Enhancing Creditworthiness in Peer-to-Peer Lending Using Human Centered Artificial Intelligence

W. L. P. M. Wijetunga
Index No : 19001942

Supervisor: Dr. Thilina Halloluwa

April 2024

Submitted in partial fulfillment of the requirements of the
BSc in Computer Science Final Year Project (SCS4224)

Enhancing Creditworthiness in Peer-to-Peer Lending Using Human Centered Artificial Intelligence

W. L. P. M. Wijetunga

Declaration

I certify that this dissertation does not incorporate, without acknowledgement, any material previously submitted for a degree or diploma in any university and to the best of my knowledge and belief, it does not contain any material previously published or written by another person or myself except where due reference is made in the text. I also hereby give consent for my dissertation, if accepted, be made available for photocopying and for interlibrary loans and for the title and abstract to be made available to outside organizations.

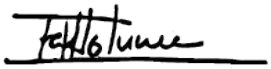
Candidate Name: W. L. P. M. Wijetunga

.....

Signature of Candidate Date: 19/04/2024

This is to certify that this dissertation is based on the work of Mr W. L. P. M. Wijetunga under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Supervisor's Name: Dr. Thilina Halloluwa



.....

Signature of Supervisor Date: 19/04/2024

Co-Supervisor's Name: Mr. Upul Rathnayake

.....

Signature of Co-Supervisor Date: 18/04/2024

Abstract

This research addresses significant gaps in the Peer-to-Peer (P2P) lending field, specifically the lack of transparency, effectiveness, human biases and fairness and high false positive rates. To tackle these issues, design science approach and research onion methodology are utilized with data from the Lending Club P2P lending company. The aim of this research is to make the process of creditworthiness in peer-to-peer lending more effective through the application of Human Centered AI. This involves identifying the most accurate Machine Learning (ML) model, determining the most interpretable eXplainable Artificial Intelligence (XAI) model, integrating both models and evaluating their effectiveness in P2P lending with a focus on interpretability and explainability. The Random Forest classifier is found to be the most accurate ML model compared to XGBoost, LLR and Classification Tree. XAI models such as SHAP, LIME and DiCE provide valuable insights into interpretability. SHAP offers global and local interpretations while LIME focuses on localized explanations. DiCE generates counterfactuals for "what-if" scenarios which help determine necessary changes to loan features. Evaluation includes quantitative metrics such as Accuracy, F1 score and AUC-ROC from ML models as well as qualitative components such as interviews and questionnaires to assess the combined ML and XAI model's effectiveness. The successful integration of an accurate ML model (Random Forest) with state-of-the-art XAI methods contributes to transparent and efficient creditworthiness assessment in P2P lending. Further research should focus on enhancing the ML and XAI framework through longitudinal studies exploring additional XAI methods across multiple P2P lending platforms. This research sets the foundation for future investigations that will advance the integration of ML and XAI in P2P lending while opening avenues for further improvement in creditworthiness assessment methodologies.

***Keywords*— Human-Centered Artificial Intelligence (HCAI), eXplainable Artificial Intelligence (XAI), AI in Finance (AIF), FinTech, Peer-to-Peer Lending (P2P)**

Acknowledgement

I would like to express my heartfelt gratitude to Dr Thilina Halloluwa, my supervisor and Mr. Upul Rathnayake, my co-supervisor, as well as Mr. Akila Gunarthna, the external advisor of the research, for their invaluable guidance and support throughout this process. Their patience and assistance have been crucial to the completion of this project and I am truly grateful for their contributions.

I also want to express my gratitude to other domain experts from the finance industry for contributing their expertise and views, which were very helpful in making this study a success.

My parents, friends and relatives have also earned my gratitude. for their unwavering love and support They helped me finish this research thesis by supporting me and believing in me. I want to thank my coworkers for their insightful criticism and assistance in improving my review.

I appreciate everyone's help and support and I hope my research will benefit the field of Human-Centered Artificial Intelligence. Thank you all for your support and guidance!

Contents

List of Figures

List of Acronyms

HCI Human Computer Interaction

AI Artificial Intelligence

HCAI Human-Centered Artificial Intelligence

XAI eXplainable Artificial Intelligence

ML Machine Learning

FinTech Financial Technology

P2P Peer-to-Peer

LR Logistic Regression

BLR Binary Logistic Regression

LRR Logistic Regression with Random coefficients

LRF Logistic Regression with Fixed coefficients

LLR Lasso-Logistic Regression

LLRE Lasso Logistic Regression Ensemble

RF Random Forest

ERT Extremely Randomized Tree

GBDT Gradient Boosting Decision Tree

XGBoost eXtreme Gradient Boosting

LightGBM Light Gradient Boosting Machine

SVM Support Vector Machine

NLP Natural Language Processing

LSTM Long-Short-Term Memory

BPSO Binary Particle Swarm Optimization

LIME Local Interpretable Model-agnostic Explanation

SHAP SHapley Additive exPlanation

SMEs Small and Medium-sized Enterprises

EECAI European External Credit Assessment Institution

1 Introduction

The financial sector has undergone a profound transformation in recent years, primarily accelerated by the emergence of Financial Technology, or FinTech (Barroso and Laborda 2022). FinTech constitutes a multifaceted and dynamic realm encompassing an array of technological innovations geared towards redefining and improving financial services. In this expansive landscape, Peer-to-Peer (P2P) lending occupies a prominent position, situated at the convergence of finance and technology (Q. Wang, X. Liu, and C. Zhang 2022). P2P lending platforms facilitate direct fund exchanges among individuals or entities, often bypassing conventional financial institutions such as traditional banks. This P2P model has gained substantial traction as an alternative financing avenue, affording borrowers access to capital while offering investors opportunities for potential returns (Lenz 2016). Some of the most popular P2P lending sites include *Lending Club* (n.d.) (USA), *Prosper* (n.d.) (USA), *Bondora* (n.d.) (European) and *Zopa* (n.d.) (European).

The advent of Artificial Intelligence (AI) has emerged as a pivotal catalyst for innovation within the P2P lending space. AI-driven algorithms analyze extensive datasets to evaluate borrower creditworthiness, streamline loan origination processes and enhance risk management practices. The integration of AI not only speeds up lending decisions but also broadens the inclusivity of financial services, reaching individuals who were previously excluded from traditional banking systems (Turiel and Aste 2020). However, this rise of P2P lending also brings forth many challenges. The utilization of Machine Learning (ML) models, though promising, has produced concerns about the "black box" issue, where decisions become complicated to interpret or justify (Dikmen and Burns 2022). Instances of discrimination and inherent biases have surfaced, underscoring the imperative of addressing ethical considerations (Wu et al. 2023). Furthermore, the prevalence of higher false positive rates in AI-driven lending decisions, coupled with a lack of accountability and transparency, has prompted scrutiny from both regulatory bodies and the research community (Haomin Wang, Kou, and Peng 2021).

Conversely, the advent of AI has also catalyzed the emergence of Human-Centered Artificial Intelligence (HCAI), focusing on developing technology that augments human capabilities and aligns with human values and requirements (Capel and Brereton 2023). A noteworthy aspect of HCAI is eXplainable Artificial Intelligence (XAI), which seeks to enhance the transparency and interpretability of AI systems, thereby fostering trust and accountability (Arrieta et al. 2020). While XAI has found application in diverse domains, such as healthcare (Gerlings, Jensen, and Shollo 2022) and agriculture (Cartolano et al. 2022), it is striking that the P2P lending sphere still needs to be addressed within this context.

This research aims to enhance the effectiveness of the creditworthiness assessment process within P2P lending. Acknowledging the pressing need for greater transparency, accountability and mitigation of human biases, we conducted a systematic literature review to assess the current state of research and practice in this domain. This literature review is a foundation for our research objectives, driven by the assessed research gaps and critical considerations.

Our research objectives have carefully been formulated in response to the research aim. The first objective is to discover the most accurate ML model that performs well concerning P2P lending decisions, thereby cultivating stakeholders' trust in the assessment results. Secondly, we strive to pinpoint transparent AI models known as XAI that provide insights and clarity into lending decisions, enhancing trust and openness. The subsequent objective of this research is to merge the most accurate ML model with the most interpretable XAI models to create an innovative and reliable creditworthiness evaluation model for P2P lending. This model is anticipated to enhance the assessment process's overall accuracy and interpretability. Within this study, we have thoroughly assessed the integrated model within the context of P2P lending, explicitly focusing on interpretability and explainability. The intention was to determine its practicality and potential in addressing gaps identified in previous studies. The objectives outlined in this study serve as a framework for exploring the complex relationship between P2P lending and Artificial Intelligence.

Our ultimate goal is to enhance credit evaluation's trust, transparency and efficacy amidst an ever-evolving financial landscape, thus enhancing its effectiveness. A comprehensive approach

has been adopted to achieve this, combining the merits of a ML model and an XAI model. Furthermore, this research is underpinned by an interview and a questionnaire, to urge and incorporate user feedback, ensuring that our model attains the highest levels of accuracy and effectiveness. Our innovative methodology holds the transformative potential to revolutionize the entire P2P lending industry, leading in an era characterized by fairness, transparency and sustainability. By placing user feedback at the forefront and harnessing the capabilities of AI, we aspire to democratize and equalize access to P2P lending, making it a more inclusive and equitable financial avenue for all.

2 Background

2.1 Peer-to-Peer (P2P) lending

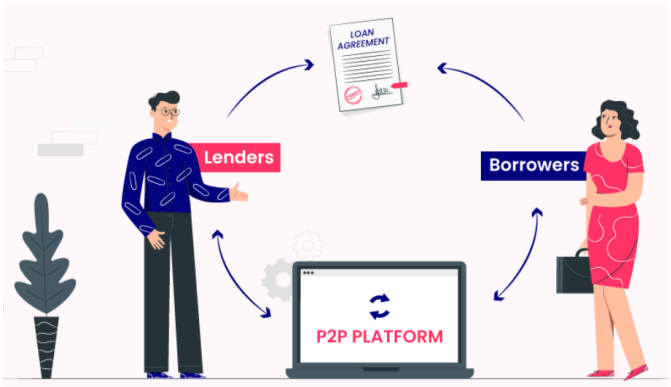


Figure 1: Peer to Peer (P2P) Lending

Peer-to-Peer (P2P) lending is a digital lending system that connects individuals seeking loans with people who want to invest or lend their money, effectively bypassing traditional banks. For those needing funds such as to start a small business, consolidate high-interest debt, make educational loan payments or make a significant purchase, P2P lending platforms provide an alternative to traditional banking. On the flip side, some individuals with spare cash are looking to make it grow by lending it to others; these lenders can be regular individuals rather than huge financial institutions (Suryono, Purwandari, and Budi 2019).

The magic of P2P lending lies in online platforms that act as intermediaries. Borrowers visit these platforms to request loans, specifying the amount they need and the purpose for which they intend to use it. Lenders browsing these platforms can then decide which loan listings they want to invest in, making lending decisions based on their preferences (Basha, Elgammal, and Abuzayed 2021). To ensure responsible lending and minimize the risk for lenders, P2P lending platforms evaluate the creditworthiness of borrowers. They analyze various data points, such as income, employment history and rental payment records. If the borrower's financial situation looks promising, the loan gets approved (Suryono, Purwandari, and Budi 2019).

Unlike traditional loans from one bank or financial institution, P2P loans are often funded by multiple lenders, contributing more minute amounts. This practice of diversifying risk among lenders facilitates borrowers' access to the necessary finances. After the approval of a loan, borrowers proceed to repay it gradually, typically accompanied by interest, in a manner similar to other conventional loans. The P2P platform facilitates the collection of payments from borrowers and subsequently allocates them to lenders, thereby assuring equitable distribution among all parties involved.

P2P financing benefits both borrowers and lenders, respectively. Borrowers can secure loans at reduced interest rates relative to those traditional financial institutions provide. Concurrently, lenders can accomplish higher returns on their assets compared to the rates achievable through traditional savings accounts or investments.

2.2 Interpretability vs. Accuracy

The tradeoff between interpretability and accuracy in Artificial Intelligence (AI) models has significant implications, particularly in critical applications such as FinTech, as highlighted by various researchers (Puzzarini et al. 2019; Uddin et al. 2022; He, Ma, and P. Wang 2020). This tradeoff reflects an inverse relationship between model accuracy and interpretability, where improving one often comes at the expense of the other, as depicted in Figure 2 below. In

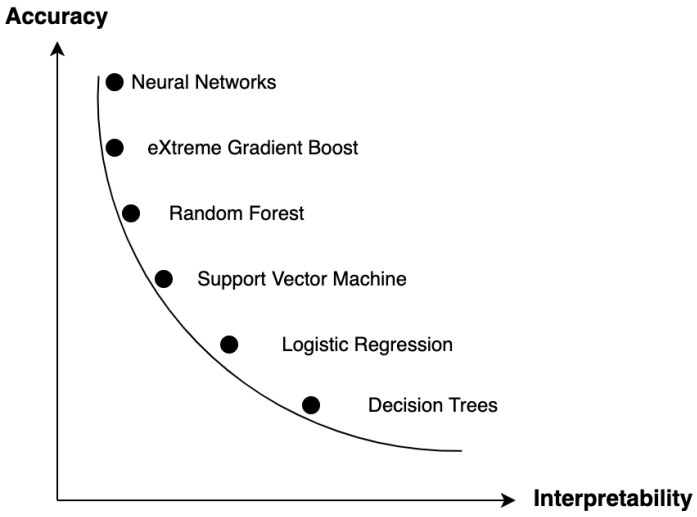


Figure 2: Accuracy vs Interpretability

practice, this means that highly accurate AI models, such as Neural Networks, eXtreme Gradient Boosting (XGBoost) and Support Vector Machines (SVM), which are employed by P2P lending platforms, tend to lack interpretability (J. Zhou et al. 2019; Coenen, Verbeke, and Guns 2022). These models surpass at making accurate predictions but can be challenging to understand and explain.

Conversely, simpler models such as Logistic Regression (LR) and decision trees may be less accurate but more transparent and easier to interpret (Hong Wang, Q. Xu, and L. Zhou 2015). In the context of P2P lending, where accuracy is crucial for minimizing risks, companies often prioritize these highly accurate yet less interpretable models. Nevertheless, this methodology may raise concerns regarding accountability and trustworthiness, especially in cases where loan applicants are rejected or presented with unfavourable conditions without transparent justifications.

In order to tackle this challenge, it is important to implement rules that guarantee adherence to principles of fairness, ethics and explainability for AI systems, particularly those employed in P2P lending (Vives 2017). Borrowers have a right to understand the basis for lending decisions and transparency is critical to building trust and fairness (Shin 2021). By requiring AI models to justify their decisions, it becomes possible to uncover and mitigate biases and ensure equitable lending practices.

Furthermore, transparency and interpretability build accountability, enabling regulatory oversight and opportunities for reform (Kochel and Skogan 2021). It is essential to recognize that the trade-off between accuracy and interpretability need not be a zero-sum game. Researchers and practitioners are actively working on methods and techniques to achieve high accuracy and meaningful interpretability, unlocking the true potential of AI in applications such as P2P lending. With proper safeguards and responsible AI practices, we can harness the power of ML to expand access and opportunities, ultimately fulfilling the promise of an inclusive and equitable financial system.

2.3 eXplainable Artificial Intelligence (XAI)

Concerns about the interpretability and transparency of AI systems' decision-making processes have surfaced recently as AI has been incorporated more and more into a variety of industries, including FinTech. To address these concerns and ensure that AI remains centred around human needs and values, a specialized field known as XAI has gained prominence. XAI aims to make AI models more transparent and interpretable, aligning them with the principles of Human-Centered AI. This approach is fundamental in promoting transparency, fairness and accountability in AI systems, particularly for those affected by these decisions.

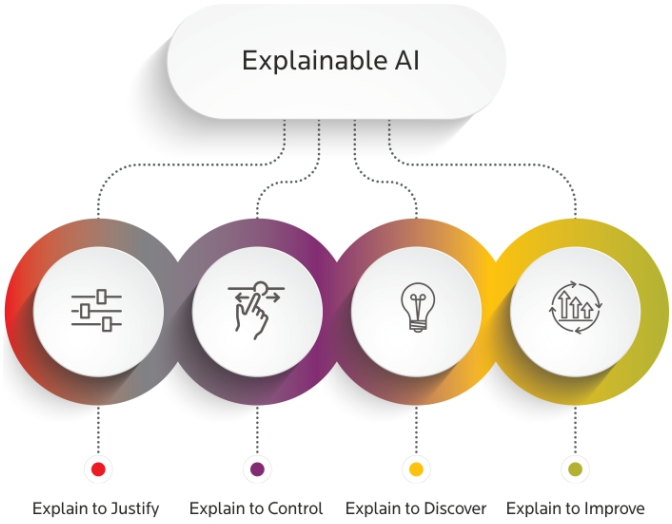
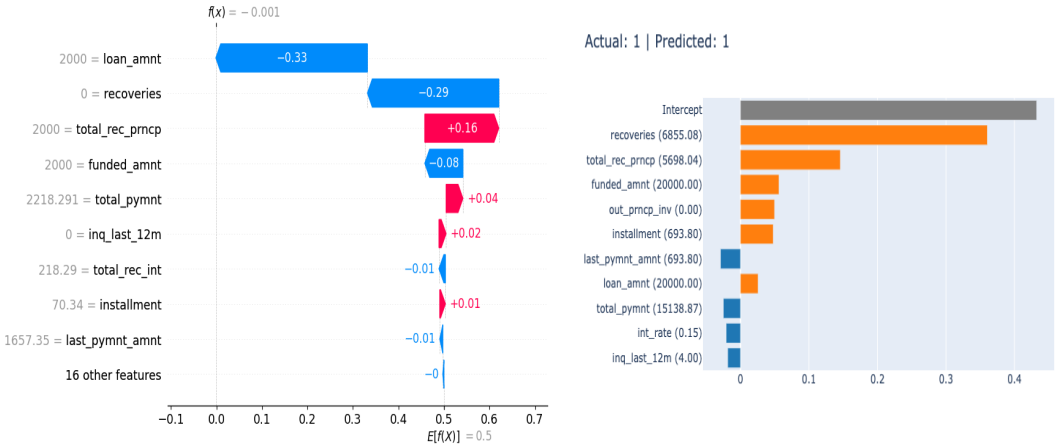


Figure 3: Reasons to have explainability

The FinTech sector has a significant need for XAI, primarily due to the crucial role that AI plays in various activities, including credit scoring and risk assessment. Although conventional AI models demonstrate solid predictive capabilities, they frequently lack the inherent transparency and interpretability of XAI models. XAI prioritizes transparency by clarifying these models' underlying mechanisms, enabling a more streamlined approach to making informed judgements. This holds great importance within financial decision-making, particularly concerning loan approvals or investment advice, because of the significant consequences these decisions can have on individuals.

Moreover, XAI plays a vital role in addressing biases inherent in AI models, hence promoting unbiased and just outcomes within the FinTech industry. XAI plays a crucial role in illuminating the underlying decision-making mechanisms, hence facilitating the identification and mitigation of biases. These biases, if left unaddressed, can have significant and wide-ranging implications on the financial prospects available to individuals. It also assures that AI models in FinTech function as intended, reassuring regulatory bodies and consumers that AI-driven financial services are reliable, ethical and consistent.

Among the various XAI models available, two prominent ones are SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) (Gramegna and Giudici 2021). These models are famous for their ability to provide interpretable explanations for AI model predictions. SHAP and LIME are widely used to help users and stakeholders understand how and why AI systems make certain decisions, further enhancing transparency and trust in AI-driven financial services.



(a) SHAP interpretation

(b) LIME interpretation

Figure 4: Sample interpretations from XAI libraries

Summarising the importance of XAI, it is pivotal in the evolution of Human-Centered AI, particularly within the FinTech industry. By enhancing transparency and interpretability, XAI models diminish concerns related to bias, ensure operational integrity and foster trust in AI-driven financial services. Researchers and experts increasingly advocate for adopting XAI in FinTech, making AI accurate but also comprehensible and accountable, ultimately serving the best interests of individuals and society.

3 Literature Review

PRISMA, an acronym representing the Preferred Reporting Items for Systematic Review and Meta Analyses, stands as an evidence-based framework designed to enhance the quality of reporting in systematic reviews and meta-analyses, as provided by Selçuk (2019). Its usefulness goes beyond these domains, offering valuable guidelines for reporting various research types, especially evaluations of interventions. The primary objective of PRISMA is to empower authors to enhance the transparency, comprehensiveness and usefulness of their systematic reviews (Rethlefsen et al. 2021). This comprehensive framework encompasses a 27-item checklist and a four-phase flow diagram. The checklist encompasses critical aspects related to the review, such as the title, abstract, introduction, methods, results, discussion and funding. Concurrently, the flow diagram visually depicts the study selection process, covering the process of record identification, inclusion criteria, exclusion criteria and rationale for exclusions.

PRISMA plays a pivotal role in mitigating research wastage by uplifting the quality and usefulness of systematic reviews. Its adoption leads to enhanced transparency and accuracy in reporting, which, in turn, enables readers to assess the strengths and limitations of the review comprehensively. This framework has been instrumental in identifying relevant research papers and facilitating a comprehensive literature review in P2P lending. A substantial collection of 198,755 sources was obtained by utilising well reputable databases including Elsevier, Springer, IEEE Xplore, and Frontiers along with a systematical search strategy that included keywords such as FinTech, HCAI, XAI, AI in FinTech, XAI in FinTech, and HCAI in FinTech. The comprehensive use of PRISMA guidelines eventually made it possible to identify 37 relevant research publications and reports related to peer-to-peer lending. Figure 5 provides a graphical representation of the explained procedure, which shows how research publications are systematically filtered out of the original source repository.

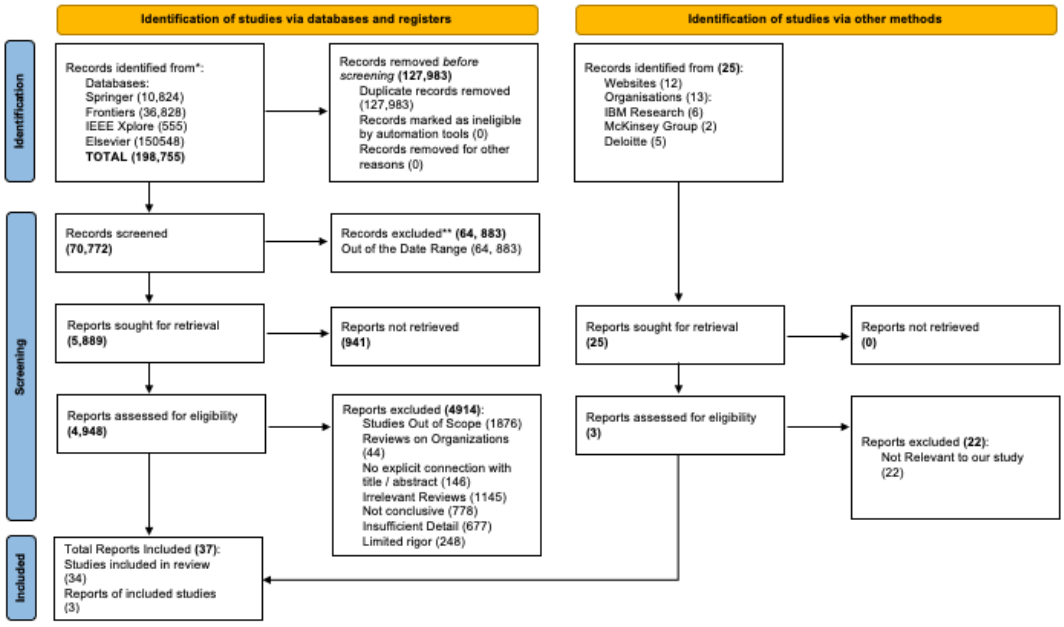


Figure 5: Visualization of the Research Paper Filtering Process using PRISMA

Based on the comprehensive literature review carried out, AI models used in existing research in P2P lending are classified into three primary categories, namely Logistic Regression models, Tree models and explainability models, as visually represented in the illustration below:

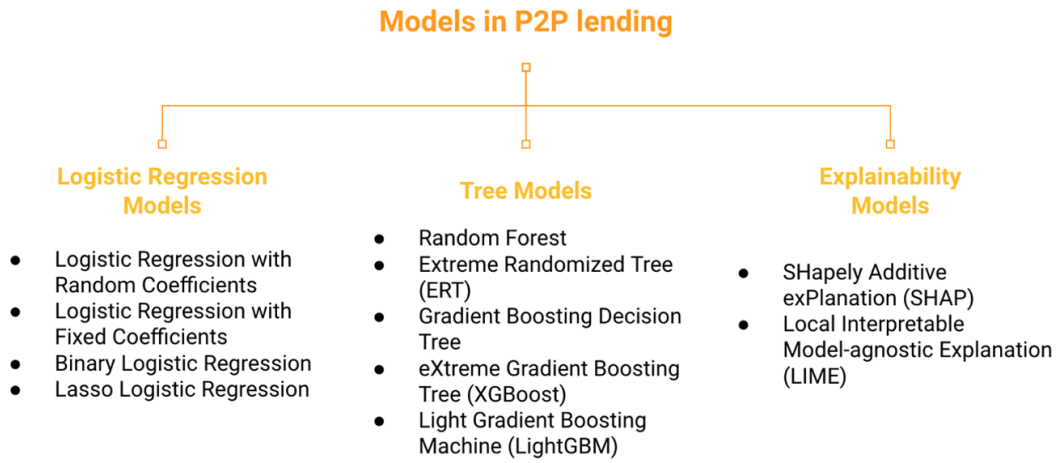


Figure 6: Categorisation of research papers

3.1 Models used in P2P lending

3.1.1 Logistic Regression (LR) Models

Within the domain of P2P lending, multiple LR models have been identified in the body of research. These encompass models include models such as Lasso-Logistic Regression (LLR), Binary Logistic Regression (BLR), Logistic Regression with Random Coefficients (LRR) and Logistic Regression with Fixed Coefficients (LRF) (Königstorfer and Thalmann 2020; Serrano-Cinca and Gutiérrez-Nieto 2016; Setiawan et al. 2019; Z. Zhang, Niu, and Y. Liu 2020; Emekter et al. 2015; Smith 2019).

LR, renowned for its efficiency in binary classification tasks, boasts a high predictor throughput and the capacity to accommodate missing data. Its predictive accuracy surpasses alternative statistical methods while providing valuable insights into predictor weights. Even with limited data, LR can rectify gaps and make accurate predictions. In the context of P2P lending platforms, LR models assume the crucial role of forecasting borrowers' likelihood of loan default (Zhu et al. 2016). This is supported by LR's ability to manage missing data, perform binary classification, and handle a large number of variables.

For instance, in evaluating Small and Medium-sized Enterprises (SMEs) credit risk in China, a study introduces a credit scorecard rooted in LRR, which extends the conventional LR model by incorporating random effects to account for unobserved variations across observations (Longford 1994). For a variety of data points, this augmentation enables LRR to catch minor differences in the relationship between independent variables and the dependent variable. When comparing LRR and LRF, it can be shown that LRR attains higher accuracy, which increases its perceived value.

Eventhough LRR seems promising, the study needs to rigorously assess its robustness or benchmark its performance against potentially superior credit scoring models, particularly where there is a lack of available data. Thus, further research is necessary to determine LRR's reliability and consistent outperformance of rival methods. In brief, LRR introduces a sophisticated and adaptable approach to binary classification tasks attained with features while accommodating missing data and heterogeneity between observations.

Similarly, BLR, a simplified LR variant, is a valuable tool for binary classification challenges (King 2008). In the context of P2P lending, BLR models categorize borrowers into 'good' and

'bad' credit risks, offering probabilities of loan repayment versus default. A study by Emekter et al. (2015) employed BLR to distinguish defaulted and non-defaulted loans, revealing that borrowers with better credit ratings exhibited a reduced likelihood of default.

Furthermore, the Lasso Logistic Regression Ensemble (LLRE) strategy integrates LLR and Ensemble techniques, presenting a particularly beneficial approach for handling high-dimensional, unbalanced datasets in binary classification tasks. Even though LLRE performed better in large, uneven datasets from Chinese banks, there are significant differences in borrower demographics, loan requirements, and contextual complexities, making it difficult to apply these findings to P2P lending platforms.

BLR, LLRE and other variants of LR models exhibit potential for credit risk assessment and decision-making within the P2P lending domain. However, their reliability and consistent superiority over alternative methods, particularly in data scarce scenarios, has to be more rigorously scrutinised. Although their general implementation awaits additional validation through comprehensive research, objective evaluations against rival models foreshadow more sophisticated and adaptive methods to binary classification difficulties important to lending platforms. Although their general implementation requires additional validation through comprehensive research, objective evaluations against alternative models indicate more sophisticated and adaptive methods to binary classification problems that are important to P2P lending platforms.

3.1.2 Tree Models

In the domain of P2P lending, a diverse array of tree-based models have been effectively harnessed, including Random Forests (RF), Extremely Randomized Trees (ERT), the eXtreme Gradient Boost (XGBoost) algorithm, Light Gradient Boosting Machine (LightGBM) and Gradient Boosting Decision Trees (GBDT) (J. Zhou et al. 2019; Setiawan et al. 2019; Kumar et al. 2016; Z. Li et al. 2021).

Complementing the application of LR models, several ML tree models have been instrumental in credit risk assessment and credit scoring on P2P lending platforms. ERT and RF have emerged as robust performers, owing to their inherent simplicity and adaptability to new data inputs. In a notable case,Setiawan et al. (2019) used Lending Club data to build a P2P lending default loan classification model, enhancing its efficacy with the use of Binary Particle Swarm Optimization (BPSO) and Support Vector Machine (SVM) techniques. During this study, they discovered that ERT performed better than RF on average.

Beyond the traditional decision tree models, we encounter modified versions utilized in boosting techniques, including the GBDT, XGBoost and LightGBM. LightGBM uses a histogram-based approach to speed up training while XGBoost uses more complex methods of training and regularisation techniques. GBDT, in contrast, trains trees sequentially and combine their predictions. J. Zhou et al. (2019) combined these models with ensemble learning techniques to improve the accuracy of default probability prediction within P2P lending networks. The result was an integrated model that outperformed the performance of its elements

Ensemble methodologies that integrate multiple tree models, such as RF and Gradient Boosting, have consistently demonstrated robust performance in classification and regression tasks within the P2P lending landscape. However, given their promising potential for credit risk assessment in the P2P lending space, a more thorough analysis is necessary to establish their reliability and consistent superiority over alternative approaches, especially when data is scarce. More advanced and flexible solutions to the primary challenges with lending platforms should be possible with a fair comparison with other models. However, adopting these techniques on a wide scale depends upon the necessity for further research and validation.

3.1.3 XAI Models

XGBoost, LLRE and BLR pose challenges for financial institutions in terms of trust due to their inherent complexity and limited explainability. Financial institutions are increasingly turning to specialized AI models that balance accuracy and explainability to address this concern and align with the stringent accountability, transparency and oversight standards mandated by regulatory bodies.

In current P2P lending research, there is a growing integration of XAI models such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) (Misheva et al. 2021; Bussmann et al. 2020; Ariza-Garzón et al. 2020). These post-hoc models are designed to provide comprehensive global and local explanations into the complex algorithms of ML. SHAP and LIME are crucial in understanding how and why ML algorithms arrive at their predictions or decisions (Gramegna and Giudici 2021). Through an in-depth analysis of feature importance, interactions and localized explanations at individual data points, researchers gain a deeper understanding of model behaviour, assumptions and constraints. With this knowledge, opportunities for enhancing model performance, fairness and robustness become more apparent and actionable.

SHAP and LIME, which provides interpretable explanations adhering to financial regulations, have been employed to assess model explainability. However, it is noteworthy that SHAP demands considerably more computational resources when generating values for numerous variables. Conversely, while LIME effectively describes models and their objects, it has certain inherent limitations.

In a separate study, Bussmann et al. (2020) applied TreeSHAP, a computationally efficient method with polynomial runtime (as opposed to exponential), to compute the Shapley Additive Explanation for tree-based models. The European External Credit Assessment Institute (EECAI), the source of the study's data, evaluated credit for P2P platforms, with a specific focus on commercial lending to SMEs. According to the researchers, network-based XAI models exhibit the potential to enhance the comprehension of factors influencing credit risk substantially. Despite utilizing SHAP and LIME in various studies, they still fail to provide the degree of explainability that end users demand.

The financial industry's desire for trustworthiness and transparency in complex AI models has prompted incorporating XAI models such as SHAP and LIME into P2P lending research. While these models provide valuable insights, there remains an ongoing pursuit of improved explainability to meet the evolving needs of stakeholders.

4 Research Gaps

The comprehensive literature review conducted has revealed a variety of research gaps within the domain of our study, such as (i) Lack of Accountability and Openness, (ii) Lack of Effectiveness, (iii) Human Biases and Fairness, (iv) High False Positive Rates, (v) Risk Profiling, (vi) Data Availability Issues and (vii) Ethical Concerns. Among these seven prominent research gaps, this research addresses the first four pivotal research gaps in the P2P lending domain with the aid of Human Centered AI.

4.1 Lack of Accountability and Openness

A key research gap in the field of peer-to-peer lending is to the appearing lack of transparency and accountability in the decision-making processes used by lending platforms. The study's steadfast dedication is to tackle this pivotal challenge by utilising XAI approaches, as demonstrated by the research conducted by Bussmann et al. (2020). The integration of XAI techniques, including SHAP and LIME, has the potential to bring about a transformation characterised by greater accountability, openness, and confidence, which will be beneficial to lenders and borrowers in both instances.

The implications of this research gap go beyond the complex workings of machine learning; they have a profound impact on the realms of trust, transparency, and accountability, which are the pillars upon which any lending ecosystem is built. Through an exploration of the complex workings of AI models and the provision of transparent, understandable explanations of how they make decisions, the application of XAI techniques has the potential to reestablish trust in the minds of both lenders and borrowers.

The explicit attempt to explain the complex decision-making processes of AI models, leading to increased transparency and trust, highlights the uncharted research landscape that awaits in the P2P lending domain. This is a landscape that requires extensive research and in-depth exploration. Its main objective is to transform the loan process and create an ecosystem that inspires the highest levels of trust, fairness, and openness in all parties involved.

4.2 Lack of Effectiveness

Effectiveness is a crucial aspect of P2P lending that requires careful examination. It should extend its influence throughout every stage of the P2P lending process, starting from the initial model development and implementation by developers, extending to borrowers applying for loans and concluding when borrowers repay their loans to investors through the platform. Effectiveness is not desired; it also essential for building trust and confidence within the entire P2P lending ecosystem.

While advanced ML models such as XGBoost, RF, Long-Short-Term Memory (LSTM) networks and ensemble learning models excel in accuracy, it is essential to recognize that high accuracy often comes at the cost of interpretability. Existing research has consistently shown that interpretability and explainability tend to decrease as accuracy increases (J. Zhou et al. 2019; Setiawan et al. 2019; Kumar et al. 2016; Z. Li et al. 2021). This creates a mystery in the P2P lending domain and raises a critical problem that hinders the effectiveness of AI models which is the lack of explainability and the well-noted "black box" problem.

When AI models operate as opaque "black boxes" and provide no insights into their decision making processes, it damages trust and undermines user confidence in these models (Havrylchyk and Verdier 2018). Real users, be they lenders or borrowers, are often reluctant to use platforms where the model's inner workings are unclear at an abstract level. This reluctance stems from users wanting to clearly understand how decisions are made, mainly regarding financial matters. Thus, the two most important components required to close this gap and guarantee the efficacy of AI models in P2P lending become transparency and interpretability.

Therefore, there exists a challenge to develop methods and techniques that enhance the explainability of high-accuracy models. In doing so, a balance between accuracy and interpretability could be compromised, ensuring that P2P lending participants can trust and understand the decisions made by AI models. This, in turn, fosters a more effective and trustworthy P2P

lending ecosystem where users are more inclined to participate when they have transparency into model operations.

4.3 Human Bias and Fairness

In the landscape of P2P lending, one of the pivotal research gaps that demands thorough investigation is the intricate relationship between human biases and the fairness of ML models (Fu, Huang, and P. V. Singh 2021). This research gap underscores the imperative need to scrutinize how inherent human biases impact the performance and fairness of ML models in the context of P2P lending platforms. The complexities of this research gap are particularly apparent in instances characterized by skewed data, where these biases can trigger judgment errors with substantial consequences.

Recognizing the significance of this issue, several set of techniques, ranging from bias correction to data augmentation and the deployment of fair ML algorithms, emerges as potential remedies. These methods are vital resources in the continuous effort to limit and mitigate the impact of human bias, ensuring that the decisions rendered by machine learning models are fair and unbiased.

Importantly, our exploration acknowledges the potential for ML models to reinforce or magnify biases present in their training datasets. We must gain a profound understanding of how these bias types impact the decision-making processes of ML models, thereby guiding us towards the formulation of effective bias reduction strategies.

An illustrative example from the P2P lending domain underscores the significance of this research gap. Consider a scenario where historical lending decisions are primarily based on data from borrowers who are predominantly white, male and possess impeccable credit ratings. In such cases, ML models trained on this skewed dataset may inadvertently perpetuate these biases, leading to unfair loan judgments. This can have further consequences, hindering access to credit for under privileged groups and undermining the essence of democratizing financial access, a core principle of P2P lending.

We must contend that ML models operating within P2P lending platforms can achieve enhanced performance and fairness when human biases are acknowledged and proactively dealt with. As such, developing tailored solutions becomes fundamental in our quest to ensure unbiased, equitable assessments within the P2P lending landscape.

4.4 High False Positive Rates

The prevalence of higher false positive rates stands out as a significant research gap in P2P lending and its implications have reverberated throughout the industry (Haomin Wang, Kou, and Peng 2021; J. J. Xu et al. 2022). In the context of credit risk assessment, false positives refer to cases where a borrower is incorrectly classified as high-risk or likely to default when they are not. These inaccurate predictions have had profound consequences for the P2P lending landscape.

One of the primary ways higher false positive rates have impacted P2P lending is by hindering the accuracy of credit risk assessment. P2P lending platforms rely heavily on the ability to accurately assess the creditworthiness of borrowers to make informed lending decisions. When false positives are prevalent, creditworthy borrowers are wrongly labelled as high-risk. This, in turn, can result in these borrowers being denied access to loans they deserve, hindering the democratization of financial access, a core principle of P2P lending.

Moreover, higher false positive rates have broader implications for the entire ecosystem. Lenders on P2P lending platforms may become concerned due to the increased risk associated with false positives. This can lead to more conservative lending practices, higher interest rates, or a reluctance to engage with P2P lending platforms altogether. For investors who provide funds on these platforms, the fear of incurring losses due to inaccurate credit risk assessments can deter their participation, affecting the availability of funds for borrowers.

4.5 Risk Profiling

This research gap raises essential questions about how different risk management techniques may affect the likelihood of defaults occurring within P2P lending systems (Yoon, Y. Li, and Feng 2019; Zhao et al. 2014; Serrano-Cinca and Gutiérrez-Nieto 2016). With comprehensive studies on this topic, it remains unclear how portfolio diversification, risk-based pricing and borrower screening strategies may influence the overall default risk in these lending platforms. Future research should identify the relationship between various risk management strategies and default risk in P2P lending systems.

One avenue for investigation is examining how different risk management strategies impact distinct borrower types. This involves considering the varying risk profiles of borrowers, such as high-risk and low-risk individuals. For instance, rigorous borrower eligibility criteria and higher interest rates might effectively reduce default risk for high-risk borrowers. In contrast, portfolio diversification across numerous low risk loans could serve as a risk mitigation strategy for more creditworthy borrowers.

Systematically evaluating the effectiveness of risk management strategies across different borrower segments would provide invaluable insights for P2P lending platforms. By understanding which techniques or combinations optimize default prevention while maintaining a desirable borrower base, these platforms can craft policies that balance risk aversion and accessibility. In essence, the studies could help lending platforms design risk management strategies that effectively mitigate defaults and are conducive to sustainable growth.

4.6 Data Availability Issues

In the context of P2P lending, a pressing research gap centres around the challenge of scarcity of available information that significantly impacts the ability of lending platforms to assess borrower credit risk accurately and make informed lending decisions while reducing default probability.

One potential avenue of research in addressing this gap involves exploring the benefits of incorporating additional data sources, such as financial transaction data and social media data, to enhance risk prediction models. These supplementary data sources provide valuable insights into borrower behaviour, including their character, beliefs, spending habits and prudent financial literacy. By investigating the potential contributions of these additional data streams, researchers can explore how they can bolster the accuracy and reliability of credit risk assessments within P2P lending platforms.

Furthermore, applying Natural Language Processing (NLP) techniques in analyzing borrower ratings and comments offers an intriguing solution to this research gap. NLP, a branch of Artificial Intelligence, excels at extracting insightful information from text-based data, such as borrower reviews and feedback. This information can be leveraged to improve the overall effectiveness and accuracy of P2P lending platforms. Potential areas of exploration within this research gap include assessing various NLP techniques, such as sentiment analysis, topic modelling and named entity identification, in the context of borrower assessments and comments. These techniques promise to yield valuable insights from the qualitative data borrowers provide in their reviews and feedback on P2P lending websites.

4.7 Ethical Concerns

AI's involvement in P2P lending brings to the forefront the critical issue of potential bias within lending decisions, which can have further consequences. The primary ethical concern is that AI models may unintentionally discriminate against certain borrower groups and promote biases. This risk is particularly prominent when biased data is used for model training or when algorithms are designed to favour specific borrower profiles over others. For instance, if the training data primarily represents borrowers who are predominantly of one race or gender or possess exceptional credit ratings, the AI model may exhibit bias against those who do not conform to this profile. This bias can lead to unfair loan decisions, making it harder for disadvantaged people to get credit and defeating the main goal of democratising financial access.

Addressing this research gap requires a multifaceted approach. Researchers and developers must proactively acknowledge and mitigate potential biases within AI models by ensuring that training data sets encompass the full diversity of borrowers, including various demographics and credit profiles. Additionally, conducting rigorous bias audits at every stage of AI model development is crucial to identify and rectify discriminatory patterns in lending decisions. Enforcing policies within P2P lending platforms that actively work to mitigate discrimination, adhere to ethical norms, uphold civil rights and promote fairness in lending practices is equally essential. Furthermore, involving key stakeholders in developing and applying HCAI models, such as ethicists, legal experts, community representatives and borrowers themselves, helps ensure that AI models not only avoid disadvantaging any particular group but also uphold ethical values and foster inclusivity.

5 Research Objectives

Below are the research objectives derived from the systematic literature review, aligning with the research aim of enhancing the effectiveness of the creditworthiness assessment process in P2P lending.

1. To identify the most accurate Machine Learning (ML) model in Peer-to-Peer (P2P) lending.
2. To identify the most interpretable eXplainable Artificial Intelligence model in Peer-to-Peer (P2P) lending.
3. To integrate the most accurate Machine Learning (ML) model with the most interpretable eXplainable Artificial Intelligence model in Peer-to-Peer (P2P) lending.
4. To evaluate the novel model's effectiveness in Peer-to-Peer (P2P) lending with a focus on interpretability and explainability.

6 Research Questions

Based on the research objectives derived from the systematic literature review, research questions are formulated to advance the research aim of improving the creditworthiness assessment process in P2P lending.

1. What is the most accurate Machine Learning (ML) model in P2P lending?

   This research question is fundamental because ML models' accuracy directly influences lending decisions' quality in P2P platforms. Identifying the most accurate model is essential for minimizing credit risk, reducing defaults and ensuring the financial sustainability of the lending platform. Accurate models can significantly impact the platform's performance by improving the precision of credit risk assessments. However, the issue of higher false positive rates, which affects the accuracy of ML models, underscores the need to delve deeper. False positives can lead to unnecessarily cautious lending decisions and hinder the lending process's effectiveness. By identifying the most accurate model, this research aims to mitigate this issue, contributing to trust and confidence among lenders and borrowers.

   In addressing this research question a thorough analysis was conducted to minimize false positives, which is crucial for enhancing lending decisions' quality and platform performance. Various ML models were compared, including Logistic Regression, Tree-based models such as Lasson logistic Regression with Random Co-efficients, Random Forest, Classification Tree, and XGBoost. These models were chosen based on prior research and their suitability to the lending context. By systematically comparing their accuracies, the study aimed to identify models that effectively mitigate false positives, thereby bolstering trust among lenders and borrowers and advancing credit risk assessment in P2P lending platforms.

2. What is the most interpretable eXplainable Artificial Intelligence (XAI) model that could be used in Peer-to-Peer (P2P) lending?

   Extensive research has consistently demonstrated that interpretability and explainability tend to decrease as accuracy increases. This challenge is particularly pertinent in the P2P lending domain, which raises a critical issue, the lack of explainability and the famous "black box" problem. When AI models operate as opaque "black boxes" and provide no insights into their decision-making processes, it erodes trust and undermines user confidence in these models. Real users, whether lenders or borrowers are understandably reluctant to engage with platforms where the model's inner workings remain mysterious. This reluctance is rooted in the desire to understand how decisions, especially those involving finances, are made. Therefore, transparency and interpretability become pivotal elements to bridge this gap and ensure the effectiveness of AI models in P2P lending.

   In addressing the above research question, the study recognized the crucial need for transparency and interpretability within lending platforms. Realizing the importance of bridging this gap, three XAI models—SHAP, LIME, and DiCE were evaluated to unravel hidden factors impacting model behavior. SHAP offers global explanations, shedding light on model behaviour and trends, while LIME provides local explanations, unravelling hidden factors that affect a specific loan decisions. DiCE, on the other hand, offers counterfactuals, offering insights relevant to end consumers' needs. By systematically comparing these XAI models, the study aimed to identify the most interpretable model, crucial for enhancing user confidence and trust in P2P lending platforms.

3. What are the results of integrating the most accurate Machine Learning (ML) model with most interpretable eXplainable Artificial Intelligence (XAI) model in Peer-to-Peer (P2P) lending?

   Integrating accurate yet complex ML models with interpretable XAI models is instrumental in achieving our core research aim which is enhancing creditworthiness assessment in P2P lending. Accurate machine learning models enhance credit risk assessments and loan performance forecasts by utilising complex feature relationships. Interpretable XAI

models, on the other hand, provide clarity by outlining the decision-making process. P2P lending platforms are empowered by this harmony between transparency and correctness, which is in line with our study goal of increasing P2P lending's effectiveness, reliability, and accessibility for all participants. By integrating user friendly transparency with accurate forecasts, we close the gap that has impeded the effectiveness of peer-to-peer lending, thereby benefiting all parties by enhancing P2P lending ecosystem.

In order to address the above research question, the study involves in integrating the most accurate Machine Learning model with most interpretable XAI model(s). By offering user friendly transparency alongside accurate forecasts, we sought to enhance the overall ecosystem of P2P lending, benefiting lenders, borrowers and the other stakeholders. This synergy between accuracy and interpretability not only fosters trust and confidence among users but also strengthens the foundation of P2P lending as a viable and transparent financial alternative. .

4. How to conduct an evaluation on the results of the novel model made using the most accurate Machine Learning (ML) model with most interpretable eXplainable Artificial Intelligence (XAI) model in Peer-to-Peer (P2P) lending?

   This research question is pivotal in assessing the novel model's effectiveness in P2P lending, explicitly focusing on interpretability and explainability. It directly contributes to our overarching research aim of enhancing the creditworthiness assessment process in P2P lending. By evaluating the model's real-world applicability, this objective validate the practicality of our integrated approach, which combines accurate yet complex ML models with interpretable XAI models. The emphasis on interpretability and explainability directly addresses the persistent issue of "black box" AI models. Understanding the rationale behind lending decisions is essential in the context of P2P lending, where trust and transparency are paramount. This evaluation guarantees that the model performs well and is understandable to users, boosting its trust and interpretability.

   In addressing the research question, our study focused on validating its effectiveness and practicality. This evaluation is pivotal in assessing the model's real-world applicability and directly contributes to our overarching aim of enhancing the creditworthiness assessment process in P2P lending. To conduct this evaluation, a qualitative approach was adopted, involving a questionnaire provided to ML experts in a P2P lending company and interviews conducted with key personnel in the P2P lending domain. Through these interactions, we sought to gather insights into the model's performance, explainability, interpretability and usability validating its practicality and effectiveness in real-world lending scenarios.

7 Methodology

The research framework introduced by Saunders, Lewis, and Thornhill (2007), widely adopted in social sciences and business studies, is valuable for constructing the theoretical foundation of research endeavours. Comprising multiple layers, ranging from the outermost to the innermost, this framework encompasses philosophy, approach, strategy, choice, time horizon, data collection and data analysis. It proves equally instrumental in structuring the methodology of systematic scientific investigations, offering a methodical and comprehensive approach to its development. The research onion, by guiding the selection of appropriate options within each layer, facilitates the creation of a robust research methodology.

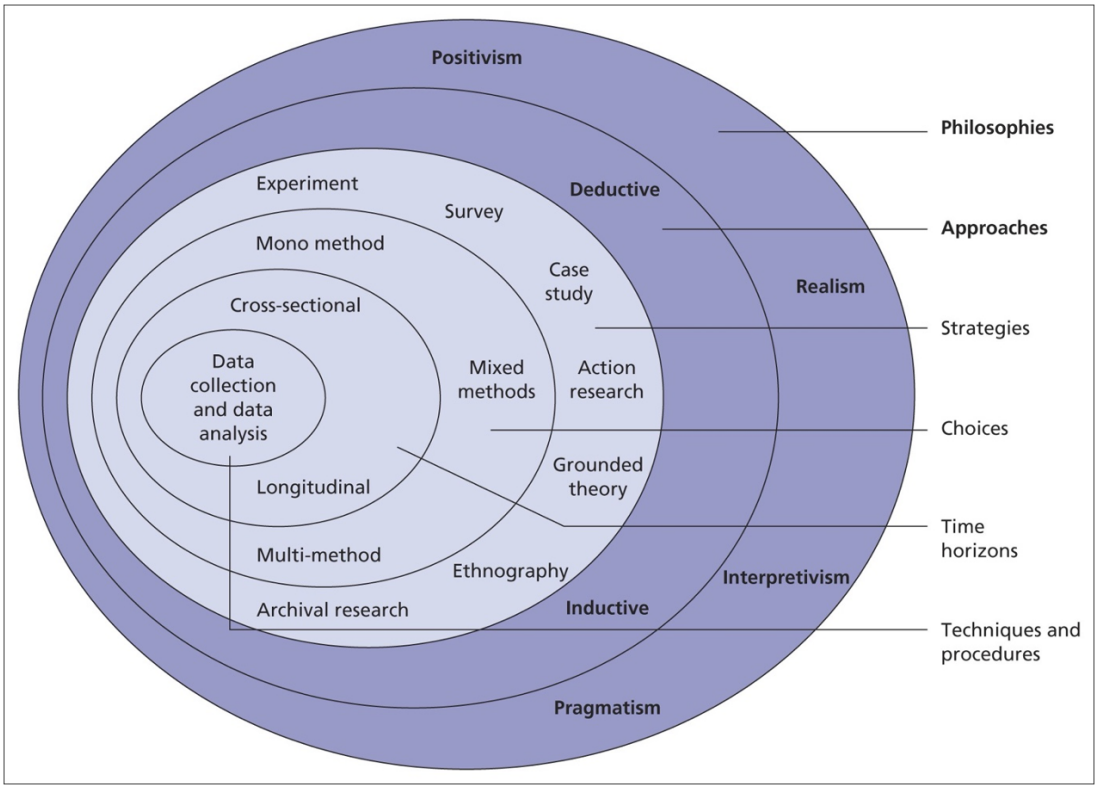


Figure 7: Research Onion

According to Saunders, Lewis, and Thornhill (ibid.)'s research onion, this study's methodology is meticulously structured, moving from the outermost layer to the innermost. Pragmatism is the chosen research philosophy, emphasizing practical solutions to real-world challenges. The research has adopted an inductive approach, commencing with observations to formulate theories. We have employed design science strategy introduced by Dresch et al. (2015) as our research strategy because we aim to generate innovative artefacts, such as use-cases of explainability mechanisms, to enhance both the creditworthiness and explainability of AI models. A mixed-method approach, incorporating both qualitative and quantitative data, ensures a comprehensive comprehension of the research domain. With a longitudinal temporal horizon spanning over five years, the study seeks to identify temporal changes and patterns. Lending-related datasets had served as the foundation for data collection and subsequent stages of exploratory analysis, dimension reduction and model creation, all aimed at evaluating effectiveness and providing well-founded justifications.

7.1 Research Philosophy

This research embraces pragmatism as its underlying philosophical framework. As a philosophy, pragmatism strongly emphasizes the accomplishment of practical and valuable outcomes, urging researchers to employ diverse methodologies that effectively address real-world issues. In the context of this study, which revolves around evaluating creditworthiness and enhancing AI model explainability in P2P lending, a pragmatic philosophy is well-suited. This is because the study's primary objective is to identify practical solutions that function effectively in practice, prioritizing tangible results over abstract or theoretical principles.

By applying pragmatism to this research, we can focus on developing pragmatic strategies to improve both creditworthiness evaluation process and AI model explainability within P2P lending. This approach is instrumental in ensuring that the conclusions and recommendations drawn from this study are relevant and genuinely helpful for P2P lending practitioners. In contrast, other philosophical orientations, such as realism, interpretivism and positivism, are less suitable for this research endeavour. Realism, for instance, delves into the perception and explanation of reality as it exists beyond human observation, making it more theoretically oriented than required for this practical issue. Interpretivism, on the other hand, seeks to understand social phenomena by interpreting the meanings humans attribute to them, adopting a subjective perspective that may not align with the evaluation of creditworthiness and AI models. Positivism, known for its scientific, evidence-based approach using observable and quantifiable data, while relevant in many research contexts, may prove too rigid and quantitative for a study primarily focused on applied problem-solving. Consequently, pragmatism has been judiciously selected as the research philosophy for this study, striking a harmonious balance between practical applicability and theoretical rigour, which perfectly complements the solution-focused nature of this research endeavour.

7.2 Research Approach

In this research, an inductive approach is a more fitting choice, whereas a deductive approach would be less effective. A deductive approach typically begins with well established theories and hypothesis, which are then tested by gathering data. However, when dealing with the complex challenge of evaluating AI models in P2P lending and enhancing their explainability, adopting deductive reasoning could result in the omission of significant details or the creation of unduly simplistic assumptions.

In contrast, an inductive approach begins the research with open ended observations, permitting theories and insights to emerge organically from the collected data. This method facilitates the acquisition of a profound and thorough understanding of the issues. By employing inductive inquiry, we open doors to uncovering essential factors that influence creditworthiness, recognizing the hurdles associated with model explainability and identifying opportunities for enhancement that might remain hidden within the confines of a deductive approach. Furthermore, deductive reasoning requires the formulation of hypotheses and theories prior to data collection and analysis. In the context of a relatively new domain, it proves challenging to construct comprehensive theories and establish well-founded hypotheses without initially observing the data itself. An inductive approach overcomes this challenge by allowing insights to evolve gradually throughout this research journey.

7.3 Research Strategy

Design Science research represents a strategy squarely focused on devising innovative solutions to real-world challenges. As elucidated by Dresch et al. (2015), this approach is structured around a 12-step framework encompassing activities ranging from issue definition to artefact creation, evaluation and eventual implementation.

Notably, the introduction of the research onion framework predates the incorporation of the Design Science research strategy. However, in the context of this research aimed at enhancing the creditworthiness and explainability of AI models in P2P lending, we find it pertinent to introduce the Design Science approach. This strategic addition allows us to forge novel artefacts, including

explainability mechanisms tailored to AI models, potentially revolutionizing how we address the challenges within P2P lending.

7.3.1 How Design Science Research is used:

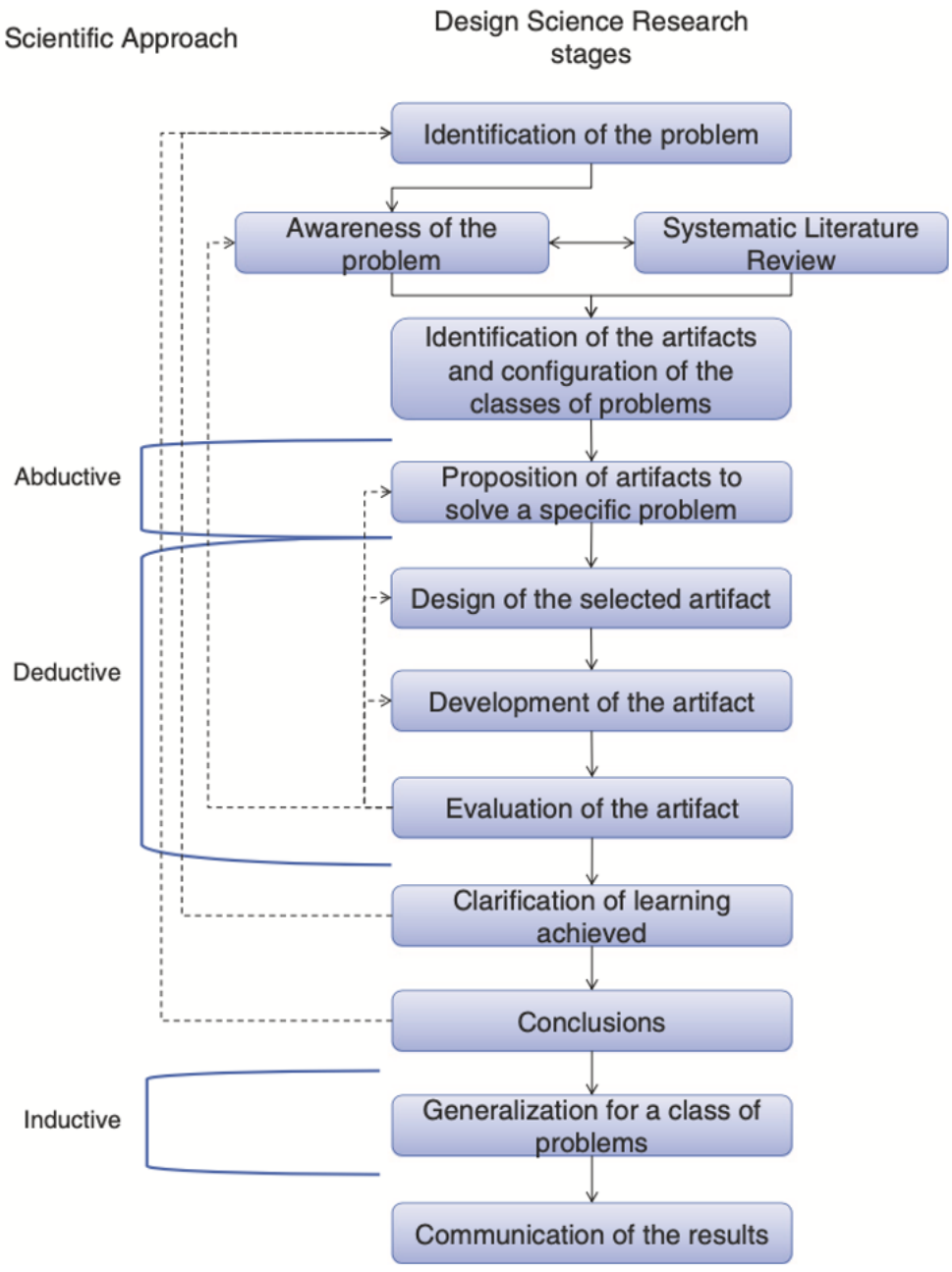


Figure 8: Research strategy based on Design Science Research

Applying the design science approach in this research follows a systematic and pragmatic process (Dresch et al. 2015). Initially, it involves identifying the issue at hand, which entails a comprehensive exploration of the current landscape of creditworthiness and explainability within P2P lending. Once a thorough understanding is established, the approach shifts towards proposing a viable solution, exemplified by developing an explainability mechanism tailored for

integration into AI models utilized in P2P lending.

The subsequent steps involve carefully refining and testing this explainability mechanism, ensuring its effectiveness and functionality before incorporating it into the artefact. The artefact's impact on enhancing creditworthiness and explainability within the P2P lending context is meticulously assessed through experiments and rigorous evaluations. This empirical validation serves as a critical validation of the research's practical applicability and potential to address existing challenges.

Finally, the research results in the artefact being built, making it accessible to practitioners operating in the field of P2P lending. This collaborative approach encourages the utilization of the developed solution in real-world scenarios, fostering practical advancements in the domain. Ultimately, this design science-driven research endeavour offers innovative, real-world solutions and contributes to the broader knowledge base of P2P lending by applying a systematic and solution-oriented strategy.

7.4 Research Choice

In this research, we have opted for a mixed-methods approach, a strategic choice with significant promise. The mixed-methods paradigm offers a comprehensive and multifaceted exploration of the research problem by integrating qualitative and quantitative data collection and analysis techniques.

Qualitative methodologies, such as expert interviews and questionnaires are instrumental in delving deeply into the nuances of creditworthiness and explainability within the P2P lending landscape. They provide valuable insights into the current state of affairs and serve as a means to identify critical issues, opportunities and potential solutions. The open-ended nature of qualitative inquiries allows for a holistic understanding of the domain. On the other hand, the quantitative facet, characterized by the analysis of historical P2P loan transactions, introduces an objective dimension to our research. We can uncover hidden patterns and trends within the data through rigorous statistical analysis, shedding light on the intricate relationships between variables and a borrower's creditworthiness. Furthermore, this quantitative approach enables us to evaluate the accuracy of various explainability mechanisms employed in AI models.

We create a synergistic relationship between qualitative and quantitative methodologies by embracing this mixed-methods strategy. The collaboration between these two methodologies enables us to analyze and evaluate the outcomes of each strategy, thus enhancing our comprehension of the research aim. The above-mentioned approach establishes a solid basis for producing recommendations and formulating solutions, improving our final results' overall transparency and accuracy.

7.5 Time Horizons

Our research of improving creditworthiness using XAI models in P2P lending involves adopting a longitudinal temporal horizon. This specific temporal framework entails the systematic accumulation of data over a prolonged duration, presenting the opportunity for a comprehensive and intricate comprehension of the research being examined.

The dataset included in our study comprises of a comprehensive collection of P2P loan transactions that span twelve years. The longitudinal technique demonstrates its significance by effectively uncovering temporal changes and trends within the dataset. This enables the monitoring of various shifts, such as fluctuations in interest rates and the progression of AI model explainability. It facilitates the identification of patterns and potential factors influencing these shifts. This holistic viewpoint enhances our understanding of the study problem, resulting in more accurate and dependable outcomes.

In addition, using a longitudinal temporal horizon provides us with an added advantage in assessing the potential long-term consequences of different options. Analyzing long-term data trends makes it possible to accurately identify potential dangers, evaluate the advantages of various approaches and make informed decisions regarding the optimal course of action. However,

it is essential to acknowledge that a longitudinal timeframe may also present challenges, including the potential for economic fluctuations and changes in financial laws and regulations.

7.6 Data collection and Data Analysis

Various lending-related datasets from diverse sources were identified, including *Bondora* (n.d.), *Lending Club* (n.d.), and *Prosper* (n.d.). These datasets encompassed a wealth of information concerning loan applications, credit scores and other pertinent factors that held immense potential for creating prediction models and assessing creditworthiness.

After conducting a detailed exploratory statistical analysis of these datasets, we decided to determine the most appropriate dataset for our research. Ultimately, we chose to utilize the *Lending Club* (n.d.) dataset, which has been widely accepted and employed in previous research. This led to an increase in the credibility of our research outcomes compared to other available datasets. As a result, our endeavours to identify significant factors associated with loan repayment and other crucial outcomes were substantially aided by this selection.

Before utilizing the dataset in our investigation, a thorough data pre-processing phase was executed. This crucial pre-processing stage ensured that the data was suitably prepared for our research objectives and elaborated upon in the implementation segment. The *Lending Club* (n.d.) dataset that had been processed beforehand served as the cornerstone for devising models and assessing creditworthiness and interpretability. Such measures were implemented to guarantee the durability and dependability of our research results, enabling us to deliver significant contributions to the domain of P2P lending by leveraging HCAI.

Lending Club dataset can be found through this link:
`https://www.kaggle.com/datasets/ethon0426/lending-club-20072020q1`

8 Implementation

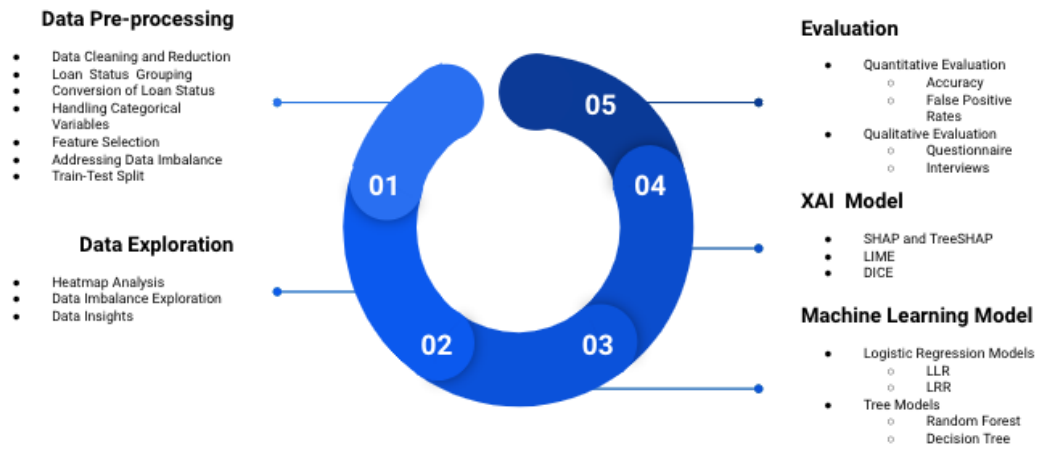


Figure 9: Implementation Process

8.1 Data Pre-processing

In the next section, we outline the measures taken for data pre-processing on the original dataset from *Lending Club* (n.d.), which covers a period from 2007 to 2020Q3. The primary dataset was composed of 800,000 records and encompassed 142 factors. Several pre-processing steps were executed to ensure that the data is suitable to train our ML models and tackle any resource constraints.

8.1.1 Data Cleaning and Reduction

The initial stage involved addressing the dataset's null values (N/A). All records that included missing values were eliminated, leading to a reduction in observations, leaving 457,824 entries in the dataset. Consequently, a randomly shuffled subset of 100,000 records was extracted from these 457,824 records to work within the resource constraints while preserving the randomness and representativeness of the data.

8.1.2 Loan Status Grouping

The "loan_status" feature is the dependent variable and it has multiple categories in the current dataset, such as "Charged-Off", "Default", "Fully Paid", "Issued" and "Current". To simplify the analysis, the "Charged-Off" and "Default" categories were merged and labelled as "Default". Similarly, the "Issued" and "Current" categories were merged and considered as "Current". We decided to retain only two main categories, "Fully Paid" and "Default", as they align with the research objective of predicting loan defaults.

8.1.3 Conversion of Loan Status

In order to facilitate the binary classification task, the "Fully Paid" category was encoded as 0, representing fully paid loans and the "Default" category was encoded as 1, representing loans that resulted in default.

8.1.4 Handling Categorical Variables

Since ML models, such as the RF model, classification tree and LR model, are utilized in this research only to support numerical inputs, the categorical variables were converted into numerical format through one-hot encoding. However, specific categorical attributes were deemed irrelevant or carried numerous unique values, potentially leading to overfitting. Therefore, the following attributes were dropped from the dataset: 'id', 'sub_grade', 'emp_title', 'issue_d', 'pymnt_plan', 'url', 'title', 'zip_code', 'addr_state', 'earliest_cr_line', 'revol_util', 'initial_list_status', 'last_pymnt_d', 'last_credit_pull_d'.

8.1.5 Feature Selection

Upon completing the data pre-processing phase, the dataset retained 100 numerical features. An additional step of dimensionality reduction was carried out using the K Best method to enhance model performance and reduce model complexity. This method effectively identified the 23 most pertinent independent features and the dependent feature "loan_status." Consequently, the final dataset size was reduced to (63448 rows x 23 columns). The relevance of these identified features was further corroborated by comparing them with findings from past research that utilized the same *Lending Club* (n.d.) dataset and insights gained from domain expertise. Below is the list of features identified as inputs:

Feature	Description
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
term	The number of months that the borrower will be on the settlement plan.
installment	The monthly payment owed by the borrower if the loan originates.
grade	LC assigned loan grade.
home_ownership	The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER.
annual_inc	The self-reported annual income provided by the borrower during registration.
verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified.
purpose	A category provided by the borrower for the loan request.
delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years.
fico_range_low	The lower boundary range the borrower's FICO at loan origination belongs to.
fico_range_high	The upper boundary range the borrower's FICO at loan origination belongs to.
out_prncp_inv	Remaining outstanding principal for portion of total amount funded by investors.
total_pymnt	Payments received to date for total amount funded.
total_rec_prncp	Principal received to date.
total_rec_int	Interest received to date.
total_rec_late_fee	Late fees received to date.
last_pymnt_amnt	Last total payment amount received.
tot_coll_amt	Total collection amounts ever owed.
il_util	Ratio of total current balance to high credit/credit limit on all install acct.
open_rv_12m	Number of revolving trades opened in past 12 months.
max_bal_bc	Maximum current balance owed on all revolving accounts.
inq_last_12m	Number of credit inquiries in past 12 months.
int_rate	Interest rate.

8.1.6 Addressing Data Imbalance

One critical issue in the dataset was the class imbalance in the "loan_status" column, as the number of default instances was likely much smaller than the fully paid instances. To mitigate the impact of this problem, we employed oversampling techniques, which involved creating synthetic samples of the minority class (default) to balance the dataset. As a result, the dataset was oversampled and the final dataset used for input data in the subsequent black-box AI models contained 81934 rows x 23 columns.

Following these data pre-processing steps ensured the dataset was well-structured and suitable for ML models. We addressed missing values, data imbalance and categorical variables, allowing us to train the AI models and obtain meaningful insights from the research data.

8.2 Train-Test Split and Cross-Validation

With the pre-processing steps completed, the decision to employ an 80-20 split for the dataset division into a training set and a testing set is supported by several key considerations. This widely adopted practice, known as the "Pareto Principle" (Backhaus 1980) ensures a judicious balance between the amount of data used for training and testing in our machine learning models. Allocating 80% of the pre-processed data to the training set provides a substantial volume for the model to learn and generalize effectively, contributing to the robustness of the final model. Simultaneously, reserving the remaining 20% for the testing set enables a comprehensive evaluation of the model's performance on unseen data, simulating real-world scenarios. This strategic allocation helps mitigate overfitting risks, enhances the model's generalization capabilities and strikes a practical balance in terms of computational resources and time constraints. The 80-20 split, therefore, serves as a well-founded approach, contributing to the reliability and validity of our research outcomes.

The choice of using 100,000 records as a validation set from the initial pre-processed dataset of 457,824 records is grounded in the pursuit of rigorously evaluating the generalization capability of our machine learning models while mitigating the risk of overfitting. By selecting a substantial yet independent subset, we aim to provide a robust validation mechanism during the training process. This ensures that the models are not overly tailored to the nuances of a specific set of data, promoting a more generalized performance. The 100,000-record subset serves as a representative sample, enabling effective cross-validation. During training, this validation set becomes instrumental in fine-tuning hyperparameters and gaining valuable insights into the model's performance across different data partitions. Consequently, this meticulous approach enhances the reliability of our models by subjecting them to a rigorous evaluation process, fostering their adaptability to diverse datasets and real-world scenarios.

By employing this train-test split and cross-validation strategy, we aimed to strike a balance between training our models on a sufficiently large dataset while ensuring robust evaluation metrics on unseen data. This approach enhances the reliability of our research findings and strengthens the predictive power of our ML models.

In the next sections, we present the details of the ML algorithms employed, model training procedures, evaluation metrics and the results of our predictive analysis based on this carefully pre-processed and partitioned dataset.

8.3 Data Exploration

Data exploration is critical in understanding the dataset's underlying patterns, relationships and potential imbalances. This section presents the findings obtained through visualization techniques, including heatmap analysis and bar charts.

8.3.1 Heatmap Analysis

A heatmap was constructed to identify the correlation between the features in the dataset. The heatmap visualizes the pairwise correlations between numerical features, providing insights into the strength and direction of relationships. Figure 9 showcases the heatmap representation.

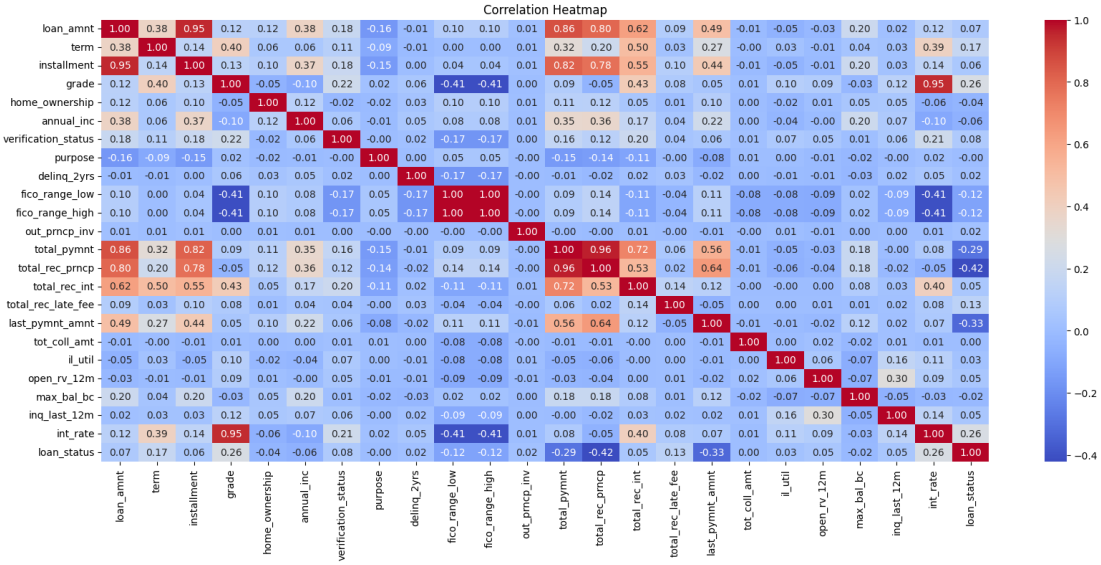


Figure 10: Heatmap

By representing the correlation between various features and creditworthiness scores, the heat map provides a comprehensive overview of the underlying relationships within the dataset. Through color gradients, it highlights the strength and direction of correlations, thereby enabling researchers to identify key factors influencing creditworthiness. This visualization facilitates the identification of important predictors and aids in uncovering potential patterns or anomalies that may inform model development and decision-making processes.

8.3.2 Data Imbalance Exploration using Bar Charts

Another important aspect of data exploration is evaluating potential data imbalances, especially in the target variable "loan_status." Bar charts were employed to visualize the distribution of loan statuses within the dataset.

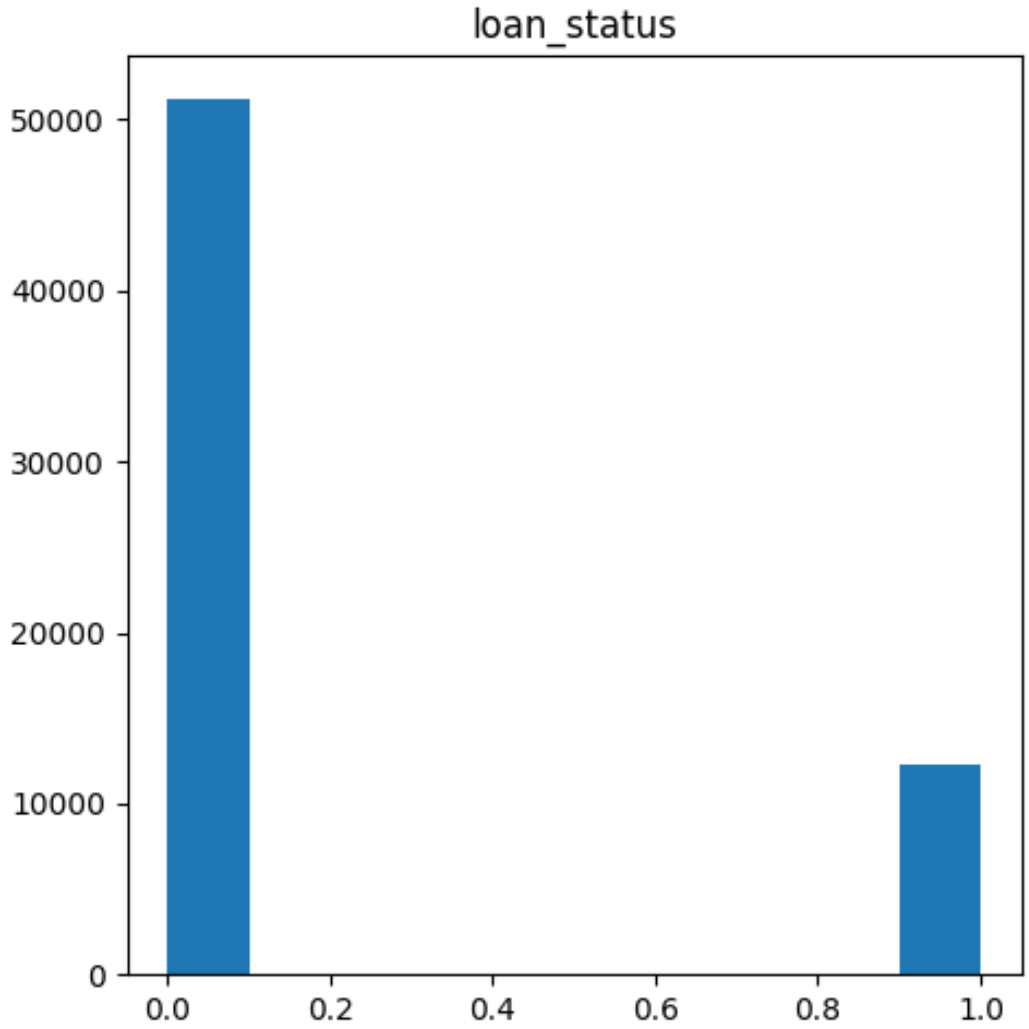


Figure 11: Loan Status Distribution

Figure 10 presents the bar chart illustrating the distribution of loan statuses. The dataset suffers from class imbalance, as the number of fully paid loans (Class 0) significantly outweighs the number of defaulted loans (Class 1). This imbalance may pose challenges during model training, potentially leading to biased predictions. To address this issue, we previously applied oversampling techniques to create synthetic samples of the minority class, thereby reducing the imbalance in the dataset and improving the performance of subsequent ML models.

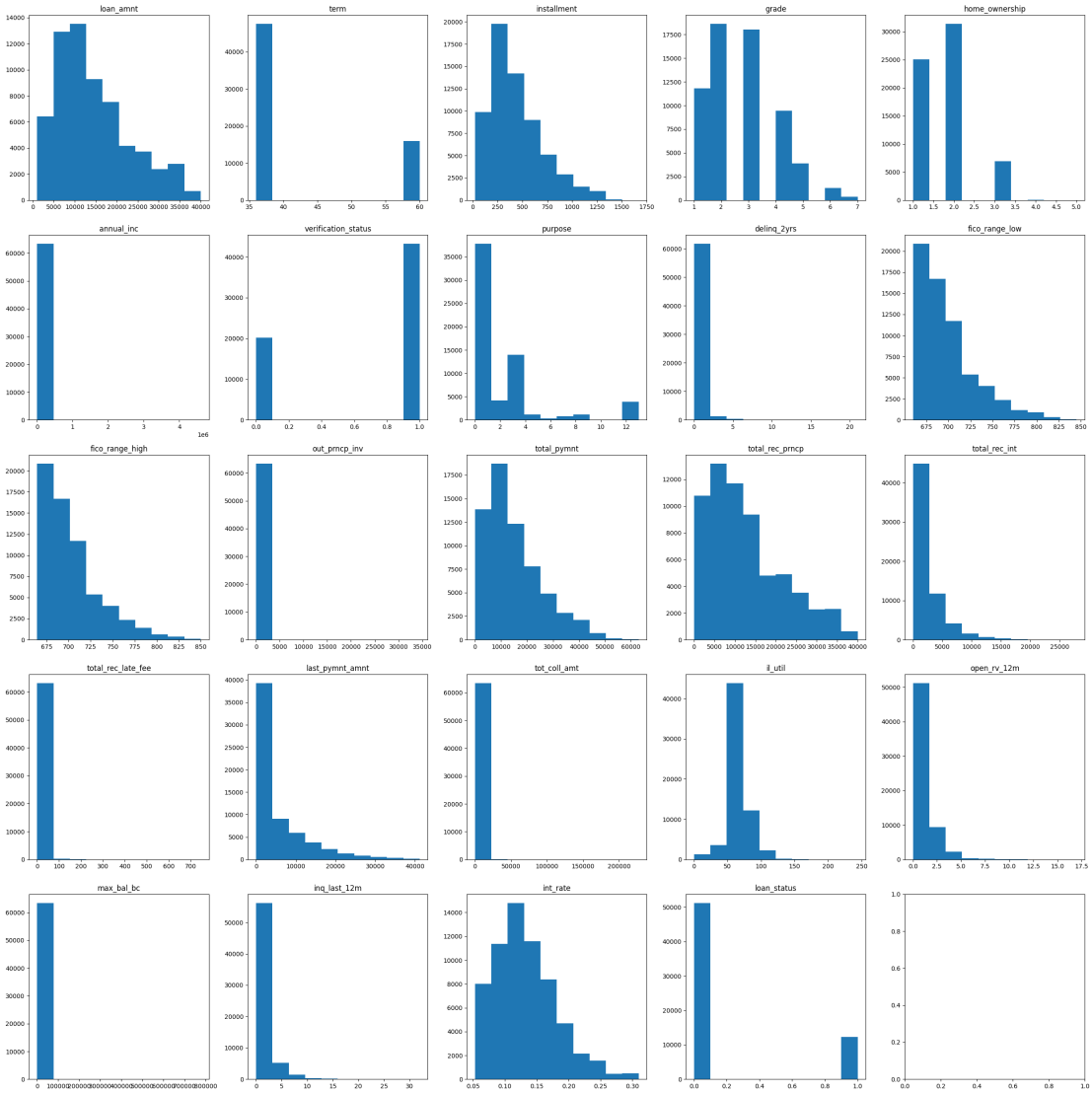


Figure 12: Input Feature Distribution

Figure 12 provides a visual representation of the data distribution across each feature, offering valuable insights into potential biases within the dataset. With a total of 24 features, including the dependent variable (loan_status), the illustration allows for a comprehensive examination of key attributes such as loan amount, term, installment, grade, and home ownership. By analyzing the distribution patterns, researchers can effectively identify any discrepancies or imbalances that may exist within the data, thus enabling proactive measures to mitigate biases and ensure the integrity and fairness of subsequent analyses and model development processes.

8.3.3 Data Insights from Exploration

Data exploration through visualization techniques has provided valuable insights into the dataset. The heatmap analysis revealed correlations among numerical features, aiding in feature selection and understanding potential dependencies. Additionally, the bar chart highlighted the data imbalance in the target variable "loan_status", which was mitigated through oversampling techniques to ensure unbiased model training and accurate predictions.

These initial data exploration steps are a foundation for subsequent analyses and model building. By gaining a deeper understanding of the data distribution and inter-feature relationships, we can proceed with greater confidence in generating meaningful outcomes and contributing to the overall objectives of this research.

8.4 Machine Learning (ML) models

This section aims to select the most accurate ML model from a pool of black-box AI models, including logistic regression and tree models.

8.4.1 Logistic Regression (LR) Models

Following a comprehensive literature review, we identified four variants of logistic regression models: Binary Logistic Regression (BLR), Logistic Regression with Fixed coefficients (LRF), Logistic Regression with Random coefficients (LRR) and Lasso-Logistic Regression (LLR) (Königstorfer and Thalmann 2020; Smith 2019; Serrano-Cinca and Gutiérrez-Nieto 2016; Setiawan et al. 2019; Z. Zhang, Niu, and Y. Liu 2020). Drawing from previous research findings, we observed that LLR models and regression with random coefficients exhibited superior performance compared to other logistic regression models across diverse scenarios (Hong Wang, Q. Xu, and L. Zhou 2015). As a result, for our research purposes, we have chosen to Lasso Logistic Regression models with random coefficients, which had served as our primary candidate model within the logistic regression category.

8.4.2 Tree Models

In the context of tree models, we have opted to utilize the RF, XGBoost and Classification tree models. The RF model is renowned for its resilience and proficiency in handling intricate relationships within the data, while classification trees offer a straightforward and transparent structure for the model. Our research has chosen to incorporate both RF and classification tree models within the field of tree models. The RF model is recognized for its robustness and effectiveness in managing complex connections within the data, whereas classification trees present a clear and transparent framework for the model. XGBoost, renowned for its high predictive performance, excels in capturing non-linear relationships within the data and incorporates regularization techniques to prevent overfitting. The algorithm's ability to provide feature importance scores enhances interpretability, crucial in the context of loan default prediction. Additionally, XGBoost's scalability makes it well-suited for handling large datasets, aligning with the demands of the P2P lending industry.

These choices were also influenced by pertinent literature on forecasting loan defaults in P2P lending, confirming their appropriateness for our research goals (Setiawan et al. 2019; J. Zhou et al. 2019). The subsequent step is to identify the most appropriate ML model with regard to interpretability. This will be achieved by applying XAI models to gain more profound insights into each of the previously mentioned models' decision-making processes. XAI techniques will enable us to uncover rationales behind predictions and understand individual feature impact on outcomes. After identifying the most interpretable model, we will select an effective XAI model that provides transparency and helps ensure that generated explanations are clear and applicable for lenders and borrowers in the P2P lending process. The following section comprehensively describes the XAI models used for this purpose. Using innovative methodologies, our goal is to enhance trust and transparency in loan default prediction's decision-making process within the P2P lending industry.

8.4.3 XAI Models

In order to improve the comprehensibility of our ML algorithms and acquire significant knowledge regarding their decision-making abilities, we utilized three contemporary XAI models, namely SHAP, LIME and DiCE. These models come with unique benefits that enable the delivery of interpretations locally and globally.

1. SHAP (SHapley Additive exPlanations)

   The SHAP model proved to be a precious asset in the domain of XAI, as it could provide local and global interpretability. When examining local interpretability, SHAP offered thorough insights into individual predictions, thereby facilitating a deeper comprehension of the contribution made by each feature in a particular prediction (Lundberg et al. 2020). This attribute enabled us to discern the pivotal factors that influenced the model's decision-making process concerning individual loan applications.

   In global explanations, SHAP furnished all-inclusive perspectives on the overall influence of each characteristic throughout the entire dataset. This holistic understanding of feature importance assisted us in identifying critical attributes influencing loan default predictions consistently throughout the dataset.

2. LIME (Local Interpretable Model-agnostic Explanations)

   LIME was instrumental in providing local interpretations of our ML models. LIME generated simple and transparent explanations for individual predictions by approximating the models locally (Ribeiro, S. Singh, and Guestrin 2016). This allowed us to grasp the reasons behind specific loan approval or rejection decisions on a case-by-case basis.

   While LIME excelled in producing interpretable explanations for individual instances, it demonstrated limitations in understanding the model's global behaviour comprehensively. However, its localized interpretability proved valuable when focusing on individual predictions was essential.

3. DiCE (Diverse Counterfactual Explanations)

   DiCE provided a unique and insightful perspective by offering local interpretations in terms of counterfactuals. These counterfactuals guided the changes required in specific features to alter the final prediction outcome (Mothilal, Sharma, and Tan 2020). For example, DiCE enabled us to determine the precise adjustments needed to increase the chances of loan approval for a given loan application, transitioning it from a predicted default to a projected payable loan. Such actionable insights made possible by DiCE allowed us to understand the "what-if" scenarios, offering crucial guidance to borrowers and lenders in improving loan applications and reducing potential default risks.

9 Results

9.1 Results of Machine Learning (ML) Models

The results of our study on loan default prediction in P2P lending through ML models are presented in this section. The performance of each model was assessed using three vital metrics, namely F1 Score, Accuracy and AUC. Among the LLR models with random coefficients that were tested, one achieved an F1 Score of 0.9908, an Accuracy of 0.9943 and an AUC of 0.9855, indicating its high level of efficacy. Additionally, the Classification Tree model was found to have an F1 Score of 0.7541, an Accuracy of 0.7978 and an AUC score reading of 0.8648. At the same time, the RF Classifier excelled amongst all models with a precise F1 Score value of 0.9952 and highly accurate values for accuracy (at 0.9970) and AUC (at 0.9916). Notably, the XGBoost model demonstrated exceptional performance, with an F1 Score of 0.9612, Accuracy of 0.9769 and AUC of 0.9402. These results highlight the predictive capabilities of the XGBoost algorithm in the context of loan default prediction within the P2P lending landscape.

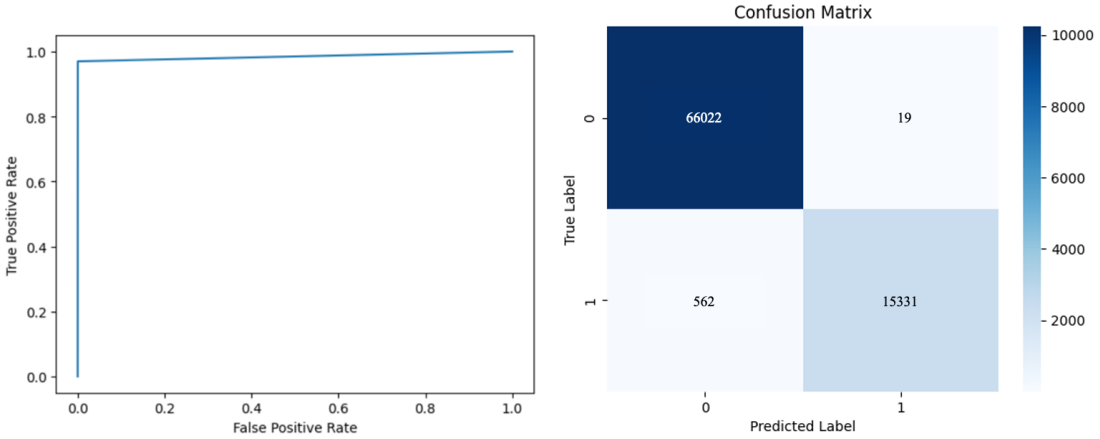


Figure 13: Result of LLR Ensemble Model

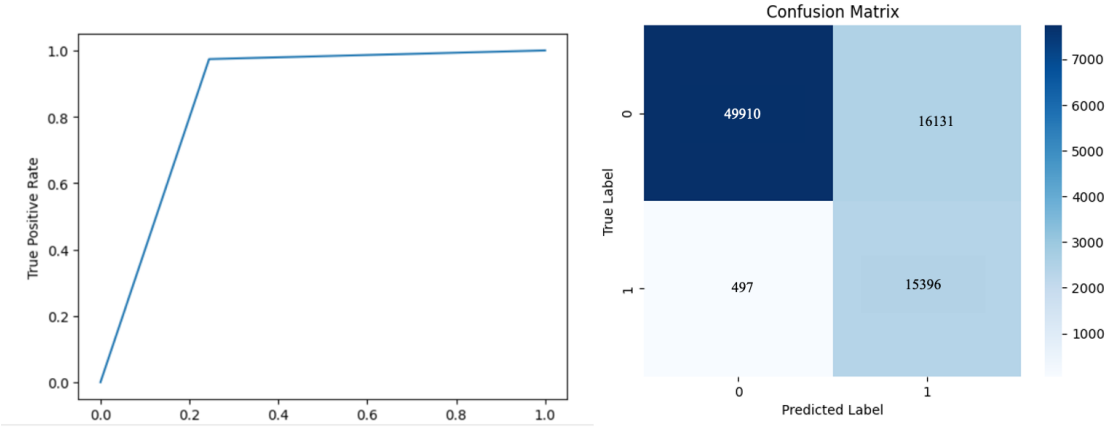


Figure 14: Result of Classification Tree Model

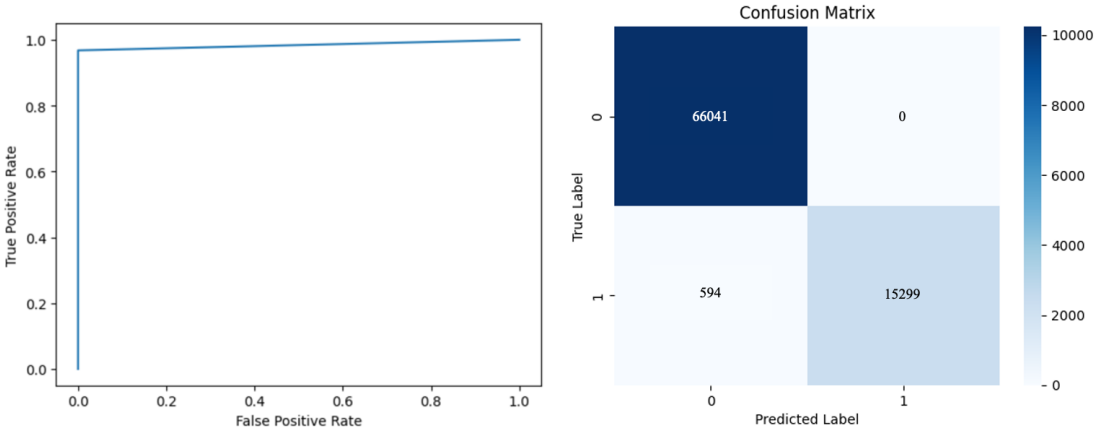


Figure 15: Result of Random Forest Model

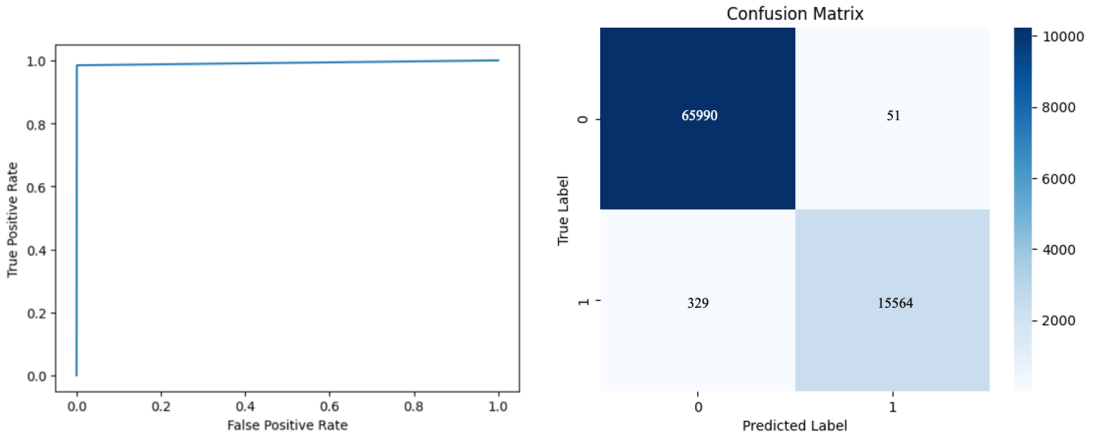


Figure 16: Result of XGBoost Model

The RF classifier was the most accurate of all the analyzed ML models. It exhibited remarkable accuracy and discriminatory power, rendering it a dependable option for forecasting loan defaults. Its prowess in managing intricate data relationships and utilizing diverse decision trees contributed significantly to its superiority over other models. Moreover, due to its transparent construct, stakeholders found it easier to comprehend factors that impact loan default predictions.

9.2 Results of XAI Models

While exploring XAI models, we utilized three primary models: SHAP, LIME and DiCE. Each method furnished valuable insights into the interpretability of ML models. In particular, SHAP emerged as an influential XAI model that provides global and local interpretations. Conversely, LIME concentrated on delivering localized interpretations while providing straightforward explanations for individual predictions. Though helpful in comprehending specific cases, its limited scope in capturing overall model behaviour made it less appropriate for obtaining a comprehensive view of feature importance. DiCE offers unique features, such as generating counterfactuals to create "what-if" scenarios that modify prediction outcomes. These counterfactuals proved highly insightful in determining necessary changes to loan features required to transform default predictions into payable loans assisting borrowers; and lenders in improving loan applications.

9.2.1 Global level explanation using SHAP

This provides a global level explanation by assigning a value to each feature in a model, indicating its contribution to the model's output across all instances.

The below beeswarm plot offers an informative summary of how the top features in a dataset influence the model's output. In this plot, each instance is represented by a single dot on each feature row. The position of the dot along the x-axis corresponds to the SHAP value of that feature for the given instance, with dots accumulating to show density along each feature row. Additionally, color is utilized to indicate the original value of a feature. This visualization provides a comprehensive understanding of the impact of various features on model predictions, allowing for insightful analysis of feature importance and model behavior across different instances in the dataset.

As per the above beeswarm plot, when total principal amount received to date is high, the possibility of the person getting being able to not pay the loan is low (Negative SHAP value). Another insights is consideration is that most people who have total principal amount recieved to date as low, thus their loan applications won't get approved. Similarly, it is evident that when loan amount is high, there is a higher possibility of the borrower not being able to pay the loan.

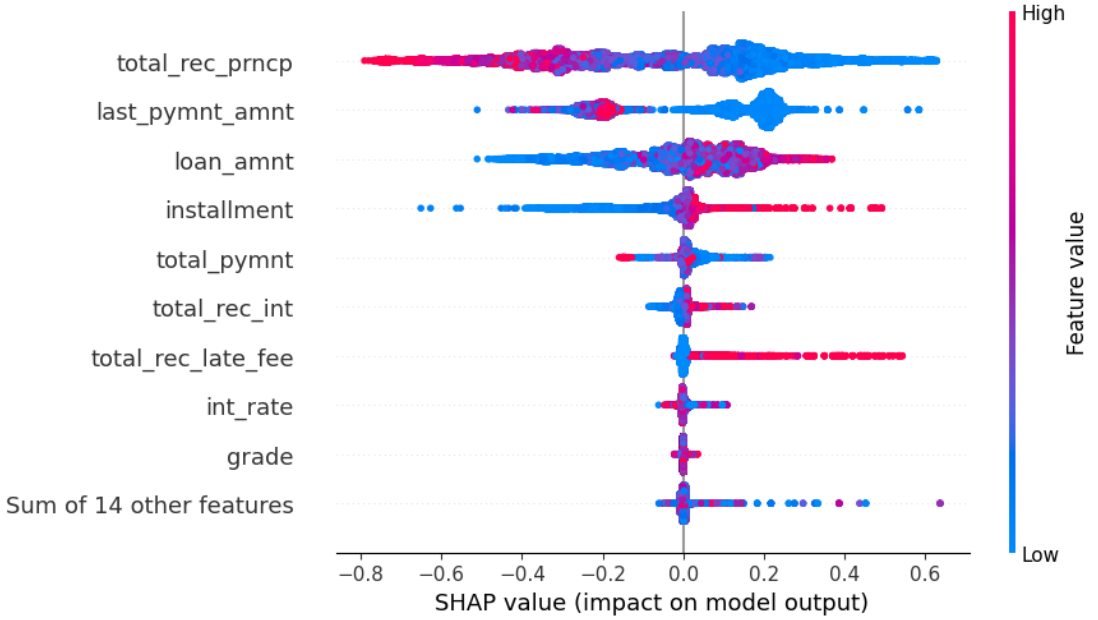


Figure 17: Global level explanation using SHAP - Beeswarm Plot

Following is a global feature importance plot generated by passing a matrix of SHAP values to the bar plot function. This plot showcases the mean absolute value for each feature across all samples, offering insights into the relative significance of different features in influencing model predictions. It serves as a valuable tool for understanding the key drivers behind the model's behavior and informing decision making processes such as model evaluation and feature selection in machine learning applications.

As per the global bar plot, 4 highest impacting factors for the loan decisions in the descending order are total received principal amount, last payment amount of the loan, loan amount and installment amount. And it is also evident that only the top 7 factors are impacting more than 98% of the loan decisions as per the selected dataset.

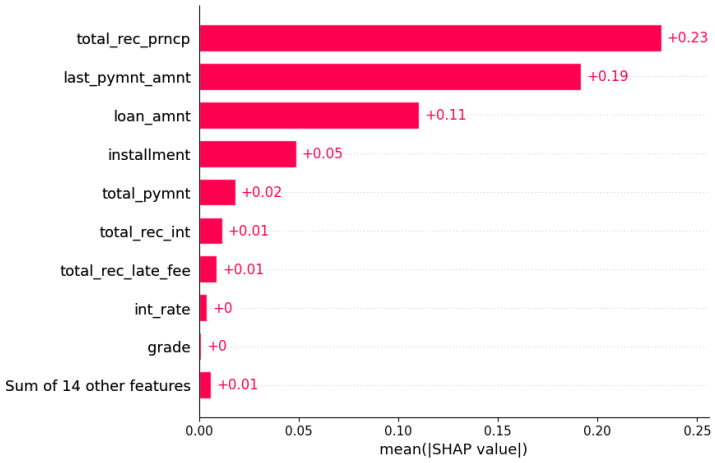


Figure 18: Global level explanation using SHAP - Global Bar Plot

Cohort bar charts offer a comprehensive view of how different classes, such as default and non-default classes, contribute to various factors influencing loan decisions. Each bar in the chart represents a specific factor or feature, segmented to show the distribution of classes within that feature. By examining these contributions, analysts can discern the relative impact of each class across different factors. This visualization not only highlights the importance of each feature but also provides insights into how default and non-default classes influence loan decisions differently. Such analysis enables stakeholders to make informed decisions regarding risk assessment, loan approval processes and mitigation strategies.

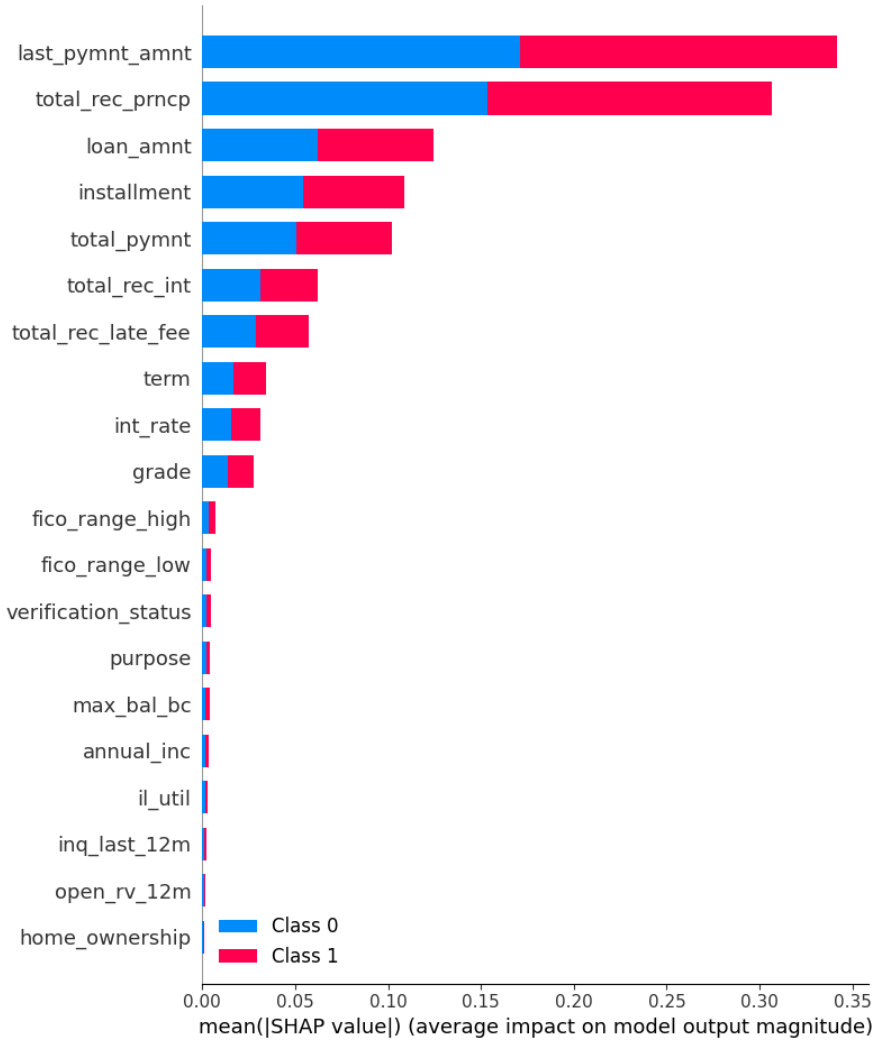


Figure 19: Global level explanation using SHAP - Cohort Bar Chart

Understanding global feature importance using SHAP in lending can reveal which factors, such as total received principal amount, last payment amount, funded amount and installment have the most significant impact on loan application decisions across the entire dataset.

9.2.2 Local level explanation using SHAP

At the local level, SHAP provides insights into the decision-making process for individual instances by explaining the contribution of each feature to a specific prediction. This examination of local explanations with SHAP is particularly valuable for lenders seeking to understand the rationale behind specific loan application decisions, thereby enhancing transparency and fostering trust in the lending process.

For investors, the application of SHAP at the local level allows them to comprehend why a particular borrower is categorized as capable of paying back the loan or potentially defaulting. This transparency in decision-making, backed by detailed explanations, serves as valuable insights into the software application, aiding investors in making informed lending decisions.

Additionally, for borrowers, SHAP's local explanations offer insights into why their loan application might be rejected. This information empowers borrowers by providing a clear understanding of the factors influencing their creditworthiness, enabling them to take targeted actions to improve their credit rating for future loan applications.

The below force plot dynamically illustrate the contributions of individual features to a model's predictions for a specific instance. Each feature is represented as a bar, with the length indicating the magnitude and direction of its influence on the prediction. The plot begins with the baseline prediction at the center and the contributions of each feature are depicted as arrows pushing the prediction up (positive influence) or down (negative influence). As the features interact, their combined effects are reflected in the final prediction. Force plots offer an intuitive way to interpret how different features drive individual predictions, aiding in model understanding and decision-making processes.

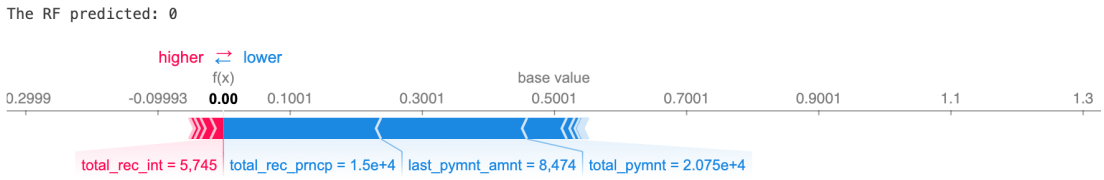


Figure 20: Local level explanation using SHAP - Force Plot

As per the above Force plot, the borrower will get defaulted (Class 0). The primary factors affecting this decision are loan amount, recovered amount, total received principal amount and total received interest. Out of the aforementioned 4 factors, loan amount, recovered amount and total received principal amount are pushing the loan decision towards default (class 0), which means the loan application will get rejected, while total received interest amount is pushing it towards class 1, which means loan decision will get approved. However, since the combined impact of loan amount, recovered amount and total received principal amount is greater than that of total received interest. Hence, the ultimate loan decision on the above prediction is that the loan application will get rejected.

Waterfall plots provide detailed explanations for individual predictions, requiring a single row of an Explanation object as input. The plot begins at the expected value of the model output, with each subsequent row illustrating how the positive (red) or negative (blue) contribution of each feature influences the output from the expected model output over the background dataset to the model output for the specific prediction. This visualization offers a clear depiction of the impact of each feature on the prediction, facilitating the interpretation of model decisions and enhancing transparency in the prediction process.

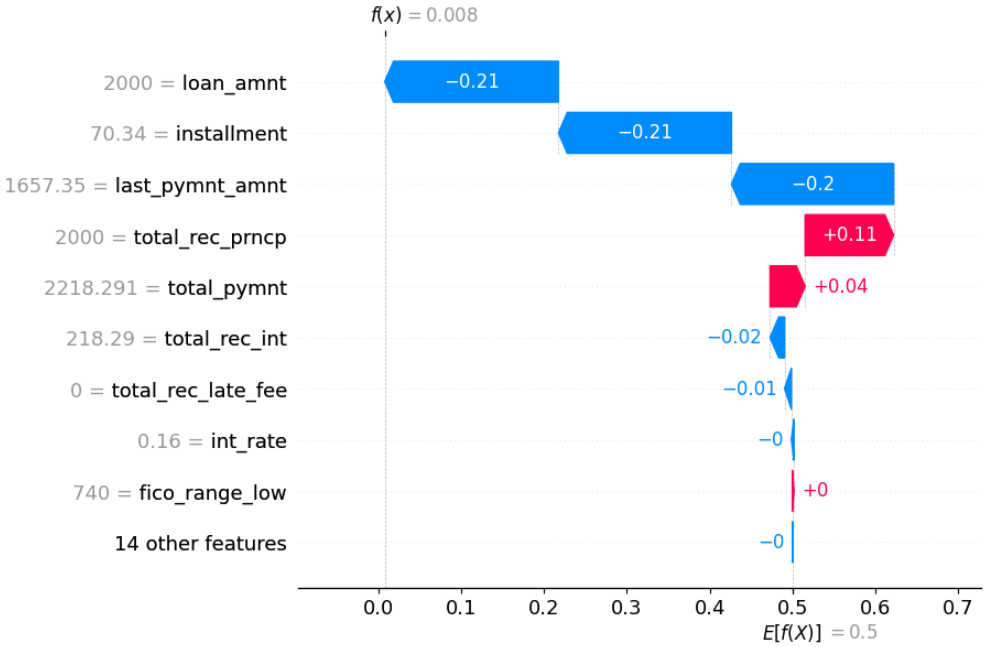


Figure 21: Local level explanation using SHAP - Waterfall Plot

As per the above waterfall plot, loan amount, installment, last payment amount, total received interest, late fees received to date, interest rate and 14 other features are pushing the loan decision towards Class 0, while factors such as total received principal amount, total payment and FICO range low are pushing the loan decision towards Class 1. However, the impact of the factors pushing the loan decision towards Class 0 from base value 0.5 is greater than that of the factors that push the loan decision towards Class 1. Therefore, as per the above waterfall plot, it is evident that the prediction is Class 0, means the loan decision will get rejected.

9.2.3 Local level explanation using TreeSHAP (Only for Tree models)

TreeSHAP provides local explanations specifically tailored for tree-based models, breaking down the contribution of each feature in decision paths. In P2P lending, where decision trees and random forests may be used, TreeSHAP helps in understanding the detailed rationale behind individual loan application decisions, aiding in model interpretation and validation.

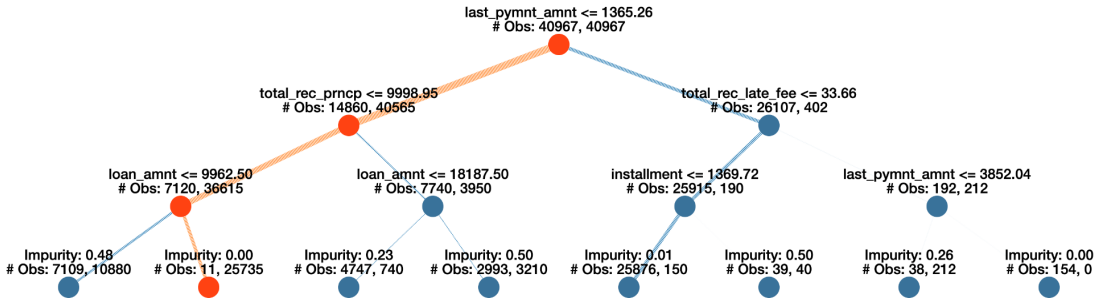


Figure 22: Local level explanation using TreeSHAP - 1

As per the above decision tree, the loan decision will get approved and the set of decision points are highlighted in orange color. The loan decision are affected by 3 main factors:last payment amount, total received principal amount and loan amount. The rest of the factors are considered as impurities as their impact is significantly low. Decision tree also depict how decisions are made in each step of the tree starting from its root (most impacting factor) towards the leaf nodes (least impacting factors).

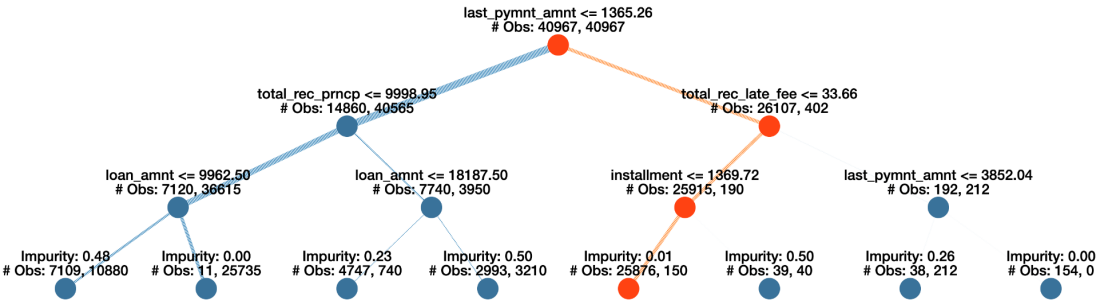


Figure 23: Local level explanation using TreeSHAP - 2

Similar to the previous decision tree, in this decision tree also it highlights the path in which the decision were made. However, in this prediction, the loan application will get rejected and it is affected by 3 main factors: last payment amount, total late fees received to date and installment amount.

9.2.4 Local level explanation using LIME

LIME generates locally faithful explanations for complex models by approximating them with simpler, interpretable models for individual instances. LIME can be applied in P2P lending to provide understandable, local explanations for specific loan decisions, even when using intricate machine learning models, improving interpretability for stakeholders.

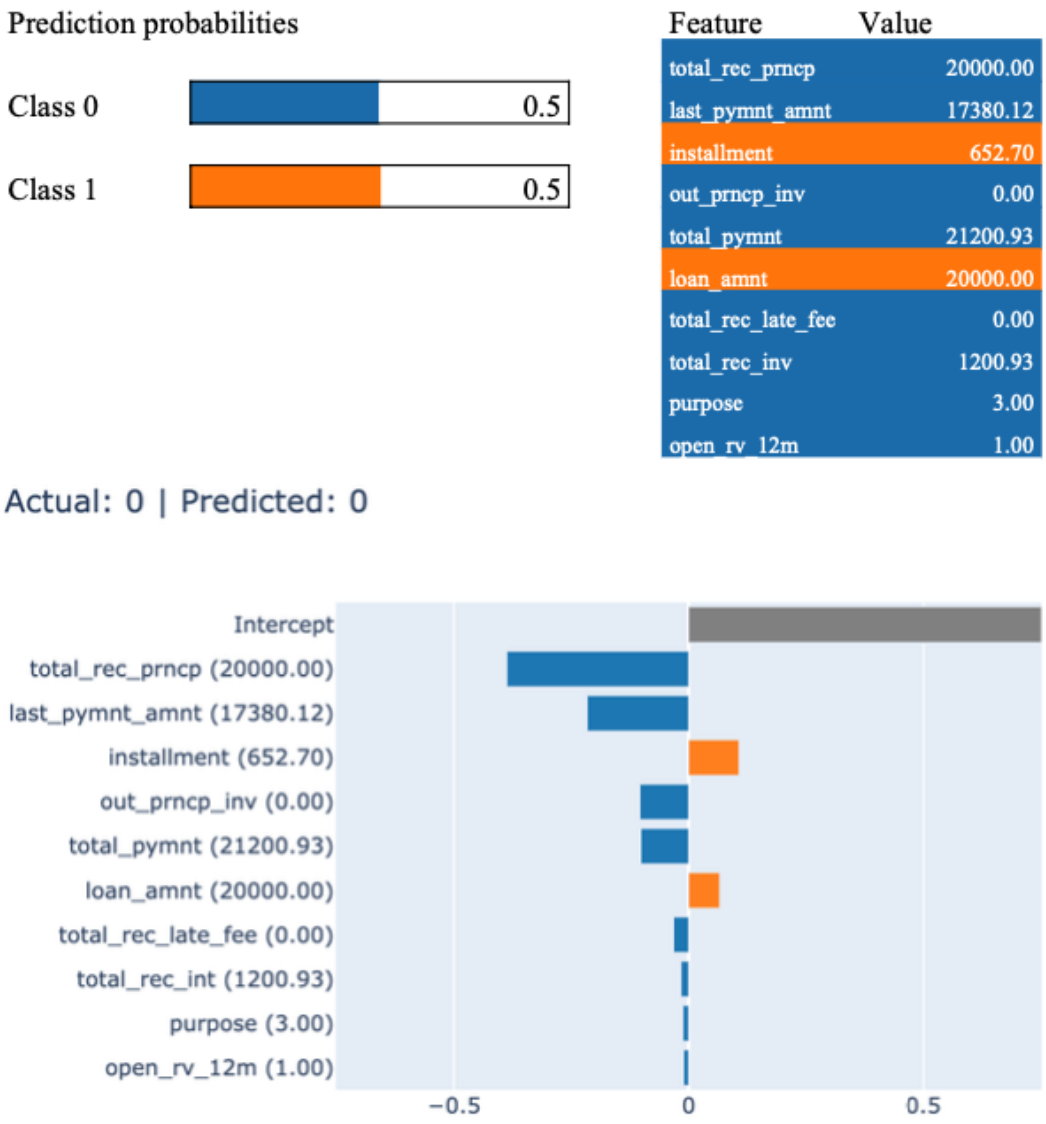


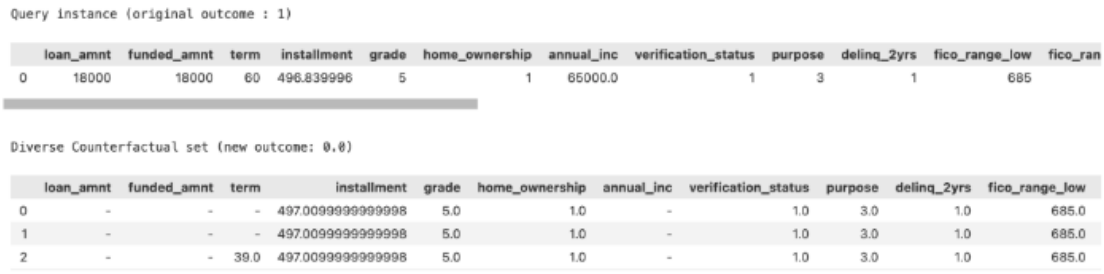
Figure 24: Local level explanation using LIME

First diagram out of the above 3 diagrams depict the biases of the model. Since this is a binary class classification problem, ideally the biases of the model towards each class should be 0.5. And with the first diagram, it depicts that the model is not biased towards either of the classes. Second diagram depicts the lime values as well as the top 10 factors that affected this specific loan decision. In the third diagram, it visually represents how each factor is affecting the loan decision to be categorised into Class 0 with their respective impacting lime values.

9.2.5 Counterfactuals using DiCE

DiCE generates diverse counterfactual instances, representing alternative scenarios that lead to a different model prediction while maintaining realism. DiCE can be used to generate realistic counterfactual scenarios for loan applications, helping lenders understand how changing specific features might alter the decision, thereby aiding in sensitivity analysis and model fairness assessments.

Take a look at the following two illustrations:



Query instance (original outcome : 1)

	loan_amnt	funded_amnt	term	installment	grade	home_ownership	annual_inc	verification_status	purpose	delinq_2yrs	fico_range_low	fico_ran
0	18000	18000	60	496.839996	5	1	65000.0	1	3	1	685	

Diverse Counterfactual set (new outcome: 0.0)

	loan_amnt	funded_amnt	term	installment	grade	home_ownership	annual_inc	verification_status	purpose	delinq_2yrs	fico_range_low
0	-	-	-	497.0099999999998	5.0	1.0	-	1.0	3.0	1.0	685.0
1	-	-	-	497.0099999999998	5.0	1.0	-	1.0	3.0	1.0	685.0
2	-	-	39.0	497.0099999999998	5.0	1.0	-	1.0	3.0	1.0	685.0

Figure 25: Counterfactuals using DiCE - 1

Query instance (original outcome : 1)

fico_range_high	out_prncp_inv	total_pymnt	total_rec_prncp	total_rec_int	total_rec_late_fee	recoveries	last_pymnt_amnt	tot_coll_amt	il_util	open_rv_12m	ma
689	0.0	14013.450195	6030.37	7958.24	24.84	0.0	496.84	0.0	98.0	7.0	

Diverse Counterfactual set (new outcome: 0.0)

fico_range_high	out_prncp_inv	total_pymnt	total_rec_prncp	total_rec_int	total_rec_late_fee	recoveries	last_pymnt_amnt	tot_coll_amt	il_util	open_rv_12
689.0	0.0	14014.050000000003	6030.37	7958.24	24.84	0.0	12401.77	0.0	98.0	7
689.0	0.0	14014.050000000003	6030.37	7958.24	24.84	0.0	35098.76	0.0	98.0	12
689.0	0.0	14014.050000000003	6030.37	7958.24	24.84	0.0	20396.35	0.0	98.0	7
689.0	0.0	14014.050000000003	6030.37	7958.24	24.84	0.0	19720.32	0.0	98.0	7
689.0	0.0	14014.050000000003	6030.37	7958.24	24.84	0.0	7570.55	0.0	98.0	7

Figure 26: Counterfactuals using DiCE - 2

In the first illustration, we explore the impact on a loan decision that is initially predicted to default. I have focused on altering only two variables—specifically, the last payment amount and the open revolving amount in the last 12 months. By manipulating these variables, I generated five potential counterfactual scenarios, transforming the initial prediction of a defaulted loan into one of non-default.

The practical application of this approach is to offer borrowers insights into improving the chances of their loan application being approved, especially after a rejection. For investors, it provides valuable insights into how a person initially deemed as non-defaulting could potentially default if certain factors or a combination of factors undergo changes.

10 Evaluation

10.1 Evaluation of Machine Learning (ML) Models (RQ 1)

In order to improve the effectiveness of credit evaluation in P2P lending, this study conducted a thorough examination of ML models. The primary objective was to predict loan defaults, a crucial component of lending platforms. We extensively compared three significant performance metrics to achieve this goal: F1 Score, Accuracy and AUC (Area Under the Receiver Operating Characteristic Curve). The F1 Score, a harmonic mean of precision and recall, provides a balanced measure of a model's accuracy in binary classification tasks, highlighting both false positive and false negative rates. Accuracy, offers a straightforward assessment of a model's overall correctness in classifying instances but can be misleading in imbalanced datasets. AUC, derived from the Receiver Operating Characteristic (ROC) curve, quantifies a model's discrimination ability, with higher values indicating superior performance in distinguishing between classes. These metrics were crucial in accurately evaluating the models' capacity to classify loan applications as default or payable categories. Therefore, our findings contribute significantly towards ensuring the dependability of lending decisions.

10.1.1 Lasso Logistic Regression Model with Random Coefficients

During our analysis, one model that particularly stood out was the LLR with random coefficients. Its performance was remarkable, producing notable results such as an F1 Score of 0.9908, an Accuracy of 0.9943 and an AUC of 0.9855. These outstanding metrics demonstrate its ability to achieve a harmonious balance between precision and recall in order to classify loan applications accurately. Additionally, its high accuracy and substantial AUC value confirm its effectiveness in differentiating between positive and negative instances, making it a highly suitable candidate for predicting loan defaults within the domain of P2P lending.

10.1.2 Classification Tree Model

The Classification Tree model exhibited impressive efficiency, exhibiting an F1 Score of 0.7541, an Accuracy of 0.7978 and an AUC of 0.8648. Although these metrics were slightly inferior to the LLR model with random coefficients, the model effectively demonstrated its ability to classify instances and distinguish between the two classes accurately. This level of performance highlighted its potential as a feasible choice for predicting loan defaults, presenting a distinct approach for enhancing credit assessment procedures.

10.1.3 Random Forest Model (RF)

The RF Classifier stands out as the top-performing model among all evaluated models. It demonstrated unparalleled predictive capabilities, achieving an exceptional F1 Score of 0.9952, Accuracy of 0.9970 and AUC of 0.9916. Its unique capacity to handle intricate relationships within the data and utilize the collective intelligence of numerous decision trees resulted in outstanding accuracy and discriminative power. These results clearly establish the RF Classifier as the most suitable and reliable model for predicting loan default in the P2P lending field.

The exceptional model performance's consequences are profound and can potentially revolutionize the P2P lending sector. The RF Classifier's capability to enhance credit assessment effectiveness by considerably enhancing default prediction accuracy and minimizing potential losses is significant. Its resilience and comprehensibility make it an invaluable resource for all P2P lending stakeholders, allowing them to make informed decisions and proficiently alleviate risks.

10.1.4 XGBoost Model

The XGBoost model demonstrated its predictive capabilities with an F1 Score of 0.9612, Accuracy of 0.9769 and an AUC of 0.9402. These metrics highlight the model's superior performance in accurately predicting loan defaults within the P2P lending context. The noteworthy F1 Score

emphasizes the balance between precision and recall, while the high Accuracy and AUC values underscore the model's overall effectiveness and ability to discriminate between default and non-default instances. These results position XGBoost as a robust and reliable algorithm for enhancing the accuracy of loan default predictions in the dynamic landscape of P2P lending.

10.2 Evaluation of XAI Models (RQ 2)

In order to enhance the interpretability of ML models for predicting loan defaults within the P2P lending sector, a thorough assessment was conducted on several XAI techniques. This involved investigating the efficiency of three notable XAI models: SHAP, LIME and DiCE. The objective was to ascertain which models provided greater interpretability and transparency appropriate for our research scenario.

10.2.1 SHAP Model

The SHAP methodology thoroughly evaluates the importance of features in ML models. Examining feature significance at both the local and global levels provides valuable insights on specific instances and across the entire dataset. This dual perspective is crucial for comprehending model decisions that are made for each loan application and discovering broader trends and patterns that affect those decisions. The contribution of SHAP is highly beneficial in promoting transparency, trustworthy and interpretable decision-making regarding model predictions such as loan application approvals.

10.2.2 LIME Model

In contrast, LIME demonstrated exceptional proficiency in providing clear explanations for individual predictions. This enabled us to understand better the reasoning behind classifying a specific loan application as either default or payable, thus furnishing beneficial insights at a local level.

10.2.3 DiCE Model

DiCE expanded upon the concept of interpretability by integrating counterfactuals into their analysis. This novel approach facilitated the identification of factors that impacted loan application outcomes, providing valuable insights into achieving alternative results and enriching our comprehension of model decision-making.

10.2.4 Combined Synergy

Our evaluation revealed that all three XAI models, SHAP, LIME and DiCE, brought unique strengths. While SHAP provided a comprehensive view of feature importance, LIME excelled in transparent interpretations for individual predictions and DiCE offered actionable insights through counterfactuals.

Therefore, it is challenging to definitively determine which XAI model is the most interpretable, as each has its applicability and advantages. Instead, we conclude that all three models have their merits and can be employed in different contexts to enhance the effectiveness of the P2P lending process. The synergy of SHAP, LIME and DiCE models enriched our understanding of ML model decisions, contributing to transparency and applicability in the context of loan default prediction in P2P lending. This combined approach empowers stakeholders to better comprehend model behaviour and make well-informed decisions, ultimately fostering trust and transparency in lending.

10.3 Integration of ML and XAI Models (RQ3)

In our aim to enhance the interpretability and effectiveness of creditworthiness assessment in P2P lending, the integration of Machine Learning (ML) and eXplainable Artificial Intelligence (XAI) models emerged as a critical stage. Initially, our research led us to identify the Random Forest model as the most accurate ML model, effectively addressing our first research question. Next, we investigated the integration of XAI models, including SHAP, LIME, and DiCE, with our ML model and we discovered that each of these models have unique sense of interpretability that would ultimately support our research aim. Hence, three combined models were produced as a result of this integration, which we will be using in this section to obtain the results.

Upon integration, we discovered the varied capabilities of different XAI models. While SHAP demonstrated proficiency in furnishing both global and local explanations, facilitating model refinement and regulatory adjustments, LIME predominantly offered local explanations, beneficial for developers and regulators in model fine-tuning and discrepancy identification. However, these combined models (ML + SHAP, ML + LIME) fell short in catering to the needs of end consumers, particularly lenders and borrowers. This realization underscored the necessity for a consumer-centric approach to interpretability, prompting us to explore the integration of the DiCE model.

The fusion of the DiCE model with our ML framework ushered in a new dimension of interpretability, enabling the provision of counterfactual explanations tailored to the needs of lenders and borrowers. By offering insights into post rejection loan approval enhancement for borrowers and highlighting potential default risks for investors, DiCE enriched the decision-making process with actionable insights pertinent to end-user requirements. This comprehensive approach to interpretability not only fosters transparency and trustworthiness but also empowers stakeholders with valuable information to make informed financial decisions within the dynamic landscape of P2P lending.

10.4 Evaluation of the Combined Model (RQ 4)

In our pursuit of creating an effective and interpretable AI model for loan default prediction in the P2P lending domain, we recognized the importance of interviewing and gathering feedback from experts in ML in the P2P lending domain. This comprehensive approach allowed us to assess the model's performance and interpretability from multiple perspectives.

10.4.1 Questionnaire

This questionnaire is designed to evaluate the effectiveness of the suggested combined model in the context of Peer-to-Peer (P2P) lending, an industry where machine learning (ML) models significantly influence decision-making. The suggested model, incorporating various ML approaches, aims to offer a sophisticated solution for enhancing lending decisions. The survey's purpose is to empirically assess the real-world implications of this combined model and bridge the gap between theory and practical application.

Gathering insights from ML experts in P2P lending holds paramount importance. These experts bring a deep understanding of technical intricacies and challenges within the domain, allowing for a complex evaluation of the suggested model's effectiveness. Furthermore, their expertise offers valuable guidance on adaptability and scalability, ensuring the model remains relevant amid the evolving landscape of P2P lending. By tapping into their perspectives, this survey seeks to validate the practical utility of the suggested combined model.

ML experts, being at the forefront of innovation in P2P lending, provide critical insights for a comprehensive assessment. Their perspectives enrich our understanding of the model's effectiveness, ensuring it aligns with the industry's unique characteristics and complexities. Through this survey, we aim to enhance the model's applicability and impact, validating its utility from the viewpoint of those actively engaged in leveraging ML for lending decisions in the dynamic context of P2P lending.

10.4.2 Questionnaire Design

The questionnaire design for the evaluation of the suggested combined model in P2P lending was thoughtfully constructed to ensure a comprehensive assessment. A balanced mix of closed-ended and open-ended questions was employed, allowing for both quantitative and qualitative insights. The closed-ended questions provided structured data for analysis, while the open-ended ones allowed ML experts from a P2P lending platform to share their opinions and suggestions. Key themes covered in the questionnaire included the perceived effectiveness of the combined model, challenges faced during implementation and recommendations for improvement. This approach aimed to yield a holistic understanding of the model's practical utility, potential shortcomings and areas for enhancement.

In terms of sampling and participants, a diverse and knowledgeable pool of ML experts was selected from a P2P lending platform to ensure varied perspectives. The 27 participants included in the survey possessed diverse backgrounds in P2P lending, contributing to a comprehensive range of insights. Background information on participants, such as their roles and experience in ML applications for P2P lending, was collected initially to provide context to their responses. The survey was distributed using a Google Form, offering a convenient and standardized platform for participants to share their insights. Measures were implemented to enhance the validity and reliability of the collected data, including clear and unbiased question formulations to encourage honest and thoughtful responses. This approach sought to capture a well-rounded dataset, reflecting the expertise and experiences of ML experts from different roles within a P2P lending platform, thereby enhancing the credibility and applicability of the survey results.

The survey was designed to gather insights from ML experts on the effectiveness of the suggested combined model in P2P lending. The questions aimed to elicit comprehensive responses, covering various aspects of the model and the incorporation of eXplainable Artificial Intelligence (XAI) libraries. Key questions included awareness of XAI libraries such as SHAP and LIME, perceived value and effectiveness of these libraries, preferences between them and the impact on decision-making processes. Additionally, participants were queried on the frequency and methods of using SHAP or LIME and their satisfaction levels with these tools and observed impacts on the decision-making process.

To analyze the survey data, a combination of qualitative and quantitative methods was employed. Quantitative analysis involved categorizing responses to closed-ended questions, computing averages for satisfaction ratings and utilizing Likert scales for perceived effectiveness. Qualitative analysis focused on extracting insights from open-ended responses regarding participants' preferences, challenges faced and specific examples of positive contributions to decision-making. The survey responses were subjected to statistical measures, including percentages for categorical responses and mean values for Likert scale questions. Additionally, visual representations such as charts and graphs were generated to provide a clear and concise interpretation of the survey results.

The survey data was analyzed using google forms itself with its inbuilt statistical tool in order to draw meaningful conclusions. Key findings include ML experts' high awareness of XAI libraries, a preference for SHAP over LIME, a positive impact on decision-making processes and the perceived effectiveness of XAI libraries in enhancing the analysis of P2P lending datasets. The following sections delve into more detail on the specific responses and trends observed during the analysis.

10.4.3 Survey Questions and Responses

The survey yielded valuable insights from ML experts operating within the P2P lending sector. Participants exhibited diverse levels of experience, ranging from 1 to 5 years in their respective roles, contributing to a rich spectrum of perspectives. While the awareness of the SHAP library was generally high, there was varying familiarity with LIME, attributed to sources such as personal research, conference presentations, colleague recommendations and community discussions.

Notably, SHAP received widespread acknowledgment for its effectiveness in understanding feature importance, as 24 out of 27 respondents deemed it very valuable. Conversely, opinions on LIME's effectiveness in gaining insights into specific instances were mixed, with 15 participants indicating moderate effectiveness and 5 expressing that it was not very effective.

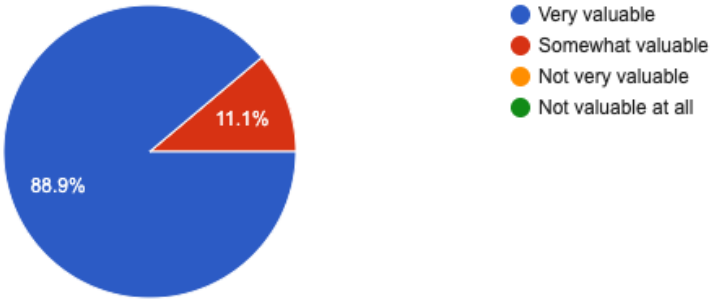


Figure 27: Valuability of SHAP in understanding feature importance

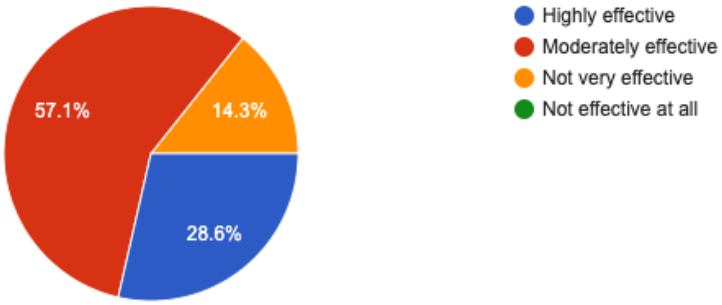


Figure 28: Effectiveness of LIME

The preference for SHAP over LIME emerged prominently, with 24 out of 27 participants favoring SHAP. This preference was justified by SHAP's adept handling of high-dimensional data, compatibility with visualization tools, graceful management of non-linearity and its versatile, model-agnostic nature. Satisfaction levels with SHAP were generally high, with 15 participants rating it as 4 and 12 participants as very satisfied. In contrast, satisfaction with LIME was comparatively lower, with 18 out of 27 respondents providing ratings less than or equal to 3.

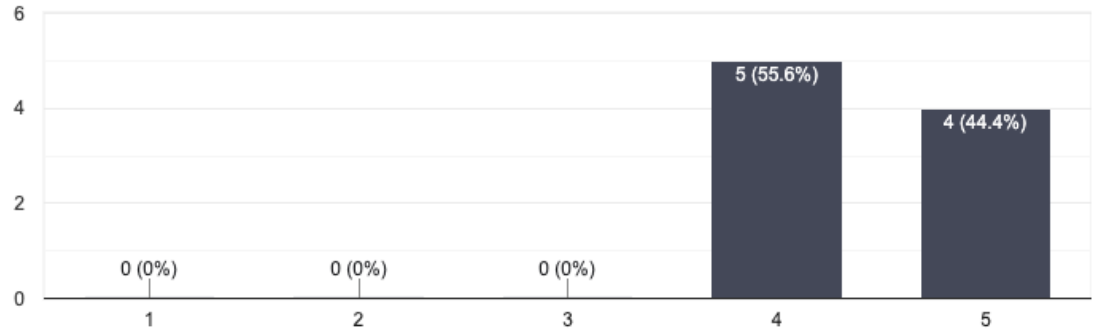


Figure 29: Satisfaction levels with SHAP

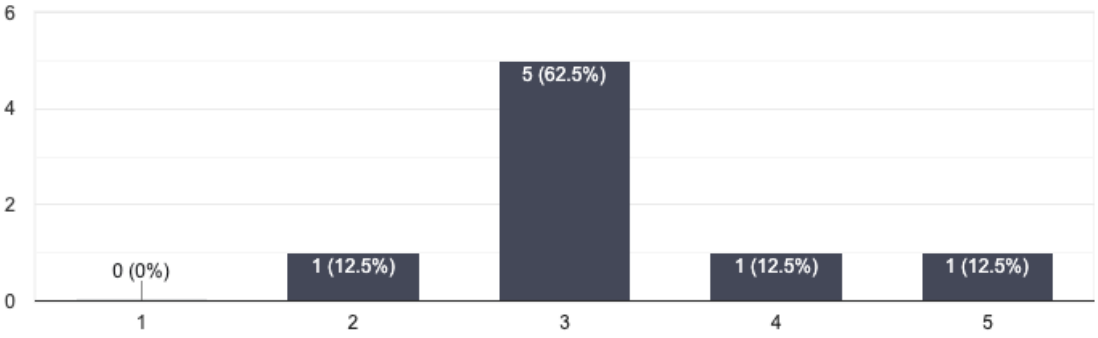


Figure 30: Satisfaction levels with LIME

Participants reported a quarterly utilization of SHAP, employing it for diverse purposes, including enhancing interpretability, improving model understanding, clarifying feature importance, uncovering influencing factors, creating explanatory visualizations, debugging, model validation and continuous model improvement.

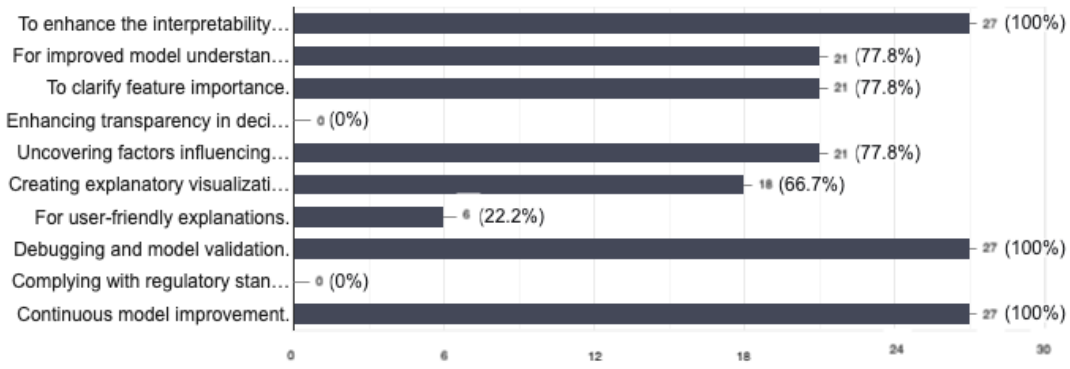


Figure 31: Utilization of SHAP and LIME

The impact of incorporating SHAP or LIME on the decision-making process was predominantly positive, as reported by 24 out of 27 participants. These individuals found the explanations provided by XAI libraries, particularly SHAP and LIME, highly useful in enhancing their understanding of decisions made by the platform. All 27 participants unanimously agreed that XAI libraries were very effective in enhancing the analysis of the P2P lending dataset.

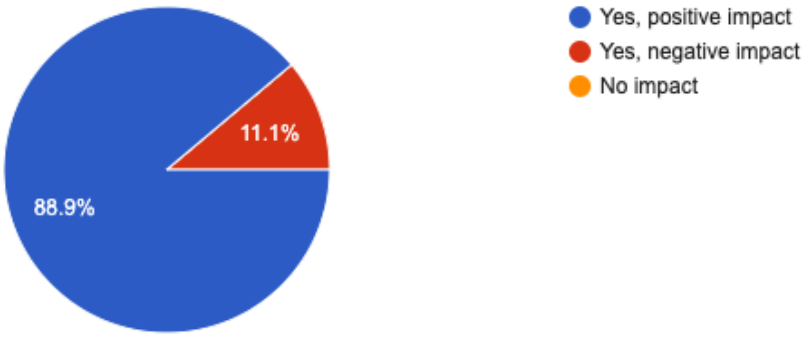


Figure 32: Impact on the decision-making process

Figure 33: Effectiveness of SHAP and LIME

Approximately 18 out of 27 participants disclosed that they had altered their decisions or behaviors based on information provided by XAI, underscoring instances where these tools positively contributed to decision-making. While only 6 participants reported challenges in incorporating SHAP or LIME, the majority did not face any hindrances. Furthermore, suggestions for improvements included incorporating alternative data sources, providing clear explanations to rejected loan applicants, integrating social media data, enhancing tools for explaining decision-making processes and considering local economic factors in credit scoring.

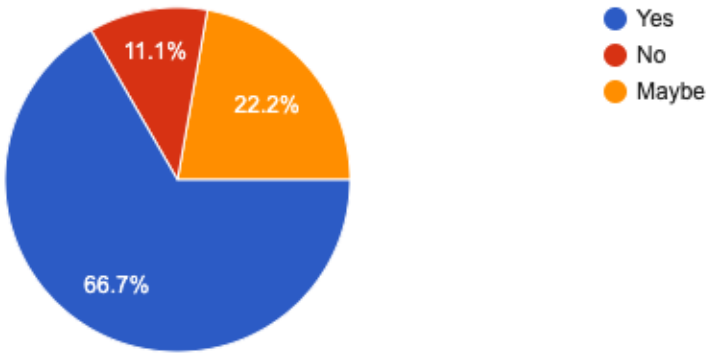


Figure 34: Decision Alteration due to XAI information

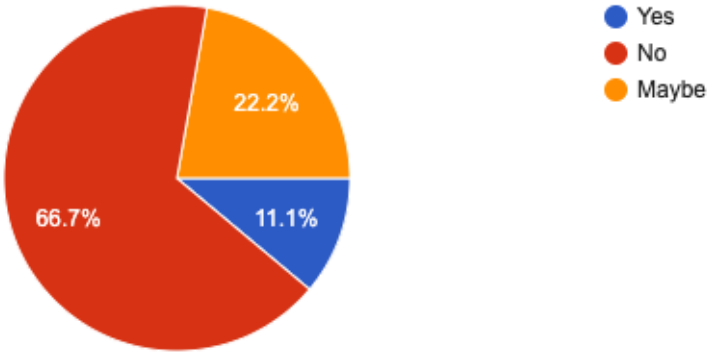


Figure 35: Challenges or Limitations of incorporating XAI in the process

10.4.4 Interview Design

The interview questions were designed to adhere to the rigorous standards of design science research, serving as a crucial tool for assessing the practicality of the proposed solution within the domain of P2P lending platforms. Each question was designed to delve into the workings of credit scoring systems in order to extract valuable knowledge about the underlying systems that the platforms use. Through an exploration of the complex workings of credit scoring processes and the widely used models (e.g., machine learning versus deep learning paradigms), the interviews aimed to reveal present methods while also pointing out possible directions for improvement and innovation.

In addition, the questions concerning the integration of eXplainable AI (XAI) models, which specifically make reference to Diverse Counterfactual Explanations symbolize a proactive approach to improving openness and user understanding in lending decisions. By requesting feedback on the practicality and usefulness of integrating these methods into existing models, the interviews stimulated discussions on innovative approaches to credit evaluation and risk reduction. Furthermore, the purposeful formulation of questions aimed at gathering input on potential improvements, in the P2P lending platforms themselves as well as in the study methods, demonstrated a dedication to incremental development.

10.4.5 Interview Response

In illuminating conversations with key personnel from a leading P2P lending platform, crucial insights into their credit scoring practices, model choices and considerations for future enhancements were unveiled. Regarding credit scoring, the platform adopts a sophisticated approach, generating credit scores based on feature probabilities and categorizing individuals into five risk levels. The dynamic nature of risk profiles, fine-tuned quarterly based on evolving data, underscores the commitment to precision in their credit assessment. In terms of models, the platform utilizes a machine learning (ML) model for its operations, underscoring the pivotal role of ML in their credit scoring mechanisms. The potential of transitioning to deep learning models was also discussed, limited due to factors such as the nature of the data and return on investment, with a recognition of the current data volume constraints for deep learning applications.

In the domain of Explainable AI (XAI), the platform leverages SHAP values for fine-tuning at the end of each quarter, enhancing model interpretability. However, the adoption of more complex XAI models such as LIME or DiCE within the application is tempered by the need for simplicity in user interfaces for borrowers and investors. The potential implementation of DiCE for end-user insights, especially for borrowers facing loan rejection, was considered an enhancement, though not an immediate priority due to volume considerations and the need to build an API.

Beyond model behavior, the interviewee emphasized the importance of incorporating macroeconomic variables into the credit scoring model, recognizing the impact of sector-specific economic conditions on borrowers' ability to repay loans. This holistic perspective suggests an approach to model enhancement, potentially incorporating rule-based models.

Addressing the evaluation of model performance, the interviewee acknowledged the analysis favoring the Random Forest model and highlighted the importance of addressing model drift over time. Continuous training and fine-tuning over new datasets were emphasized to ensure sustained model effectiveness, particularly in the face of changing economic factors. These substantial responses provide insight into the platform's thorough approach to credit rating, its deliberate application of ML models, and its plans for further improvements and modifications.

11 Research Contribution

Previous research studies in the field of P2P lending have faced with several limitations, primarily centered around accuracy, high false positive rates, biasedness and lack of consideration from a human-centred perspective that address the ethical, accountability and explainability perspective of models. Some of these studies have relied solely on SHAP for explainability, neglecting the exploration of other valuable explainability techniques such as DiCE . This narrow focus restricted the depth of interpretability and hindered the holistic understanding of model behaviors.

In addressing these limitations, this research adopts a novel approach by integrating eXplainable Artificial Intelligence (XAI) techniques, incorporating both global and local explanations in the context of P2P lending under FinTech. By leveraging global explanations, this research boosts the overall behavior of P2P lending models, providing insights into overarching trends and patterns. Additionally, the use of local explanations offers granular insights into individual predictions, enhancing transparency and trust in model decisions. Furthermore, the incorporation of DiCE into the P2P lending domain represents a pioneering step, as it enables the generation of diverse counterfactual explanations, thereby facilitating a deeper understanding of model decision-making processes for the end users such as lenders and borrowers.

Considering the incorporated research methodologically, this research contributes with a unique blend of design science and the research onion framework, traditionally utilized in social science research. This methodological innovation highlights the rigor and efficacy of research in the computer science domain, offering a structured yet flexible approach to knowledge creation and validation with practical applications. By seamlessly integrating these frameworks, my research transcends disciplinary boundaries, paving the way for interdisciplinary collaboration and knowledge exchange.

The results of this study have broader practical implications. By offering borrowers insights into improving their loan application approval chances, especially following a rejection, this approach empowers individuals to make informed financial decisions. Conversely, for investors, it provides invaluable insights into the potential default risks associated with seemingly non-defaulting borrowers, thus enabling more prudent investment strategies. Ultimately, the practical application of this research holds the potential to enhance the effectiveness and transparency of P2P lending platforms, benefiting both borrowers and investors.

This research makes significant contributions to the field of P2P lending by addressing the limitations of previous studies, introducing novel research methodologies for computer science research and offering practical insights with real-world implications. Through the integration of XAI techniques and methodological innovations, this work not only advances theoretical understanding but also provides actionable solutions to current challenges in the P2P lending domain.

12 Conclusions

The successful completion of this research has accomplished its main objectives, which center around discovering and incorporating the most effective Machine Learning (ML) model combined with eXplainable Artificial Intelligence (XAI) models in the field of Peer-to-Peer (P2P) lending. The first phase involved a thorough examination of different ML models, with a particular emphasis on accuracy. Among these models, the Random Forest model emerged as the prime example of meeting both criteria, demonstrating strong performance in the complex domain of P2P lending.

In order to enhance the comprehensibility and openness, eXplainable Artificial Intelligence (XAI) models such as SHAP, LIME and DiCE were purposefully utilized. These libraries for explainable AI not only offered comprehensive insights into the decision-making process of the models, but also provided specific explanations and conterfactual scenarios. The visual representations generated by these libraries significantly contributed to improving the evaluation of creditworthiness in peer-to-peer lending platforms, enabling a better understanding of the factors that impact individual predictions.

The innovative fusion of the ML model (Random Forest) with the XAI models introduced a holistic method, merging accuracy with interpretability. The assessment of this merged ML+XAI model involved traditional metrics associated with accuracy and was further enhanced by a qualitative approach. A thoughtfully crafted survey, aimed at gathering insights from experts in the P2P lending field, along with an interview conducted with an individual from a prominent P2P lending platform, yielded a comprehensive evaluation of the model's efficacy.

The research was built upon a comprehensive dataset obtained from Lending Club, which effectively represented the complexities of P2P lending transactions in real-world scenarios. The combination of quantitative and qualitative assessments further enhanced the reliability and validity of the suggested model. This research not only discovered an accurate and interpretable machine learning model for P2P lending but also successfully incorporated it with state-of-the-art XAI methods, resulting in advancements in transparent and trustworthy creditworthiness assessment process within the P2P lending sector.

13 Limitations

Certain limitations should be acknowledged despite the significant progress made by this research in enhancing the comprehension of creditworthiness evaluation in Peer-to-Peer (P2P) lending by incorporating Machine Learning (ML) and eXplainable Artificial Intelligence (XAI) models.

The utilization of the dataset from Lending Club poses limitations on the generalizability of the results. P2P lending platforms can have subtle differences in their data structures, user behaviors in different regions and risk factors, which could affect the model's relevance to other platforms. Additionally, the temporal aspect of the dataset may not fully capture the changing dynamics of the P2P lending industry.

The study's dataset does not take into account any economic factors that may affect the borrower's ability to repay their loan due to external circumstances. It is crucial to consider such changing economic conditions as they may directly impact the borrower's ability to repay the loan. For example, during a recession in the country where the borrower resides, their living expenses may increase due to inflation, while their disposable income will decrease, making it more difficult for them to pay back the loan. Although most P2P lending platforms do not use such factors to assess borrower creditworthiness, we believe that incorporating them is essential for a better outcome.

While the current study focused primarily on assessing the effectiveness and transparency of the model, it's important to recognize that the analysis may be perceived as limited in scope. A notable gap exists in the exploration of broader practical implications, including considerations of financial viability, risk management and regulatory compliance. Addressing these areas of concern will contribute to a more robust evaluation of the model's feasibility and potential impact in practical settings.

14 Future Directions

The findings of this study establish the groundwork for various areas of further investigation in the field of peer-to-peer lending and the integration of models.

In the first place, forthcoming investigations could pursue the improvement of the ML+XAI framework by integrating supplementary characteristics and enhancing those already in existence. The examination of alternative data sources, such as social media platforms and macroeconomic indicators, has the potential to augment the model's predictive capacities and offer a more exhaustive insight into borrower conduct.

Furthermore, it would be possible to conduct longitudinal research in order to observe the progress of the integrated model over a period of time. This approach would address any concerns about model drift and enable adjustments to be made in response to changing economic conditions. It may also be beneficial to explore techniques such as ongoing model training and refining procedures to ensure that the accuracy and effectiveness of the model are maintained within the constantly evolving P2P lending environment.

In addition, there is potential to investigate the integration of further XAI methods and advanced ML algorithms, such as deep learning models, in order to improve interpretability and predictive precision. Evaluating the effects of various XAI visualizations on end-users and stakeholders could yield insightful information for optimizing user satisfaction.

Finally, it is recommended that wider collaborations within the industry and studies conducted on multiple platforms should be carried out in order to verify and apply the results of this research to various P2P lending systems. Additionally, conducting comparative analyses with alternative credit scoring approaches and industry standards could enhance our knowledge of the ML+XAI model's position within the larger financial context.

List of References

Ariza-Garzón, Miller Janny et al. (2020). "Explainability of a machine learning granting scoring model in peer-to-peer lending". In: *Ieee Access* 8, pp. 64873–64890.

Arrieta, Alejandro Barredo et al. (2020). "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information fusion* 58, pp. 82–115.

Backhaus, Jürgen (1980). "The pareto principle". In: *Analyse & Kritik* 2.2, pp. 146–171.

Barroso, Marta and Juan Laborda (2022). "Digital transformation and the emergence of the Fintech sector: Systematic literature review". In: *Digital Business* 2.2, p. 100028.

Basha, Shabeen A, Mohammed M Elgammal, and Bana M Abuzayed (2021). "Online peer-to-peer lending: A review of the literature". In: *Electronic Commerce Research and Applications* 48, p. 101069.

Bondora (n.d.). URL: `https://www.bondora.com/`. (accessed: 02.04.2023).

Bussmann, Niklas et al. (2020). "Explainable AI in fintech risk management". In: *Frontiers in Artificial Intelligence* 3, p. 26.

Capel, Tara and Margot Brereton (2023). "What is Human-Centered about Human-Centered AI? A Map of the Research Landscape". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–23.

Cartolano, Andrea et al. (2022). "Explainable AI at work! what can it do for smart agriculture?" In: *2022 IEEE Eighth International Conference on Multimedia Big Data (BigMM)*. IEEE, pp. 87–93.

Coenen, Lize, Wouter Verbeke, and Tias Guns (2022). "Machine learning methods for short-term probability of default: A comparison of classification, regression and ranking methods". In: *Journal of the Operational Research Society* 73.1, pp. 191–206.

Dikmen, Murat and Catherine Burns (2022). "The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to-peer lending". In: *International Journal of Human-Computer Studies* 162, p. 102792.

Dresch, Aline et al. (2015). *Design science research*. Springer.

Emekter, Riza et al. (2015). "Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending". In: *Applied Economics* 47.1, pp. 54–70.

Fu, Runshan, Yan Huang, and Param Vir Singh (2021). "Crowds, lending, machine, and bias". In: *Information Systems Research* 32.1, pp. 72–92.

Gerlings, Julie, Millie Søndergaard Jensen, and Arisa Shollo (2022). "Explainable ai, but explainable to whom? an exploratory case study of xai in healthcare". In: *Handbook of Artificial Intelligence in Healthcare: Vol 2: Practicalities and Prospects*, pp. 169–198.

Gramegna, Alex and Paolo Giudici (2021). "SHAP and LIME: an evaluation of discriminative power in credit risk". In: *Frontiers in Artificial Intelligence* 4, p. 752558.

Havrylchyk, Olena and Marianne Verdier (2018). "The financial intermediation role of the P2P lending platforms". In: *Comparative Economic Studies* 60, pp. 115–130.

He, Congjie, Meng Ma, and Ping Wang (2020). "Extract interpretability-accuracy balanced rules from artificial neural networks: A review". In: *Neurocomputing* 387, pp. 346–358.

King, Jason E (2008). "Binary logistic regression". In: *Best practices in quantitative methods*, pp. 358–384.

Kochel, Tammy Rinehart and Wesley G Skogan (2021). "Accountability and transparency as levers to promote public trust and police legitimacy: findings from a natural experiment". In: *Policing: an international journal* 44.6, pp. 1046–1059.

Königstorfer, Florian and Stefan Thalmann (2020). "Applications of Artificial Intelligence in commercial banks–A research agenda for behavioral finance". In: *Journal of behavioral and experimental finance* 27, p. 100352.

Kumar, Vinod et al. (2016). "Credit risk analysis in peer-to-peer lending system". In: *2016 IEEE international conference on knowledge engineering and applications (ICKEA)*. IEEE, pp. 193–196.

Lending Club (n.d.). URL: `https://www.lendingclub.com/`. (accessed: 02.04.2023).

Lenz, Rainer (2016). "Peer-to-peer lending: Opportunities and risks". In: *European Journal of Risk Regulation* 7.4, pp. 688–700.

Li, Zhiqiang et al. (2021). "Application of XGBoost in P2P default prediction". In: *Journal of Physics: Conference Series*. Vol. 1871. 1. IOP Publishing, p. 012115.

Longford, Nicholas T (1994). "Logistic regression with random coefficients". In: *Computational Statistics & Data Analysis* 17.1, pp. 1–15.

Lundberg, Scott M et al. (2020). "From local explanations to global understanding with explainable AI for trees". In: *Nature machine intelligence* 2.1, pp. 56–67.

Misheva, Branka Hadji et al. (2021). "Explainable AI in credit risk management". In: *arXiv preprint arXiv:2103.00949*.

Mothilal, Ramaravind K, Amit Sharma, and Chenhao Tan (2020). "Explaining machine learning classifiers through diverse counterfactual explanations". In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 607–617.

Prosper (n.d.). URL: `https://www.prosper.com/`. (accessed: 02.04.2023).

Puzzarini, Cristina et al. (2019). "Accuracy and interpretability: The devil and the holy grail. New routes across old boundaries in computational spectroscopy". In: *Chemical reviews* 119.13, pp. 8131–8191.

Rethlefsen, Melissa L et al. (2021). "PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews". In: *Systematic reviews* 10.1, pp. 1–19.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). ""Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144.

Saunders, Mark, PHILIP Lewis, and ADRIAN Thornhill (2007). "Research methods". In: *Business Students 4th edition Pearson Education Limited, England* 6.3, pp. 1–268.

Selçuk, Ayşe Adin (2019). "A guide for systematic reviews: PRISMA". In: *Turkish archives of otorhinolaryngology* 57.1, p. 57.

Serrano-Cinca, Carlos and Begoña Gutiérrez-Nieto (2016). "The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending". In: *Decision Support Systems* 89, pp. 113–122.

Setiawan, Netty et al. (2019). "A comparison of prediction methods for credit default on peer to peer lending using machine learning". In: *Procedia Computer Science* 157, pp. 38–45.

Shin, Donghee (2021). "The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI". In: *International Journal of Human-Computer Studies* 146, p. 102551.

Smith, Sean Stein (2019). *Blockchain, artificial intelligence and financial services: Implications and applications for finance and accounting professionals*. Springer.

Suryono, Ryan Randy, Betty Purwandari, and Indra Budi (2019). "Peer to peer (P2P) lending problems and potential solutions: A systematic literature review". In: *Procedia Computer Science* 161, pp. 204–214.

Turiel, JD and T Aste (2020). "Peer-to-peer loan acceptance and default prediction with artificial intelligence". In: *Royal Society open science* 7.6, p. 191649.

Uddin, Mohammad S et al. (2022). "Leveraging random forest in micro-enterprises credit risk modelling for accuracy and interpretability". In: *International Journal of Finance & Economics* 27.3, pp. 3713–3729.

Vives, Xavier (2017). "The impact of FinTech on banking". In: *European Economy* 2, pp. 97–105.

Wang, Haomin, Gang Kou, and Yi Peng (2021). "Multi-class misclassification cost matrix for credit ratings in peer-to-peer lending". In: *Journal of the Operational Research Society* 72.4, pp. 923–934.

Wang, Hong, Qingsong Xu, and Lifeng Zhou (2015). "Large unbalanced credit scoring using lasso-logistic regression ensemble". In: *PloS one* 10.2, e0117844.

Wang, Qi, Xin Liu, and Chenghu Zhang (2022). "Evolutionary game analysis of FinTech transformation: A social co-governance pattern of peer-to-peer lending market in China". In: *Frontiers in Psychology* 13, p. 954132.

Wu, Bao et al. (2023). "Underdog mentality, identity discrimination and access to peer-to-peer lending market: Exploring effects of digital authentication". In: *Journal of International Financial Markets, Institutions and Money* 83, p. 101714.

Xu, Jennifer J et al. (2022). "PEER-TO-PEER LOAN FRAUD DETECTION: CONSTRUCTING FEATURES FROM TRANSACTION DATA." In: *MIS quarterly* 46.3.

Yoon, Yeujun, Yu Li, and Yan Feng (2019). "Factors affecting platform default risk in online peer-to-peer (P2P) lending business: an empirical study using Chinese online P2P platform data". In: *Electronic Commerce Research* 19, pp. 131–158.

Zhang, Zaimei, Kun Niu, and Yan Liu (2020). "A deep learning based online credit scoring model for P2P lending". In: *IEEE Access* 8, pp. 177307–177317.

Zhao, Hongke et al. (2014). "Investment recommendation in p2p lending: A portfolio perspective with risk management". In: *2014 IEEE International Conference on Data Mining*. IEEE, pp. 1109–1114.

Zhou, Jing et al. (2019). "Default prediction in P2P lending from high-dimensional data based on machine learning". In: *Physica A: Statistical Mechanics and its Applications* 534, p. 122370.

Zhu, You et al. (2016). "Predicting China's SME credit risk in supply chain financing by logistic regression, artificial neural network and hybrid models". In: *Sustainability* 8.5, p. 433.

Zopa (n.d.). URL: `https://www.zopa.com/`. (accessed: 02.04.2023).

A Appendix

A.1 Questionnaire

Questionnaire for P2P lending developers

Introduction

Welcome to our survey on the use of SHAP and LIME in P2P lending process. Your insights are invaluable in helping us assess the impact of these tools on the overall effectiveness of our P2P lending process.

This questionnaire aims to gather your perspectives on how SHAP and LIME contribute to the effectiveness of P2P lending process. We are eager to understand your experiences, use cases, and overall perceptions regarding the role of these tools in enhancing our analytical capabilities.

We recognize the sensitivity of the information you provide and assure you that your responses will be treated with the utmost confidentiality. Your feedback is crucial for our research purposes, and we want to emphasize that your individual responses will remain confidential and will only be used within the scope of this study.

Your participation is voluntary, and we appreciate your candid and thoughtful responses. Thank you for taking the time to contribute to our ongoing efforts to improve the effectiveness of P2P lending processes.

Role

1. What is your primary role within the development team of the P2P lending platform?

   Mark only one oval.

   - ◯ Data Scientist
   - ◯ Machine Learning Engineer
   - ◯ Developer
   - ◯ Other: _____

2. How many years of experience do you have in your current role within the development team of the P2P lending platform?

   _____

Awareness of XAI libraries

Please find the explanations derived from loan status predictions using XAI libraries in the following link: https://drive.google.com/file/d/193ZdEXX9fVUilQNiq-cfOx2EGuUtDkz6/view?usp=sharing

3. Were you aware of the XAI library SHAP (SHapley Additive exPlanations) ?

Mark only one oval.

- ◯ Yes
- ◯ No
- ◯ Maybe

4. Were you aware of the XAI library LIME (Local Interpretable Model-Agnostic Explanation)?

Mark only one oval.

- ◯ Yes
- ◯ No
- ◯ Maybe

5. If yes, how did you become aware of them?

6. Considering SHAP, which provides both global and local interpretations, how valuable do you find this approach in understanding feature importance across the entire dataset and individual predictions?

   Mark only one oval.

   - ◯ Very valuable
   - ◯ Somewhat valuable
   - ◯ Not very valuable
   - ◯ Not valuable at all

7. LIME was used to provide localised interpretations for individual predictions. In your opinion, how effective do you think this approach is in gaining insights into specific instances while overlooking the overall model behaviour?

   Mark only one oval.

   - ◯ Highly effective
   - ◯ Moderately effective
   - ◯ Not very effective
   - ◯ Not effective at all

8. Do you have a preference between SHAP and LIME for providing explanations within the P2P lending platform?

   Mark only one oval.

   - ◯ Yes
   - ◯ No
   - ◯ Maybe

9. If yes, could you briefly explain why you prefer one over the other?

10. In your opinion, which tool (SHAP or LIME) provides clearer and more understandable explanations for model predictions?

Check all that apply.

- [ ] SHAP
- [ ] LIME

11. On a scale of 1 to 5, where 1 is "Not satisfied at all" and 5 is "Very satisfied," how satisfied are you with the overall performance and capabilities of SHAP for model explanations?

Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

12. Similarly, rate your overall satisfaction with LIME for model explanations.

Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

Usage

13. Do you use either of those XAI libraries in your P2P lending platform?

Mark only one oval.

- ◯ Yes
- ◯ No
- ◯ Maybe

14. How frequently do you currently use SHAP or LIME or both in your analysis on the P2P lending platform?

Mark only one oval.

- ◯ Monthly
- ◯ Quarterly
- ◯ Semi-Annually
- ◯ Annually
- ◯ Rarely
- ◯ Never
- ◯ Other: _____

15. How do you utilize SHAP or LIME or both in your P2P lending model?

Check all that apply.

- ☐ To enhance the interpretability of our model.
- ☐ For improved model understanding.
- ☐ To clarify feature importance.
- ☐ Enhancing transparency in decision-making.
- ☐ Uncovering factors influencing predictions.
- ☐ Creating explanatory visualizations.
- ☐ For user-friendly explanations.
- ☐ Debugging and model validation.
- ☐ Complying with regulatory standards.
- ☐ Continuous model improvement.
- ☐ Other: _____

Impact of using XAI

16. Have you observed any impacts on the decision-making process related to P2P lending since incorporating SHAP or LIME?

Mark only one oval.

- ◯ Yes, positive impact
- ◯ Yes, negative impact
- ◯ No impact

17. Are the explanations provided by XAI are understandable?

Mark only one oval.

- ◯ Yes
- ◯ No
- ◯ Maybe

18. How clear are the explanations provided by XAI to you?

Mark only one oval.

- ◯ Not clear at all
- ◯ Slightly clear
- ◯ Moderately clear
- ◯ Very clear
- ◯ Extremely clear

19. To what extent do you feel the explanations provided by XAI enhance your understanding of the decisions made by the platform?

Mark only one oval.

- Not useful at all
- Slightly useful
- Moderately useful
- Very useful
- Extremely useful

20. How effective do you find XAI libraries (SHAP or LIME) in enhancing the analysis of the P2P lending dataset?

Mark only one oval.

- Not effective at all
- Slightly effective
- Moderately effective
- Very effective
- Extremely effective

21. Have you ever changed your decision or behavior based on the information provided by XAI?

Mark only one oval.

- Yes
- No
- Maybe

22. If so, can you briefly state a few specific examples or instances where these tools have contributed positively to decision-making?

23. Have you faced any challenges or limitations when incorporating SHAP or LIME into the analysis process on the P2P lending platform?

Mark only one oval.

- Yes
- No
- Maybe

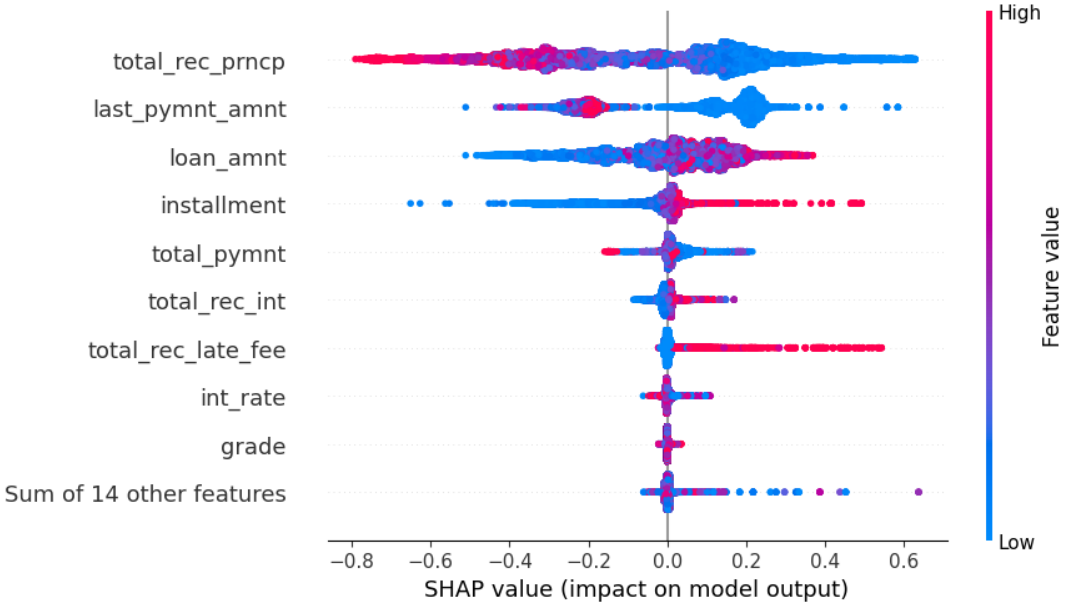
24. If yes, could you briefly state few such challenges?

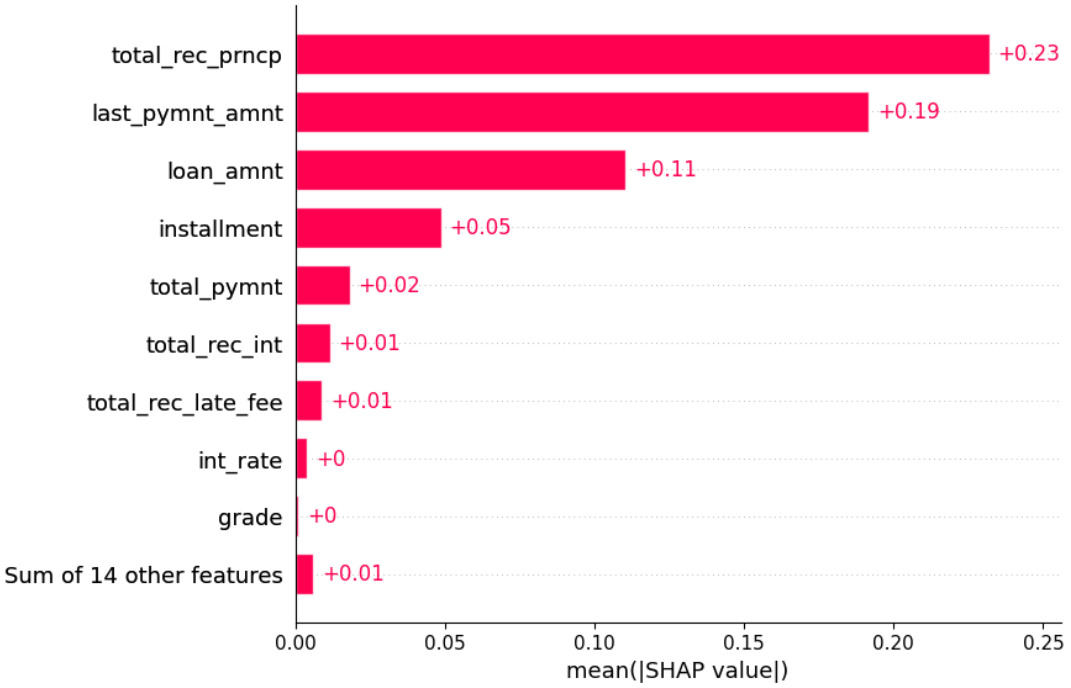
25. Are there any additional features or improvements you would like to see to enhance the P2P lending process?

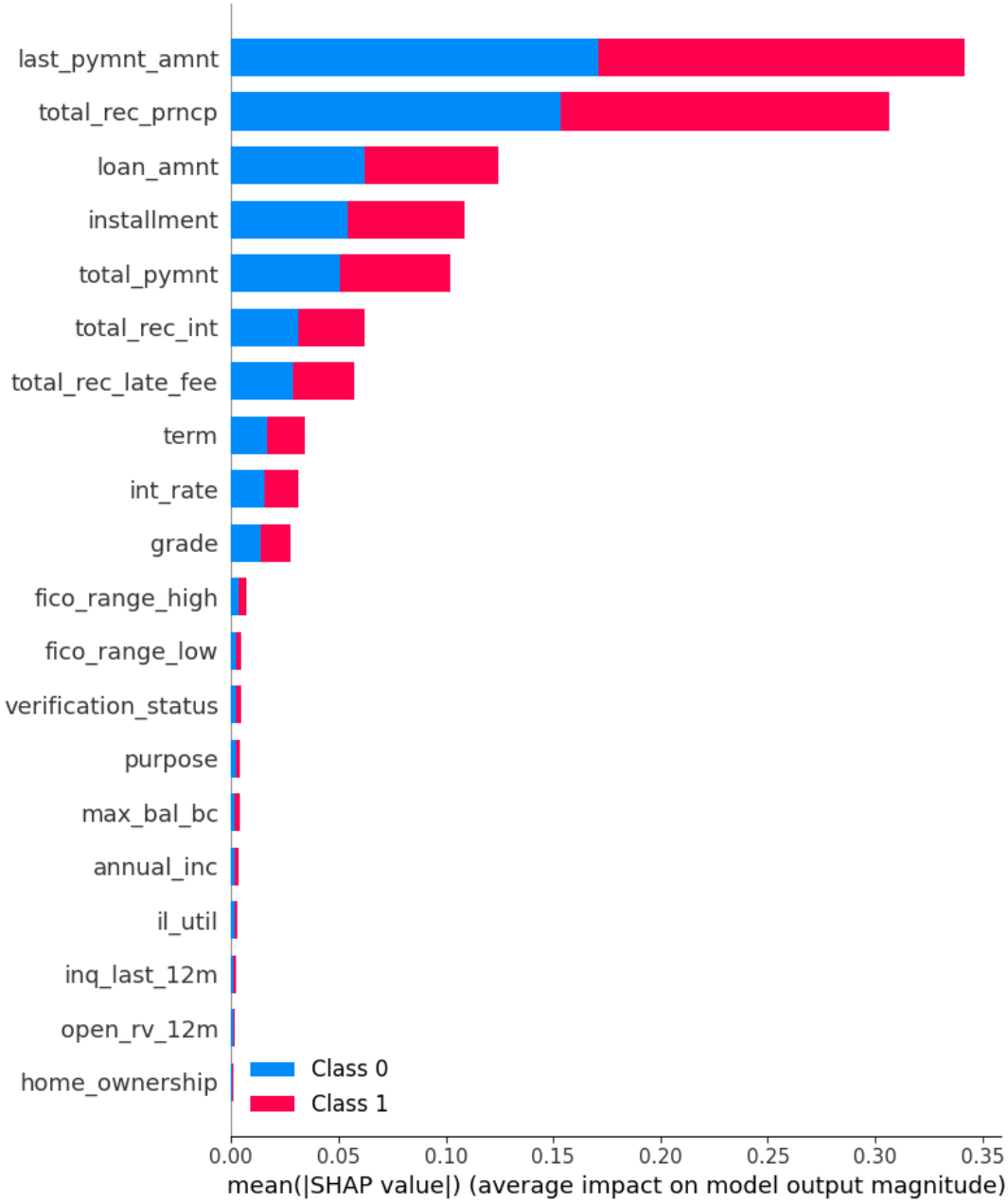
Thank you for taking the time to participate in this survey. Your valuable insights contribute significantly to our understanding of the preferences and experiences related to model explanations in the context of the P2P lending platform.

Explanations derived from XAI

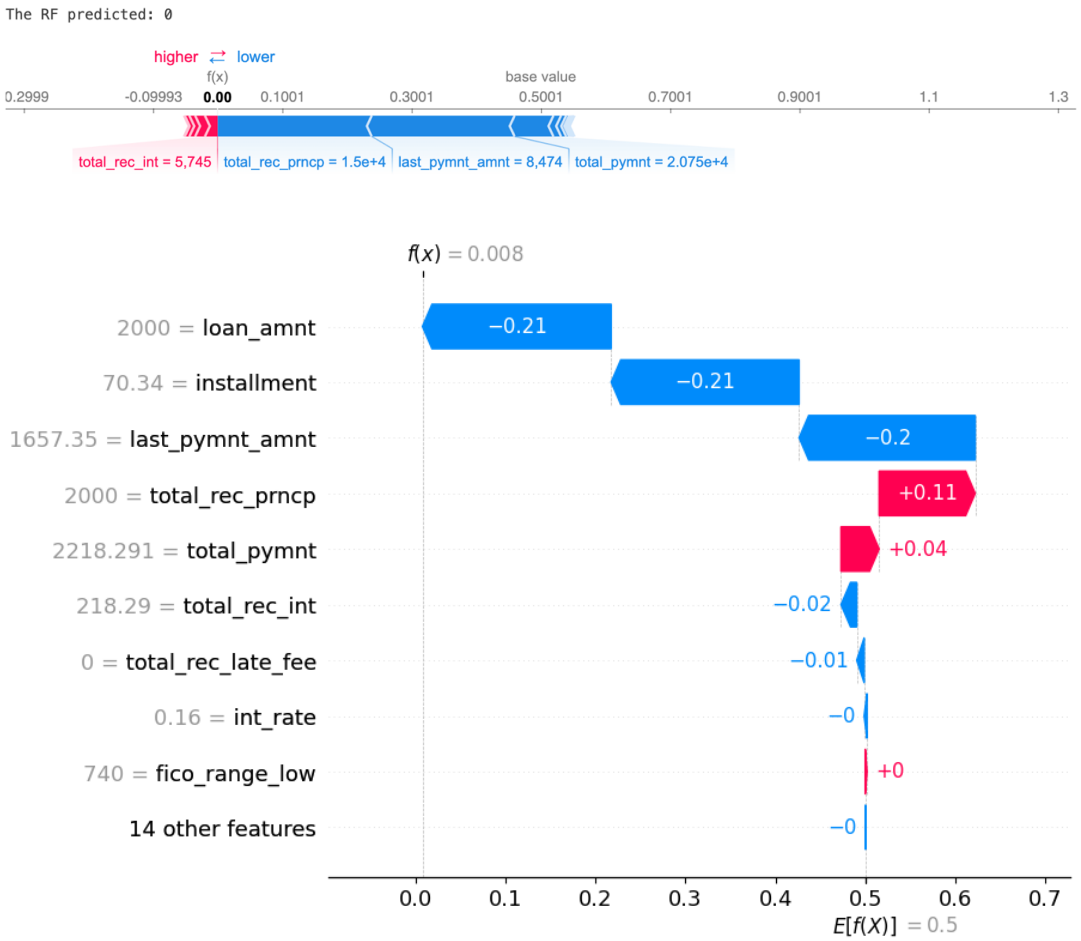
1. Global level explanation using SHAP



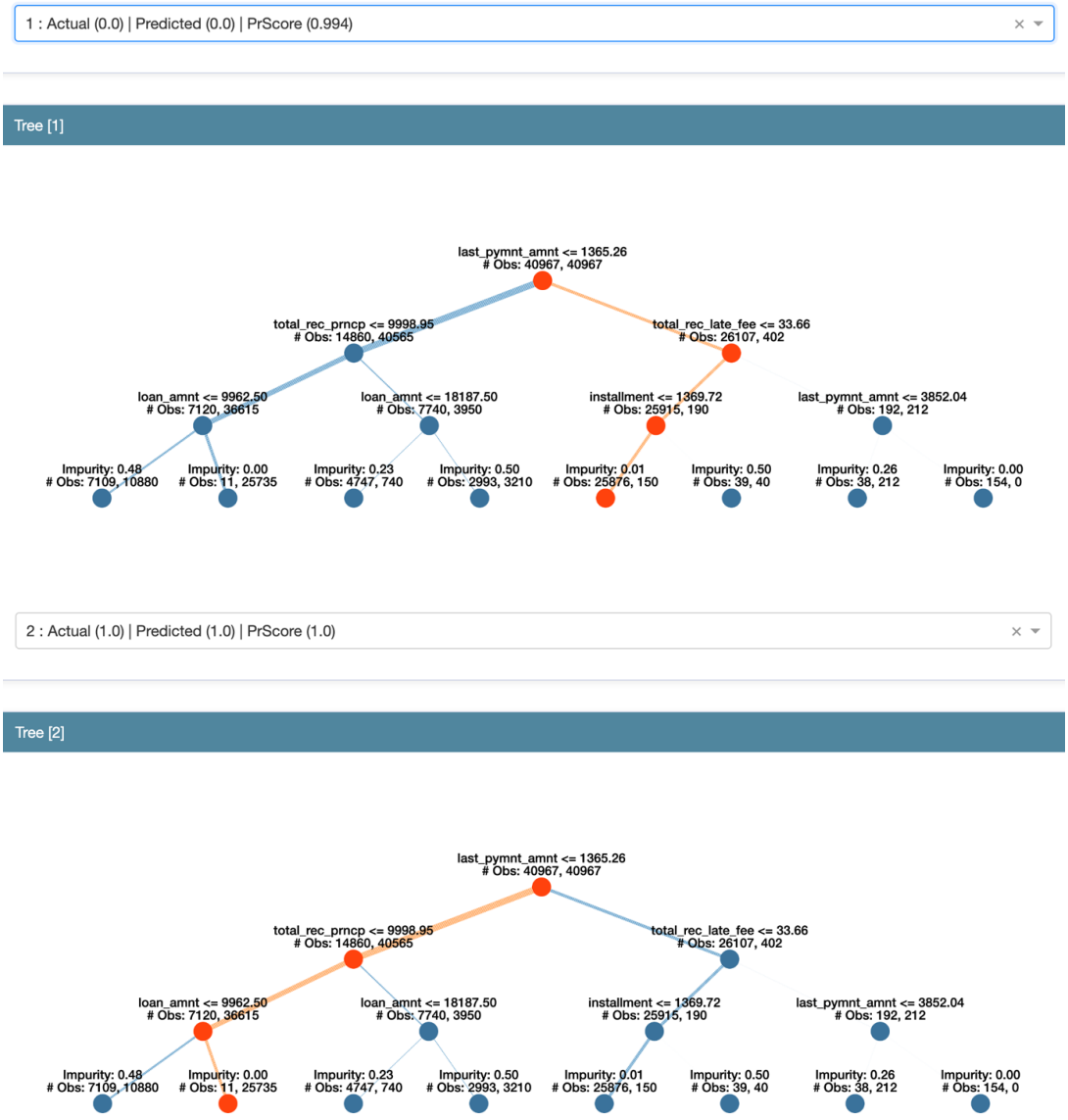



last_pymnt_amnt
total_rec_prncp
loan_amnt
installment
total_pymnt
total_rec_int
total_rec_late_fee
term
int_rate
grade
fico_range_high
fico_range_low
verification_status
purpose
max_bal_bc
annual_inc
il_util
inq_last_12m
open_rv_12m
home_ownership
Class 0
Class 1
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
mean(|SHAP value|) (average impact on model output magnitude)

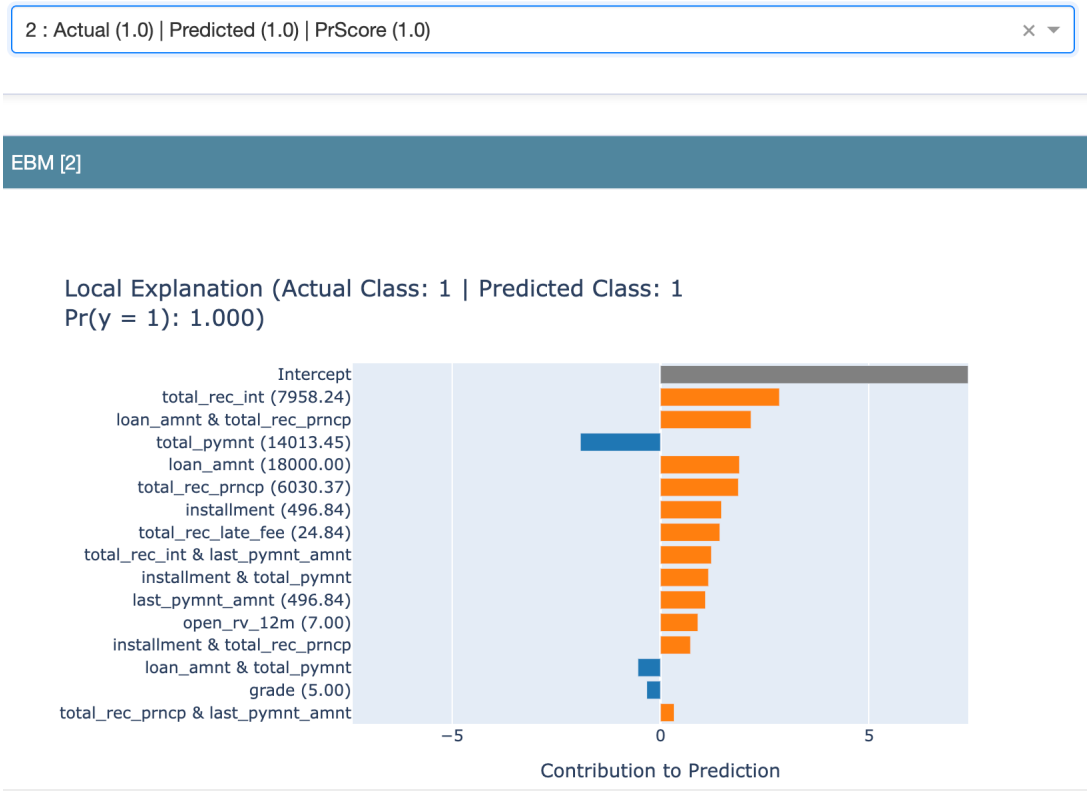
2. Local level explanation using SHAP

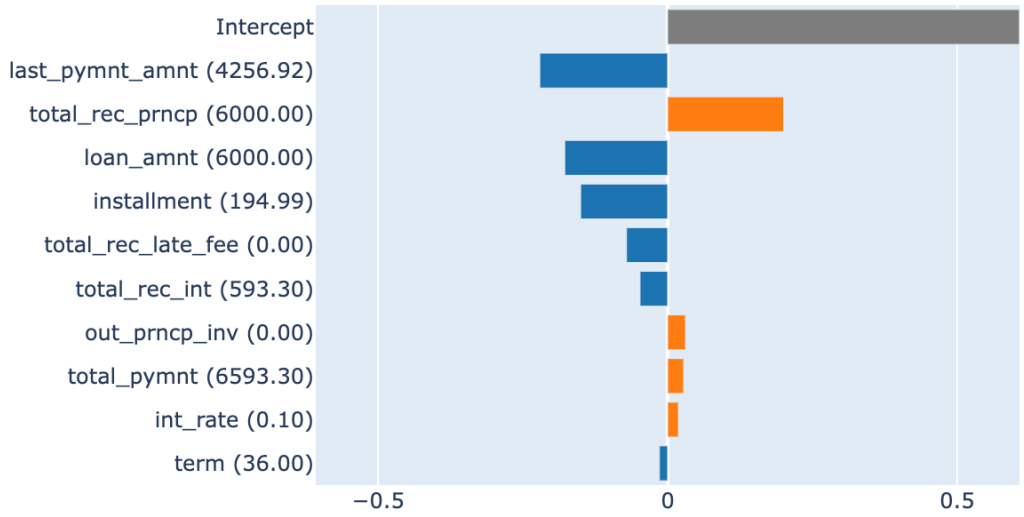


3. Local level explanation using TreeSHAP (Only for Tree models)



4. Local level explanation using LIME





A.2 Interview Response

Interview response from P2P lending platform CEO:

1. How does your plaorm conduct credit scoring? Does it generate a credit score, or does it categorize individuals directly based on specific factors? Could you explain the process?

We come up with a credit score based on the probabilities of features and categorise them into 5 categories based on the risk level. These risk profiles are find tuned at the end of each quarter based on the data we get.

2. What kind of model is employed in your P2P lending platform?

We have a ML model

3. Do you see any potential to move to a deep learning model rather than using a machine learning model?

At this point in time, what we have is mostly tabular data. We are not pulling off any senCment or any unstructured sort of text data currently. We might pull in meta data from documents, going forward to include it. But I'm not sure at that point whether we switch to a Neural Networks or use the ML models by summarizing those features by conducCng sort of feature engineering.
And we don't have huge volumes of data to go forward with Deep learning models as of now is that of data for that.
Also, it depends upon the ROI. Whether we want to invest in that kind of thing for a marginal outcome, then we might not move to deep learning.

4. Do you use XAI models in your P2P lending Platform?

We use SHAP values at the end of every quarter to fine tune our model, which means from developers' end.
We haven't used LIME or any sort of DICE yet.
But we don't use it within the application and show anything to the end users (borrowers and investors) because we can't complicate it too much for investors and we have to provide a simplified version to them whether it is institutional investors or retail investors.
To be honest, SHAP values is a little bit of a technical analysis, but we don't want to provide the technical analysis and make it complicated for the end users.

5. What do you think of using DICE for your model to provide a set of insights for the end users? As an example, we can provide a set of insights for the borrower when their loan gets rejected, what are the reasons for that and what improvements that he can made to get his loan approved.

Possibly that can be done. But looking at the volumes we are looking at we have to build an API. But it is not high priority for our company at the moment.
We should also consider which variables to change and their upper and lower limits to change.
For sure it would be an enhancement that we could add to our platform as well.

6. How can your model be enhanced or improved?

To be honest, it is not always the model behaviour we should consider. In an industry like this, we should also consider the macro-economic situation of that country as variables within. Because in certain times, certain sectors might perform better and certain sectors might not. For example, textile and leisure. Leisure now is doing good, but in 2020 during the global pandemic it didn't perform. It would have a direct impact on the ability of an employee of such a sector to pay back their loan.

Even we haven't brought in those variables at the moment, but could be done. We could even use rule based model for that rather than a ML model by defining macro-economic variables at a more higher level. We could identify the borrower's employment sector and identify which of those sectors are more prone to kind of economic viability.

7. What enhancements can I implement in my research?

You included a lot of behvioural sort of variables. If you want to, think of the markets that the P2P lending platgorm operates in, their economic variables or rather economic conditions. May be on the individual borrower, what industry they are in. Because it will be used in your DICE and would give you more value. As an example, construction is not doing well at the moment. So you can think of using such variables.

8. As per my analysis, random forest model performed better, but as per your experience what are your thoughts on this?

Your analysis make sense. One thing to notice is that there can be model drif when the economic factors change from time to time. So need to keep training it over the new dataset to fine tune your model.