ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master Thesis Data Science and Marketing Analytics

Unlocking Investment Potential in Peer-to-Peer Lending: Integrating Machine Learning Insights for Informed Decision Making

Student name: Harjiv Singh

Student ID number: 616925

Supervisor: Bas Donkers

Second assessor:

Date Final Version:

Abstract

This study presents a comprehensive exploration of the intricate dynamics within the Peer-to-Peer (P2P) lending landscape, with a focus on enhancing the predictive power of machine learning models for loan repayment classification. Rooted in real-world data sourced from lendingclub.com, the research endeavours to establish a robust predictive model capable of effectively categorizing loan repayment probabilities. The core objective of this thesis centres around the facilitation of informed investment strategies by delicately balancing the interplay between risk and return.

Central to this research is the development of a sophisticated predictive model that amalgamates loan-specific attributes and borrower characteristics to refine the precision and reliability of loan repayment classification, specifically within the context of risky loan categories. Through the harmonization of these multifaceted variables, the study aspires to construct a nuanced framework for gauging the likelihood of loan repayment, thus furnishing a platform for astute investment decision-making.

Concomitantly, the investigation delves into the depths of machine learning methodologies, serving as guiding compasses for the analysis process. The research navigates through critical methodologies tailor-fitted to the intricate challenges posed by loan classification.

Furthermore, this research assumes a dual mantle by not only exploring predictive modelling but also aiming to bridge the chasm of interpretability inherent in enigmatic black box models. Employing techniques such as partial dependence plots, the inquiry strives to illuminate the intricate nexus between influential attributes and model predictions. Through the revelation of these concealed relationships, the research seeks to distil invaluable insights to foster potent investment strategies.

Ultimately, the research endeavours to discern optimal investment strategies by enhancing the interpretability of black box models, allowing the method most aligned with analysis objectives to emerge as the preferred choice. This study thus contributes to the burgeoning field of investment decision-making within the dynamic domain of P2P lending, setting the stage for future scholars to delve deeper into the multifaceted landscape of investment strategies and P2P lending dynamics.

Table of Contents

Introduction

The emergence of peer-to-peer (P2P) lending represents a contemporary financial phenomenon that has redefined traditional borrowing and lending dynamics. Operating on a premise of direct connection between borrowers and individual investors through online platforms, this innovative approach circumvents the conventional financial intermediary, offering a more accessible funding solution for borrowers while concurrently presenting novel investment diversification avenues for investors.

In recent years, the steadfast establishment of P2P lending within the transformative financial paradigm has revolutionized traditional lending dynamics by harnessing the capabilities of digital platforms. This evolution has been significantly expedited by technological advancements, resulting in decentralized systems that prioritize transparency—a quality both borrowers and investors value. Notably, the effective utilization of diverse datasets for assessing borrower creditworthiness has propelled these platforms, catalysing investment opportunities for stakeholders.

Nevertheless, P2P lending platforms have encountered persistent challenges attributed to information asymmetry restrictions inherent to their environments. Elevated information asymmetry within the P2P lending market emanates from multifaceted factors, including privacy constraints concerning lenders and borrowers. This information disparity impedes both parties; lenders remain unaware of investor information requisites, while investors face obstacles accessing and leveraging relevant data. Empirical research by Serrano-Cinca et al. (2015) further substantiates this assertion, underscoring the scarcity of information within platforms such as lendingclub.com.

Oftentimes, investors encounter detours due to inadequate information and uncertainty regarding its utilization. This thesis assumes significance as it endeavours to bridge this gap by discerning the borrower and loan characteristics influencing loan repayment probabilities in the context of P2P loans on the Lending Club platform. Given Lending Club's prominent position within the North American P2P market, this research contributes to a comprehensive understanding of the broader P2P lending landscape.

As a relatively nascent financial domain, P2P lending is characterized by dynamic evolution influenced by multifarious factors often overlooked. This thesis seeks to mitigate this gap by intertwining investor participation within P2P lending, addressing the alignment of investment strategies with the distinctive dynamics of this industry. In this context, balancing the inherent risk-return trade-off acquires paramount importance for investors navigating the intricate P2P lending environment. This study also aims to elucidate the intricate connection between risk, return, and investor strategies within the P2P lending realm, thereby enriching the understanding of this evolving financial landscape.

Research Objective:

In view of the above introduction of existing challenges within the P2P lending landscape, this study embarks on an endeavour to analyse potential pathways for enhancing the efficacy of machine learning models in predicting loans with a high likelihood of repayment. Grounded in real-world data sourced from lendingclub.com, the fundamental aspiration is to cultivate a robust predictive model capable of effectively classifying loan repayment probabilities. The overarching purpose of this thesis lies in facilitating the formulation of sound investment strategies through the discernment of a comprehensive risk-return trade-off.

Central to this research is the development of an intricate predictive model that synthesizes both loan-specific attributes and borrower characteristics, thereby elevating the precision and dependability of loan repayment classification, particularly within the ambit of risky loan categories. By harmonizing these diverse variables, the study strives to construct a more intricate and efficacious framework for evaluating the prospects of loan repayment, thereby furnishing a platform for well-informed investment decisions.

In parallel, this inquiry endeavours to delve into the depths of machine learning methodologies, serving as guiding tools for the analysis. Specifically, the study seeks to navigate and apply pivotal methodologies that resonate with the challenges of loan classification.

Furthermore, this study assumes a dual role in not only exploring predictive modelling but also aiming to bridge the interpretability gap inherent in enigmatic black box models. Through the employment of techniques such as partial dependence plots, the research seeks to illuminate the intricate interplay between influential attributes and model predictions. By

unravelling these concealed relationships, the study aspires to distil invaluable insights for devising potent investment strategies.

Ultimately, the final objective is to identify a profound understanding of optimal investment strategies, aided in the interpretability of black box models, by allowing the method most aligned with the analysis objectives to emerge as the preferred choice.

Literature Review

Information Asymmetry in P2P Market

In the ever-evolving landscape of peer-to-peer (P2P) lending, the phenomena of information asymmetry and the dearth of comprehensive information have emerged as formidable obstacles. Information asymmetry creates a scenario wherein one party possesses a greater or more superior body of knowledge compared to the other party involved in a transaction. Within the world of P2P lending, lenders frequently encounter restricted access to complete and dependable information about borrowers, thereby impeding their ability to accurately evaluate creditworthiness and make judicious investment choices. In a study conducted by (Chen et al., 2021) it was found that information asymmetry represents a fundamental challenge in P2P lending, as it may yield inaccurate estimations of default risk. The paucity of information gives rise to uncertainties, heightens risk levels, and necessitates the formulation of effective strategies to mitigate potential adverse outcomes. Notably, (Serrano-Cinca et al., 2015)reinforces the criticality of information asymmetry within the P2P lending landscape and emphasize the imperative for P2P lending platforms to furnish prospective lenders with comprehensive borrower information, encompassing pertinent details regarding loan purpose. The present master thesis aims to delve into the assessment of risk factors and the development of strategies aimed at addressing the challenges posed by information asymmetry and information scarcity in the P2P lending market.

The success of the P2P lending market hinges upon the efficient utilization of diverse data points, enabling lenders to assess borrower creditworthiness and make well-informed investment decisions. However, the market is plagued by a deficiency of comprehensive and transparent information, impeding lenders' ability to accurately evaluate risk and mitigate potential losses. In a study conducted by (Serrano-Cinca et al., 2015), it was found that factors such as loan purpose, annual income, current housing situation, credit history, and indebtedness significantly impact default in P2P lending, further underscoring the importance of information utilization for informed investment decisions. This master thesis seeks to explore the utilization of available information within the P2P lending market and propose strategies to mitigate the adverse effects of information gaps. Through an examination of existing research, analysis of data patterns, and the adoption of innovative approaches, this study aims to contribute to the understanding of how information can be effectively leveraged and how the detrimental consequences of information asymmetry can be minimized.

Ultimately, this research endeavours to provide valuable insights for lenders, borrowers, and industry practitioners alike.

Tackling research Objectives

The most important thing to consider here is the utilization of relevant literature and how it could aid our research inquiry. In order to get a holistic view of our problem we need to consider the following dimensions

How does the predictive accuracy and robustness of traditional logistic regression models compare to that of advanced machine learning methods, specifically random forests, and Gradient Boosting Machines (GBM), in the context of classifying loan repayment within the domain of risky loan categories?

Numerous studies have been undertaken that employ machine learning and logistic regression techniques in the realm of peer-to-peer (P2P) lending, with the overarching objective of enhancing risk assessment and guiding investment decisions. Notably, one such investigation conducted by (Serrano-Cinca et al., 2015) employs a multifaceted approach encompassing hypothesis testing, survival analysis, and logistic regression analysis to develop a comprehensive understanding of default probability in P2P lending. The study successfully identifies a range of factors that significantly contribute to loan default in this context, encompassing both borrower-specific characteristics, loan-related attributes, and prevailing macroeconomic conditions. Furthermore, corroborating these findings, (Anand et al., 2022) conducted a separate study and identified similar key variables, including credit score, loan amount, interest rate, loan duration, and employment status, as pivotal predictors of loan behaviour in P2P lending settings by utilising decision trees, random forest, and bosting techniques. They found that machine learning models outperform simple logistic regression. These research findings collectively shed light on the intricate dynamics of default risk established with the help of a more nuanced machine learning model and providing valuable insights for lenders and investors operating within the P2P lending domain.

Meanwhile, the research conducted by (Serrano-Cinca et al., 2015) unveils the practice of assigning a grade to each loan within peer-to-peer (P2P) lending platforms, drawing upon third-party information such as the widely utilized FICO score employed by conventional banks and credit grantors. This grading system adopted by P2P lending platforms serves as an

evaluative measure for the borrower's creditworthiness and associated risk level. The grading process typically incorporates a multifaceted analysis of various factors, including credit score, debt-to-income ratio, employment status, loan amount, interest rate, and loan purpose. It is noteworthy that these grading systems will be further explored in-depth in relation to the second research question. Notably, a higher grade assigned to a loan signifies a lower risk of default. Although the grade assigned by the P2P lending site stands as the most predictive factor of default, the accuracy of the model is enhanced by the inclusion of additional information, with particular emphasis on the borrower's debt level. Echoing these findings, (Emekter et al., 2015) also highlight the crucial role played by the credit grade assigned by the platform as the primary determinant in predicting loan performance, closely followed by the borrower's debt-to-income ratio and the loan amount. Moreover, their study underscores the significance of credit grade, debt-to-income ratio, FICO score, and revolving line utilization as influential factors in loan defaults. These researches highlight the aspect of high risk loans accentuated by their specific credit grading

In a study conducted by (Lin et al., 2017), it was observed that certain borrower characteristics have a significant influence on loan behaviour in peer-to-peer (P2P) lending. Specifically, borrowers with a higher credit score, higher income, and longer loan duration were found to have a lower risk of default. Conversely, borrowers with a higher loan amount and higher debt-to-income ratio were associated with a higher default risk. Notably, the borrower's age and gender were found to be insignificant factors in determining default risk. (Lin et al., 2017) employed a logistic regression model to quantitatively examine the impact of various characteristics on default risks within the Chinese P2P lending market.

However, another study conducted by (Ma et al., 2018) yielded contrasting findings which investigated the prediction of defaults in P2P lending using different machine learning algorithms, specifically LightGBM and XGboost. They concluded that loan details and financial status emerged as the two most crucial factors in predicting defaults, highlighting the importance of machine learning methods comparing to logistic regression.

The research aims to enhance loan repayment classification by integrating loan and borrower features, facilitating more informed investment decisions.

Given the disparity in findings between studies employing different methodologies, the present thesis aims to contribute to the existing literature by conducting a quantitative assessment of the impact of loan and borrower characteristics on successful probability of loan repayment in P2P lending. The objective is to identify the underlying factors that influence loan repayment in a comprehensive manner, incorporating a combination of personal and loan-related variables. By utilizing machine learning methods similar to (Anand et al., 2022; Chen et al., 2021; Lin et al., 2017). Through this analysis, valuable insights will be gained regarding the interplay between borrowers' individual attributes and loan-specific factors, ultimately advancing our knowledge in this field.

The (IBISWorld, 2023) report has identified the peer-to-peer (P2P) lending industry as the fastest-growing sector in the United States, highlighting its significance for research purposes. Consequently, this thesis will focus on investigating the US market to contribute to the existing body of literature. To achieve this objective, the study will utilize data from LendingClub.com, one of the largest P2P lending platforms in the US. By leveraging this dataset, which serves as a valuable source of information, the research aims to advance our understanding of risk assessment practices within the context of P2P lending, with a specific emphasis on the US market.

Through the analysis of the LendingClub.com data, this study endeavours to address the informational limitations encountered in less developed countries by providing insights derived from a prominent platform operating within a highly dynamic and rapidly growing market. By focusing on the US market, where the P2P lending industry has demonstrated remarkable growth, the research aims to contribute to the existing literature by exploring risk assessment practices, identifying key factors influencing default risk, and shedding light on the mechanisms underlying the success of P2P lending platforms in the US context.

Prior studies have underscored the significance of borrower characteristics, loan details, and grading systems in the prediction of default risk within the peer-to-peer (P2P) lending context. Nevertheless, the literature presents inconsistent findings regarding the specific risk factors and their relative importance. To achieve effective risk mitigation, it is evident that a hybrid analytical approach combining various techniques is essential. This notion is supported by the investigation conducted (de Castro Vieira et al., 2019), wherein the authors propose the adoption of both statistical and machine learning-based methods by banks for predicting

customer defaults in their regular business operations. On the contrary, (Kim & Cho, 2019b) focuses on prediction of successful repayment of loans accentuated by machine learning based methods. This research centres on the extraction of effective features derived from borrower information and loan product characteristics. To achieve this objective, a deep dense convolutional network is employed as the foundational model.

The aforementioned studies have highlighted intricate details involved in predictions of successful repayment based of highly sophisticated machine learning models. However, there exists a gap in the literature regarding the comprehensive assessment random forest, GBM models which can predict loan repayment, meanwhile they are successfully adopted in default prediction by (Anand et al., 2022; Chen et al., 2021). Meanwhile (de Castro Vieira et al., 2019) has demonstrated the efficacy of random forests, decision trees feature elimination, and random oversampling in predicting creditworthiness within a retail credit bank dataset. Nonetheless, our investigation is directed towards the domain of loan repayment, aligning with the objectives pursued by researchers such as (Brown & Zehnder, 2007; Kim & Cho, 2019b), who similarly sought to predict the likelihood of successful loan repayment. In this endeavour, we will adopt the underlying framework utilized by these scholars for predicting loan repayment outcomes. It's noteworthy that the methodologies employed in their studies exhibit a degree of complexity beyond the scope of our analysis. Therefore, we will leverage the machine learning approaches recommended by researchers who have focused on default prediction.

Through the utilization of comparable classification techniques such as logistic regression, random forest and gradient boosting method, this research seeks to identify the key risk factors and determine their relative importance in P2P loan investments. The findings of this study will make a valuable contribution to the existing literature by offering insights that can aid investors in effectively mitigating risks and enhancing the overall stability and sustainability of the P2P lending market. By comprehensively assessing risk factors and developing investment strategies for risk mitigation, this research endeavours to improve the decision-making process for P2P loan investments and provide valuable guidance for industry stakeholders.

The research aims to enhance the interpretability of black box models by employing methods like partial dependence plots, thus facilitating a deeper comprehension of the connections between influential features and model predictions.

Concurrently, ascertaining the most suitable models to underpin our analysis assumes paramount significance. However, our exploration does not halt at the model selection stage; rather, it extends to the investigation of the intricate connection between pivotal features and the predictive likelihood of loan repayment. The work (Ariza-Garzon et al., 2020), who ventured into the realm of partial dependence plots as a tool to unravel the opacity shrouding black box models and their interpretability. In their study, partial dependence plots emerged as a potent mechanism for unearthing features of heightened significance within the predictive framework. This methodological trajectory aligns with our pursuit of comprehending the nuanced interplay between influential features and the probability of successful loan repayment.

This thesis endeavours to comprehensively examine and analyse the credit scoring mechanism implemented by LendingClub, an online peer-to-peer lending platform. LendingClub employs a distinctive grading system to assess the inherent risks associated with loans, assigning distinct grades spanning from "A" (reflecting minimal risk) to "E" (signifying elevated risk) to individual loans. Integral to this grading system is the concurrent determination of corresponding interest rates, which align with the assessed risk level. In this context, loans categorized with lower risk grades receive preferential lower interest rates, while loans associated with higher risk grades incur comparatively higher interest rates. This dual approach not only communicates the level of inherent risk but also tailors the interest rates in line with the assigned risk grade, enhancing the transparency of the lending process.

At the heart of our investigation lies a pivotal endeavour involving the implementation of a predictive model tailored to discerning the viability of loans situated within the spectrum of high-risk lending, distinguished by elevated interest rates. This undertaking assumes a paramount significance, primarily attributable to its resonance with the strategic pursuit of investors to secure favourable returns on their investment. In their exploration of default prediction, (Polena & Regner, 2018) have consolidated the loan categories E, F, and G, which fall under the classification of very high-risk loans. Within this context, the predictive model serves as a potent tool, equipping investors with the capability to methodically assess the risk-

return interplay inherent to viable loan propositions. This assessment fundamentally empowers investors to orchestrate judicious investment decisions, wherein the inherent risk profile of the prospective viable loan is meticulously evaluated against the backdrop of the envisaged returns. In essence, this predictive model assumes a pivotal role in engendering an informed and dynamic investment landscape within the realm of high-risk loans.

The primary objective entails the development of a robust predictive framework capable of discerning between loans poised for successful repayment and those fraught with potential default. By harnessing sophisticated machine learning techniques including logistic regression and random forest, this research aspires to construct a predictive paradigm that augments the decision-making processes of investors while mitigating potential financial setbacks. Through a thorough analysis of diverse loan attributes and risk parameters, the model endeavours to pinpoint pivotal indicators instrumental in distinguishing high-risk loans with favourable repayment prospects from those warranting heightened cautionary measures.

Data

The chapter presents an inquiry into the data followed to generate findings aligned with our research inquiries. This segment encompasses aspects like research design, data collection, and data preparation. Integrating these components cohesively will steer us towards addressing the central research questions.

The significance of this chapter will manifest distinctly, as it not only underscores the vital role it plays but also serves as an initial insight into the methodical revelation of the approaches and strategies employed to unravel the intricacies within the research expedition.

LendingClub.Com Data

As previously delineated, our analysis will harness data procured from LendingClub, an online platform renowned for facilitating the convergence of borrowers seeking loans and investors seeking lending opportunities. Operating as a virtual marketplace for loans, LendingClub.com affords borrowers the ability to submit loan applications for an array of purposes, spanning debt consolidation, home enhancement, and personal expenditures. In parallel, investors wield discretion to selectively allocate investments in loans, contingent upon factors such as risk assessment and projected returns. Through seamless technological integration, LendingClub.com establishes an efficient conduit between borrowers and investors, engendering a heightened level of efficacy and accessibility within the lending process.

The dataset furnished by LendingClub possesses a notable degree of intricacy, rendering it a conspicuous entity within the domain of peer-to-peer (P2P) platforms. This dataset, publicly available and notably prevalent within the North American context, assumes prominence within our purview, considering our focus on the LendingClub market. The utilization of this comprehensive dataset is of paramount significance, given its encompassment of a diverse array of variables spanning both the attributes of loans and the characteristics of borrowers. The strategic aggregation of this dataset underscores its immediate pertinence to our research inquiry, thereby facilitating the identification of loans poised for repayment. This attempt, in turn, substantiates our overarching aim of elucidating the mechanisms that underscore optimal profit generation from P2P loans within the contextual framework of LendingClub.

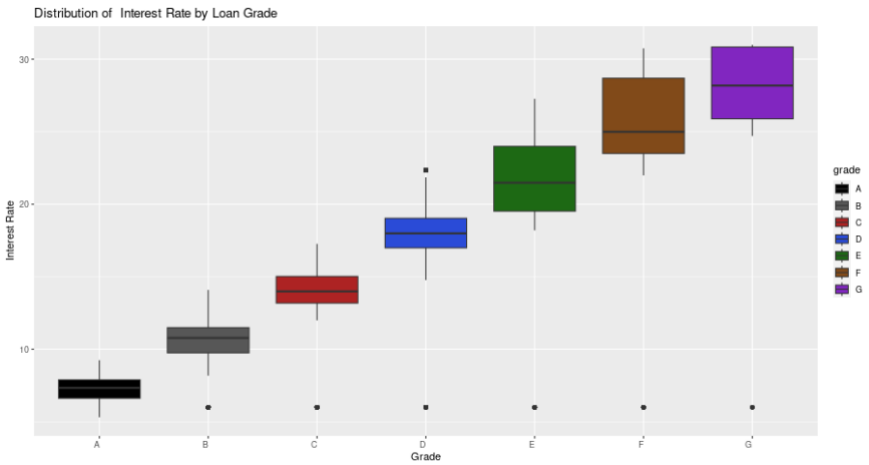
The Lending Club dataset encompasses a substantial volume of 2,260,701 entries, attesting to its considerable scale. However, certain variables within this dataset exhibit conspicuous gaps due to missing values that lack pertinence to our defined research objectives and fail to impart a quantifiable assessment of the loans under investigation. In response, our data pre-processing endeavours commence by expunging these superfluous missing values, thus aligning the dataset with the tenets of our analytical pursuit.

In tandem with these considerations, an exigent factor necessitates deliberation - the existence of outliers within the quantitative variables. Given the plausible incongruity between certain values and those of their analogously characterized counterparts, the dataset is predisposed to harbouring outliers. Mitigating the influence of these outliers is an imperative exercise. To this end, we opt for the Z-score method, a robust approach that gauges the extent of a data point's divergence from the mean. Those data points that manifest significant deviations from the mean are recognized as outliers. Our attempt to curate the dataset involves the strategic removal of these outliers, a pursuit that augments the viability of our predictive models while simultaneously amplifying the precision and reliability of our analytical assessments.

Managing extensive datasets, such as the Lending Club dataset, poses considerable complexity due to the substantial volume of information present. This abundance of data can often include elements that might be perceived as noise, consequently complicating the identification of pertinent factors aligned with our analytical approach.

In pursuit of efficiency, a viable strategy involves initially encompassing all loan and borrower characteristics that hold relevance to our analysis. After this comprehensive inclusion, undertaking exploratory data analysis across these variables is imperative. This process aids in discerning variables that contribute meaningfully to our research objectives. Consequently, a critical step involves curating a refined list of variables by eliminating those deemed irrelevant for our study.

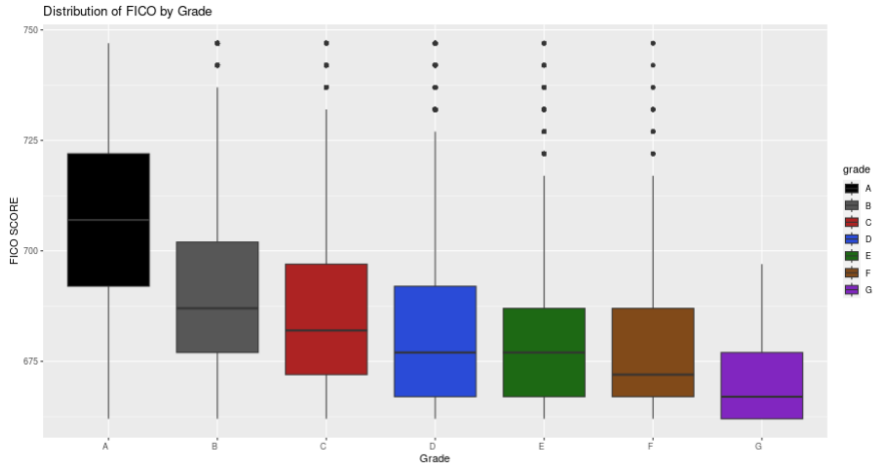
Deep dive into the features:



As previously elucidated, the incorporation of loan grade assumes a foundational role deeply enmeshed within our analytical paradigm. Evident within this context is the emergence of a discernible pattern, wherein lower loan grades consistently correspond with elevated interest rates. Consequently, the prospect emerges wherein loans of a lower grade classification may potentially engender more favourable outcomes, in contrast to their counterparts positioned within higher grade strata.

The graphical exposition furnished above substantiates our cognizance of the explicit correlation between loan grades and the accompanying interest rates. Notably discernible is a coherent trajectory evident as one traverses the hierarchy of loan grades, wherein interest rates exhibit a persistent ascendant trend.

Considering these discernments, the strategic leverage of this inference becomes an exigent imperative within our analytical construct, particularly in the context of contemplating investments within the ambit of more precarious loans. Such strategic integration enables us to harness the established interplay between loan grades and interest rates, thereby augmenting the efficacy of our decision-making endeavours within the realm of ventures characterized by higher risk-associated loans.



Based on our research findings, we have identified Loan Grade Category E as a suitable choice, effectively engaging with loans of heightened risk while maintaining a degree of stability by avoiding the extreme volatility associated with highly precarious loans. LendingClub employs a comprehensive approach in assigning grades to loans, leveraging an extensive historical dataset encompassing diverse loan categories. This practice facilitates a nuanced evaluation of repayment patterns. The platform's utilization of a proprietary credit rating system underscores their commitment to optimizing loan classification, aligning with underlying risk considerations.

The graphical representation provided above encapsulates the distribution of loan grades in relation to FICO scores. This graphical exposition facilitates a discerning differentiation between the creditworthiness attributed to LendingClub and the corresponding profile of the borrower. Echoing the findings of (Emekter et al., 2015), the inverse relationship between FICO Score and risk propensity becomes evident, with lower FICO Scores aligning with heightened risk. This alignment is validated by the observable descending trajectory of the loan grades as associated with varying FICO scores. The segmentation of loans based on their respective grades is clearly depicted in the graph above. Evidently, loan grades A through C belong to a more stable category, exhibiting relatively lower levels of risk. Separating loans of higher risk from those of lower risk can be depicted by using grade D as a demarcation point. This observation reinforces our belief that Loan Grade E represents an optimal segment to target for our analysis.

Moreover, lending platforms and financial institutions typically utilize extensive historical data to determine the performance of loans across various risk categories. By analysing the

repayment behaviour of loans categorized as A to C over time, institutions can discern patterns that suggest lower default rates and higher rates of successful repayment. This empirical analysis forms the foundation for asserting that loans falling within the A to C range are relatively less risky compared to higher-risk categories.

Analysing the feature of Loan Status provides insights into the loan's current state, encompassing loans that have been charged off, defaulted, as well as those that are currently being repaid. Our research places a substantial emphasis on loans that exhibit a likelihood of repayment. This focus serves as a foundational framework for our modelling attempt and facilitates the extraction of actionable business insights for informed investment decision-making.

Feature Engineering and Cleaning

Processing the Lending Club dataset demands meticulous data cleansing and strategic engineering efforts to attain optimal outcomes. Initial pre-processing involves translating features such as FICO score, spanning 650 to 850. To facilitate this transformation, we refer to the tabulated data presented by (Bev O'Shea & Amanda Barroso, 2023), which serves as a key resource for converting and comprehending FICO scores effectively. A similar approach was applied for "Loan Status" variable.

FICO Range	FICO Score Band
300-629	Poor
630-689	Fair
690-719	Good
720-850	Excellent

Subsequent to initial data pre-processing, the subsequent phase of data cleaning assumes pivotal significance, aiming to ensure the preservation of essential data. In this phase, a threshold of 40% data absence is set, leading to the identification and subsequent elimination of features that exceed this threshold. By attending to the inner data values prior to this elimination, the objective of minimizing data loss is upheld. As an illustrative example, the "desc" feature, accounting for loan descriptions, is identified with a substantial 92% data

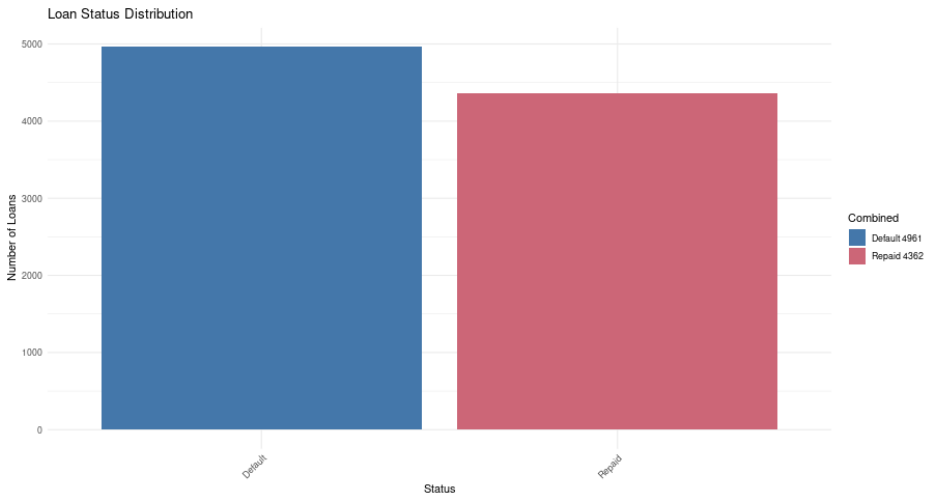
absence, prompting its removal. Following this, attention is directed towards the exclusion of annual values across all features.

The conclusive stage of data preparation is anchored in the meticulous selection of features aligned with the focal scope of our analysis, emphasizing loan and borrower characteristics of relevance. As established in our earlier review of literature, our analysis will encompass the utilization of loan and borrower characteristics previously identified in extant articles and reports. This methodical selection ensures the alignment of our investigation with established insights within the domain.

Identifying the response variable:

In our final analysis, our attention will be exclusively directed toward loans that have been successfully repaid and charged off loans. The extensive dataset encompasses merely 35 instances of defaults, thus prompting us to consider the "charged off" category of loans as the negative class. This methodological approach aids in distinguishing between successful loan repayment and potential losses. (Dorfleitner & Oswald, 2016; Emekter et al., 2015) both studies engage in a binary classification task, where the binary labels represent whether a loan has been charged off or successfully repaid. However, they operationalize the concept of "charged off" as a default category in their analyses. We have discussed a similar approach to be employed above. Similarly, (Krishna Menon & Williamson, 2018) adopt a comparable methodology by examining a binary target variable, distinguishing loans that have been successfully repaid and the ones that have defaulted.

The core premise underpinning this section is to discern predictive patterns that significantly impact the likelihood of loan repayment. Moreover, the inclusion of charged off loans enable us to make a distinction establish a good risk-return trade-off. Notably, the challenge encountered in our analysis pertains to unravelling and comprehending these intricate patterns, thus posing a significant aspect of our investigative process. A similar strategy was adopted in the study conducted by (Kim & Cho, 2019a) In their research, they focused on predicting the probability of successful loan repayment. Their chosen methodology involved utilizing a binary target variable, wherein a value of 1 denoted successful loan repayment, while a value of 0 if it’s default. This methodological choice resonates with the approach we are employing in our current study.



Within this specific classification segment, an observed distribution emerges, comprising 4362 loans that have successfully reached a state of full repayment. Concurrently, a collection of 4961 loans is classified as being in a default status. Notably, the prevalence of loans achieving complete repayment stands out in contrast to those that have defaulted. However, it is crucial to emphasize that our analytical focus remains directed solely towards loans categorized as either fully repaid or charged-off. Upon completing the data cleaning process and addressing outliers, our refined subset of loans has notably diminished in comparison to the original dataset, thus accentuating the information gaps inherent in P2P lending. However, the resultant dataset now presents a more intricate view, comprising 9324 loans that exhibit a relatively balanced distribution between charged-off and successfully repaid categories.

Identifying explanatory variables

Subsequent to the identification of the response variable, the ensuing crux of paramount significance resides in the meticulous curation of discerning loan and borrower attributes. The deliberate selection of these salient attributes is poised to underpin our analytical scaffold, facilitating a lucid exposition of the optimal investment trajectory to be pursued. By delving into the profundities of these noteworthy variables, our analytical continuum is poised to attain a heightened degree of enrichment, affording perspicuous insights into the most propitious conduits for investment to be navigated.

Loan Characteristics	Borrower Characteristics
Grade – Lending Club assigns borrower loan to seven loan grades, ranging from A to G, with A-grade indicating the lowest risk.	**FICO Range** - The FICO score range consists of categories that span from Poor to Exceptional credit.
Sub Grade - are 35 loan subgrades ranging from A1 (safest) to G5 for borrowers.	**Employment Length** - Employees' current tenure (years) with employer
Term – Duration of loan either 36 or 60 months.	**Housing Situation** - Own, rent and mortgage.
Purpose - 14 purposes for loans: wedding, credit card, car, major purchase, home improvement, debt consolidation, housing, vacation, medical needs, relocation, renewable energy, education, small business, and miscellaneous.	**Debt to Income Ratio** - Borrower's DTI ratio, monthly non-mortgage debt payments divided by reported monthly income.
Loan Amount – The listed amount of the applied loan.	**Revolving Utilisation** - the borrower's credit usage compared to available credit.
Interest Rate - Interest Rate of the loan	**Annual Income** - The yearly earnings given by the borrower upon registration.

(Lin et al., 2017) conducted an investigation aimed at discerning the influential loan characteristics shaping loan behaviour. Their findings elucidated those borrowers possessing higher credit scores, elevated income levels, and extended loan durations exhibit diminished default risk. Conversely, borrowers with larger loan amounts and heightened debt-to-income ratios are associated with heightened default risk. Additionally, (Ma et al., 2018) underscored the pivotal role played by credit scores and interest rates in determining the likelihood of default in a loan.

Diverse strategies have been adopted by researchers for feature selection and extraction prior to model training. Notably, (Polena & Regner, 2018) implemented a comprehensive feature selection approach, categorizing features into two primary information sources: Borrowers' Self-Provided Information, encompassing attributes like Annual Income, Housing Situation, Length of Employment, Loan Amount, and Loan Purpose; and Borrowers' Credit File, involving factors like Debt-to-Income ratio and Delinquency in the Past 2 Years. Similarly,

the works of (Serrano-Cinca et al., 2015) and (Emekter et al., 2015) initially employed a comparable approach of shortlisting features based on loan and borrower characteristics to predict default determinants.

A noteworthy observation is that (Polena & Regner, 2018) have included features previously discussed in the aforementioned studies for their own analysis based on their inclusion for the respective analysis. This study follows a similar approach in terms of feature selection, drawing inspiration from the methodologies employed in these research ventures..

After review and analysis of the expansive literature and scholarly articles that delve into the intricate facets characterizing both loans and borrowers, a pivotal necessity emerged – that of pinpointing the above attributes as pivotal foundation for our analytical pursuit. These attributes have been discerned to play an indispensable role, acting as navigational compasses to guide us toward the culmination of our results in a manner that is not only substantively potent but also endowed with robustness and steadfastness. In our quest to navigate the landscape of loan repayment prediction, these meticulously selected attributes have garnered prominence as foundational pillars upon which our analytical framework rests. Their strategic inclusion ensures that our analytical trajectory is imbued with the essential elements to generate outcomes that are grounded and firmly substantiated. This discerning identification process accentuates the coherence of our analytical approach, reinforcing its underpinning rationale and resolute methodology.

Methodology

In this part of the paper, we will explain the different tools and techniques we used for our study. This includes the machine learning models, the method we used to balance the data, ways to fine-tune our results, and how we measured the performance. After that, we'll outline the specific steps we followed to carry out our research. We'll make sure each step makes sense based on the methods we discussed earlier, as well as what other researchers have done in similar studies, as mentioned in the literature review section.

Logistic Regression

The statistical model known as logistic regression, also referred to as the logit model, is frequently employed for classification and predictive analytics tasks. This technique facilitates the estimation of the probability associated with a particular event occurring, such as whether an individual voted or didn't vote, based on a set of independent variables present in the dataset. Notably, the outcome is expressed as a probability, resulting in the dependent variable being constrained within the range of 0 to 1.

Logistic regression operates by applying a logit transformation to the odds ratio, which represents the likelihood of success divided by the likelihood of failure. In our research success here equates to the repayment of loan. This transformation is often called the log odds or the natural logarithm of odds.

Logistic Regression is chosen as the starting point due to its simplicity, transparency, and computational efficiency. To guard against overfitting, we employ techniques like feature selection and regularization. Feature selection helps to focus on the most relevant variables, preventing the model from fitting noise. Regularization methods, such as L1 (Lasso) and L2 (Ridge) regularization, introduce penalty terms to the model's cost function, discouraging overly complex models and thus mitigating overfitting. This ensures that the model generalizes well to new, unseen data.

Machine Learning Methods

To ensure justification to the literature review we will be employing two ensemble methods stemming from machine learning.

In the realm of predictive modelling for binary outcomes like loan repayment, decision trees stand as a foundational tool that unveils the intricate patterns hidden within the data. As we delve into the complexity of this technique, we'll explore its inner workings and relevance in the context of our loan repayment prediction thesis.

At its core, a decision tree is a hierarchical structure that simulates a series of decisions, leading to a classification or prediction. In our case, it's the binary outcome of loan repayment - whether the loan will be repaid or not.

Each decision tree consists of nodes and edges. The top node, known as the root, represents the initial decision. Subsequent nodes, called internal nodes, split the data based on a chosen feature. The leaf nodes at the bottom of the tree represents the final classifications.
They thrive on the concept of entropy, a measure of impurity within a node. The goal is to minimize entropy as we traverse down the tree. The splitting process involves selecting the feature that best separates the data, reducing uncertainty.

The Gini Impurity and Information Gain are two commonly used metrics for evaluating feature splits. The Gini Impurity measures the probability of misclassifying a randomly chosen element, while Information Gain calculates the reduction in entropy achieved by a split.

Mathematically, Gini Impurity (Gini (p)) is expressed as:

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

Where p_i is the probability of an instance belonging to class i in node p.

Information Gain $\left(IG(D_p, f)\right)$ is defined as the difference in entropy before and after the split on feature f :

$$IG(D_p, f) = \text{Entropy}\,(D_p) - \sum_{j=1}^{v} \frac{|D_{pf}|}{|D_p|} \cdot \text{Entropy}\,(D_{pf})$$

Here, D_p is the parent node, v is the number of values that feature f can take, D_{pf} is the subset of data for which feature f takes value v, and $\text{Entropy}\,(D)$ is the entropy of dataset D.

In our thesis focused on predicting loan repayment, decision trees play a pivotal role in understanding the most influential factors. For instance, a decision tree might first split based on a borrower's credit score, followed by subsequent splits concerning income, employment history, and loan amount. The tree's structure serves as an intuitive representation of how different attributes impact the likelihood of loan repayment. This interpretable model provides actionable insights for decision-makers, allowing them to tailor their strategies based on the branching paths of the tree.

In conclusion, decision trees empower us to decipher complex loan repayment dynamics. By systematically evaluating attributes and making informed splits, these trees pave the way for actionable strategies in our pursuit of accurately predicting loan repayment outcomes for an effective investment strategy.

Random Forest

Random Forest, a prominent ensemble learning technique, holds substantial significance in the realm of predictive modelling. In the context of binary loan repayment prediction, Random Forest orchestrates a collective intelligence of decision trees, embodying the fusion of predictive power and robustness.

The ensemble nature of Random Forest is rooted in its construction of multiple decision trees, each grown on a bootstrapped sample of the training data. This random sampling, along with the utilization of a subset of features for each tree split, engenders diversity among trees, thereby mitigating overfitting. The aggregation of predictions from these diverse trees culminates in a robust, well-generalizing model.

Mathematically, the prediction of a Random Forest model for a binary outcome, like loan repayment, can be succinctly captured through averaging:

$$RF_{\text{prediction}} = \frac{1}{N}\sum_{i=1}^{N} \text{Tree}_i\ (x)$$

Where N is the number of trees and $\text{Tree}_i\ (x)$ is the prediction of the i-th tree for input x.
Gradient Boosting Machine (GBM):

Random Forest is employed after Logistic Regression to harness its predictive power and address overfitting concerns. It excels at capturing complex interactions in data while reducing the risk of overfitting. During tree construction, the algorithm randomly selects a subset of features at each node, ensuring that no single feature dominates the decision-making process. By aggregating predictions from multiple trees, Random Forest effectively averages out the noise present in individual trees, resulting in a more stable and accurate model.

Random Forest's ensemble approach accommodates complex relationships by combining diverse trees' insights, collectively yielding a comprehensive understanding of these interactions.

Gradient Boosting Machine (GBM)

Gradient Boosting Machine (GBM) emerges as a powerful ensemble method, tailored for crafting predictive models with unparalleled accuracy. Embedded within the principles of boosting, GBM harnesses the cumulative potency of individual trees to establish a formidable predictive framework, particularly within the context of binary loan repayment prediction. GBM iteratively refines its predictions by sequentially constructing decision trees that focus on minimizing prediction errors from preceding trees. This iterative nature inherently enables GBM to rectify the deficiencies of prior predictions, progressively approximating the underlying data distribution. Such progressive learning is encapsulated in the formulation:

$$GBM_{\text{prediction}} = \sum_{i=1}^{N} \text{Tree}_i\ (x)$$

Where N is the number of trees and $\text{Tree}_i\ (x)$ is the prediction of the i-th tree for input x.

Research Design

In view of the methodology presented above, as well as the approaches taken by previous studies in the field, this paper will conduct the investigations our final dataset using the framework presented while describing the feature attributes

The first step of our model begins by splitting final data set into training test and validation set. our final data set contains 63952 entries which shall be divided in the following way:
Training set (80%): 7,457 data points
Validation set (10%): 933 data points.
Test set (10%): 933 data points

Splitting the dataset into training, test, and validation sets aids hyperparameter tuning for decision tree models by allowing us to iteratively adjust parameters on the training set and validate their performance on the validation set, preventing overfitting. Additionally, this splitting facilitates estimating the model's generalization performance on unseen data, ensuring reliable performance in real-world scenarios.

The initial stage of our analysis commences by implementing logistic regression, alongside ensemble methodologies such as the decision trees' random forest and Gradient Boosting Machine (GBM). This process entails the classification of loans into categories of repayment outcomes. Notably, logistic regression stands out for its computational efficiency, allowing for quicker processing of data. On the other hand, machine learning models, as highlighted by (Munkhdalai et al., 2019), yield more favourable outcomes. To streamline our analysis, we will primarily focus on comparing the accuracy achieved by logistic regression against that of our machine learning models. This comparative assessment will determine the feasibility of employing logistic regression as a suitable candidate for further exploration in subsequent stages.

The latter phase of our analysis entails the application of ensemble methods, intended to forecast loan repayment outcomes based on the attributes previously identified through logistic regression. A pivotal attempt within this phase involves a comparative assessment of the performance of these ensemble methods. This imperative comparative analysis seeks to discern which of the ensemble methods demonstrates superior predictive capabilities. By adjudicating the performance metrics, we can efficaciously extract optimal business insights

to inform our decision-making processes. This evaluative undertaking is essential in securing robust and well-informed strategic inferences, crucial for prudent and effective decision-making endeavours.

In our thesis focused on binary prediction, tuning the hyperparameters of Random Forest and Gradient Boosting Machine (GBM) involves a systematic approach. For Random Forest, parameters like the number of trees, maximum depth, and the number of features considered at each split can be adjusted using techniques like grid search or random search.

Similarly, for GBM, parameters like the learning rate, number of boosting stages, and tree depth can be fine-tuned. Utilizing techniques like cross-validation, we iteratively explore different parameter combinations, evaluating their impact on the model's performance through metrics like accuracy, precision, and recall. This meticulous hyperparameter tuning process ensures that our models are optimized for predicting binary outcomes accurately, thereby contributing to the efficacy of our loan repayment prediction strategies.

The refinement of hyperparameters within our model warrants a comprehensive comparison. In this evaluative process, we scrutinize the metrics of accuracy, precision, recall, ROC AUC score, and F1 Score. By discerning the model that yields the most favourable performance across these critical measures, we ascertain the definitive configuration of our model.

After the thorough comparison between the random forest models and the Gradient Boosting Machine (GBM) model, it becomes crucial to choose one ultimate model. This final model's assessment revolves around its ability to enhance predictive accuracy by utilizing relevant features. This, in turn, leads to the extraction of a crucial balance between risk and potential gain, which plays a pivotal role in guiding well-informed investment strategy decisions.

Evaluation Metrics

In our binary classification prediction analysis, the evaluation of the model's performance is paramount. We will utilize key metrics including accuracy, recall, F1 score, and ROC AUC Score to comprehensively assess our models' effectiveness in predicting loan repayment outcomes.

Accuracy

Accuracy measures the proportion of correctly classified instances out of the total instances. It's calculated as:

$$Accuracy = \frac{True\ Postives + True\ Negatives}{Total\ Instances}$$

It provides an overall assessment of the model's correctness in classifying both positive and negative outcomes. However, it might not be ideal for imbalanced datasets where one class significantly outweighs the other, which is often the case in binary classification.

Recall (Sensitivity or True Positive Rate)

Recall quantifies the model's ability to correctly identify positive instances among the actual positive instances. It's calculated as:

$$Recall = \frac{True\ Postives}{True\ Postives + False\ Negatives}$$

it is particularly significant when the cost of missing positive instances is high, as in loan repayment prediction. High recall implies the model effectively captures most actual positive cases, minimizing the risk of false negatives, which could lead to potential financial losses.

F1 Score

The F1 score combines precision and recall into a single metric to balance the trade-off between false positives and false negatives. It's calculated as:

$$F1\ Score = 2\ \times \frac{Precison\ \times\ Recall}{Precision + Recall}$$

F1 score accounts for both false positives and false negatives, making it an excellent metric for maintaining a balance between precision and recall. This is crucial for our analysis as we aim to optimize our model for accurate loan repayment predictions while minimizing potential errors.

True positives indicate correctly predicted instances of loan repayment, while true negatives signify accurately identified instances of non-repayment. False positives represent instances

falsely identified as repayment, and false negatives denote instances inaccurately predicted as non-repayment. This holistic perspective aids in gauging the model's precision, recall, and overall predictive effectiveness, facilitating sound investment decisions.

ROC AUC

The Receiver Operating Characteristic – Area Under the Curve (ROC-AUC) score quantifies the model's ability to differentiate between the positive and negative classes across various threshold settings. It computes the area under the ROC curve, where the ROC curve plots the True Positive Rate (Recall) against the False Positive Rate (1 - Specificity) for different classification thresholds. The AUC ROC score ranges from 0 to 1, with higher values indicating better discriminatory power. A score of 0.5 represents random guessing, while a score of 1 signifies perfect separation between the classes.

In our analysis of loan repayment prediction, the AUC ROC score holds particular significance. It provides insights into how well the model ranks positive instances above negative instances, aiding us in selecting the most appropriate threshold for classifying loans as likely to be repaid or not. By optimizing the AUC ROC score, we enhance our model's ability to differentiate between these two classes, which is of paramount importance in making informed investment decisions and managing financial risks effectively.
In essence, accuracy, recall, F1 score, and the confusion matrix collectively empower us to gauge the effectiveness of our binary classification models for loan repayment prediction. By evaluating these metrics, we can tailor our models to strike an optimal balance between minimizing financial risks and ensuring accurate predictions.

In the culmination of our analysis geared towards the identification of optimal investment strategies for loan repayment prediction, we shall employ Partial Dependence Plots (PDPs) on the predictions of our selected model. This approach holds considerable importance in deciphering trends and patterns within the variables encompassing both loan and borrower characteristics, consequently facilitating the formulation of insightful investment strategies.

Sharpe Ratio

The Sharpe Ratio, a fundamental tool in financial analysis established by renowned investment experts, plays a pivotal role in moulding our comprehension of the intricate interplay between risk and return in investment choices. Within this framework, the Sharpe Ratio functions as an encompassing perspective through which we can dissect the delicate equilibrium between investment gains and the linked risks.

At its essence, the Sharpe Ratio captures this equilibrium through mathematical representation, illustrating the surplus return garnered for each unit of risk embraced:

$$Sharpe\ Ratio = \frac{E[R_p - R_f]}{\sigma_p}$$

In this equation:

- $E\left[R_p - R_f\right]$ denotes the anticipated excess return of the investment, considering in the risk-free rate (R_f) subtracted from the projected return of the investment (R_p). According to (Roger Wohlner, 2023) the 3-month Treasury Bill Rate is often used as a proxy for risk-free rate as per the industry standard. (Board of Governors of the Federal Reserve System (US), 2023) have stipulated 5.28% as the 3-month treasury bill rate as of August 2023 which will be our risk-free rate.
- (σ_p) is the standard deviation of the investment's returns, which reflects the inherent risk of the investment.

The Sharpe Ratio, a pivotal metric in portfolio evaluation, finds application as a tool of comparison in assessing the performance of diverse investment portfolios. In the context of our analysis, this metric assumes a role of significance by offering a means to evaluate the efficacy of predictive power conferred by our machine learning models. Through the prism of the risk-return trade-off, the Sharpe Ratio allows for a systematic appraisal of the trade-offs between anticipated returns and associated risks. In our analytical context, this metric is poised to unravel distinctive vistas of investment prospects.

For the investor, the Sharpe Ratio emerges as a compass guiding toward informed decision-making. By casting a quantitative light on the balance between returns and risks, it provides the means to discern the model endowed with the most propitious risk-return equilibrium. This dual role of the Sharpe Ratio, as both an evaluator and a navigator, resonates with the

discerning investor's quest to unearth investments that manifest an optimal interplay between potential rewards and prudent risk management.

Accordingly, our methodology involves treating the predictive outcomes generated by our machine learning models as a representative portfolio which is also the test set. This approach aims to gauge the impact and viability of employing the Sharpe ratio as an evaluative metric. By integrating the Sharpe ratio within our portfolio analysis, we seek to gain a more lucid comprehension of the effectiveness of our predictions within the context of real-world borrower and loan behaviours. A heightened Sharpe ratio would signify the robustness of our predictions, indicating their alignment with actual observations.

Our fundamental aim is to replicate the machine learning predictions, rooted in loan repayment or default scenarios. To achieve this, we categorize loans into those correctly predicted as successfully repaid (True Positives) and those inaccurately labelled as successes but are, in fact, defaults (False Positives).

From an investor's perspective, the Test set serves as our portfolio. Our focus lies in investing in loans predicted as successful, which entails gains for True Positives and losses for False Positives representing the return of the portfolio (R_p). The balance between these outcomes gauges the portfolio's effectiveness, revealing how well our model performs under simulation. This exercise aids in discerning which model delivers superior returns.

The assessment of Standard Deviation (σ_p) involves these steps:

1. Calculating Individual Returns/Losses: Incorporating both True Positives and False Positives to evaluate each loan's outcome.
2. Mean Return Calculation: Determining the average return across all loans in the portfolio.
3. Variance Computation: Establishing variance by averaging the squared deviations between each loan's return and the mean return.
4. Deriving Standard Deviation: Obtaining standard deviation by taking the square root of the calculated variance.

Collectively, these steps enable us to assess portfolio performance and risk, providing insight into the effectiveness of the machine learning models.

Partial Dependence Plots

PDPs provide a comprehensive visual representation of the relationship between a specific feature and the model's predictions, while keeping other features fixed at certain values. By varying the selected feature's values and observing resulting changes in predictions, PDPs elucidate the influence of that feature on the model's outputs. These plots offer a holistic view of how specific variables impact the prediction outcomes, enabling us to comprehend the intricate interactions and trends within the dataset.

In our analysis, PDPs serve as a powerful tool for uncovering latent insights that guide our investment strategy formulation. As experts in the field, we understand that loan repayment prediction involves multifaceted variables, including borrower characteristics, loan attributes, and economic factors. By discerning how variations in these features impact the model's predictions, PDPs equip us to tailor our investment strategies to capitalize on favourable trends and mitigate risks.

Moreover, the significance of PDPs lies in their capacity to provide an interpretable bridge between complex model outputs and actionable investment decisions. Their graphical nature offers clarity to stakeholders, enabling them to grasp the implications of variables in a pragmatic manner. Consequently, we can harness these insights to refine our investment approach, optimizing our strategies for loan repayment prediction, and ultimately enhancing the efficacy of our decision-making processes.

Results

This chapter elucidates the methodologies and research design outlined in the preceding section, further divided into three distinct segments, as previously discussed. The initial segment delves into a comparative analysis between logistic regression and machine learning techniques. Subsequently, the subsequent stage entails the process of hyperparameter tuning, executed via the outlined methodology. Consequent to this, the ensuing phase embarks upon an appraisal of the models' effectiveness, culminating in the selection of an optimal model. The final phase involves a comprehensive scrutiny of the behaviours and importance of featured variables.

Logistic Regression V/s Machine Learning

In accordance with the aforementioned methodology, the ensuing section undertakes a comparative examination of logistic regression and machine learning methodologies. Specifically, we scrutinize a refined logistic regression model, emphasizing the regularization parameter (λ). Through a systematic calibration process, the optimal value of λ is established at 0.036.

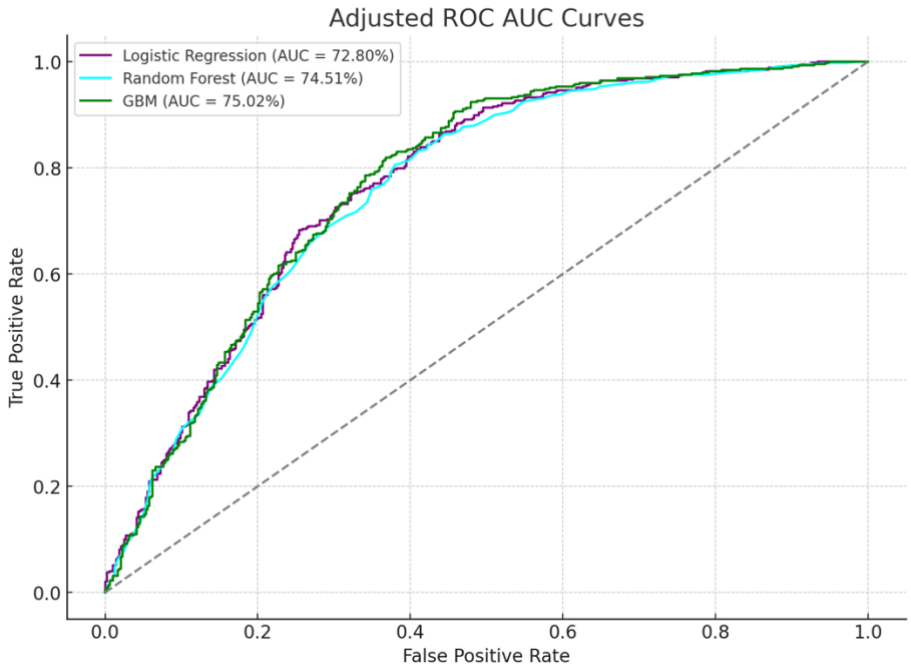
In the subsequent illustration, the subsequent table encapsulates the performance metrics of the logistic regression model, contextualized within a comparative framework vis-à-vis the random forest and GBM models. It is worth noting that the random forest and GBM models have yet to undergo hyperparameter tuning in this juncture.

	Logit	**Random Forest**	**GBM**
Accuracy	72.78%	74.38%	74.92%
Recall	74.24%	81.66%	80.57%
F1 Score	72.81%	75.79%	75.93%
ROC AUC	72.80%	74.51%	75.02%

The comparative analysis between our logistic regression model and the machine learning approach reveals notable disparities in performance. Specifically, the logistic regression model demonstrates marginally inferior performance across various evaluation metrics. Consequently, for the sake of conciseness in our analysis, we have decided to exclude this model from further consideration.

The logistic regression model exhibits shortcomings across multiple fronts, encompassing accuracy, recall, F1 score, and AUC-ROC metrics. The diminished accuracy score indicates the model's challenges in effectively distinguishing between loans that have been repaid and those that have not, considerably underperforming in relation to alternative models.

Of particular significance is the recall metric, which gauges the proportion of repaid loans correctly identified. In this aspect, the logistic regression model yields a lower score, indicating that it’s performance is not that viable in comparison to the machine learning models,



The culminating aspect of our evaluation lies in the assessment of the ROC AUC plot, a pivotal measure that comprehensively gauges model performance across an array of classification thresholds. It aids in discerning the differentiation between the probabilities of successful loan repayment and defaults. A heightened ROC AUC curve signifies the efficacy

of a model in effectively distinguishing between positive and negative classes. Upon examination, it becomes evident that the AUC value stands at a 72.80%, marginally lower than that of the random forest model (74.51%) and GBM Model (75.02%). This observation exemplifies the comparative performance of machine learning models against the conventional logistic regression technique in the context of classification.

Hyperparameter tuning the selected models:

The next section deals with the hyperparameter tuning of our models and seeing whether any major differences could be observed we shall be incorporating Sharpe ratio based on our predictions for both our models. Our Gradient Boosting Machine (GBM) model demonstrated a slightly higher Sharpe ratio of 1.2015, surpassing that of the random forest model at 1.1986.

Both models exhibit similar accuracy, with the GBM model (74.92%) slightly outperforming the Random Forest model (74.38%). Here, the GBM model is slightly better are predicting whether a borrower will repay a loan or not.

The Random Forest model has a recall of 80.57%, which is marginally lower than the GBM's recall of 81.66%, indicating GBM model identifies a marginally larger proportion of loans that were repaid.

The F1 scores are closely matched, with the Random Forest model at 75.79% and the GBM model at 75.93% suggesting both models maintain a balance between precision and recall.

Both models have a marginally similar AUC ROC score of 74.51% for the RF model and 75.02% for the GBM model showcasing their strong ability to differentiate between the successful loan repayment and default.

The GBM model's Sharpe ratio (1.2015) is marginally higher than the Random Forest model's (1.1986), indicating a slightly better risk-adjusted performance. (Nick Lioudis, 2023) explains the usage of different loan characteristics for the calculation of Sharpe Ratio which we have already discussed and explained. According to (Roger Wohlner, 2023), a Sharpe ratio below 1 is deemed unfavourable. Ratios between 1 and 1.99 are considered satisfactory or favourable, those between 2 and 2.99 are categorized as highly favourable, while ratios exceeding 3 are regarded as exceptional. As a result, both of our simulated portfolios fall within the satisfactory to favourable range. This outcome suggests that our models offer investors accurate predictions that generate commendable returns.

The models discussed above exhibit a remarkable degree of similarity in terms of their performance. Marginally, the GBM model demonstrates a slightly superior performance in

comparison to the random forest model. Yet, our focus now shifts to the exploration of hyperparameter-tuned models. The ensuing parameters emerged through training the models on the validation set for both the random forest and GBM models.

Random Forest	**GBM**
n_estimators: 100	n_estimators: 100
max_features: ‘sqrt’	learning_rate: 0.05
max_depth: 10	subsample: 0.9
min_samples_split: 5	max_features: ‘sqrt’
min_samples: 2	max_depth: 50
bootstrap: True	

Based on the aforementioned hyperparameters, the following performance metrics were derived from our models:

	Random Forest	**GBM**
Accuracy	75.67%	75.03%
Recall	83.84%	81.44%
F1 Score	77.19%	76.20%
ROC AUC	75.82%	75.14%
Sharpe Ratio	1.2405	1.2387

The presented table highlights that the GBM model's metrics exhibit minimal improvement over the default model, whereas the random forest model showcases enhancements with a raised accuracy rate (75.67%) and improved Recall (83.84%) and F1 Score (77.19%) and ROC AUC Score (75.82%). It is intriguing to observe that the finely tuned random forest model exhibits superior performance across all metrics compared to both the default GBM model and the optimized GBM model, with a marginal advantage. In the context of the risk-

return trade-off, it is noteworthy that the Sharpe Ratio of the Random Forest model (1.2405) exhibits a marginal superiority over the GBM model (1.2387). Although the discrepancy is minimal, this comparison underscores the commendable performance of both models in optimizing the balance between risk and return.

It should be noted that while both these models are very similar, we shall be picking the tuned Random Forest Model due to its higher metrics and Sharpe ratio. The utilisation of evaluation metrics offers a quantitative assessment of the model's efficacy, with the Sharpe ratio facilitating the simulation of predictions as an investment portfolio. This underscores the model's capability to enable investors to achieve favourable returns while effectively managing the trade-off between risk and reward.

A Sharpe ratio between 1 and 1.99 is an indicator of satisfactory or favourable performance, signifying an advantageous balance between risk and return. Given the commendable performance of both models, we find contentment in endorsing their consideration for further analysis. Of notable significance is the nuanced interplay between risk and return, which serves as a pivotal foundation for investment decisions. This trade-off equips investors with invaluable insights, aiding in prudent decision-making by comprehending the potential return in relation to the associated risks.

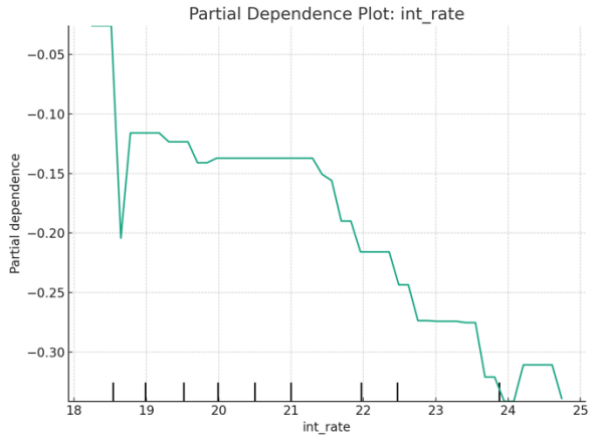
As the next phase of our analysis unfolds, we pivot our attention towards the Random Forest model, affirming its status as the final model of choice. In this ensuing stage, we delve into the exploration of future behaviours and variable importance, unravelling the intricate web of relationships between features and their influence on predictions. This investigative endeavour aims to unearth the pivotal attributes that inform our investment strategy, consequently solidifying our understanding of the mechanisms that drive optimal decisions in the realm of loan repayment classification.

Interpreting our final machine learning model:

In this segment, we delve into the exploration of feature behaviour and significance, a focal point propelled by our Random Forest model. Within this specialized phase, our focus converges on the examination of partial dependence plots, a pivotal analytical tool. As elucidated in the earlier methodology, these plots unravel the nuanced behaviours of crucial features, fostering a deeper understanding of their contributions to prediction outcomes.

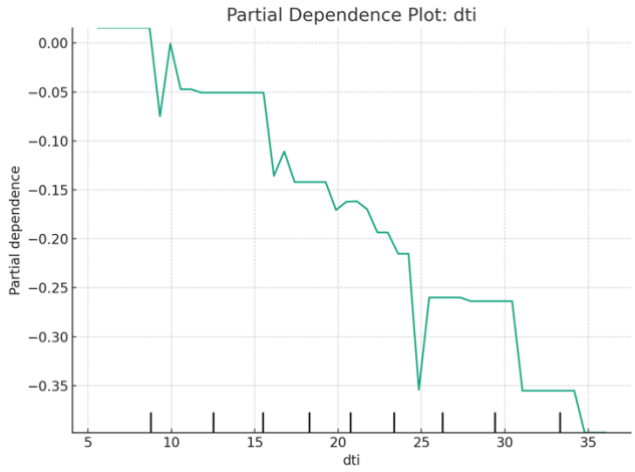
The application of partial dependence plots unveils recurrent patterns within variables or features, potentially paving the way for their strategic utilization. This process attempts to extract contextual insights from the machine learning model, thereby enriching our comprehension of the underlying relationships inherent in the data.

From our meticulous analysis, a collection of paramount variables that exert substantial influence on the likelihood of loan repayment emerges. These variables exert either a positive or negative impact, thus warranting focused attention. Below, we present the partially dependent plots pertaining to these pivotal variables, further elucidating their intricate characteristics.



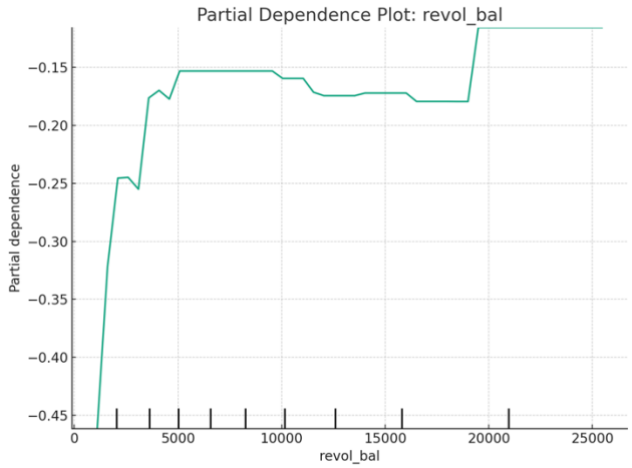
Interest Rate Interval	No. of Loans
15-20	2
20-25	4058
25-30	5263

The aforementioned partial dependence plot of Interest Rate recognises the x axis as the predicted probability of the positive class i.e., successful repayment of loan. Evident within the domain of interest rates is a notable pattern, as the probability of loan repayment consistently diminishes. This observation underscores a distinct inverse correlation between interest rates and the likelihood of accomplishing successful loan repayment. Upon investigating the relationship between interest rates and loan frequency within specific intervals, it becomes apparent that nearly all loans possess interest rates exceeding 20%, with only two exceptions observed in the 15-20% interval. Furthermore, a consistent downward trajectory is observed beyond the 21% interest rate threshold. This trend underscores the notion that elevated interest rates correlate with a higher probability of unsuccessful loan repayment.



DTI Interval	No. of Loans
0-10	1228
10-20	3141
20-30	3255
30-40	1675
40-50	24

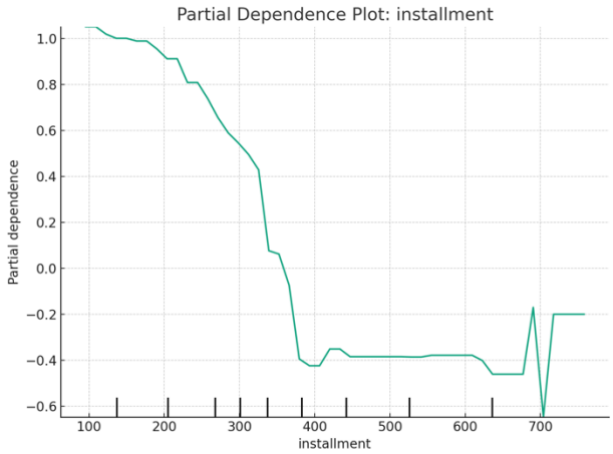
Parallel to this, an analogous trend unfolds within the debt-to-income ratio. As the ratio increases, a coherent pattern emerges, signalling a sharp decline in the likelihood of loan repayment. A sharp decline in the debt-to-income ratio is discernible until the 15% mark indicating a negative relationship with successful probability of loan repayment. It is worth noting that a notable majority of loans fall within the specified interval, thereby underscoring the substantial range that effectively contributes to the assessment of the likelihood of successful loan repayment. This distinctive pattern accentuates the profound impact exerted by entries with a debt-to-income ratio exceeding 15% on the probability of loan repayment, signifying an inverse relationship.



Revolving Utilization Interval	No. of Loans
0-5000	2789
5000-10000	2785
10000-15000	1705
15000-20000	998
20000-25000	549
25000-30000	276
30000-35000	221

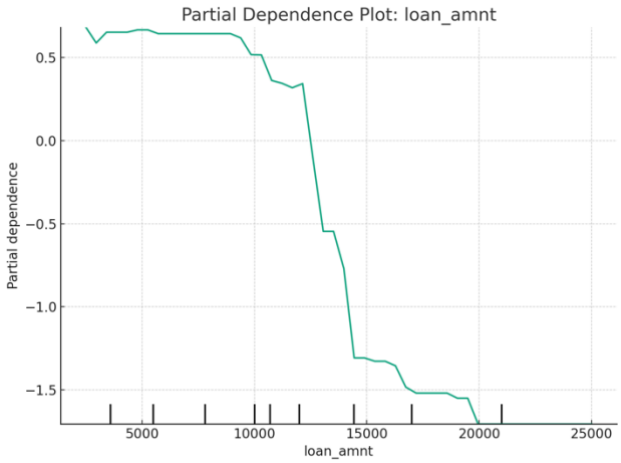
The presented partial dependence plot underscores the role of revolving utilization within the context of loan entries. This feature intertwines with the borrower's creditworthiness and their balance of available credit and demonstrates comparatively modest influence on the likelihood of loan repayment which is attributed due to its flatter curve. The interaction

between the Revolving Utilization feature and the likelihood of successful loan repayment exhibits a predominantly linear trend, primarily situated within the range of the highest loan frequency. Nevertheless, a notable proportion of loans are present within the 0-5000 interval, where a pronounced surge in the probability of successful loan repayment is evident. Aside from this significant surge, limited distinct observations can be discerned within the data.



Loan Instalment Interval	No. of Loans
0-200	1861
200-400	4135
400-600	2196
600-800	798
800-1000	315
1000-1200	18

Turning attention to instalments, a distinct pattern emerges. Upon initial inspection, the influence of instalments on the likelihood of successful loan repayment might appear minimal. Yet, a more detailed analysis of loan distribution within the 0-400 range reveals a pronounced concentration of loans in this interval. This concentration implies a distinct negative correlation between instalments and loan repayment within this interval. Thus, as the number of instalments increases, a corresponding decrease in the probability of successful loan repayment becomes evident.



Loan Amount Interval	No. of Loans
0-5000	1783
5000-10000	2594
10000-15000	2551
15000-20000	1339
20000-25000	647
25000-30000	278
30000-35000	131

The conclusive partial dependence plot centres around the listed amount of the applied loan. A salient observation within this plot underscores a distinct pattern: a pronounced decline at an intermediate point, amidst a backdrop of relatively lower loan amounts before this juncture. Based on the distribution of data points, a significant portion of loans is observed in instances where the loan amount below 15000. This observation highlights a relatively linear association between this feature and the desired probability of loan repayment. However, the abrupt decline in the curve also indicates that loans with values exceeding 15000 exert a negative influence on the successful repayment of loans. This trend sheds light on the strategic considerations that need to be factored into investment decisions, considering the pivotal role played by the magnitude of the applied loan in shaping its prospects of repayment.

Conclusion

This concluding chapter serves as a platform for comprehensive deliberation, encompassing an analysis of the results, delineating the inherent limitations of this study, and charting potential avenues for future research within the purview of the scrutinized topics.

The focal objective of this study revolved around the classification of loan repayment within the domain of risky loan categories. The investigation commenced by elucidating the challenges inherent in peer-to-peer (P2P) lending, specifically centred on the scarcity of information that both investors and lenders grapple with. In response to these challenges, this study sought to leverage the repository of information available on lendingclub.com to construct classification models. While prior research has predominantly concentrated on risk assessment and the identification of influential variables, a notable gap persisted in the form of an integrated investigation that holistically addressed both aspects. It is within this lacuna that the current study positioned itself, aiming to bridge the existing divide and provide a comprehensive analysis.

The initial facet of our analysis was dedicated to discerning the distinctions between decision tree models and logistic regression. This pursuit aimed to underscore the potential and robustness inherent in machine learning models. Our investigation effectively demonstrated that machine learning methods offer a potent alternative to conventional analytical approaches in addressing investment decisions. This affirmation of the superiority of machine learning methods emerged as a significant outcome, advocating their adoption for resolving investment-related quandaries.

The focal crux of our analytical ventures within this study revolved around the prediction of loan repayment, hinging upon the selection of the most fitting model. Through meticulous exploration, we established both the random forests and Gradient Boosting Machine (GBM) models to exhibit comparable performance. However, our preference gravitated towards the finely tuned Random Forest model, propelled by its notable achievements in terms of the Sharpe ratio, recall, and accuracy. This methodical selection attested to the rigorous evaluation process undertaken to discern the optimal model for the purpose of predicting loan repayment, a pivotal pursuit within our investigation.

The evaluation of metrics emerged as a pivotal aspect in the identification of the most suitable machine learning approach for this study, a necessity stemming from the pivotal role of metrics in validating the relevance of our models. The robustness of our chosen model hinged upon the proficient performance of these metrics, substantiating the model's efficacy in solving our research query. This performance attainment materialized under carefully defined circumstances, signifying the model's viability in practical applications. An imperative consideration in this assessment was the risk-return trade-off, a quintessential factor instrumental in illuminating loan patterns. These patterns, once deciphered, could be effectively leveraged by investors to maintain a balanced portfolio, exemplifying the synergy between predictive modelling and investment strategies.

The culminating phase of our analysis encompassed the translation of machine learning models into tangible business insights, fostering meaningful inferences. This progression bore essential significance for investors seeking to navigate the dynamic landscape of the Peer-to-Peer (P2P) lending market. Central to this pursuit was the formulation of a machine learning model capable of effectively guiding investment decisions, offering strategic guidance and promoting informed choices that aimed for favourable returns. Notably, this strategic orientation was intrinsically entwined with risk considerations, reflecting the acute awareness of the delicate balance between risk and reward in the realm of investment.

The utilization of partial dependence plots has proven to be a discerning tool in our analytical toolkit, enabling a comprehensive understanding of the underlying dynamics within our models. By examining the critical attributes such as the interest rate tied to the loan, the borrower's debt-to-income ratio, revolving credit utilization, the initial loan amount, and the associated instalments, we have deciphered their behavioural impact on the probability of loan repayment. The illumination of these feature behaviours holds immense value for investors, offering insights into maintaining an optimal risk-return equilibrium. These patterns, extracted from the partial dependence plots, hold the potential to yield favourable outcomes and enhance investment strategies.

The culmination of our study lies in its capacity to confidently guide investment endeavours in the P2P lending market. This final phase synthesizes the insights gleaned from data exploration, modelling, and interpretation. It has culminated in a robust understanding of the

investment landscape, particularly addressing the information gaps that once posed challenges. Our study's significance lies in its ability to bridge this knowledge void, providing a reliable framework for confidently navigating investment efforts in the P2P lending realm.

In summary, our study serves as an exemplar of a methodological fusion that amalgamates diverse analytical tools, encompassing intricate data analysis, rigorous modelling techniques, and nuanced interpretation. This multifaceted approach has culminated in a comprehensive investigation that has not only illuminated the central query of our research but has also positioned itself as a catalyst for future explorations. By deftly navigating the intricate landscape of investment and the P2P lending market, our study offers a comprehensive lens through which the intricate interplay between predictive modelling and investment decisions can be examined.

Our study's significance lies in its discerning ability to unravel the intricate interplay between data-driven analyses and investment strategies, thereby bridging the divide between theoretical constructs and practical implementation. The culmination of diverse resources has enabled us to probe beyond surface-level insights, delving deep into the underlying mechanics that govern the behaviour of loan repayment. By successfully addressing our study's central inquiry, our findings resonate beyond the confines of the present analysis, setting a precedent for further investigations. The pathways that have emerged from this research horizon beckon future scholars to delve deeper into the intricate landscape of investment decisions, lending a renewed impetus to empirical inquiries within the dynamic sphere of investment and the burgeoning P2P lending market.

Limitations

Notwithstanding the achievements of our study, several considerations deserve contemplation, aiming to provide a balanced assessment that does not overshadow the attained success. Our study admirably elucidates the functioning of the P2P lending platform lendingclub.com; however, it is crucial to acknowledge its specific focus, which does not encapsulate the entirety of the broader market landscape. Various factors, such as the socioeconomic context of lenders and borrowers, market regulations, and information transparency, significantly diverge across different geographic regions. Consequently, our study's findings might not seamlessly translate to encompass the multifaceted intricacies inherent in various markets.

The interpretability of machine learning models heralds a notable achievement within our study, yet it is important to recognize that such interpretability might not always furnish a comprehensive panorama of the underlying mechanisms. While we diligently harnessed the best available resources to unravel the nuanced relationships between features and predictions, it remains challenging to offer precise elucidation of the intricate dynamics governing these interactions. Indeed, our study diligently aims to present the optimal inferences drawn from our models, but the inherent complexity of predictive modelling necessitates a measure of caution in overgeneralizing interpretations.

Concurrently, our study treads lightly in terms of commenting on the causal relationships emanating from our predictive models, particularly concerning loan repayment and the associated features. A deeper exploration into causality would considerably enhance our understanding of the P2P lending market, shedding light on the causal mechanisms underpinning the observed predictions. This endeavour, albeit complex, holds the promise of yielding insights that surpass mere correlation, paving the way for a more nuanced comprehension of the underlying dynamics shaping the behaviour of loan repayment within the P2P lending ecosystem.

Future Work

The realm of the P2P lending market is notably beset with the challenge of information asymmetry. However, it is worth considering that the mitigation of this issue in the future could potentially yield improved outcomes. Subsequent analyses have the potential to unveil and establish an optimal equilibrium between adept risk management practices and strategic investment approaches.

Furthermore, the exploration of novel methodologies for data simulation, meticulously tailored for the nuances of the P2P lending market, stands as a promising avenue for future research ventures. The inherent complexity of P2P lending data, characterized by its density and susceptibility to noise, underscores the significance of new methodologies that attempts to eliminate extraneous noise while effectively retaining essential information. The culmination of such efforts could potentially lead to the development of near-ideal predictive models, subsequently unlocking new realms of analysis and insight within the domain of P2P lending.

References:

Anand, M., Velu, A., & Whig, P. (2022). Prediction of Loan Behaviour with Machine Learning Models for Secure Banking. *Journal of Computer Science and Engineering (JCSE)*, *3*(1), 1–13. https://doi.org/10.36596/jcse.v3i1.237

Ariza-Garzon, M. J., Arroyo, J., Caparrini, A., & Segovia-Vargas, M.-J. (2020). Explainability of a Machine Learning Granting Scoring Model in Peer-to-Peer Lending. *IEEE Access*, *8*, 64873–64890. https://doi.org/10.1109/ACCESS.2020.2984412

Bev O'Shea, & Amanda Barroso. (2023, June 5). *What Is a Good Credit Score and How Do I Get One?* NerdWallet. https://www.nerdwallet.com/article/finance/what-is-a-good-credit-score

Board of Governors of the Federal Reserve System (US). (2023, August 21). *3-Month Treasury Bill Secondary Market Rate, Discount Basis [DTB3], retrieved from FRED*. Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/DTB3

Brown, M., & Zehnder, C. (2007). Credit reporting, relationship banking, and loan repayment. *Journal of Money, Credit and Banking*, *39*(8), 1883–1918. https://doi.org/10.1111/j.1538-4616.2007.00092.x

Chen, Y. R., Leu, J. S., Huang, S. A., Wang, J. T., & Takada, J. I. (2021). Predicting Default Risk on Peer-to-Peer Lending Imbalanced Datasets. *IEEE Access*, *9*, 73103–73109. https://doi.org/10.1109/ACCESS.2021.3079701

de Castro Vieira, J. R., Barboza, F., Sobreiro, V. A., & Kimura, H. (2019). Machine learning models for credit analysis improvements: Predicting low-income families' default. *Applied Soft Computing Journal*, *83*. https://doi.org/10.1016/j.asoc.2019.105640

Dorfleitner, G., & Oswald, E. M. (2016). Repayment behavior in peer-to-peer microfinancing: Empirical evidence from Kiva. *Review of Financial Economics*, *30*, 45–59. https://doi.org/10.1016/j.rfe.2016.05.005

Emekter, R., Tu, Y., Jirasakuldech, B., & Lu, M. (2015). Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. *Applied Economics*, *47*(1), 54–70. https://doi.org/10.1080/00036846.2014.962222

IBISWorld. (2023). *Fastest Growing Industries in the US by Revenue Growth (%) in 2023*. IBISWorld. https://www.ibisworld.com/united-states/industry-trends/fastest-growing-industries/

Kim, J. Y., & Cho, S. B. (2019a). Predicting repayment of borrows in peer-to-peer social lending with deep dense convolutional network. *Expert Systems*, *36*(4). https://doi.org/10.1111/exsy.12403

Kim, J. Y., & Cho, S. B. (2019b). Towards repayment prediction in Peer-to-Peer social lending using deep learning. *Mathematics*, *7*(11). https://doi.org/10.3390/math7111041

Krishna Menon, A., & Williamson, R. C. (2018). The Cost of Fairness in Binary Classification. In *Proceedings of Machine Learning Research* (Vol. 81).

Lin, X., Li, X., & Zheng, Z. (2017). Evaluating borrower's default risk in peer-to-peer lending: evidence from a lending platform in China. *Applied Economics*, *49*(35), 3538–3545. https://doi.org/10.1080/00036846.2016.1262526

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. https://doi.org/10.1016/j.elerap.2018.08.002

Munkhdalai, L., Munkhdalai, T., Namsrai, O. E., Lee, J. Y., & Ryu, K. H. (2019). An empirical comparison of machine-learning methods on bank client credit assessments. *Sustainability (Switzerland)*, *11*(3). https://doi.org/10.3390/su11030699

Nick Lioudis. (2023, June 3). *Understanding the Sharpe Ratio*. Investopedia. https://www.investopedia.com/articles/07/sharpe_ratio.asp

Polena, M., & Regner, T. (2018). Determinants of borrowers' default in P2P lending under consideration of the loan risk class. *Games*, *9*(4). https://doi.org/10.3390/g9040082

Roger Wohlner. (2023, February 7). *What Is the Sharpe Ratio? Definition & Formula*. TheStreet. https://www.thestreet.com/dictionary/s/sharpe-ratio#

Serrano-Cinca, C., Gutiérrez-Nieto, B., & López-Palacios, L. (2015). Determinants of default in P2P lending. *PLoS ONE, 10*(10). https://doi.org/10.1371/journal.pone.0139427