (2005). The first two models are computed based on residual income valuation, while the last two rely on abnormal earning growth. The model description is summarised in the appendix. However, each of these models can be biased, resulting in a misleading conclusion about the ESG effects on the equity cost. To address this concern, we compute the final cost of equity by averaging the four (Boubakri et al., 2012; Dhaliwal et al., 2014; Ke, 2021). The averaging approach is likely to remove a portion of such noise. We reckon this implied cost of equity approach still has drawbacks. Hence, a robustness check is performed in later sections.

ASSET4 started in 2002. Its coverage universe comprises listed companies from well-known indexes, such as S&P 500 and Russell 1000. It evaluates firms' ESG from three perspectives: environment, social, and government. They lead to three principal ESG scores, where the higher, the better. An overall score, termed ESG Combined Scores, is further computed based on the three metrics. ASSET4 also provides ESG score components describing a firm's specific ESG attributes. For example, the Human Rights Score screens the employee rights, and the Emission Score monitors the greenhouse produced by the firm. These scores are also included in the study for the robust test.

To single out the incremental value that ESG adds to the cost of equity, we include a group of variables that determine equity pricing (El Ghoul et al., 2011; Chava, 2014; Al Guindy, 2021; Rjiba et al., 2021). They are fundamental firm-level variables (such as market beta and book-to-market ratio) to estimate firms' systematic and idiosyncratic risk, long-term growth, and profits. We also include analysts' forecast basis and dispersion because the cost of equity is computed based on analysts' forecasts which can be biased. The two variables model the forecasting uncertainty in the cost of equity. Additionally, the 48-industry classification (Fama and French, 1997) is included to modify industry fixed effects, and the year dummies are included to address potential time-series variations in the cost of equity. They also monitor macroeconomic condition changes through time. We control for country-fixed effects by introducing country-level data such as GDP, GDP growth, inflation, and country dummies fixed effects.

4.4.2. Statistical Description

We start with ESG data from ASSET4. It contains ESG scores for 8,716 firms worldwide from 2002, the inception of ASSET4 ESG scores, to 2020. Among these firms, we are able to compute financial indicators and the cost of equity for 3,055 unique firms from 51 different countries over 18 years. They compose 15,229 firm-year observations with 34 independent variables and 15 ESG variables (treatment variables). All the financial data are winsorised at 1% and 99% to handle outliers. We list our variables in the appendix.

The sample in our study offers a rich country coverage compared to the mainstream of the literature. Previous studies tend to focus on datasets with less country diversity (Breuer et al., 2018), such as a single country (Gupta^{a}, et al., 2018; Ke, 2021). The sample in our study covers advanced, emerging, and frontier regions[12], including counties that are rarely examined, such

[12] The country classifying is done according to https://www.msci.com/market-classification.

as Peru. Country composition is presented in the appendix. Statistics are summarised in Table 4-1, where significant gaps across regions are observed. Most observations come from advanced regions, while few firms from frontier regions report ESG scores. Contrary to our expectation, the frontier subgroup with the lowest average cost and standard error among the three presents absolute strength in the cost of equity. It may be caused by the inadequate sample, as only strong and healthy frontier-region firms' data is available. Nevertheless, firms listed in advanced regions are more social-responsible in general, in terms of the relatively high ESG scores. We also note disadvantages in emerging regions compared to the advanced regions, as firms experience high equity, poor ESG performance, and significant heterogeneity within the group.

Table 4-1: Region Statistics

		All	Frontier Regions	Emerging Regions	Advanced Regions
Number of firms		3,055	4	799	2252
Number of observations		15,229	7	3121	12101
Cost of Equity	Mean	0.08	**0.07**	0.10	0.08
	Standard Error	0.59	**0.02**	1.28	0.12
ESG Combined Score	Mean	45.69	42.51	45.16	**45.83**
	Standard Error	19.03	**11.87**	20.08	18.75
Environment Score	Mean	41.25	**45.57**	37.72	42.16
	Standard Error	29.77	**26.58**	28.09	30.12
Social Score	Mean	47.53	36.86	46.78	**47.73**
	Standard Error	24.12	**22.10**	25.25	23.82
Government Score	Mean	52.16	50.87	51.09	**52.43**
	Standard Error	22.23	**16.33**	22.08	22.27

Note: The table presents sample statistics by region. It summarises the mean and the standard error of the cost of equity and ESG performance in each region subgroup. The best value in each column is in bold.

Table 4-2 displays the year composition where growth is shown in both the number and the ESG performance. More firms are willing to disclose ESG information, and firms have acted more socially- responsibly in recent years. This pattern is consistent with the ESG trend in the last two decades.

Table 4-2: Sample Breakdown by Year

Year	N	%	Mean ESG	Mean R_{AVE}
2002	114	1.15	37.91	0.09
2003	168	1.69	37.05	0.09
2004	354	3.57	34.78	0.08
2006	536	5.40	37.56	0.07
2007	629	6.34	40.78	0.06
2008	687	6.93	42.34	0.08

2009	693	6.99	43.68	0.10
2010	969	9.77	44.04	0.09
2011	1160	11.69	43.79	0.09
2012	1077	10.86	44.47	0.09
2013	1055	10.64	45.10	0.08
2014	1151	11.60	45.16	0.07
2015	1230	12.40	47.10	0.07
2016	1172	11.82	47.84	0.08
2017	**1279**	**12.89**	48.54	0.13
2018	1233	12.43	50.18	0.07
2019	1060	10.69	51.60	0.07
2020	662	6.67	51.68	0.09

Note: The table summarises the sample by year. The columns N and % are the number and the percentage of the observations. The last two columns are the mean value for the ESG Combined Score and the mean value for the cost of equity.

4.5. Empirical Results

4.5.1. Nonlinear ESG Effects

We use four ESG scores to evaluate ESG in the main study, namely Environment Score (E), Social Score (S), Government Score (G), and ESG Combined Score. The first three assess the principal ESG aspects, and ESG Combined Score summarises the overall performance. The ESG score components are introduced in the robustness test.

Khalifa et al. (2019) point out the nonlinear effect of conditional conservatism on the cost of equity capital, and Farah et al. (2021) prove the inverted U-shaped effect of ESG on systematic risk. They verify the nonlinearity by adding the quadratic form of the treatment variables. Following them, we study the nonlinearity between the ESG scores and the cost of equity. We also explore the quadratic forms of our variables. Table 4-3 presents the estimated ESG effects on the cost of equity.

Table 4-3: Estimated ESG Effects on the Cost of Equity

Treatment Variables	Dependent variable: Cost of Equity	
	Linear Effects	Quadratic Effects
ESG Combined	-0.0075***	-0.0077***
	(-0.0002)	(-0.0002)
E	-0.0060***	-0.0065***
	(-0.0002)	(-0.0002)
S	-0.0066***	-0.0069***
	(-0.0002)	(-0.0002)
G	-0.0063***	-0.0067***
	(-0.0001)	(-0.0001)

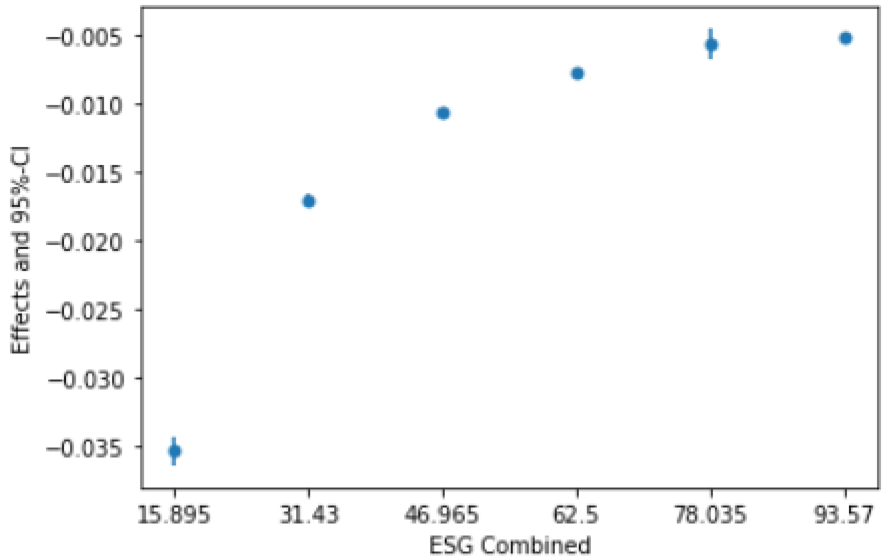
ESG Combined^2	-0.0001*** (-0.0000)
E^2	-0.0001*** (-0.0000)
S^2	-0.0001*** (-0.0000)
G^2	-0.0001*** (-0.0000)

Note: The table presents the average causal effects of ESG on the cost of equity estimated by the Double Machine Learning model with XGBoost. ESG Combined^2, S^2, E^2, and G^2 are the quadratic terms of ESG Combined Score, E Score, S score, and G Score, respectively. The standard errors are reported in brackets. *** denotes significant at the 99% level.

Our model documents the adverse average causal effects of ESG variables on the equity cost. These negative effects are subtle but statistically significant both for the level and the quadratic terms of the ESG variables. Firms that are operating responsibly can benefit from the lower cost of equity. The negative effects we note are in line with the literature. For example, El Goul et al. (2011) use a sizeable US-firm sample to show that responsible employee relations, environmental policies, and product strategies effectively influence the financing cost. Chava (2014) and Ng and Rezaee (2015) examine firms' social responsibility and note that investors expect higher returns on firms with environmental issues than those without such concerns.

This finding suggests that the relationship between firms' social behaviour and the equity capital cost may be curvilinear. To further identify the nonlinearity, we break the sample into six subgroups according to their ESG Combined Score, where most firms' ESG Combined Scores lay in the two middle subgroups. Subgroups presenting the lowest and the highest ESG Combined Score contain the fewest observations. We re-visit XGBoost-baed DML to estimate ESG effects in the subgroups via the quadratic form. Figure 1 plots these estimated effects and the related 95% confidence intervals. The figure strongly indicates the nonlinear relationship, suggesting the margin of ESG effects decreases with firms' social responsibility enhancement. Subgroups with better social-responsible performance experience less ESG effects on the equity cost, while social-irresponsible firms can significantly reduce the cost of equity by improving their ESG scores.

Figure 4-1: ESG effects change with the change of ESG Combined Score



Note: The figure illustrates the trend that ESG effects change along with the change of ESG Combined Score.

In regard to the robustness of our findings, we train the XGBoost-based DML model using ESG individual ESG components. These scores describe firms' social responsibility performance in terms of specific aspects. Table 4-4 presents the results.

Table 4-4: Estimated ESG Components Effects on the Cost of Equity

Treatment Variables	Dependent variable: Cost of Equity	
	Linear Effects	Quadratic Effects
Emissions Score	-0.0052*** (-0.0001)	-0.0057*** (-0.0001)
Environmental Innovation Score	-0.0044*** (-0.0001)	-0.0060*** (-0.0002)
ESG Controversies Score	-0.0041*** (-0.0001)	-0.0043*** (-0.0001)
Human Rights Score	-0.0048*** (-0.0001)	-0.0061*** (-0.0001)
Management Score	-0.0050*** (-0.0001)	-0.0060*** (-0.0001)
Product Responsibility Score	-0.0048*** (-0.0002)	-0.0059*** (-0.0002)
Resource Use Score	-0.0052*** (-0.0002)	-0.0057*** (-0.0002)
Shareholders Score	-0.0051*** (-0.0001)	-0.0061*** (-0.0001)
Workforce Score	-0.0053*** (-0.0001)	-0.0057*** (-0.0001)

CSR Strategy Score	-0.0052*** (-0.0001)	-0.0059*** (-0.0001)
Community Score	-0.0051*** (-0.0002)	-0.0058*** (-0.0002)
Emissions Score ^{2}		-0.0001*** (-0.0000)
Environmental Innovation Score ^{2}		-0.0001*** (-0.0000)
ESG Controversies Score ^{2}		0.0000*** (-0.0000)
Human Rights Score^2		-0.0001*** (-0.0000)
Management Score ^{2}		-0.0001*** (-0.0000)
Product Responsibility Score ^{2}		-0.0001*** (-0.0000)
Resource Use Score ^{2}		-0.0001*** (-0.0000)
Shareholders Score ^{2}		-0.0001*** (-0.0000)
Workforce Score ^{2}		-0.0001*** (-0.0000)
CSR Strategy Score ^{2}		-0.0001*** (-0.0000)
Community Score ^{2}		-0.0001*** (-0.0000)

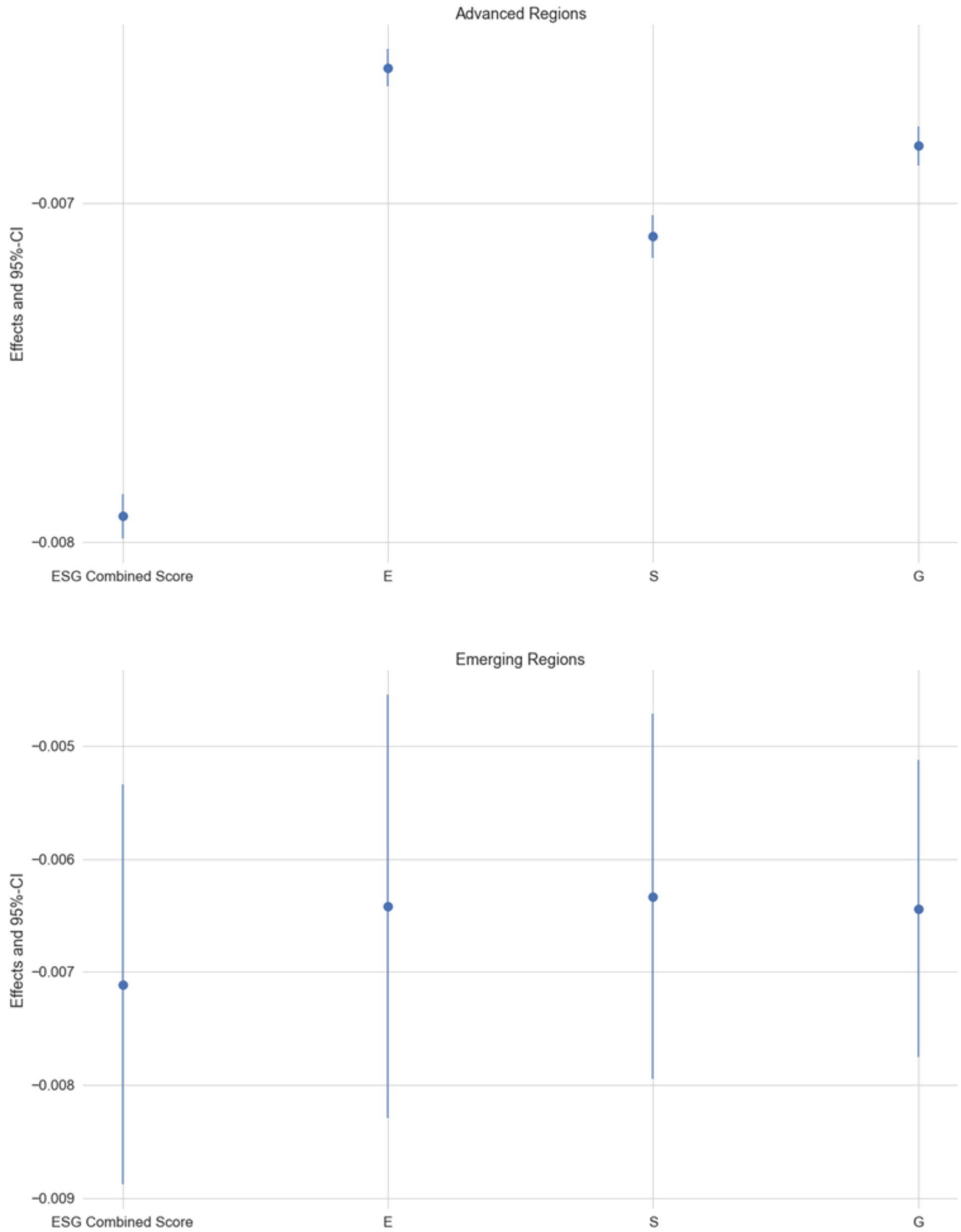
Note: The table presents the average causal effects of ESG components on the cost of equity estimated by the Double Machine Learning model with XGBoost. The standard errors are reported in brackets. *** denotes significant at the 99% level.

Our model notes the adverse causal effects of every single ESG individual component. Both linear and quadratic forms provide statistical-significant estimation, suggesting a nonlinear relationship. The effects are subtle but significant, without any ESG variables acting extremely strongly or weakly. These findings are consistent with our main results.

4.5.2. Heterogeneity across Regions

ESG effects may vary in different countries. The cross-country studies by Dhaliwal et al. (2014) and Breuer et al. (2018) suggest that though disclosure on social issues negatively influences equity cost, this correlation is stronger in some countries. Thus, our interest is to investigate heterogeneity among regions. In Section 4, we classify the firms based on their country of origin into advanced, emerging, and frontier. Due to inadequate samples of frontier countries, we do not consider this group in this section. We recall our DML model to compare the rest two groups. Results are illustrated in Figure 4-2, where the estimated effects of ESG Combined Score, E Score, S score, G Score, and 95% confidence intervals are marked.

Figure 4-2: Average ESG effects in advanced and emerging countries

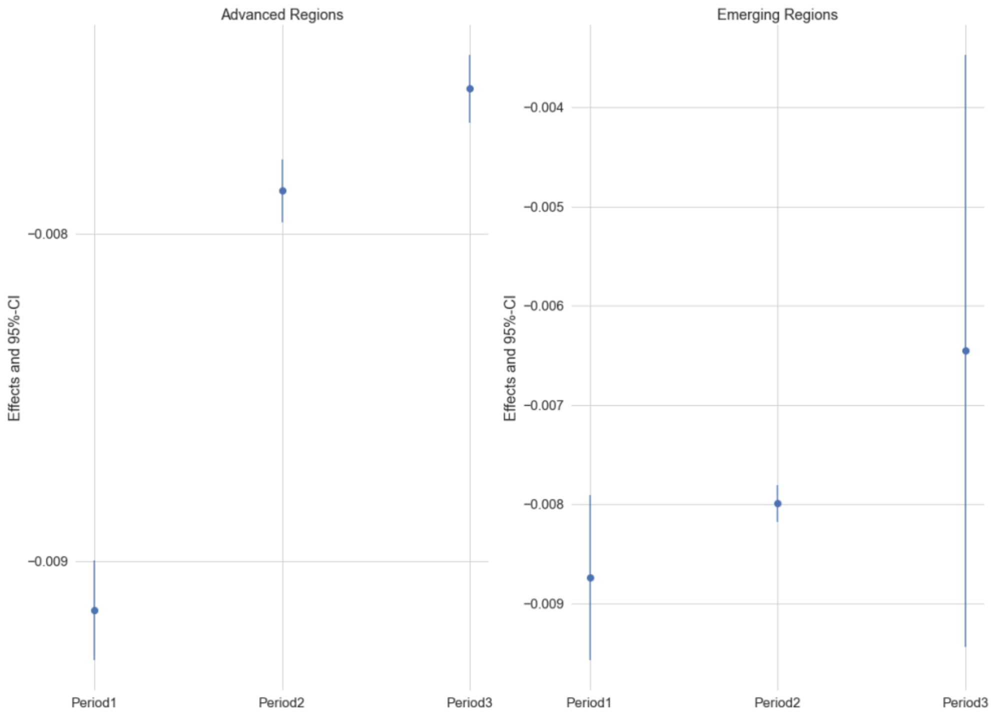


Note: The figure illustrates the average ESG effects in advanced and emerging countries, estimated by XGBoost-based DML with quadratic terms. Effects refer to the estimated effects, and 95% CI stands for 95% confidence intervals.

Consistent with the main results, ESG effects are estimated to be negative in both advanced and emerging regions regarding ESG Combined Score, E Score, S Score, and G Score. Nevertheless, a notable gap can be seen. Firms in advanced countries face a more significant impact on the cost of equity than firms in emerging countries. Investors tend to find ESG activities promising in advanced regions and suppose higher expected returns on socially responsible firms listed on developed stock markets. But in emerging regions, the cost of equity is more likely to be determined by non-ESG factors. Additionally, we notice that the 95% confidence interval is much narrower in advanced countries. It means less heterogeneity among the firms in advanced regions, while the ESG effects are likely to vary from firm to firm in emerging countries.

The last two decades have witnessed ESG investment soar. We learn from Table 4-2 that the number of firms who report ESG scores is increasing over time and that firms tend to have better ESG performance in recent years. This section tests if investors' expectation of ESG investment has a similar pattern. The sample is split into three periods: 2002 to 2008, 2009 to 2014, and 2015 to 2020. We employ our model with quadratic terms in each group. ESG Combined Score results are listed in Figure 3 as a summary of the overall ESG effects. The rest estimations for E, S, G scores are illustrated in Appendix Figure 4-6.

Figure 4-3: ESG Combined Score effects on the cost of equity over time in firms in advanced and emerging countries

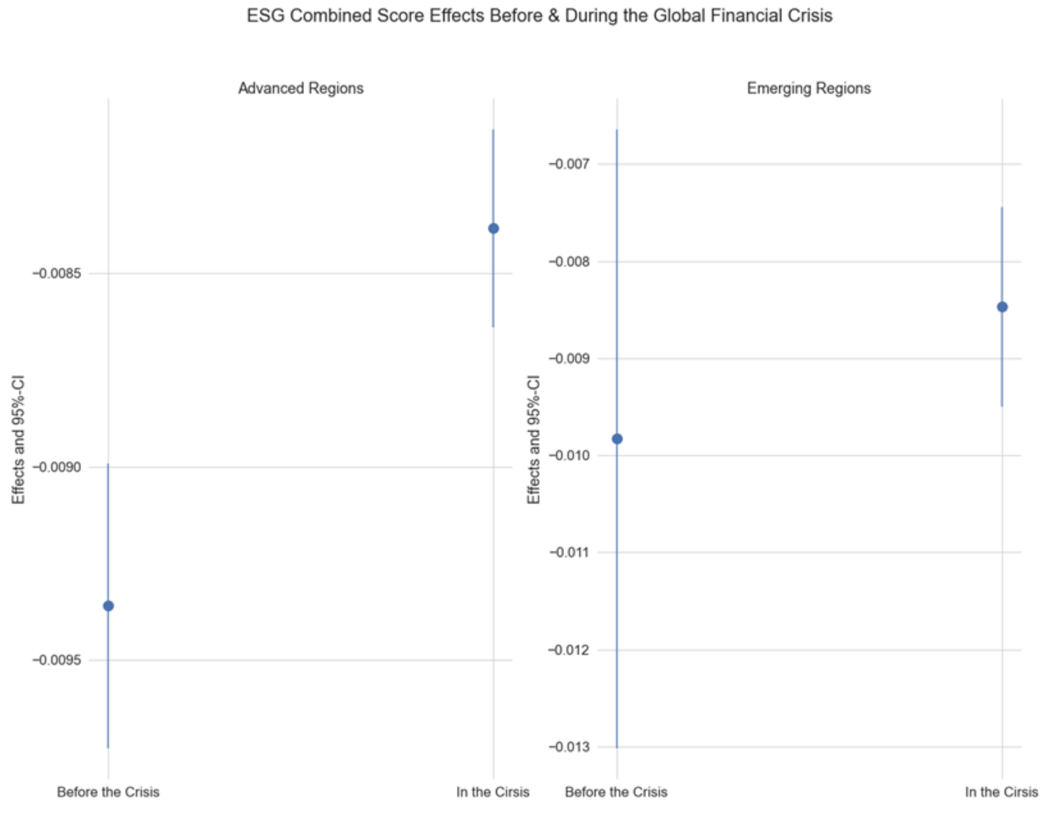


Note: The figure illustrates ESG Combined Score effects on the cost of equity over time in firms in advanced and emerging countries, estimated by XGBoost-based DML with quadratic terms. Period 1 refers to 2002 to 2008, Period 2 to 2009 to 2014 and Period 3 to 2015 to 2020. Effects refer to the estimated effects, and 95% CI stands for 95% confidence intervals.

ESG affects the cost of equity negatively in different periods. We notice a decrease in ESG impact over time. With the popularity of ESG investment getting higher and investing assets in ESG increasing, investors' expectation of the EGS return is moving in the opposite direction. Firms do not benefit from being socially responsible as much as they used to in terms of equity cost. It seems the novelty has gradually been taken away as more firms actively participate in the ESG activities.

In addition to the contributors mentioned above, we explore the relationship between ESG scores and the cost of equity over two external shocks, the global financial crisis of 2007-2008 and the COVID-19 pandemic. The worldwide financial crisis in 2007-2008 has been widely studied as an external shock. It brings our interest to investigate its role in between the ESG effects and the cost of equity. To perform the comparison, we use firm observations in 2006(before the crisis) and 2008 (during the crisis). Figure 4-4 presents estimated effects for ESG Combined Score and corresponding 95% confidence intervals. The rest results for E, S, and G Scores are summarised in Appendix Figure 4-7.

Figure 4-4: ESG Combined Score effects on the cost of equity before the global financial crisis and during the crisis time in firms in advanced and emerging countries

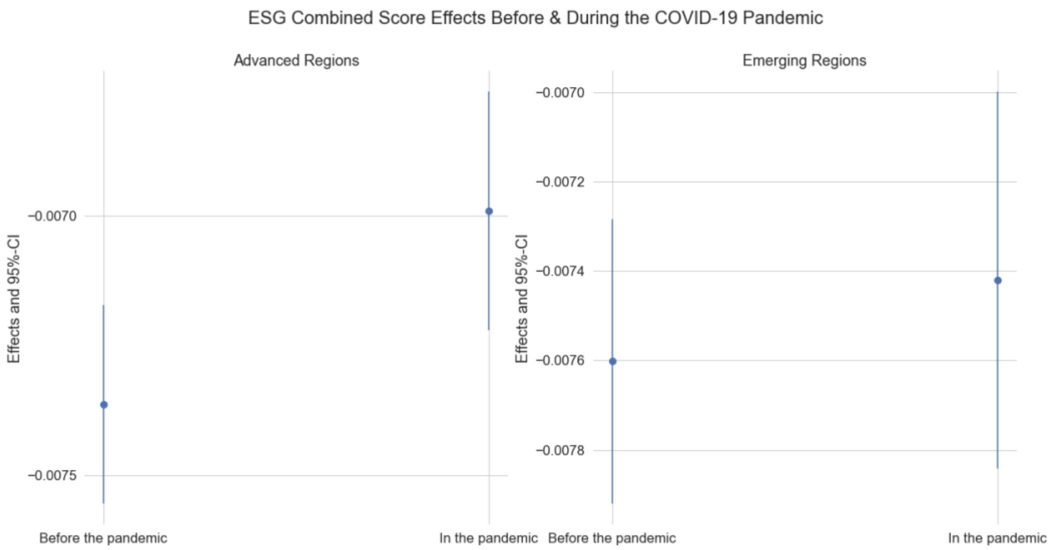


Note: The figure shows ESG Combined Score effects on the cost of equity before the global financial crisis and during the crisis time in firms in advanced and emerging countries, estimated by XGBoost-based DML with quadratic terms. Before the crisis refers to the firms in 2006 and during the crisis to the firms in 2008. Effects refer to the estimated effects, and 95% CI stands for 95% confidence intervals.

Notable gaps are presented in the figure. ESG effects on the cost of equity before the global financial crisis are stronger than during the crisis. The social responsibility of the firms seems to weigh more in investors' evaluations before the crisis. The powder of ESG scores drops during the crisis. This drop is greater in advanced regions than emerging regions, which may be caused by the fact that financial markets in advanced countries such as the U.S. were more affected by the financial crisis. Firms struggle to survive. Good ESG performance leads to potential growth in the long term but has less value in the context of a financial crisis. Investors may acknowledge that and require the risk premium based on firms' financial performance rather than ESG behaviours.

More recently, the COVID-19 pandemic has been a great shock to all aspects of our life. Ke (2021) examines the cost of equity change and notes a 172 basis-point increase during the pandemic. Inspired by it, we also test the role of the pandemic in our study. We investigate it by considering two samples, the year 2019 (before the pandemic) and the year 2020 (during the pandemic). Figure 4-5 presents the casual effects of the ESG Combined Score and the corresponding 95% confidence intervals of our firms over these two years. The related results for the E, S, and G scores are shown in Appendix Figure 4-8.

Figure 4-5: ESG Combined Score effects on the cost of equity before the COVID19 pandemic and during the pandemic in advanced and emerging regions



Note: The figure shows ESG Combined Score effects on the cost of equity before the COVID19 pandemic and during the pandemic in advanced and emerging regions, estimated by XGBoost-based DML with quadratic terms. Before the pandemic refers to the year 2019 and in the pandemic refers to the year 2020. Effects refer to the estimated effects, and 95% CI stands for 95% confidence intervals.

In general, a responsible firm tends to enjoy a lower cost of equity. But similar to the financial crisis situation, the sensitivity of the equity cost towards social responsibility awareness is weaker during the pandemic than before the pandemic. Because firms struggle with uncertainties during the pandemic, which is out of the help of great ESG strategies. Again, this drop in ESG effects is more significant in advanced regions. This difference may be caused by various measures towards the pandemic among countries. Many emerging countries have imposed strict public health control measures, and social risk evaluation is vital for firms' COVID-19 response. It is likely to lead investors' attention to ESG performance. Meanwhile, the hardest-hit companies are mainly from advanced countries, though the Covid-19 pandemic has devastated the global economy[13]. Thus, investors may prefer good financial performance in advanced countries rather than responsible ESG behaviours in terms of the risk premium.

4.6. Conclusion

Environment, social, and governance investment plays an increasingly crucial role in the world economy. This paper examines the link between firms' social responsibility and the cost of equity implied in stock prices and analysts' earnings forecasts. In particular, we use an up-to-date dataset consisting of 3,055 unique firms from 51 countries over the last 18 years. We collect ESG data from ASSET4. Firm-level financial indicators, including consensus forecasts on earnings and dividends, are sourced from Thomson Reuters I/B/E/S and Worldscope. We also employ country-specific data from World Bank Open Data Project. Our choice of variable closely follows the literature. The choice of computing the equity cost closely follows the literature. We average the results from four widely used models, Gebhardt et al. (2001), Claus and Thomas (2001), Easton (2004) and Ohlson and Juettner-Nauroth (2005).

Armed with the Xgboost-based double machine learning technique, we document negative and nonlinear effects from every ESG dimension and component on the firms' cost of equity. Investors tend to require a higher premium from socially irresponsible firms. But the margin of ESG effects decreases with firms' social responsibility enhancement. Firms that operate better in terms of social responsibility experience less ESG effects on the equity cost. Social-irresponsible firms can greatly reduce the cost of equity by improving their ESG scores. This result is robust on the battery of robustness tests. ESG effects also tend to vary in regions. To further analyse the causal effects, we break down the sample by regions. We note gaps in the

[13] https://www.investmentmonitor.ai/analysis/companies-hardest-hit-covid-19-pandemic

emerging and advanced regions, where the equity financing cost is less heterogeneous and more sensitive to ESG performance in the advanced countries. We also note that the ESG effects on the cost of equity weaken over time, along with more socially responsible firms appearing in the last decade. Additionally, we test the role of external shocks by considering the global financial crisis and the COVID-19 pandemic. We find a notable decrease in the ESG effects on the cost of equity during the financial crisis and the pandemic. Still, investors in advanced regions are more sensitive in the response of external shock toward ESG effects on the financing cost.

The findings in this paper also have practical implications. On the one hand, the results may encourage firms to take socially responsible strategies and increase the confidence of the socially responsible firms, as they enjoy the lower equity cost. On the other hand, the power of ESG effects on the equity financing cost does not keep growing regardless, and it can get weakened in different settings.

Appendix

4.A.1. Cost of Equity Computation

Following Ding, Ni, Rahman & Saadi (2015), Ferris, Javakhadze & Rajkovic (2017), and Gupta[b], Krishnamurti & Tourani-Rad (2018), we utilise four widely used approaches to estimate the implied cost of equity. The first two are computed based on residual income valuation, while the last two are computed based on abnormal earning growth.

Variable definitions:

P_t Market trading price of a firm's shares at time t as reported by I/B/E/S

BV_t Book value per share of a firm at time t

BV_{t+i} Expected book value per share for year i at time t.

EPS_{t+i} Forecasted EPS from I/B/E/S for the next i-th year at time t.

ROE_{t+i} Forecasted return on equity.

DPS_0 Dividends per share at year t-1

EPS_0 Actual earnings per share reported by I/B/E/S for year t-1

POUT Expected dividend payout ratio, assumed to be constant. It is computed as DPS_0/EPS_0

LTG Long-term growth forecast reported in I/B/E/S

Model 1 Gebhardt et al. (2001)

$$P_t = BV_t + \sum_{i=1}^{12} \frac{(ROE_{t+i} - R_{GLS}) * BV_{t+i-1}}{(1 + R_{GLS})^i} + \frac{(ROE_{t+12} - R_{GLS}) * BV_{t+11}}{R_{GLS} * (1 + R_{GLS})^{12}}$$

where $BV_{t+i} = BV_{t+i-1} + EPS_{t+i}\ (1 + POUT)$. This is based on the clean surplus relation. For the first three periods, ROEs are collected from I/B/E/S. From 4-th to 12-th, the ROE forecasts are assumed to linearly approach the industry ROE, with industries defined using the 48-industry classification in Fama and French (1997). The industry ROE is calculated as an average of historical 10-year industry-specific median returns. R_{GLS} is the solution for the equation and serves as the estimation of the implied cost of capital.

Model 2 Claus and Thomas (2001)

$$P_t = BV_t + \sum_{i=1}^{5} \frac{EPS_{t+i} - (R_{CT} * BV_{t+i-1})}{(1 + R_{CT})^i} + \frac{(EPS_{t+5} - R_{CT} * BV_{t+4}) * (1 + LTG)}{(R_{CT} - LTG) * (1 + R_{CT})^5}$$

where $BV_{t+i} = BV_{t+i-1} + EPS_{t+i}\ (1 + POUT)$. This is based on the clean surplus relation. For the first two periods, EPS are collected from I/B/E/S. They equal to the one, two -year-ahead I/B/E/S analyst median forecast at time t. For years three, four, and five: $EPS_{t+i} = EPS_{t+i-1} * (1 + LGT)$. R_{CT} is the solution for the equation and serves as the estimation of the implied cost of capital.

Model 3 Ohlson and Juettner-Nauroth (2005)

$$R_{OJ} = A + \sqrt{A^2 + \frac{EPS_{t+1}}{P_t}\left(\frac{EPS_{t+2} - EPS_{t+1}}{EPS_{t+1}} - LTG\right)}$$

Where $A = \frac{1}{2} * (LTG + \frac{POUT*EPS_{t+1}}{P_t})$

$EPS_{t+2} > 0\ , and\ \ EPS_{t+1} > 0$

R_{OJ} estimates the implied cost of capital.

Model 4 Easton's (2004)

$$P_t = \frac{EPS_{t+2} - EPS_{T+1} + (R_E * EPS_{T+1} * POUT)}{R_E^2}$$

$where\ EPS_{t+2} \geq EPS_{t+1} \geq 0$

This model is a particular case of Ohlson and Juettner-Nauroth (2005). R_E is the solution for the equation and serves as the estimation of the implied cost of capital.

Model 5 Averaging approach

$$R_{AVE} = \frac{1}{4} \times (R_{GLS} + R_{CT} + R_{OJ} + R_E)$$

R_{AVE} is the final proxy for the cost of equity in this study.

4.A.2. Variable Description

Our dataset consists of 15,229 firm-year observations with 34 control variables and 15 ESG/CSR variables. All the financial data are winsorised at 1% and 99% to handle outliers. Table 4-5 lists all the variables under study.

Table 4-5: Variable Description

Firm-level Variables	**Country-level Variables**
Analyst Forecast Dispersion	Country dummies
Analyst forecast error	GDP Growth

Book-value	GDP billion$ (2010 constant)
Debt/assets	Inflation
Earnings per share (a year forward forecast)	
Fama/French 48 industry code	
Financial leverage	
Free cash flow per share	
Long term debt	**ESG Variables**
Long term debt/total assets	ESG Combined Score
Long term growth rate (median)	Environment Pillar Score (E)
Market Beta (Computed via CAPM)	Social Pillar Score (S)
Market Capitalisation	Governance Pillar Score (G)
Market-to-book	Community Score
Natural logarithm of firm market value	CSR Strategy Score
Natural logarithm of total assets	Emissions Score
Net Profit (margin)	Environmental Innovation Score
Operating income growth	ESG Controversies Score
Return On Assets	Human Rights Score
Sale per share growth	Management Score
Sales revenue	Product Responsibility Score
Sales revenue/total assets	Resource Use Score
Size factor dummies	Shareholders Score
Stock Price (median)	Workforce Score
Stock Price volitivity	
The number of analysts following the firm	
Total debt	
Total debt/Market Value	**Outcome Variable**
Turnover	Cost of equity
Year dummies	

4.A.3. Sample Breakdown by Country

Table 4-1 presents that our data offers a rich country coverage compared to the mainstream of previous studies focusing on a single country (Gupta et al., 2018; Ke, 2021) or datasets with less country diversity (Breuer et al., 2018). The sample in our study covers both developed regions and developing regions, including counties that are rarely examined, such as Peru. We observe significant heterogeneity in the cost of equity and ESG performance across countries. Unlike the expectation, the advanced countries do not present absolute strength in social responsibility or cost of equity. For example, Turkey is doing better than Japan and the US in terms of ESG Combined Score, and investors in Chile require a lower premium than investors from most countries.

Table 4-6: Sample Breakdown by Country

Country	Firms	N	%	Mean ESG	Mean R_{AVE}	Country	Firms	N	%	Mean ESG	Mean R_{AVE}
Australia	51	186	2.12	44.28	0.14	Japan	348	2618	29.80	44.76	0.08
New Zealand	12	39	0.44	35.29	0.07	Korea	103	431	4.91	47.03	0.09
Canada	99	412	4.69	49.00	0.08	Indonesia	37	219	2.49	46.82	0.06
Oman	1	1	0.01	23.06	**0.05**	Malaysia	31	109	1.24	44.17	0.07
Austria	15	66	0.75	46.81	0.09	Kuwait	4	13	0.15	31.63	0.08
Belgium	15	91	1.04	51.72	0.10	Philippines	13	86	0.98	40.72	0.05
Czech Republic	3	10	0.11	44.76	0.06	Singapore	33	153	1.74	36.10	0.07
Belgium	98	577	6.57	50.69	0.10	Thailand	38	159	1.81	51.22	0.07
Spain	35	160	1.82	60.80	0.08	United Arab Emirates	5	7	0.08	30.63	0.08
France	78	535	6.09	54.48	0.09	Saudi Arabia	16	40	0.46	25.05	0.07
Poland	17	72	0.82	43.96	0.07	Israel	13	56	0.64	41.50	0.08
Greece	13	55	0.63	49.61	1.66	Qatar	7	16	0.18	30.69	0.08
Italy	37	156	1.78	50.59	0.12	Vietnam	2	3	0.03	49.47	0.06
Hungary	3	26	0.30	45.72	0.09	Egypt	7	20	0.23	32.38	0.10
Netherlands	26	109	1.24	53.34	0.11	Kenya	1	3	0.03	42.04	0.08
Portugal	6	34	0.39	**61.64**	0.15	South Africa	53	268	3.05	53.32	0.08
Russia	20	52	0.59	45.85	0.06	Argentina	6	12	0.14	50.14	0.08
Switzerland	72	480	5.46	45.92	0.08	Brazil	67	212	2.41	46.25	0.09
Turkey	24	97	1.10	49.69	0.09	Chile	17	60	0.68	37.00	**0.05**
United Kingdom	218	1583	18.02	51.41	0.06	Colombia	5	17	0.19	60.46	0.06
Ireland	15	50	0.57	53.73	0.41	Mexico	27	81	0.92	52.71	0.06
China	168	467	5.32	33.20	0.07	Peru	1	3	0.03	41.03	0.16
Hong Kong	177	827	9.41	36.39	0.07	Denmark	28	183	2.08	50.50	0.07
India	114	589	6.71	48.59	0.09	Finland	20	127	1.45	56.41	0.10
Japan	348	2618	29.80	44.76	0.08	Norway	22	131	1.49	51.68	0.07

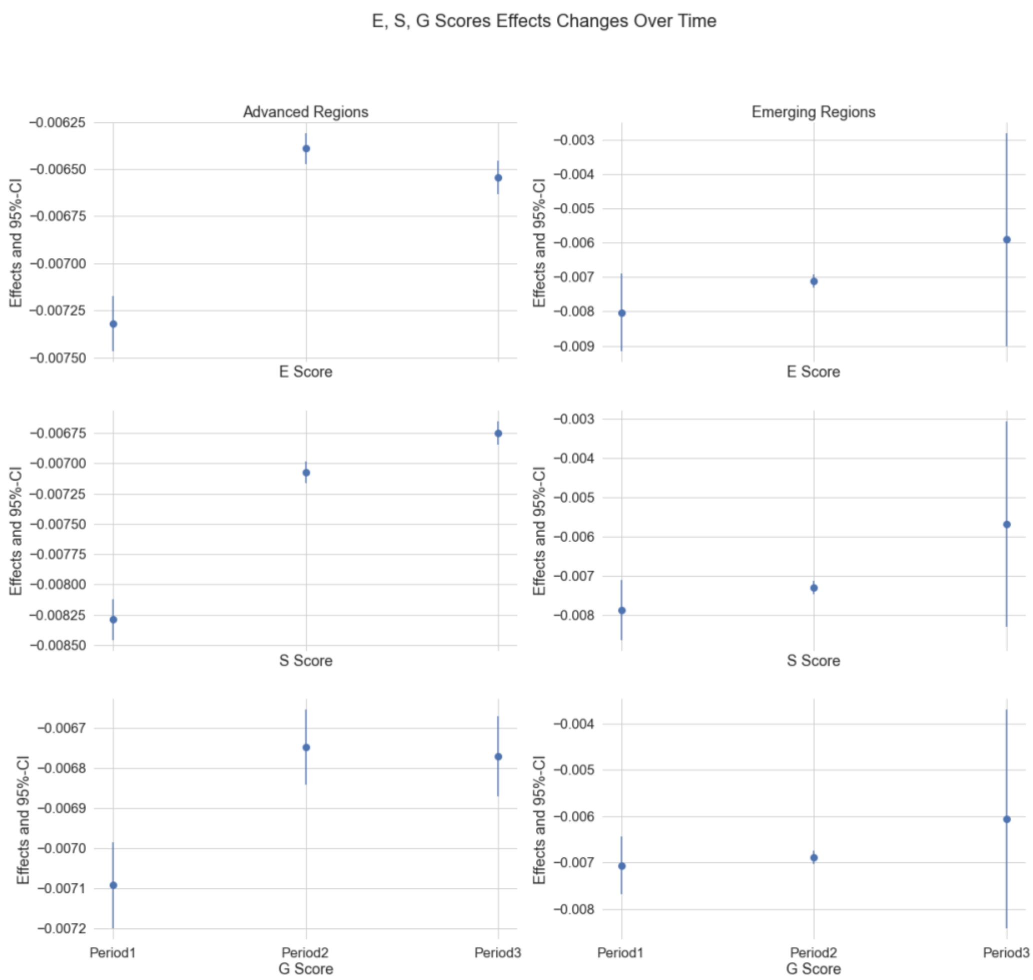
Korea	103	431	4.91	47.03	0.09	Sweden	47	278	3.16	54.72	0.09
Indonesia	37	219	2.49	46.82	0.06	United States	**787**	**3250**	**37.00**	41.05	0.08

Note: The table summarises the sample by the country where the firm is listed. The column Firms counts the number of firms filed in that category. N, and % columns are the numbers of the observations and the percentages against the whole sample. The last two columns are the mean value for the ESG Combined Score and the mean value for the cost of equity. The best value in each column is in bold.

4.A.4. ESG Effects Changes over Time

The last two decades have witnessed a soar in ESG investment. In section 4.5, we show that the ESG Combined Score negatively affects the equity cost, and the effects have gotten weaker in recent years. Here we present the rest results for E, S, and G Scores' effects on the cost of equity in advanced and emerging regions.

Figure 4-6: E, S, G Score affect the cost of equity over time in advanced and emerging regions



Note: The figure shows E, S, G Score affect the cost of equity over time in advanced and emerging regions, estimated by XGBoost-based DML with quadratic terms. Effects refer to the estimated effects, and 95% CI stands for 95% confidence intervals.

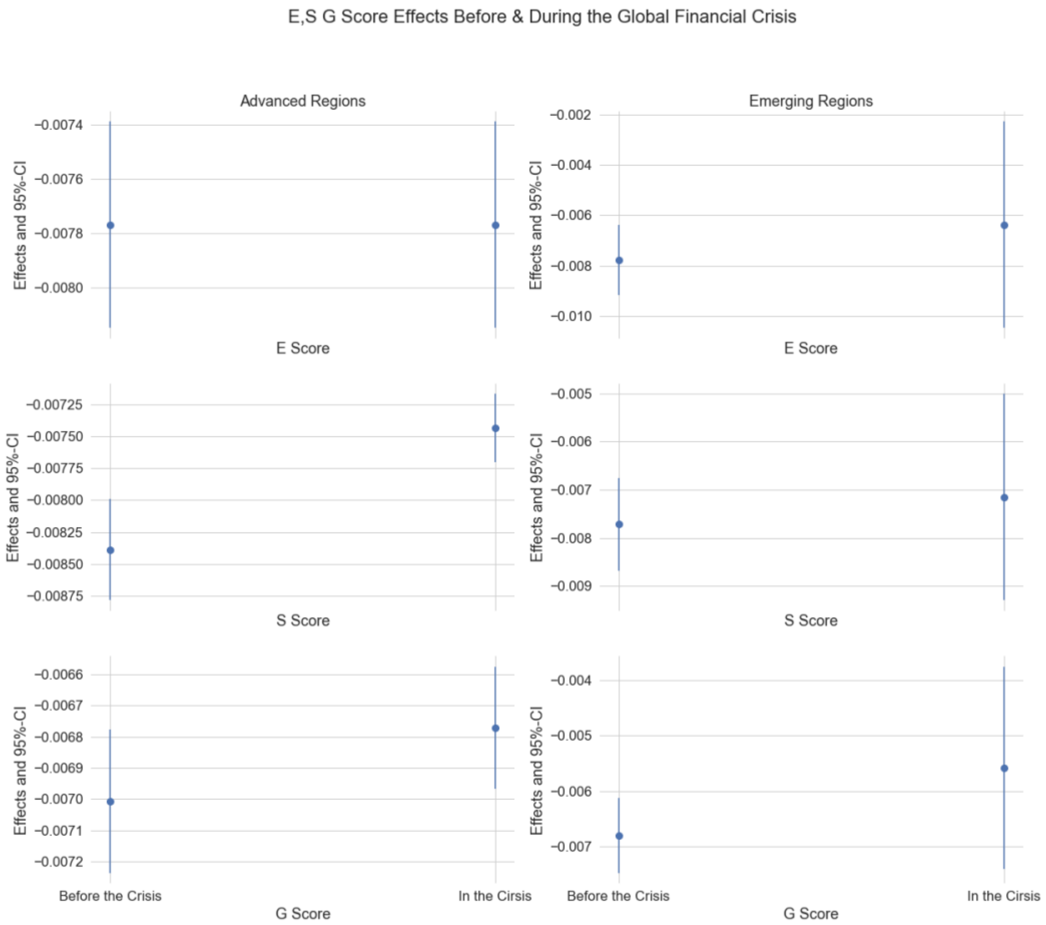
The results confirm the negative relationship between ESG and equity cost in both advanced regions and emerging regions. In general, investors value firms' ESG behaviour less in recent years when it comes to the cost of equity. However, investors' attention to Environment Score or Government Score does not decrease in advanced regions.

4.A.5. The External Shocks

We also consider the response of ESG effects to external shocks such as the global financial crisis and the COVID-19 pandemic.

In section 4.5, we show that the global financial crisis weakens the ESG effect on the cost of equity in advanced and emerging regions by reporting the estimated effects of the ESG Combined Score. Here we present the rest results for E, S, and G Scores.

Figure 4-7: E, S, G Score affect the cost of equity in response to the global financial crisis in advanced and emerging regions

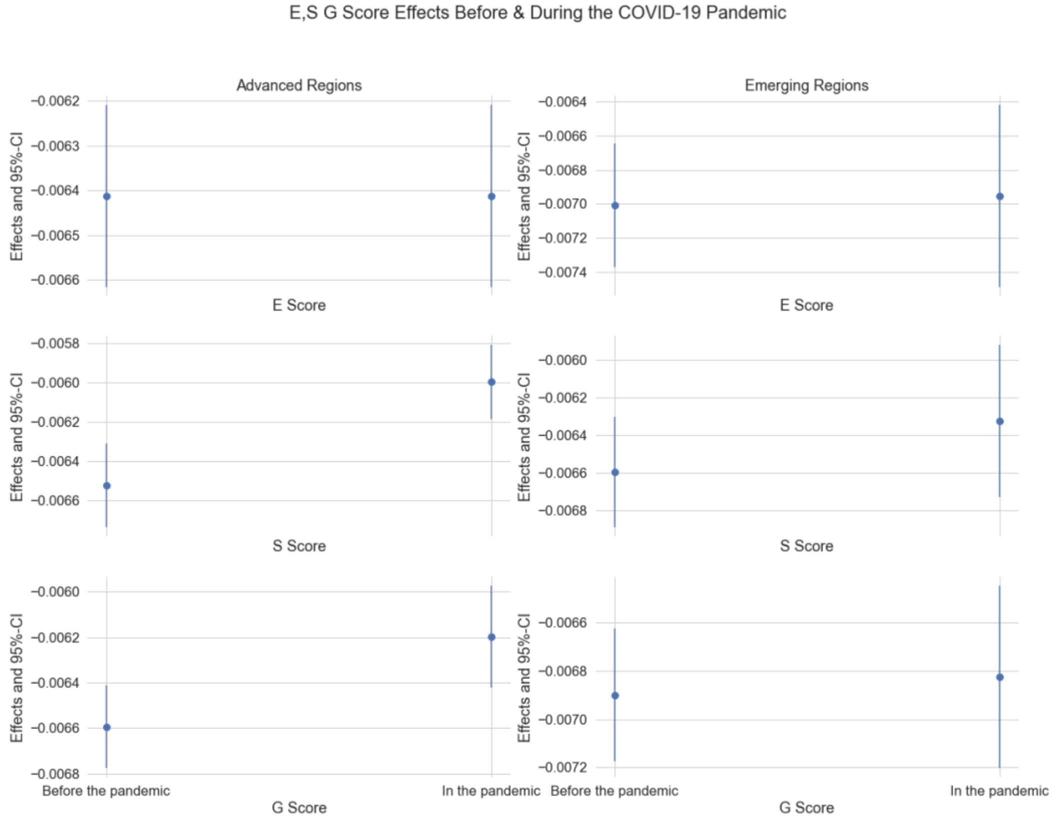


Note: the figure shows E, S, G Score affect the cost of equity in response to the global financial crisis in advanced and emerging regions, estimated by XGBoost-based DML with quadratic terms. Effects refer to the estimated effects, and 95% CI stands for 95% confidence intervals.

Similar to our main results, the results for E, S, G Score indicate that firms' social-responsible behaviour may matter more in investors' evaluations before the crisis. The drop in ESG effects is more extraordinary in advanced regions than emerging regions, which may be caused by the fact that financial markets in advanced countries such as the US were worse damaged by the crisis.

Figure 4-8 illustrates the response of ESG effect on the cost of equity to the COVID 19 pandemic in advanced and emerging regions using E, S, G Score. the Same pattern as we present in Section 4.5 can be seen. In general, investors value less of firms' social responsibility when considering the risk premium during the pandemic.

Figure 4-8: E, S, G Score affect the cost of equity in response to the COVID-19 pandemic in advanced and emerging regions



Note: The figure shows E, S, G Score affect the cost of equity in response to the COVID-19 pandemic in advanced and emerging regions, estimated by XGBoost-based DML with quadratic terms. Effects refer to the estimated effects, and 95% CI stands for 95% confidence intervals.

Chapter 5

Chapter 5: Conclusion

5.1. Key Findings

This thesis bundles three individual empirical studies in the credit and the equity market. Most up-to-date data and state-of-the-art machine learning techniques are intensively employed in the studies.

Powerful approaches are widely employed to examine P2P loan default risk, but few studies systematically compare these approaches in terms of credit data imbalance in the literature. Chapter 2 aims to fill the gap by proposing four presentative models from machine learning approaches. They are AdLASSO from the linear model subgroup, LightGBM from the decision-tree-based model subgroup, CNTN from the artificial neural network subgroup, and WDL which integrates linear techniques and deep learning. The classic credit analysis technique, logistic regression, serves as the benchmark in the study. 279,512 real loans (84.90% Non-default and 15.10% default) from the Lending Club with 25 feature is collected for comparison. Among the five models, LightGBM, CNTN and WDL outperform according to discrimination ability. LightGBM stands further out for its insensitivity to imbalanced data and interpretability of variable importance. It allows a deeper study on default determinants where we note that debt-to-income ratio and monthly instalments may greatly affect the default. We conduct robust tests with subgroups based on applicants' employment length, income verification, and home ownership. The results are consistent that LightGBM shows superiority over the rest four models.

We discuss P2P marketplace funding decision determinants and macroeconomic effects in chapter 3. Both loan-specific variables and macroeconomic variables are understudied. Armed with text mining techniques, we enrich our loan-level variables by analysing and classifying the textual variable Loan Title. In addition to widely studied macroeconomic variables such as CPI, GDP, FFR, and TBR, we compute Taylor residuals in the study as an indicator of monetary policy. In order to stimulate the world, we aim at big data. Over 12.4 million loan applications from the Lending Club are collected for empirical study. Using a clustering approach termed BIRCH, we collapse the original database into sub-clusters and compute dense representatives with cluster-size as the weights of each sub-cluster. In the next stage, we generate five machine learning models (Logistic Regression, Random Forest, Xgboost, Naive Bayes, and Support Vector Machine) based on the representatives. In general, SVM beats the other four models in AUC, precision, recall, and F1 score, serving the best predictive analysis for P2P loan funding decisions. We further output the importance of the variable to track the determinants and macroeconomic impact. Loan applicants' employment length tends to be the top preference when investors make funding decisions. We note that macroeconomic condition affects individuals' lending decision and risk-taking behaviour. In order to reveal the heterogeneity towards macroeconomic factors at the loan level, we also break the original database into sub-groups according to debt-to-income ratio and loan title, respectively. We document variations of macroeconomic influence between loan requests from applicants who have healthy DIR (between 0% and 50%) and applicants who have high DIR (above 50%). Furthermore, investors value DIR over employment length when it comes to an unhealthy DIR. Additionally, we notice that loan title plays an essential role in P2P funding. Applications with titles that do not state loan purposes properly are sensitive to macroeconomic conditions.

Chapter 4 investigates the link between firms' social responsibility and the cost of equity implied in stock prices and analysts' earnings forecasts. Specifically, an up-to-date dataset of 3,055 unique firms from 51 countries has been examined over the last 18 years. We employ the most up-to-date machine learning technique, double machine learning, to document adverse and nonlinear effects from every ESG dimension and component on the firms' cost of equity. According to investors ' expectations, a higher premium is assigned to socially irresponsible firms. This result is robust on the battery of robustness tests. To dive deep into these causal effects, we break down the sample into different regions. We note gaps in the emerging and advanced regions, where the equity financing cost is less heterogeneous and more sensitive to ESG performance in the advanced countries. However, with more firms joining the group of being socially responsible in the last decade, the ESG effects on the cost of equity seem to weaken over time. Most external shocks, the global financial crisis and the COVID-19 pandemic are also considered in the chapter. We find a decline in the ESG effects on the cost of equity during the financial crisis, and mixed results are in the response of ESG effects towards the pandemic.

5.2. Implication in Practice

This thesis also offers practical values to the market participants.

In the first empirical chapter, we note the superiority of LightGBM in predicting loan default risk. This technique can strengthen the current credit scoring systems of financial organisations, including traditional banks and Fintech institutions, by enhancing the accuracy of abnormal loan request detection. Our study also offers insights into the attributes of risky loan requests. Investors, especially individual investors who lack expertise, can use the feature importance output by LightGBM as guidance to optimise their P2P investment portfolios. Similarly, regulators may use this feature importance to implement rules on loan applicants. In this manner, they may reduce the potential credit risk in the market.

Furthermore, the second empirical chapter discusses the keys to successful P2P loans. Loan applicants can increase their repayability to get loan requests approved with the direction provided in this chapter. Additionally, the two chapters together reveal a gap. The attributes matter in loan granting seems to be different from those that affect default risk. Investors are likely to evaluate loan requests wrong and make risky decisions. It is also important for regulators to be aware of this gap.

The last empirical chapter is on equity cost and corporate social responsibility (CSR). The financial growth of a firm may have ethical concerns in many cases. Understanding the role of CSR in terms of equity cost is crucial because it helps firms to make developing strategies by balancing various aspects. This chapter also considers different environmental settings (the financial crisis and the COVID-19 pandemic), which offers the most up-to-date and practical insights.

5.3. Suggestions for Future Work

Based on the three empirical chapters, the thesis has some constraints that can be improved in future studies.

The main limitation comes from data limitation. Although the datasets in the thesis allow me to conduct comprehensive analyses, they are not perfect for research. Limited by the availability of P2P credit data, the Lending Club has become our choice. However, due to the Lending Club changing the data format several times since it formed, only 279,512 loans from the year 2015 are used in chapter 2 to ensure we have clean and accurate data. Variety and volume are lost in the data. Similarly, a richer dataset across countries with more loan applicants' information may boost the predictive ability of models in chapter 3. Data in chapter 4 is collected in the middle of 2021, when firms were still developing strategies to better cope with the changeable COVID-19 situation. Unluckily, the whole year's data was unavailable for us to thoroughly analyse the role of COVID-19 in ESG performance and the cost of equity. This would be an exciting topic for future studies.

References

Abellán, J. and Castellano, J. (2017). A comparative study on base classifiers in ensemble methods for credit scoring. Expert Systems with Applications, 73, pp.1-10.

Abid, L., Ouertani, M., & Zouari-Ghorbel, S. (2014). Macroeconomic and Bank-specific Determinants of Household's Non-performing Loans in Tunisia: A Dynamic Panel Data. *Procedia Economics And Finance*, 13, 58-68.

Abuka, C., Alinda, R., Minoiu, C., Peydró, J., & Presbitero, A. (2019). Monetary policy and bank lending in developing countries: Loan applications, rates, and real effects. *Journal Of Development Economics*, *139*, 185-202.

Abuka, C., Alinda, R., Minoiu, C., Peydró, J., & Presbitero, A. (2015). Monetary Policy in a Developing Country: Loan Applications and Real Effects. *IMF Working Papers*, *15*(270), 1.

Albuquerque, R., Koskinen, Y., & Zhang, C. (2019). Corporate Social Responsibility and Firm Risk: Theory and Empirical Evidence. *Management Science*, *65*(10), 4451-4469.

Al Guindy, M. (2021). Corporate Twitter use and cost of equity capital. *Journal Of Corporate Finance*, 68, 101926.

Antoniadis, A., Lambert-Lacroix, S., & Poggi, J. (2020). Random forests for global sensitivity analysis: A selective review. Reliability Engineering & System Safety, 206, 107312. doi: 10.1016/j.ress.2020.107312

Arora, N., & Kaur, P. (2020). A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment. *Applied Soft Computing*, *86*, 105936.

Badolia L, 2016. How can I get started investing in the stock market, Educreation Publishing (2016)

Balabin, R. and Lomakina, E. (2011). Support vector machine regression (SVR/LS-SVM)—an alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near infrared (NIR) spectroscopy data. The Analyst, 136(8), p.1703.

Bastani, K., Asgari, E. and Namavari, H. (2019). Wide and Deep Learning for Peer-to-Peer Lending. Expert Systems with Applications.

Becchetti, L., Ciciretti, R., & Hasan, I. (2015). Corporate social responsibility, stakeholder risk, and idiosyncratic volatility. *Journal Of Corporate Finance*, *35*, 297-309.

Bekhet, H. and Eletter, S., (2014). Credit risk assessment model for Jordanian commercial banks: Neural scoring approach. Review of Development Finance, 4(1), pp.20-28.

Blanco, A., Pino-Mejías, R., Lara, J. and Rayo, S., 2013. Credit scoring models for the microfinance industry using neural networks: Evidence from Peru. *Expert Systems with Applications*, 40(1), pp.356-364.

Bofondi, M., & Ropele, T. (2011). Macroeconomic Determinants of Bad Loans: Evidence from Italian Banks. *SSRN Electronic Journal*.

Boubakri, N., Guedhami, O., Mishra, D., & Saffar, W. (2012). Political connections and the cost of equity capital. *Journal Of Corporate Finance*, 18(3), 541-559.

Breiman, L. (2001). Random forests. *Mach Learn, 45(1):5–32.*

Breuer, W., Müller, T., Rosenbach, D., & Salzmann, A. (2018). Corporate social responsibility, investor protection, and cost of equity: A cross-country comparison. *Journal Of Banking & Finance, 96*, 34-55.

Burim, R., & Markus, P. (2019) Asymptotic Unbiasedness of the Permutation Importance Measure in Random Forest Models.

Byanjankar, A., Heikkilä, M. and Mezei, J. (2015). Predicting Credit Risk in Peer-to-Peer Lending: A Neural Network Approach. I*n: IEEE Symposium Series on Computational Intelligence*. Cape Town, South Africa, pp.719-725.

Cai, S., Lin, X., Xu, D. and Fu, X. (2016). Judging online peer-to-peer lending behavior: A comparison of first-time and repeated borrowing requests. *Information & Management, 53(7),* pp.857-867.

Casalicchio, G., Molnar, C., & Bischl, B. (2019). Visualizing the Feature Importance for Black Box Models. *Machine Learning And Knowledge Discovery In Databases*, 655-670.

Cassidy, C. (2020). Parameter tuning Naïve Bayes for automatic patent classification. *World Patent Information, 61, 101968.*

Chang, Y., Chang, K., & Wu, G. (2018). Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Applied Soft Computing, 73, 914-920.*

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

Chen, B., & Zhang, A. (2021). How does corporate social responsibility affect the cost of equity capital through operating risk?. *Borsa Istanbul Review*.

Chen, S., Härdle, W. and Moro, R. (2011). Modeling default risk with support vector machines. Quantitative Finance, 11(1), pp.135-154.

Chen, S., Zhang, Y. and Li, L., 2022. A Comparative Study of Credit Risk Infection at Domestic and Abroad Based on Knowledge Map. Procedia Computer Science, 199, pp.215-222.

Chen, T., & Guestrin, C. (2016). XGBoost. *Proceedings Of The 22Nd ACM SIGKDD International Conference On Knowledge Discovery And Data Mining*.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1-C68.

Choi, Y., Kim, H., & Ro, Y. (2016). Two-step Learning of Deep Convolutional Neural Network for Discriminative Face Recognition under Varying Illumination. Electronic Imaging, 2016(11), 1-5.

Chu, Y. and Deng, X., (2018). Risk-Taking, Reaching-for-Yield, and Monetary Policy: Evidence from Peer-to-Peer Lending. SSRN Electronic Journal.

Chu, Yongqiang and Deng, Xiaoying. (2019) The Risk-Taking Channel of Monetary Policy: Evidence from Individual Investors in the Peer-to-Peer Lending Market (March 26, 2019). *SSRN*

Claus, J., & Thomas, J. (2001). Equity Premia as Low as Three Percent? Evidence from Analysts' Earnings Forecasts for Domestic and International Stock Markets. *The Journal Of Finance*, *56*(5), 1629-1666.

D'Hondt, C., Merli, M. and Roger, T., 2022. What drives retail portfolio exposure to ESG factors?. Finance Research Letters, 46, p.102470.

Dave, M., & Gianey, H. (2016). Different clustering algorithms for Big Data analytics: A review. *2016 International Conference System Modeling & Advancement In Research Trends (SMART)*.

Delis, M., Hasan, I., & Mylonidis, N. (2017). The Risk-Taking Channel of Monetary Policy in the U.S.: Evidence from Corporate Loan Data. *Journal Of Money, Credit And Banking, 49(1), 187-213.*

Den Haan, W., Sumner, S., & Yamashiro, G. (2011). Bank Loan Components and the Time-varying Effects of Monetary Policy Shocks. *Economica*, *78*(312), 593-617.

Delong, E., Delong, D., & Clarke-Pearson, D. (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, 44(3), 837.

Devalle, A., Fiandrino, S., & Cantino, V. (2017). The Linkage between ESG Performance and Credit Ratings: A Firm-Level Perspective Analysis. *International Journal Of Business And Management*, *12*(9), 53.

Dhaliwal, D., Li, O., Tsang, A., & Yang, Y. (2014). Corporate social responsibility disclosure and the cost of equity capital: The roles of stakeholder orientation and financial transparency. *Journal Of Accounting And Public Policy*, 33(4), 328-355.

Ding, X., Ni, Y., Rahman, A., & Saadi, S. (2015). Housing price growth and the cost of equity capital. *Journal Of Banking & Finance*, 61, 283-300.

Dumitrescu, A., El Hefnawy, M., & Zakriya, M. (2020). Golden geese or black sheep: Are stakeholders the saviors or saboteurs of financial distress?. *Finance Research Letters*, *37*, 101371.

Duque-Grisales, E., & Aguilera-Caracuel, J. (2019). Environmental, Social and Governance (ESG) Scores and Financial Performance of Multilatinas: Moderating Effects of Geographic International Diversification and Financial Slack. *Journal Of Business Ethics*, 168(2), 315-334.

Easton, P. (2004). Discussion—Earnings Surprises and the Cost of Equity Capital. *Journal Of Accounting, Auditing & Finance*, *19*(4), 515-521.

El Ghoul, S., Guedhami, O., Kwok, C.Y., Mishra, D., (2011). Does corporate social responsibility affect the cost of capital? *Journal of Bank and Finance* 35 (9), 2388–2406

El Ghoul, S., Guedhami, O., Wang, H., & Kwok, C. (2016). Family control and corporate social responsibility. *Journal Of Banking & Finance*, 73, 131-146.

Emekter, R., Tu, Y., Jirasakuldech, B. and Lu, M., (2014). Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. Applied Economics, 47(1), pp.54-70.

Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., & Zomaya, A. (2014). A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis. *IEEE Transactions On Emerging Topics In Computing*, *2*(3), 267-279.

Farah, T., Li, J., Li, Z., & Shamsuddin, A. (2021). The non-linear effect of CSR on firms' systematic risk: International evidence. *Journal Of International Financial Markets, Institutions And Money*, *71*, 101288.

Fama, E., & French, K. (1997). Industry costs of equity. *Journal Of Financial Economics*, *43*(2), 153-193.

Fatemi, A., Glaum, M., & Kaiser, S. (2018). ESG performance and firm value: The moderating role of disclosure. *Global Finance Journal*, *38*, 45-64.

Ferris, S., Javakhadze, D., & Rajkovic, T. (2017). The international effect of managerial social capital on the cost of equity. *Journal Of Banking & Finance, 74*, 69-84.

Fisher, R., 1936. The Use Of Multiple Measurements In Taxonomic Problems. Annals of Eugenics, 7(2), pp.179-188.

Finlay, S. (2011). Multiple classifier architectures and their application to credit risk assessment. European Journal of Operational Research, 210(2), pp.368-378.

Florez-Lopez, R., & Ramon-Jeronimo, J. (2014). Modelling credit risk with scarce default data: on the suitability of cooperative bootstrapped strategies for small low-default portfolios. Journal Of The Operational Research Society, 65(3), 416-434.

Friedman, M. , (1970). The Social Responsibility of Business is to Increase its Profits. The New York Times Magazine 1970.

Gambacorta, L., and Marques-Ibanez, D. (2011). The bank lending channel: lessons from the crisis. *Economic Policy, 26(66):135–182.*

Gavurova, B., Dujcak, M., Kovac, V., & Kotásková, A. (2018). Determinants of Successful Loan Application on Peer-to-Peer Lending Market. *Economics & Sociology*, *11*(1), 85-99.

Gebhardt, W., Lee, C., & Swaminathan, B. (2001). Toward an Implied Cost of Capital. *Journal Of Accounting Research*, *39*(1), 135-176.

Gete, P. (2018). Lending standards and Pedro Gete macroeconomic dynamics. *European Central Bank, (2207), 1-48.*

Gillan, S., Koch, A., & Starks, L. (2021). Firms and social responsibility: A review of ESG and CSR research in corporate finance. *Journal Of Corporate Finance*, *66*, 101889.

Goldberg Y., Levy O.(2014). word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method.

Gonzalez, L. and Komarova Loureiro, Y. (2014), "When can a photo increase credit? The impact of lender and borrower profiles on online peer-to-peer loans", *Journal of Behavioral and Experimental Finance,* Vol. 2 No. 2, pp. 1-14.

Goss, A., & Roberts, G. (2011). The impact of corporate social responsibility on the cost of bank loans. *Journal Of Banking & Finance*, *35*(7), 1794-1810.

Grace, O., Robert, P., & Eric, O. (2016). Macroeconomic Factors and the Performance of Loans of Commercial Banks in Ghana: A Case Study of HFC Bank. *European Journal Of Economics, Finance And Administrative Sciences, (81), 38-46.*

Gu, J., & Lu, S. (2020). An effective intrusion detection approach using SVM with naïve Bayes feature embedding. *Computers & Security*, 102158.

Guo, Y. (2019). Credit Risk Assessment of P2P Lending Platform towards Big Data based on BP Neural Network. *Journal of Visual Communication and Image Representation*, p.102730.

Guo, Y., Zhou, W., Luo, C., Liu, C. and Xiong, H. (2016). Instance-based credit risk assessment for investment decisions in P2P lending. *European Journal of Operational Research,* 249(2), pp.417-426.

Gupta^{a}, A., Raman, K., & Shang, C. (2018). Social capital and the cost of equity. *Journal Of Banking & Finance*, *87*, 102-117.

Gupta^{b}, K., Krishnamurti, C., & Tourani-Rad, A. (2018). Financial development, corporate governance and cost of equity capital. *Journal Of Contemporary Accounting & Economics, 14(1),* 65-82.

Harjoto, M. A., & Jo, H. (2015). Legal vs. normative CSR: Differential impact on analyst dispersion, stock return volatility, cost of capital, and firm value. Journal of Business Ethics, 128(1), 1e20

Harrington, J. and Salibián-Barrera, M. (2008). Finding approximate solutions to combinatorial problems with very large data sets using BIRCH. *Computational Statistics & Data Analysis*, 54(3), pp.655-667.

Harris, T. (2013). Quantitative credit risk assessment using support vector machines: Broad versus Narrow default definitions. *Expert Systems with Applications*, 40(11), pp.4404-4413.

Hasan, M., Hossain, M., Cheung, A. and Habib, A., (2015). Corporate life cycle and cost of equity capital. *Journal of Contemporary Accounting & Economics*, 11(1), pp.46-60.

He, H., Zhang, W., & Zhang, S. (2018). A novel ensemble method for credit scoring: Adaption of different imbalance ratios. *Expert Systems With Applications*, 98, 105-117.

Hens, A. and Tiwari, M. (2012). Computational time reduction for credit scoring: An integrated approach based on support vector machine and stratified sampling method. *Expert Systems with Applications, 39(8),* pp.6774-6781.

He, W., Lepone, A., & Leung, H. (2013). Information asymmetry and the cost of equity capital. *International Review Of Economics & Finance*, 27, 611-620.

Hoepner, A., Oikonomou, I., Sautner, Z., Starks, L., & Zhou, X. (2021). ESG Shareholder Engagement and Downside Risk. *SSRN Electronic Journal*.

Hong, H., & Kacperczyk, M. (2009). The price of sin: The effects of social norms on markets. *Journal Of Financial Economics*, 93(1), 15-36.

Huang, Y., Li, X. and Wang, C., (2019). What Does Peer-to-Peer Lending Evidence Say About the Risk-Taking Channel of Monetary Policy?. *SSRN Electronic Journal*.

Ibrahim, M., & Shah, M. (2012). Bank lending, macroeconomic conditions and financial uncertainty: Evidence from Malaysia. *Review Of Development Finance, 2*(3-4), 156-164.

Ince, H. and Aktan, B., (2009). A Comparison of data mining techniques for credit scoring in banking: A managerial perspective. *Journal of Business Economics and Management*, 10(3), pp.233-240.

Ip, C. and Lam, F., (2020). Funding Decisions in Online Marketplace Lending. *SSRN Electronic Journal*.

Jiang, C., Wang, Z., Wang, R., & Ding, Y. (2017). Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. *Annals Of Operations Research*, 266(1-2), 511-529.

Jiménez, G., Ongena, S., Peydró, J. and Saurina, J., (2014). Hazardous Times for Monetary Policy: What Do Twenty-Three Million Bank Loans Say About the Effects of Monetary Policy on Credit Risk-Taking. *Econometrica, 82(2), pp.463-505.*

Jiraporn, P., Jiraporn, N., Boeprasert, A., & Chang, K. (2014). Does Corporate Social Responsibility (CSR) Improve Credit Ratings? Evidence from Geographic Identification. *Financial Management, 43*(3), 505-531.

Ke, Y. (2021). The impact of COVID-19 on firms' cost of equity capital: Early evidence from U.S. public firms. *Finance Research Letters*, 102242.

Kerkemeier, M. and Kruse-Becher, R., 2022. Join the club! Dynamics of global ESG indices convergence. *Finance Research Letters*, p.103085.

Khalifa, M., Zouaoui, H., Ben Othman, H., & Hussainey, K. (2019). Exploring the nonlinear effect of conditional conservatism on the cost of equity capital: Evidence from emerging markets. *Journal Of International Accounting, Auditing And Taxation*, *36*, 100272.

Khashman, A. (2010). Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. *Expert Systems with Applications*, 37(9), pp.6233-6239.

Khashman, A. (2011). Credit risk evaluation using neural networks: Emotional versus conventional models. *Applied Soft Computing*, 11(8), pp.5477-5484.

Khandani, A., Kim, A. and Lo, A. (2010). Consumer Credit Risk Models Via Machine-Learning Algorithms. *SSRN Electronic Journal*.

Kim, Y., Park, M., Wier, B., 2012. Is earnings quality associated with corporate social responsibility? *Account Review*. 87 (3), 761–796.

Kvamme, H., Sellereite, N., Aas, K., & Sjursen, S. (2018). Predicting mortgage default using convolutional neural networks. *Expert Systems With Applications*, 102, 207-217.

Komarova Loureiro, Y. and Gonzalez, L. (2015). Competition against common sense. *International Journal of Bank Marketing*, 33(5), pp.605-623.

Koju, L., Koju, R., & Wang, S. (2018). Macroeconomic and Bank-Specific Determinants of Non-Performing Loans: Evidence from Nepalese Banking System. *Journal Of Central Banking Theory And Practice, 7(3), 111-138.*

Kruppa, J., Liu, Y., Biau, G., Kohler, M., König, I. R., Malley, J. D. and Ziegler, A. (2014). Probability estimation with machine learning methods for dichotomous and multi-category outcome: theory. *Biometrical Journal* **56**, 534– 563.

Lacy, P., Hayward, R., (2011). A new era of sustainability in emerging markets? Insights from a global CEO study by the United Nations Global Compact and Accenture. *Corp. Gov*. 11, 348–357.

Lenz, R. (2016). Peer-to-Peer Lending: Opportunities and Risks. *European Journal of Risk Regulation*, 7(04), pp.688-700.

Li, H., Zhang, Y., Zhang, N. and Jia, H., (2016). Detecting the Abnormal Lenders from P2P Lending Data. *Procedia Computer Science*, 91, pp.357-361.

Li, F., Morgan, K. and Zaslavsky, A., (2017). Balancing Covariates via Propensity Score Weighting. *Journal of the American Statistical Association. 113(521), pp.390-400.*

LI, S., SHIUE, W. and HUANG, M. (2006). The evaluation of consumer loans using support vector machines. *Expert Systems with Applications*, 30(4), pp.772-782.

Li, X., Luo, J., & Chan, K. (2018). Political uncertainty and the cost of equity capital. *Finance Research Letters*, *26*, 215-222.

Li, Z., Li, K., Yao, X., & Wen, Q. (2018). Predicting Prepayment and Default Risks of Unsecured Consumer Loans in Online Lending. *Emerging Markets Finance And Trade*, *55*(1), 118-132.

Li, Z., Tian, Y., Li, K., Zhou, F., & Yang, W. (2017). Reject inference in credit scoring using Semi-supervised Support Vector Machines. *Expert Systems With Applications*, 74, 105-114.

Lin, K., & Dong, X. (2018). Corporate social responsibility engagement of financially distressed firms and their bankruptcy likelihood. *Advances In Accounting*, *43*, 32-45.

Liu, J., Zhang, S. and Fan, H., (2022). A two-stage hybrid credit risk prediction model based on XGBoost and graph-based deep neural network. *Expert Systems with Applications*, 195, p.116624.

Lundberg, S. and Lee, S., (2017). A Unified Approach to Interpreting Model Predictions.

Luo, D., 2022. ESG, liquidity, and stock returns. *Journal of International Financial Markets, Institutions and Money*, 78, p.101526.

Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q., & Niu, X. (2018). Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. *Electronic Commerce Research And Applications*, 31, 24-39.

Mahmoud, M., Blankson, C., & Hinson, R. (2017). Market orientation and corporate social responsibility: towards an integrated conceptual framework. *International Journal Of Corporate Social Responsibility*, *2*(1).

Malekipirbazari, M. and Aksakalli, V. (2015). Risk assessment in social lending via random forests. *Expert Systems with Applications,* 42(10), pp.4621-4631.

Michael Funke, Xiang Li, Andrew Tsang, (2019), Monetary policy shocks and peer-to-peer lending in China. *Bank of Finland Institution For Economies in Transition 23.2019*.

Mikolov T., Chen K., Corrado G., and Dean J., (2013), Efficient estimation of word representatives in vector space. CoRR, abs/1301.3781,.

Morse, A. (2015). Peer-to-Peer Crowdfunding: Information and the Potential for Disruption in Consumer Lending. *SSRN Electronic Journal*, 7, 463-482.

Nikolic, N., Zarkic-Joksimovic, N., Stojanovski, D., & Joksimovic, I. (2013). The application of brute force logistic regression to corporate credit scoring models: Evidence from Serbian financial statements. *Expert Systems With Applications*, 40(15), 5932-5944.

Ng, A., & Rezaee, Z. (2015). Business sustainability performance and cost of equity capital. *Journal Of Corporate Finance, 34*, 128-149.

Ohlson, J., & Juettner-Nauroth, B. (2005). Expected EPS and EPS Growth as Determinantsof Value. *Review Of Accounting Studies*, *10*(2-3), 349-365.

Oikonomou, I., Brooks, C., & Pavelin, S. (2012). The Impact of Corporate Social Performance on Financial Risk and Utility: A Longitudinal Analysis. *Financial Management*, *41*(2), 483-515.

Otchere, D., Arbi Ganat, T., Gholami, R., & Ridha, S. (2020). Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ANN and SVM models. *Journal Of Petroleum Science And Engineering, 108182.*

Othman, N., Faiz, R. and Smaïli, K., (2019). Enhancing Question Retrieval in Community Question Answering Using Word Embeddings. *Procedia Computer Science*, 159, pp.485-494.

Pal, R., Kapali, S., & Trivedi, S. (2020). A Study on Credit Scoring Models with different Feature Selection and Machine Learning Approaches. *SSRN Electronic Journal*. doi: 10.2139/ssrn.3743552

Paligorova, T. and Santos, J. (2017). Monetary policy and bank risk-taking: Evidence from the corporate loan market. J*ournal of Financial Intermediation*, 30, pp.35-49.

Pan, S. and Zhou, S., (2019). Evaluation Research of Credit Risk on P2P Lending based on Random Forest and Visual Graph Model. *Journal of Visual Communication and Image Representation*, p.102680.

Pandey, M., Mittal, M. and Subbiah, K., (2021). Optimal balancing & efficient feature ranking approach to minimize credit risk. *International Journal of Information Management Data Insights*, 1(2), p.100037.

Polena, M. and Regner, T., (2018). Determinants of Borrowers' Default in P2P Lending under Consideration of the Loan Risk Class. *Games*, 9(4), p.82.

Quercia, R., Pennington-Cross, A., & Yue Tian, C. (2012). Mortgage Default and Prepayment Risks among Moderate- and Low-Income Households. *Real Estate Economics*, 40, S159-S198.

Rivai, Veithzal dan Permata, Andria V. (2007), Credit Management Handbook: Teori, Konsep, Prosedur, dan Aplikasi Panduan Praktis Mahasiswa, Bankir, dan Nasabah (Credit

Manegement Handbook: Theory, Concept, Procedures and Application - Practical Guide for Students Banks and Customers), PT RajaGrafindo Persada, Jakarta, Indonesia.

Rjiba, H., Saadi, S., Boubaker, S., & Ding, X. (2021). Annual report readability and the cost of equity capital. *Journal Of Corporate Finance*, *67*, 101902.

Romano, J., & Wolf, M. (2016). Resurrecting Weighted Least Squares. *SSRN Electronic Journal*.

Saci, F., & Jasimuddin, S. (2021). Does the research done by the institutional investors affect the cost of equity capital?. *Finance Research Letters*, 41, 101834.

Sanse, K. and Sharma, M. (2015). Clustering methods for Big data analysis. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 4(3), pp.642-648.*

Stellner, C., Klein, C., & Zwergel, B. (2015). Corporate social responsibility and Eurozone corporate bonds: The moderating role of country sustainability. *Journal Of Banking & Finance*, *59*, 538-549

Seltzer, L., Starks, L., & Zhu, Q. (2020). Climate Regulatory Risks and Corporate Bonds. *SSRN Electronic Journal*.

Serrano-Cinca, C., Gutiérrez-Nieto, B. and López-Palacios, L. (2015). Determinants of Default in P2P Lending. *PLOS ONE, 10(10), p.e0139427.*

Shanaev, S. and Ghimire, B., 2022. When ESG meets AAA: The effect of ESG rating changes on stock returns. *Finance Research Letters*, 46, p.102302.

Sharma, D., (2011). Improving the Art, Craft and Science of Economic Credit Risk Scorecards Using Random Forests: Why Credit Scorers and Economists Should Use Random Forests. *SSRN Electronic Journal*.

Singh, P., Uparna, J., Karampourniotis, P., Horvat, E., Szymanski, B., Korniss, G., Bakdash, J. and Uzzi, B. (2018). Peer-to-peer lending and bias in crowd decision-making. *PLOS ONE*, 13(3), p.e0193007.

Sowers, D. and Durand, D., 1942. Risk Elements in Consumer Instalment Financing. *Journal of Marketing*, 6(4), p.407.

Srivastava, A., Kumar, Y., & Singh, P. (2020). Computer aided diagnostic system based on SVM and K harmonic mean based attribute weighting method. *Obesity Medicine, 19, 100270*.

Štrumbelj, E. and Kononenko, I., (2013). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3), pp.647-665

Suki, Ryoichi P. (2010), Definisi Bank (Bank Definition), [Online], Diakses dari: http://www.scribd.com/doc/11320386/Definisi-Bank

Tanasković, S., & Jandrić, M. (2015). Macroeconomic and Institutional Determinants of Non-performing Loans. *Journal Of Central Banking Theory And Practice, 4(1), 47-62.*

Tao, Q., & LIN, Z. (2016). Who Can Get Money? Evidence from the Chinese Peer-To-Peer Lending Platform. *SSRN Electronic Journal*.

Thomas, L., (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, 16(2), pp.149-172.

Tsai, C. and Chen, M. (2010). Credit rating by hybrid machine learning techniques. *Applied Soft Computing*, 10(2), pp.374-380.

Tuttle, J., Blackburn, L., & Powell, K. (2020). On-line classification of coal combustion quality using nonlinear SVM for improved neural network NOx emission rate prediction. *Computers & Chemical Engineering, 141, 106990*.

Van Campen, T., Hamers, H., Husslage, B. and Lindelauf, R., 2017. A new approximation method for the Shapley value applied to the WTC 9/11 terrorist attack. *Social Network Analysis and Mining*, 8(1).

Vogiazas, S., & Nikolaidou, E. (2011). Investigating the Determinants of Nonperforming Loans in the Romanian Banking System: An Empirical Study with Reference to the Greek Crisis. *Economics Research International, 2011, 1-13.*

Wang, H., Yu, M. and Zhang, L. (2019). Seeing is important: the usefulness of video information in P2P. *Accounting & Finance, 59(S2),* pp.2073-2103.

Wang, K., Kartika, F., Wang, W., & Luo, G. (2021). Corporate social responsibility, investor protection, and the cost of equity: Evidence from East Asia. *Emerging Markets Review*, *47*, 100801.

Wei, Z. and Lin, M. (2017). Market Mechanisms in Online Peer-to-Peer Lending. *Management Science*, 63(12), pp.4236-4257.

West, D. (2000). Neural network credit scoring models. *Computers & Operations Research*, 27(11-12), 1131-1152.

Wong, C. and Eng, Y., 2020. P2P finance and the effectiveness of monetary controls. *The Manchester School*, 88(4), pp.617-639.

Wong, J. and Zhang, Q., (2022). Stock market reactions to adverse ESG disclosure via media channels. *The British Accounting Review*, 54(1), p.101045.

Xia, Y., Liu, C., Li, Y., & Liu, N. (2017). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Systems With Applications*, 78, 225-241.

Xiao, H., Xiao, Z. and Wang, Y. (2016). Ensemble classification based on supervised clustering for credit scoring. *Applied Soft Computing*, 43, pp.73-86.

Yu, H., Yang, J. and Han, J. (2003). Classifying Large Data Sets Using SVMs with Hierarchical Clusters. In: *ACM SIGKDD international conference on Knowledge discovery and data mining*. New York: Association for Computing Machinery.

Yu, L., Yue, W., Wang, S. and Lai, K. (2010). Support vector machine based multiagent ensemble learning for credit risk evaluation. *Expert Systems with Applications*, 37(2), pp.1351-1360.

Zainol, J., Nor, A., Ibrahim, S., & Daud, S. (2018). Macroeconomics Determinants of Non-Performing Loans in Malaysia: An ARDL Approach. *International Journal Of Academic Research In Business And Social Sciences, 8(10).*

Zhang, D., Bhandari, B. and Black, D., (2020). Credit Card Fraud Detection Using Weighted Support Vector Machine. *Applied Mathematics*, 11(12), pp.1275-1291.

Zhang, H., Jiang, L., & Yu, L. (2021). Attribute and instance weighted naive Bayes. *Pattern Recognition, 111, 107674*.

Zhang, H., Mao, J., Qi, H., Xie, H., Shen, C., Liu, C., & Ding, L. (2020). Developing novel computational prediction models for assessing chemical-induced neurotoxicity using naïve Bayes classifier technique. *Food And Chemical Toxicology, 143, 111513*.

Zhang, T., Ramakrishnan, R. and Livny, M. (1996). BIRCH: An efficient data clustering method for very large database. *ACM SIGMOD Record*, 25(2), pp.103-114.

Zhang, T., Ramakrushnan, R. and Livny, M. (1997). BIRCH: A New Data Clustering Algorithm and Its Applications. *Data Mining and Knowledge Discovery*, 1, pp.141–182 (1997).

Zhang, Y., Li, H., Hai, M., Li, J. and Li, A., (2017). Determinants of loan funded successful in online P2P Lending. *Procedia Computer Science*, 122, pp.896-901.

Zhao, Z., Xu, S., Kang, B., Kabir, M., Liu, Y. and Wasinger, R. (2015). Investigation and improvement of multi-layer perceptron neural networks for credit scoring. *Expert Systems with Applications, 42(7),* pp.3508-3516.