Artificial Intelligence Driven Personalized Loan Pricing in Peer to Peer Lending Finance

MSc Research Project
Data Analytics

Parth Patel
Student ID: X23166029

School of Computing
National College of Ireland

Supervisor: Noel Cosgrave

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Parth Patel
Student ID:	X23166029
Programme:	Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Noel Cosgrave
Submission Due Date:	12/12/2024
Project Title:	Artificial Intelligence Driven Personalized Loan Pricing in Peer to Peer Lending Finance
Word Count:	10322
Page Count:	25

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	25th January 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Artificial Intelligence Driven Personalized Loan Pricing in Peer to Peer Lending Finance

Parth Patel
X23166029

Abstract

This study analyses the use of AI in creating individualized loan pricing models for peer-to-peer (P2P) lending platforms, using the Lending Club Corporation dataset, and solving problems in credit risk management, borrower-lender linearisation and transparency, and financial inclusion. AI-driven data processing can boost the creation of individualized loan pricing models and allow for risk stratification of borrowers by adding extra insights to lenders and thus ameliorate performance. The dataset was pretreated and had the following techniques for preprocessing steps, such as oversampling to counteract class imbalance and feature importance analysis to detect the main predictor variables affecting the study of creditworthiness. The research used a broad range of methods that include Logistic Regression, Random Forest, Gradient Boosting, and Neural Networks to estimate the default probabilities and develop more elaborate loan pricing plans. The effectiveness of displayed high Neural Networks has been found to be a 92% accuracy rate and 95% recall due to the method's ability to precisely distinguish high-risk borrowers. The AI-powered personalized pricing schemes grounded in these models were very successful in this regard and therefore, it is possible to set competitive interest rates while reducing risks, and thus contribute to fair lending practices. Through Lending Club's realistic dataset, the research conducts a comprehensive study of the potential of AI to solve the real-world problems that are caused by the P2P lending platforms. In addition, the article expounds on the bigger issues such as the use of power for the development of people without banking services and on ethical issues of data privacy and algorithms. The research is bringing newness to P2P lending ecosystems and thus the creation of a scalable and data-driven framework for personalized loan pricing, which is a breakthrough in financial technology and is a new definition of the traditional lending practices.

1 Introduction

Peer-to-peer lending has greatly influenced the finance sector by the development of the peer-to-peer model through which borrowers can directly communicate with lenders instead of the touch of banks. This model gives the possibility to the lenders to get higher returns and for the borrowers to access credit. Nevertheless, even though P2P lending platforms face enormous issues such as the exact requirement of credit rating and loan pricing, they are still the best. As such, traditional credit scoring techniques are often insufficient in this situation causing efficiencies and increasing default risks. The use of

AI in P2P lending platforms acts as a viable tool to the problems faced, by improving the credit risk assessment and providing personalized loan pricing.

Personalized loan pricing is a process in which interest rates are set accordingly to various borrower profiles by adopting thorough examinations of the creditworthiness of those consumers, their financial behavior, and the concomitant risks. Click to Peer-to-Peer lending, which consists of various categories of borrowers that have different credit histories and financial necessities, this personified method is an essential aspect. Moreover, these non dynamic credit systems, which most of the time are based on inadequate information, and use static models that do not always reflect the uniqueness of borrowers and thus the mispricing can occur. Such mispricing can either overburden borrowers with higher interest rates or expose lenders to elevated default risks.

The Lending Club Corporation dataset provides a complete set of data that aids in the study of lending in the real world. Among the features that are included are comprehensive borrower details which include credit scores, income levels, loan amounts as well as payment patterns. This data set allows to gain better insight into borrower profiles and also offers the possibility for the estimation of customer AI-based pricing models. Still, data class imbalance is known as one of the problems that can appear if proper measures are not taken to ensure robust model performance. The use of methods such as oversampling in order to rectify class imbalances results in the reliability of the predictions being tested more.

Aside from efficiency improvement, AI involvement in the pricing of loans aids in bringing about financial inclusion through proven credit providers to a large population that is not adequately served. Rigid credit scoring systems, which are normally used in the old-style lending institutions, have the effect of mostly excluding those who have little credit history or those who have inconsistent income sources. Contrary to the AI-enabled models that test the worthiness of a credit applicant by the consumption patterns data, other alternative data sources are also added, thus a more accurate credit rating. This generous approach, thus, enables P2P platforms to lend money to numerous people yet upholding low default rates.

According to recent research, a number of different machine learning strategies have been examined to bring about improvements in the field of loan pricing and credit risk assessment in P2P lending. Among them, Logistic Regression, Random Forest, Gradient Boosting are machine learning algorithms that have most been used to build predictive models. The superiority of such techniques lies in the fact that they are accurate, scalable, and even capable of handling non-linear data relationships. Neural Networks, for example, are particularly adept at the identification of complex patterns, thus allowing for the dynamic and personalized loan pricing strategies. Methods of feature selection like feature importance ranking make it possible to find the most significant variables in terms of their contribution to loan outcomes, and thus, the creation of more targeted and efficient models.

Even with the successes, implementing AI personalized loan pricing is still problematic to some degree in P2P lending. The issues surrounding data quality and availability seem to be the main barriers to the adoption of algorithms. On the other hand, privacy concerns and the lack of transparency in the algorithms compound the reluctance to the adoption. Besides, biases in AI algorithms that have their roots either in partial data or a deficient model make ethical matters like justice and accountability important. The solution to these problems calls for a comprehensive tactic that merges durable data preprocessing, introduction of algorithmic fairness strategies, and introduction of legal criteria for the safe use of AI.

This research is a follow-up to the previous studies conducted on the existing literature by developing and testing AI-driven personalized loan pricing models using the LendingClub Corporation dataset. The research strategy includes data preprocessing to deal with class imbalances, feature selection, predictors selection to enhance predictive accuracy, and machine learning algorithms to create high-rate real-time and reliable models. The assessment of the models is performed by using sticky bugs learned on accuracy, precision, recall, and AUC - ROC in order to guarantee the models effective application in the real world Feng et al. (2018).

This study makes a big difference for P2P lending platforms and the financial tech industry with a wider reach. The results are the benefits of personalized prices in the loaning process, the study helps to bring out more efficient, transparent, and easily accessible credit ecosystems. The research also introduces the fact that we need to abide by regulators and be ethical while designing and implementing artificial intelligence.

In conclusion, AI-driven personalized loan pricing in P2P lending is a huge shift towards data-driven solutions which can effectively solve the problem of credit risk estimation and loan prices that have been around for ages. The data is based on real-world data sets and the machine learning algorithms used are top-notch, hence this research will be able to show the path for actionable insights. As a result, P2P platforms will not only improve their operations but also contribute to financial inclusion whereby all people can access fair and affordable credit. This research through its key features of scalability, and ethical, and most importantly innovation has provided the groundwork for the future of personalized financial services in the digital era.

1.1 Research Question

One of the primary goals of this research is to enhance the precision and efficiency of credit risk evaluation in P2P lending platforms by introducing AI-assisted personalized loan pricing models. This can be obtained by constructing models that collect all the details such as historical data, financial behavior, and credit profile of a borrower to forecast default probabilities and keep them on line with respect to the individual risk levels. Through these methods, an AI-based loan pricing system is introduced that balances the creditworthiness of a borrower and the lender's risks. For this, the LendingClub Corporation database was used to train and test the different machine learning models. The models chosen and implemented were formulated from both the best-performing ones and the ones most relevant in the literature.

RQ - ***"How are the accuracy and efficiency of credit risk assessment in P2P lending platforms enhanced when AI-driven personalized loan pricing models are integrated into them, and what are the measurable impacts on borrower default rates and lender returns?"***

To solve the first research question, classification and regression algorithms which are Logistic Regression, Random Forest, Gradient Boosting, Neural Networks, and Support Vector Machines were applied to the dataset. These models were selected for their ability to deal with complex financial datasets and provide precise estimations. The efficacy of each model was assessed through the use of metrics like accuracy, precision, recall, and AUC-ROC, which guarantee reliable evaluation of credit risk and loan pricing outcomes.

The models accuracy and effectiveness impact the decision-making process in P2P lending directly. The greater accuracy of the model in predicting borrower default rate leads to a reduction in the default rate, on the other hand, efficient loan pricing increases the lender's return which is made possible through the proper allocation of the interest rate to the borrower's risk. Through the delivery of these AI-driven models, this research promises to offer a solution to major challenges in P2P lending that can be scaled and implemented, thus, the benefits of borrowers and lenders become factual.

1.2 Research Objectives

In order to answer the research question and to fulfill the aims of this study, the following objectives will be followed as well:

- A thorough and significant review should be conducted which consists of all the existing literature such as articles, conference papers, and well-known journals which concentrate on AI-supported credit risk management, personalized loan pricing, and machine learning applications in peer-to-peer lending.
- The LendingClub Corporation dataset should be the main source of data, which has the historical borrower's transaction data and credit profiles.
- Get the data analysis ready by data preprocessing steps like dealing with missing values, normalizing variables, and encoding categorical features.
- Conduct Exploratory Data Analysis (EDA) to come up with provisional insights, reveal trends, and comprehend the distribution and relationships of variables in the dataset.
- The balance between the classes within the dataset can be restored through oversampling methods, for example Resampling, we can use such a methodology to bring in the necessary samples of key groups, which in this case are defaulters and non-defaulters.
- Train and test on the processed dataset for supervised machine learning models, including the use of Logistic Regression, Random Forest, Gradient Boosting, Neural Networks, and Support Vector Machines.
- Measurement of initial performance of the models through evaluation metrics such as Accuracy, Precision, Recall, F1-Score, and AUC-ROC in terms of their power to predict the likelihood of a borrower defaulting is a key parameter.
- Use feature selection techniques to ascertain the most important parameters affecting risk borrowing and pricing of the loans within the data set.
- Compress the dataset and select features to update the learning models which, as a result, minimize the operational cost and raise predictive performance.
- Using GridSearch or other similar optimization methods, the hyperparameters of the models can be optimized to reach the best possible outcomes for each of the algorithms.

- Repeat the models on the improved datasets over the same evaluation metrics to give and identify the most effective AI driven personal loan pricing model.

- Evaluate the significance of AI-powered personalized loan pricing in the context of borrower defaults and lender returns, by measuring the improvements in accuracy and efficiency.

2 Related Work

The assessment of credit risk is an integral part of peer-to-peer lending procedures, as it defines the borrower's creditworthiness and decides the loan pricing policies. By basing their decisions on static rules and limited datasets, traditional credit evaluating methods sometimes become inadequate to reveal the hidden complexities of borrowers behavior in dynamic financial settings. That has led to the appearance of inefficiencies, and to the introduction of higher risks for lenders. The use of AI and ML, for instance, makes it possible to see the whole picture and derive the actual loan profile based on different indicators and on the relationship of these indicators with default probability. Logistics regression, Random Forest, Gradient Boosting, and Neural Networks are some of the models that have developed due to the developments of data-driven methods, which are more accurate and efficient, and thus, provide more reliable data on credit risk and facilitate personalized loan pricing. This personalized approach, which entails dynamically adjusting the interest rates, is based on the individual borrower's profile and, therefore, risk is correlated to pricing and a greater fairness is achieved in lending Zhang et al. (2019).

However, challenges remain in implementing AI-driven credit risk models. Class imbalance, where defaulters are underrepresented compared to non-defaulters, often skews machine learning models, leading to biased predictions and misclassification. Oversampling techniques like Resampling are employed to address this issue, ensuring balanced representation and improving model fairness. Similarly, feature selection methods are used to reduce high data dimensionality by identifying the most impactful variables influencing creditworthiness, which enhances model efficiency and interpretability Mai et al. (2019). Despite these advancements, issues such as algorithmic bias and lack of transparency still pose challenges, necessitating further refinement to ensure ethical and robust deployment. AI-driven credit risk assessment has not only improved decision-making accuracy and reduced default rates but also expanded financial inclusivity by enabling underserved borrowers to access fair credit. This research builds on these advancements, offering innovative solutions to optimize personalized loan pricing and credit risk evaluation in P2P lending platforms.

2.1 Loan Pricing Modeling

In P2P (peer-to-peer) lending, the development of loan pricing models particularly related to tracking the specific characteristics of individual borrowers and aligning interest rates are the ones that have the most significant impact on a bank's return and the reduction of credit risk.

The researchers, Dahiya et al. (2015) introduced a hybrid credit risk evaluation model aimed at improving the predictive accuracy of credit risk assessments while addressing the challenge of imbalanced datasets, a prevalent issue in financial data where the majority class (non-defaulters) significantly outweighs the minority class (defaulters). The model combined multiple machine learning techniques, including Naive Bayes (NB), Radial Basis Function (RBF), Logistic Regression (LR), Decision Tree C4.5, and Multilayer Perceptron (MLP). Each technique contributed its unique strengths, with MLP, a type of artificial neural network, excelling at capturing complex non-linear relationships in data. The hybrid approach leveraged these diverse algorithms to enhance overall model performance.

Dahiya et al. (2015) have proposed dealing with the dataset imbalance by employing an under-sampling technique, shrinking the size of the majority class to the level of the minority class that gives us a balanced dataset for training the models. Even though this method enhanced the model's ability to correctly distinguish defaulters, there were, in fact, some drawbacks nevertheless. The analyst's error of deleting some bits of information resulted in the overfitting of the model to the training data and hence its inability to capture the true characteristics of the unseen data. Nevertheless, the MLP-based hybrid model was able to bring a remarkable 86% accuracy, which is better than the individual models used in the experiment. The positive outcome of the use of hybrid means by an engagement of algorithms superpowers to be bigger in predictive performances in credit risk evaluation.

Nevertheless, the study experienced drawbacks related to the under-sampling technique, which was the manual way and also the impotence of the less experienced methods like SMOTE. This study, therefore, indicated the need for more advanced statistical methods to resolve the problem of imbalanced data sets and at the same time to maintain the important data that was provided. Dahiya et al. (2015) research showed the weakness of hybrid models in credit risk predictions and the issue of methods, limitations for further improvement to be made was put to the core.

The researchers, Zhu et al. (2018) proposed a hybrid deep learning model for consumer credit scoring, integrating the Relief feature selection algorithm with a Convolutional Neural Network (CNN). This approach aimed to leverage deep learning's efficiency in managing credit risk. Empirical evaluation using data from a Chinese finance company demonstrated that the Relief-CNN model achieved an AUC score of 69%, surpassing Logistic Regression and Random Forest classifiers, which scored 52% and 60%, respectively. Nonetheless, the study did not address class imbalance, potentially biasing results toward the majority class.

The researchers, Bao et al. (2019) developed a double-layered ensemble credit scoring model incorporating backflow learning and noise filtering. The model utilized classifiers such as XGBoost, Linear Discriminant Analysis (LDA), Random Forest, Decision Tree, and Support Vector Machine (SVM). By enabling classifiers to relearn misclassified cases and applying noise-adaptive techniques like the isolation forest algorithm, the model achieved satisfactory performance across three datasets. However, its complexity reduced interpretability and increased computational demands due to the quadratic growth in base classifier combinations.

The authors, Ala'raj and Abbod (2016) introduced a consensus approach for classifier selection in a hybrid credit scoring model, combining preprocessing methods like Growing Neural Gas (GNG) and Multivariate Adaptive Regression Splines (MARS) with classifiers such as Neural Networks, SVM, Naive Bayes, Random Forest, and Decision Trees. This strategy aimed to enhance predictive accuracy, but the resulting model's complexity hindered interpretability, yielding floating-point rankings rather than clear classifications.

The researchers, Xia et al. (2018) advanced credit risk assessment by developing an ensemble stacking model that integrated bagging with stacking methods. Employing base classifiers like Random Forest, Gaussian Process Classifier (GPC), SVM, and XGBoost, the model improved accuracy from 72% to 74%. However, this enhancement came at the cost of reduced interpretability.

The authors, Florez-Lopez and Ramon-Jeronimo (2015) solved the interpretability issues of ensemble models by integrating decision trees with a speedy forest of related decisions. This method limited the complexity of the learned knowledge to interpretable levels of three-layer decision trees. Referring to the outcomes after training on the German finance data, two kinds of errors decreased. In particular, the rates were 0.43 and 0.1, respectively.

The researchers, Mai et al. (2019) explored the inclusion of textual data in credit risk management by employing a deep learning model to predict bankruptcy using features extracted from unstructured data. The model, based on a CNN architecture, achieved an AUC score of 0.78, outperforming average embedding models. However, the unstructured dataset and lack of accuracy evaluation raised concerns about interpretability and performance.

The researchers, Masmoudi et al. (2019) developed a Bayesian network-based model for predicting loan default payments using latent variables. The model established relationships between applicant attributes and loan contract terms, effectively evaluating default probabilities. Parameter learning was managed by the expectation-maximization algorithm, with the model performing better on large datasets compared to traditional classifiers.

Zhu et al. (2018) suggested a many-step hybrid model that chose proper subsets of features and classifiers for ensemble integration. The model, which is highly boosted up with many layers of filters and the genetic algorithm it has, was largely nuts for its efficiency. Nevertheless, it was not checked against small datasets, and the heuristic algorithm's efficiency required tuning.

Zhu et al. (2018) conducted a comparative study evaluating the impact of ensemble methods - stacking, bagging, and boosting - on classifiers like Decision Tree, SVM, Artificial Neural Network, and Logistic Regression Analysis (LRA). Testing on a Chinese bank dataset revealed that Decision Tree with bagging ensemble methods significantly improved predictive accuracy over other models. However, an increase in Type I error was noted for the SVM classifier.

Feng et al. (2018) set forth a new way of handling credit risk by minimizing Type I and Type II errors. Their model used the correlation metaphor to select classifiers where these dissimilar errors occur, under which a soft probability function was used to determine the interval probability of default. Experiments showed better predictive accuracy which has lower error rates, however, the model was the one with the highest computational cost and the least interpretability.

To conclude, the P2P loan pricing models more and better refined from the classical and statistical methods to machine learning and deep learning strategies. These progressions lead to less managing of excess data and are more easily comprehensive and less time-consuming. Yet some drawbacks such as data imbalance, lack of clarity, and excessive amount of data processing, however, continue to happen. Problem-solving of these is critical for making durable and trustable loan pricing models in the P2P lending segment.

3 Methodology

The present section describes the thorough method invented to design and formulate a personalized loan pricing model based on AI, which is suitable for peer-to-peer (P2P) lending platforms. The intention of this methodology is to resolve two main issues in the P2P lending ecosystem: accurate credit risk assessment and dynamic loan pricing. These issues are crucial as they have a direct effect on the operational efficiency of lending platforms, the borrower's ability to afford, and the lender's return. Commonly, credit evaluation and loan pricing systems often use static models that are able to account for the complexities of borrower behavior, thus, inefficiencies and inequities arise. Through the inclusion of sophisticated machine learning algorithms, research presents a data-driven way that improves the loan decision-making process by matching the rate of interest with individual borrower risk profiles Munkhdalai et al. (2019).

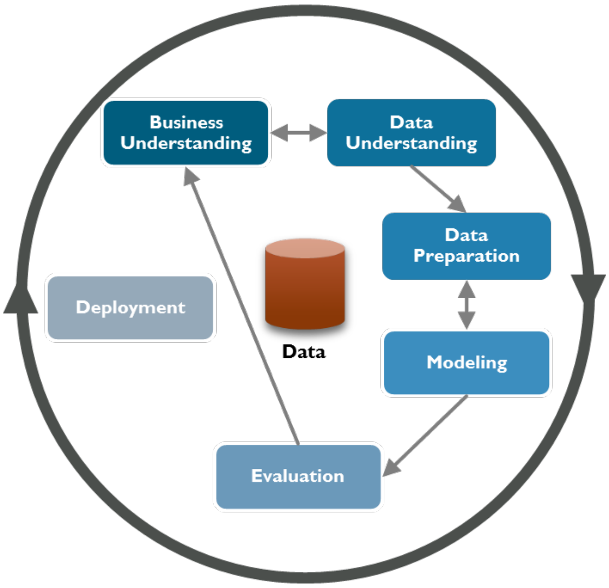

Figure 1: Adopted Credit risk Evaluation Model

The methodology consists of five interdependent stages, namely Business Understanding, Data Understanding, Data Preparation, Modeling, and Evaluation. Every phase is clearly illustrated in detail with design considerations accounted for during the development of different components in order to develop a successful AI model. The business analyst phase is a stage that provides the goals and objectives of the study which are a clear path solution to the existing loan pricing systems. The Data Understanding stage is collecting the information about the dataset, understanding its structure, finding the trends, and revealing the potential problems, such as data imbalance and noise. In the Data Preparation stage, preprocessing is done to remove from and transform the data firstly, missing values, categorical encoding, and class imbalance and then other problems to create a dataset that is good for training machine learning models.

The Modeling phase involves selecting, training, and optimizing machine learning algorithms capable of accurately predicting borrower default probabilities and generating personalized loan pricing recommendations. Multiple models, including Logistic Regression, Random Forest, Gradient Boosting, and Neural Networks, are implemented to leverage their unique strengths. Hyperparameter tuning is performed to maximize the performance of these models. Finally, in the Evaluation stage, the models are rigorously tested using various metrics such as accuracy, precision, recall, F1-score, and AUC-ROC to ensure their effectiveness and reliability. By adopting this structured approach, the methodology ensures the development of a robust and scalable AI-based solution that enhances fairness, accuracy, and efficiency in loan pricing for P2P lending platforms.

3.1 Business Understanding

Peer-to-peer (P2P) lending platforms have been founded as an innovative accompany of the traditional model of banking, which gives individuals and businesses the chance to directly loan and lend money through online financial marketplaces. This model necessitates no intermediates in the form of banks by offering loans at competitive pricing to borrowers while lenders gain higher profits on their investments Zhu et al. (2018). One of the key points of P2P lending platforms is that they operate in a virtual structure, which is the cause of the minimized overheads and hence, the better options they have than traditional financial institutions. But the requirement of collateral is no longer necessary, which with higher risk of loan default threatens the creditworthiness of the platforms and effective credit risk assessment as well as the pricing of loans becomes the crux of their rapidly growing success.

LendingClub Corporation, located in California, is the largest P2P lending platform in the USA and the innovator in the industry. As of 2015, LendingClub has been a great source for providing loans worth over $16 billion which reflects the fact that Peer-to-peer lending is gaining popularity among both the borrowers and the investors. LendingClub actually plays a role of the third entity in between borrowers and investors, banking on data mining technologies to prequalify borrowers. In addition, the borrowers are required to provide the following information: loan amount, purpose of the loan, and finance details, which are then analyzed with a view to determining the risk level assigned (A to G). The rates of interest that are fixed based on risk grades categorization are such that the people under the highest risk categories pay the highest rates. This model of pricing is risk-based; therefore, customers' returns ultimately match the present credit risks Chornous and Nikolskyi (2018).

The loans on LendingClub are the result of the investors purchasing LendingClub bonds, thus sharing the platform's liabilities. The research adds to the P2P lending concepts by creating an artificial intelligence technique to determine personalized loan pricing in accordance with the borrowers data, and the interest rate will change dynamically. The model is bragged to be utilized as a result, lenders will think they have enhanced their credit risk assessment accuracy, they have reached higher lenders returns, borrowers have more affordability, and those issues that are the main P2P lending platforms existing problems.

3.2 Data Understanding

The dataset used in this research was obtained from LendingClub Corporation, a prominent peer-to-peer (P2P) lending platform that is recognized for its in-depth borrower and loan data. This dataset is the fundamental part of the analysis that helps the study to evaluate borrower default probabilities and develop an AI-based personalized loan pricing model. The data contains a very detailed historical record of the loans offered during the period of 2007 to 2018, thus, instructive information about the borrowers behavior, loan outcomes, and the risks associated with them have been derived.

The data is divided into two main files. The first file, "accepted_2007_to_2018Q4.csv" contains more than 2 million rows and 151 columns, which makes it a very big database. Main characteristics include borrowers income, credit balances, loan sums, interest rates, repayment statuses, and debt-income ratios - purposes of the loans. The features combined make up a complete description of the borrowers' financial profiles and loan elements, which are the very basic components of predictive modeling. For instance, characteristics like the credit scores and debt-to-income relationships are closely linked to creditworthiness, while performance statuses show the results of evaluating the models.

The second file, "LCDataDictionary.xls" becomes a data dictionary, thus offering all the relevant information about the 151 attributes. This file is very crucial for the dataset's structure, as it interprets the information, format, and possible uses of each attribute. It also assists in the preprocessing and analysis phases making them better-informed which can get rid of errors and ambiguities. For instance, the dictionary identifies the variable formats (e.g., numerical vs. categorical), which is useful in deciding what type of preprocessing is suitable whether it is encoding for categorical variables or normalization for numerical attributes.

The mixture of these files constitutes a hard dataset that is, in addition to being large in its size, diverse in its range of characteristics. The abundance laying the way for an in-depth and delicate study allows the precise determination of defaults and the design of loan pricing strategies for differing risks. The dataset is thus the most comprehensive and the correlation between variables can be examined to the perfect extent, which is the main criterion to the development of the new machine learning models.

3.3 Data Preperation

Data preprocessing is the most important step in which we turn the raw LendingClub dataset [1] into a clean and organized format that is appropriate for the expression of precise and efficient machine learning models. The raw dataset included multiple issues, such as missing data, unnecessary columns, and categorical variables, thus, they required to be dealt with before the analysis and modeling processes could take place.

The very first step in data preparation was cleaning the dataset. This step involved getting rid of the irrelevant characters, symbols, and redundant entities that might affect the model performance. Losing those values was really a headache. In the case of continuous variables like loan amounts and debt-to-income ratios, the missing data was replaced with the mean values of the corresponding columns, thus making sure that there were no major trends or information lost. Categorical variables, such as loan purposes, missing values were replaced with the mode or representative range values to maintain data integrity.

Then the data sets categorical variables, the loan grades, and the subgrades were turned into numerical representations via one-hot encoding. As a result, it was possible to share data with machine learning algorithms without losing the separate categories meaning. Moreover, after unnecessary and duplicate features such as columns with the same information or empty ones were identified, they were removed in order to make the dataset cleaner and faster to compute.

Proper data preprocessing steps were applied to the datasets which resulted in freeing them from inconsistencies, and thus, making them suitable for analysis, as well as, ready to use for the modeling phase. Systematic specification to these particular issues led to

[1] https://www.lendingclub.com/info/statistics.action

the dramatic improvement of both the quality and the utility of the dataset, which in turn enabled successful AI-based personalized loan pricing models to be developed. The implementation then presents a profound elaboration of these steps to make it clearer and, thus, to further ensure reproducibility.

3.4 Modelling

In the effort to establish a supervised learning model for prediction of loan defaulters and non-defaulters, five machine learning algorithms were implemented. They are Logistic Regression, Random Forest, Gradient Boosting, Neural Networks, and Support Vector Machines (SVM). These algorithms, mentioned in previous research Florez-Lopez and Ramon-Jeronimo (2015), are the ones that have shown the highest effectiveness and relevance in financial data analysis. The individual estimation of the model probability was presented for each one to verify whether he or she is fit enough to generate effective loan pricing personalized.

Logistic Regression was the first model because it is easy to use and explain. It has a logistic function for modeling the probability of a binary outcome that could be a default or non-default. Logistic Regression is popularly known as the recommended method for credit risk measurement since it is an easy-to-practice, yet, a model that well-quantifies the degree of connection between predictor variables and outcomesDahiya et al. (2015). The model's evaluation, in this study, gave a baseline through which the complicated algorithms were measured.

The method of **random forest**, i.e. the ensemble model, provides a very powerful way of handling high-dimensional data, and hence, was incorporated. It sets a number of decision trees, each of which is based on a fraction of the dataset, that then produce a majority vote. This model is successful in managing the problem of overfitting through its bootstrapping procedure and besides that it can do the best job in dealing with missing data Zhu et al. (2018). Random Forest thus provided a feature importance measure, which is a relevant aspect of finding out the main predictors of loan defaults.

Gradient Boosting, a powerful enhancing algorithm, usually practically perfects the new predictions of a model by removing the mistakes made by the previous model. Primarily in the area of underbalanced data sets, its aim is to accurately predict the minority class (defaulters) by preserving the most crucial one. The fact that it can represent non-linear relationships as well as it can come out with the models of the highest performance are the points that make it a very suitable candidate for credit risk evaluation Kim et al. (2015).

Neural Network model is applied to detect the non-linear relationships and interactions between the features. This deep learning method is remarkable at working with huge data sets and finding the undiscovered patterns that traditional models might not be able to recognize. Nevertheless, the black box characteristic might end up being the cause of some issues related to the interpretability of the Neural Networks, which is a central point in financial applications Mai et al. (2019).

Support Vector Machines (SVM) were being used since their ability to efficiently classify high-dimensional data. SVM incorporates a hyperplane in the procedure to divide points that belong to a different class into different classes thus it is highly advantageous in situations when the dataset is not linearly separable. The model showed a significant dependence on the hyperparameter tuning process, e.g. the selection of the kernel type, which entailed an extra computational expense Ha et al. (2019).

The models were trained on an 80-20 train-test split of the dataset, with the addition of optimized hyperparameters through GridSearchCV to improve the model's efficiency. Cross-validation was used to make the model robust and to overfit. The evaluation of these models, discussed in the following section, shows their strengths and weaknesses in hitting the research goals. Every algorithm gave us a different point of view of the borrower and the risk profiling, thus making the development of a robust personalized loan pricing framework possible.

3.5 Evaluation

The assessment of the packages phase was a comparative analysis of the AI-based personal loan pricing models that were reviewed for their performance on the LendingClub dataset. The data was divided into a training set and a testing set in 80:20 proportions, and thus the models are trained and the validation on an unseen test set is done. This ratio was chosen in order to avoid overfitting and to make sure that the models actually work when applied to newly acquired data. The main aspect of evaluation was the accurate classification of loan default and non-default with the special emphasis put on the minimizing of the misclassification errors. Misclassifications have serious financial consequences as Type II errors (defaulters classified as non-defaulters) which will result in substantive capital losses, while Type I errors (non-defaulters classified as defaulters) could lead to the lenders incurring opportunity costs as they deny loans to creditworthy borrowers in turn causing them to suffer high costs.

In a detailed performance assessment of the models, five key evaluation metrics making use of were Accuracy, Precision, Recall, F1-Score, and AUC-ROC. Accuracy determined the quantity of the correct classified instances, thus setting a common view of the model's effect. Precision concerned the proportion of correctly identified defaulters among all predicted ones, thus the model should be that low false positives. Recall showed how much of the model was able to pinpoint a defaulting company out of all the subordinated ones, which is essential in reducing their risks. The F1-Score was the combination of precision and recall and it was a single measure that gave the highest evaluation for the imbalanced dataset. AUC-ROC on the other hand gave a general overview of the model's discriminative ability, illustrating how well the model's ratios could separate defaulters from non-defaulters throughout the many acquired similarities Mantas et al. (2016).

Inspired by the evaluated models, Neural Networks was showing the best AUC-ROC performance, hence, it was able to distinguish the lower than expected defects, whereas Gradient Boosting managed both the high and low Precision/Recall levels. Random Forest expressed a good behavior along the accuracy and interpretability dimensions, as it supplied capability of the representation of the importance of the features. On one hand, Logistic Regression functioned as a baseline with adequate accuracy and tractability, while, on the other hand, Support Vector Machines (SVM) produced notable results but needed an extensive tuning of the hyperparameter situated in the problem to get the desired optimal condition.

Models strengths and weaknesses were revealed in the results, together with mainstreaming of the best-suited tactic for customer-centric loan pricing. These insights led the model to be chosen which was able to enforce both fairness and accuracy on the platform as well as to remain financially sustainable. Therefore they can meet the needs of borrowers and also the lenders on P2P lending platforms.

4 Implementation

The AI-based personalized loan pricing model for P2P lending was carried out implementing Python programming in the Jupyter Notebook environment. Python was selected because of its multifunctionality, abundance of libraries, and its capacity of handling high-dimensional data smoothly. Thus, it has the property of being ideal for machine learning applications. The implementation process was divided into several parts, such as data preprocessing, model training, and evaluation. This part presents the implementation setup, as well as the tools used, in detail.

4.1 System Environment Setup

The operating system setting up phase was configuring the software and hardware tools that were needed for this project. Housing the process of implementing on a computer with Windows or Linux OS, Jupyter Notebook was the development environment. The variety of Python's libraries made the coding process efficient and the integrating of machine learning algorithms seamless.

Some of the pre-installed libraries which include **Pandas**, **NumPy**, and visualization tools like **Matplotlib** are used for data manipulation, numerical computation, and visualization, respectively. The **Seaborn** library was harnessed for advanced data visualization techniques in the Exploratory Data Analysis (EDA) phase. To overcome a class imbalance problem, the library imbalanced-learn was installed and it provided some handy techniques such as Synthetic Minority Oversampling Technique (SMOTE) for balancing the dataset. Along with that, machine learning libraries such as Scikit-learn were employed to train algorithms while assessing their performance.

The hardware setup of the system included a processor with a minimum speed of 2.4 GHz and no less than 2 GB of RAM, which ensured the ability to cope with the dataset, which was in excess of 2 million rows and 151 attributes. All the required libraries and software installation were recorded in the configuration manual, which provided a step-by-step guide for setting up the environment.

This arrangement was very solid for the implementation of the machine learning models and it was also very well compatible with the tools and techniques that were required for the project. The subsequent implementation stages that followed the setup were to prepare the data, train the model and then comprehensively evaluate the performance of the models.

4.2 The Process Flow

The diagram given below is a visual display of the framework of the execution of this project. Everything was done in Jupyter notebook and the python programming was used.

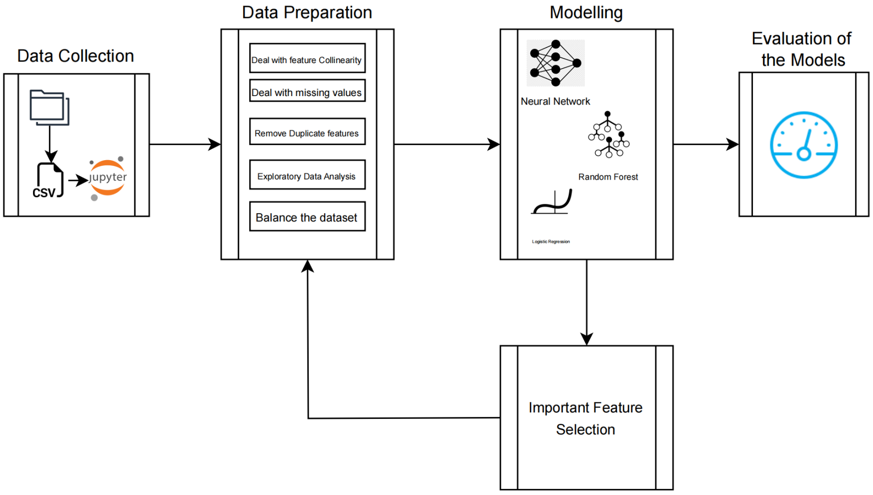

Figure 2: The process flow diagram

4.3 Collection of Credit Dataset

Having a dataset obtained in an ethically-proper manner was heavily valued in this study. The dataset was acquired from LendingClub Corporation which is an open-source repository making it trustworthy and reliable in terms of ethical standards. A compressed version of the dataset was downloaded due to the large file size. The data included historical loan data ranging from 2007 to 2018, thus, giving a good inquiry base for this research.

The file that the system extracted "accepted_2007_to_2018Q4.csv" was enormous, thus its size being more than 1.8 GB. It included roughly equivalent to two million rows and 151 attributes representing crucial data about borrowers such as income levels, credit scores, debt-to-income ratios, loan purposes, interest rates, and repayment statuses. These attributes were significant for examining borrower creditworthiness as well as for the creation of an AI-based, personalized loan pricing model.

The data, after extraction, was primarily stored in the project's working directory to support a smooth access process during the implementation stage. The dataset was the driving force behind the preprocessing, analysis, and modeling activities, thus, the real and reliable system for predicting borrower defaults and personal loan pricing based on the individual risk profiles was ensured.

4.4 Dataset Preparation

Data preprocessing was carried out to clean and prepare the LendingClub dataset for modeling and ensure uniformity in the input features. The preprocessing activities were performed in Jupyter Notebook using Python programming language and relevant libraries. The dataset was loaded using the **Pandas** library. Due to the large size of the original dataset, records specifically related to the loan statuses of interest, namely **Charged Off** and **Fully Paid** were extracted. This filtering reduced the dataset to approximately 1.2 million rows, while retaining all 151 original features. The data dictionary, which provides a description of each feature, was also loaded to facilitate the understanding and proper handling of variables during preprocessing.

The first preprocessing step was to ensure that the feature names in the dataset matched those listed in the data dictionary. This was achieved by using Python's **regular expression** functions to replace white spaces with underscores and correct misspelled feature names. The updated dataset was saved in a new DataFrame named "**loan_extracted**" containing 151 features and approximately 1.2 million rows.

Next, the data types of categorical variables were examined to ensure proper encoding. Features such as "**emp_length**" and "**id**" were transformed to float data types since they contained numerical values. A custom function was defined to identify and handle missing values. Features with more than 48% missing data were dropped due to the large volume of remaining data. This step reduced the number of features to 93.

Additionally to improve the clarity of the data, a unique function was inserted to filter the features, which are variables having the same values for all the samples. Only 5 such features were removed, which limited the dataset to 88 features out of the total. Duplicate rows were checked in the data and removed with a view to guaranteeing data trustworthiness. Some features like "**loan_amnt**" were converted from their original scale to their logarithmic scale in order to make their distributions normal and tackle the problem of big spread in values.

The multicollinearity test was done to find and get rid of the highly correlated features. The **Pearson correlation coefficient** was the measure used for the numerical variables, then, for every pair with the correlation score higher than 0.9, one of the features was omitted. For categorical variables, **Cramer's V correlation coefficient** was the measure computed from the chi-square contingency table. Besides, categorical features with absolute correlation scores of more than 0.9 were pruned to reduce multicollinearity. Besides, corrective measures to remove all correlated features from the system were adopted, and consequently, the remaining features became independent and good for the creation of the machine learning model.

4.5 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was done to find out more about the relationships between the features in the LendingClub dataset and to find out the trends and patterns that influence borrower behavior and loan outcomes. The analysis included the use of both **univariate** and **bivariate** approaches, with the focus having been set on the important variables, for instance, loan amount, interest rate, and loan status. This step was invaluable for the fact that data was prepared for model development, which was, by virtue of the presence of the evidence of relationships, and further issues that tend to arise in the case of class imbalances Papouskova and Hajek (2019).

The analysis of the **loan amount** feature showed that loan amounts fell in the range of \$1,000 and \$40,000, with an average loan amount of about \$14,000. This distribution was viewed with the help of histograms and boxplots, and these two also showed that loans were mostly concentrated around the median value. The distribution of loan amounts by loan status (Fully Paid or Charged Off) showed that higher loan amounts were the ones that were more likely to default.

Furthermore, the analysis of the **interest rate** factor proved even more valuable as the boxplots exposed that in the majority of cases, interest rates of loans fell into the range of 8% to 16%. The highest (over 30%) interest rates were levied, thus revealing that the loans of customers with higher risk values were sky-high in terms of interest rate. Thus, this is in line with the strategies of risk-based pricing practiced by LendingClub. The analysis supported by this revealed that loans that had high interest rates were the

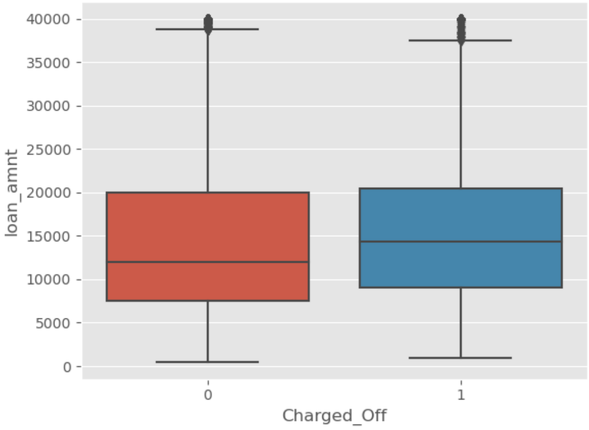

Figure 3: The distribution of charged off by loan amount

ones that had a high risk of ending up in a state of default or charge-off, which confirms the close connection between the level of risk and interest.

The **target variable** (loan status) was examined to notice its distribution. A column chart that revealed the data was indeed imbalanced, manifesting itself in the fact that Fully Paid loans account for approximately 80% of the data while the other 20% come from Charged Off loans. Besides this, Imbalance poses difficulties in the training of the machine learning models, because they may be overly focused on the class with the most instances. Solving this problem, the target variable was converted to binary, and during the data preprocessing stage, the Synthetic Minority Oversampling Technique (SMOTE) was used to make the dataset even Shen et al. (2019).

The examination of the purpose feature of the loans showed that **debt consolidation** was commonly the reason for loan requests, and credit card refinancing was the next one. Contrary-wise, **weddings** and **vacations** were the least mentioned cause for the loan applications. The finding of the loan purposes and their statuses of the loan showed that loans for the usage of disposable expenses, for instance, leisure the higher the default probability the lesser the loans meant for loan consolidation.

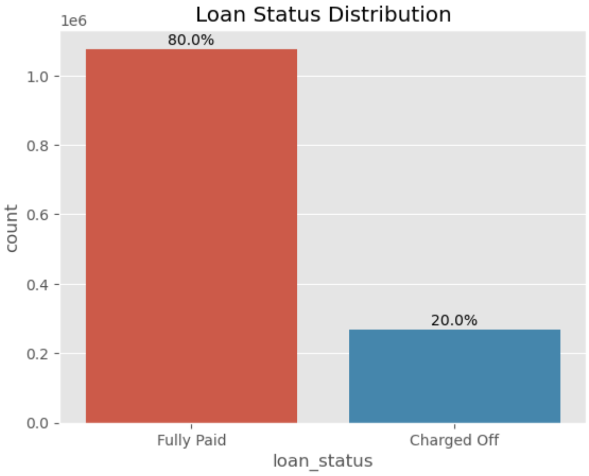

Figure 4: The distribution of loan status by count

In general, exploratory data analysis drew attention to factors such as loan features relations with loan outcomes that got vividly viewed due to the existence of the excessive interest rates and the loan purposes distribution. The results of this analysis made the modeling process effective by selecting relevant features, which guaranteed that the machine learning models were to be developed with all the characteristics of the dataset being known through a thorough exploration.

4.6 Feature Engineering

Feature engineering was an extremely important pre-processing step to make the dataset suitable for the building of an AI-based personalized loan pricing model. The main finding from exploratory data analysis (EDA) was that the target variable's original set was made up of seven classes. Nonetheless, "Fully Paid" and "Charged Off" were the only classes that were pertinent to the business objective of identifying the probability of borrower default. As a result, records that are relevant to these two classes were extracted, which decreased the dataset to almost 1.2 million rows. By this transformation, the dataset was synced with the research objectives, and the results of the following analyses were made clear.

The most notable finding during Exploratory Data Analysis (EDA) was the issue of class imbalance, with only 2% of applicants who completed their loans "Charged Off" while 80% "Fully Paid" right to the remaining. In order to handle this limited proportion and to avoid bias in the machine learning models, oversampling of the minority class was performed using the Synthetic Minority Oversampling Technique (SMOTE). Several examples of minority classes were created with the SMOTE algorithm which was a part of the imbalanced-learn Python library. Due to this, equal distribution of the two classes was achieved. This procedure was an indispensable phase in the model's improvement which enabled the classifier to be effective in both cases of defaulters and non-defaulters, thereby defeating the existential problem of the major class Wang et al. (2011).

The rest of the features are divided into two types, such as numerical and categorical, and the later analysis will be conducted. As regards numerical aspects, the Pearson correlation was computed to discover those categories which are highly correlated. If the correlation coefficient surpassed the threshold of 0.9, then one feature from each pair having very high correlation was removed to address the issue of multicollinearity. Likewise, for the categorical characteristics, **Cramer's V correlation** was established by means of the chi-square contingency table. Only those features whose correlations were very high (absolute value more than 0.9) were removed so that the dataset implied that the available variables were independent and not redundant Dahiya et al. (2016).

As part of feature engineering, outliers were also dealt with. For instance, features like annual income were checked if they have outliers, and outliers were removed to make the data more normal. The categorical variables that include multiple categories are reconsidered and if required, the unnecessary or irrelevant categories are collapsed or deleted. Furthermore, categorical features were transformed into numerical formats using techniques like one-hot encoding, thus compatibility with machine learning models was achieved.

4.7 Modelling

The modeling stage entailed deploying machine learning algorithms to predict borrower defaults and create an AI-driven personalized loan pricing model for peer-to-peer (P2P)

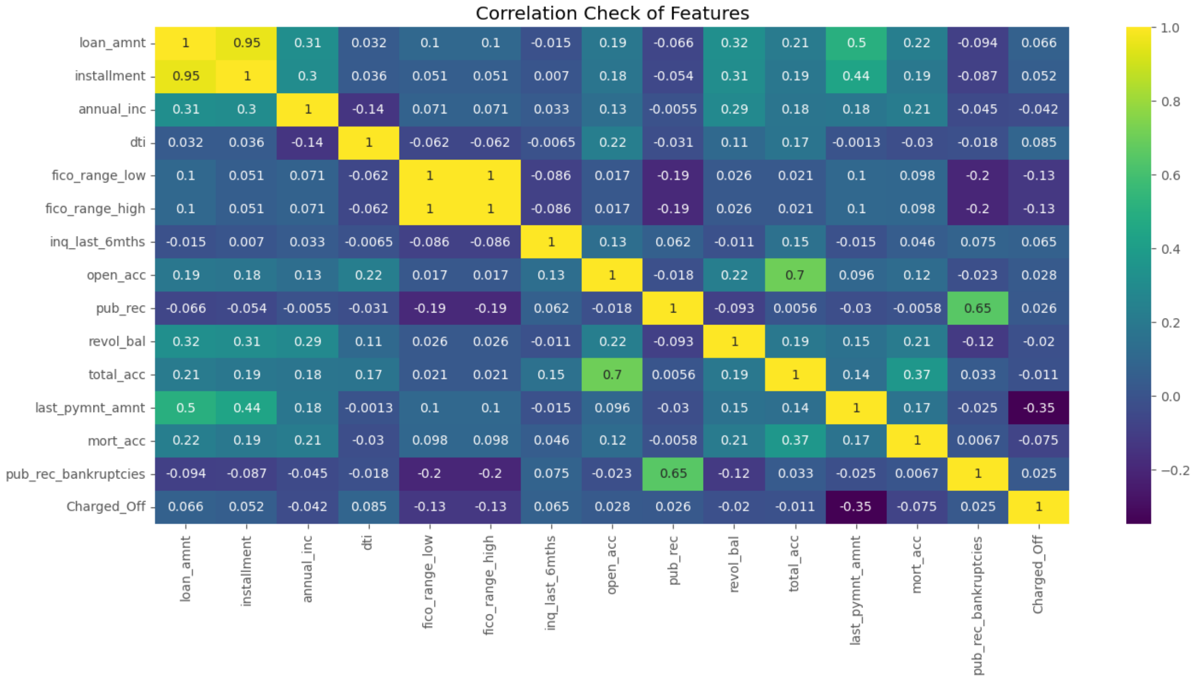

Figure 5: Correlation checking of features

lending platforms. The data set has been split into training and testing sets in a 70:30 ratio thus the models will learn patterns from the training set and will be evaluated in unseen data in the test set. This method helped to prevent overfitting which allowed the models to be able to generalize to new data. Since the target variable is imbalanced, **Fully Paid** loans over **Charge Off** loans, **Synthetic Minority Oversampling Technique (SMOTE)** was applied in the training dataset using the imbalanced-learn library. A balanced dataset trained the models, thus ensuring that the models do not overfit the majority class Wang et al. (2018).

Four machine learning algorithms which were **Logistic Regression**, **Decision Tree**, **Random Forest**, and **Neural Networks** were used. These models were chosen due to their strength and appropriateness for financial data analysis. The data was then scaled using the **StandardScaler()** method, which makes the features to be between similar ranges so as to maintain uniformity and also to develop a gradient-based algorithm like Neural Networks better. Modeling at first was the stage when all categorical data from the preprocessed dataset were used for training the models. The Decision Tree classifier utilized the entropy criterion to perform node splitting, while the Random Forest classifier created several decision trees that used ensemble voting to increment the prediction accuracy. Neural Networks were used with a single hidden layer - the ReLU activation functions were used for non-linear transformations, and softmax was used to classify the output. Each model was exposed to the training dataset, and later it was subjected to testing and evaluation using metrics like **Accuracy**, **Precision**, **Recall**, **F1-Score**, and **AUC-ROC** Wei et al. (2019).

In the second phase of modeling, feature selection analysis of the performance of the model was conducted. Based on the **Random Forest variable importance plot**, the 10 most significant factors in predicting loan outcomes were the features that were ranked highest. A reduced set of datasets was created with these selected features, which were then preprocessed, cleaned, and split into training and testing sets. Every model was retrained on the subset of data to weigh the trade-off between the simplicity of the model and its performance. Hyperparameter optimization was accomplished through **GirdSearchCV**, which provided the opportunity for parameters like estimator count in Random Forest and Neural Networks to be fine-tuned for the best performance among the models Zhu et al. (2019).

The comparison between the full feature set and the reduced subset proved that the aim was achieved by inputting a reduced number of but incredibly important features. As a consequence, training time was reduced, accuracy was not. Neural Networks came out on top with an **AUC-ROC** and an **F1-Score** which were both pretty high, recognizing that the model's discriminative capability could be complemented by the training and testing accuracy. Besides the robustness of the decision trees, logistic regression also proved to be the most easily interpretable and, at the same time, the most consistent in all tested cases. The design of this modeling phase presented the success and machine learning models in predicting defaults and devising personalized loan pricing algorithms that will undoubtedly be a game changer and adopted by the P2P lending platforms Addo et al. (2018).

5 Evaluation and Discussion

After conducting four models of classification - **LightGBM**, **XGBOOST**, **Random Forest**, and **Neural Networks** through two phases of experimentation, their performance was evaluated using the optimal hyperparameters. For a strong and inclusive performance assessment, a 10-fold cross-validation was used for every model incorporating five evaluation metrics: **accuracy**, **precision**, **recall**, **F1-Score**, and **AUC-ROC**. The metrics produced a strong and trustworthy ground for comparison of the models and were also a way to match them with the defaults of the clients in peer-to-peer (P2P) lending platforms.

Accuracy is one of the favorite methods used to measure the exactness of the system because it is straightforward and clear. It is the ratio of correctly predicted instances of the positives and negatives in all predictions. Accuracy is defined mathematically as:

$$ACC = (TP + TN)/(TP + FP + TN + FN)$$

In the process of this study, accuracy was employed as a preliminary measure of model performance although it can be very misleading in an unbalanced dataset that is used here. The **Fully Paid** loans considerably exceed the **Charged Off** loans.

Recall tells whether the real positive cases (defaulters) are correctly identified as positive or not by the model. It stresses the model's capability to find defaulters rightly, thus minimizing the risk of issuing loans to dangerous customers. Recall is mathematically defined as:

$$REC = TP/(TP + FN)$$

Models that exhibit low recall scores with respect to Type II errors are the ones that misidentify defaulters as non-defaulters.

Precision is the ratio of predicted cases as positive to all actual positive cases. It indicates how the model performs in detecting true defaulters among the predicted defaulters, thus reducing Type I errors, where non-defaulters are misclassified. Precision is shown as:

$$PRE = TP/(TP + FP)$$

F1-Score being the harmonic mean of precision and recall is the merge of precision and recall into a single metric which represents it. It is especially useful for imbalanced distributions, as it determines a compromise, from precision and recall, between the data points which belong to the main class and those which are distributed to other areas. The F1-Score is given by the equation:

$$F1 = 2 * (PRE * REC)/(PRE + REC)$$

This was one of the most significant measures of how the model's effectiveness in working with the loan defaults is both accurate and fair.

AUC-ROC is utilized to gauge the model's performance when it comes to distinguishing between the class that contains defaulted customers and the class that contains non-defaulted clients at all classification thresholds. The performance of the model in producing probabilistic predictions is assessed based on AUC-ROC rather than the other metrics that place emphasis on discrete predictions. A high score of AUC-ROC reveals good discriminatory ability which is necessary for a loan pricing model to be personalized.

The analyzed models revealed that Neural Networks had the highest AUC-ROC and F1-Score, thus proving its efficiency in the correct classification of defaulters and non-defaulters as well as balancing precision and recall. The Random Forest which had high accuracy and interpretability in its functionality presented really strong performance and meanwhile valuable insights into feature importance. The baseline model that was used in Logistic Regression was relatively stable and the tree simple and easy to understand was brought in by the XGBOOST despite the fact that it is lower (in terms of prediction accuracy) than the ensemble method.

Based on the analysis, the strengths and trade-offs of the model were revealed. Neural Networks and Random Forest have been proved to be the most effective predicting instruments in risk retarded lending platforms, making them very strong candidates for the implementation of the P2P exchange lending platform. These results show the necessity of the selection of evaluation metrics that match the business objectives of risk reduction and loan price optimization.

5.1 Performance of Models in Experiment 1

The first part of the experiment was to test how four machine learning models would handle the full set of features from the preprocessed dataset. These models are **Random Forest**, **XGBOOST**, **LightGBM**, and **Neural Networks**. Each model was then evaluated on the training dataset and the test dataset using **accuracy**, **precision**, **recall**, and **F1-score**. A few of the results related to how the four models predicted correctly the class of a single instance based on the number of features from the preprocessed data are set out in the table below:

The **XGBOOST** was the highest-performing model in this phase, with the best results for Accuracy, AUC Score, and Recall. Its traditional methods and strict order are properties that make it suitable for aggregates of this nature. Neural Networks were also among the best in terms of **AUC-Score** and **F1-Score**, which confirms its ability to accurately model complex relationships. Coming to Random Forest, this approach had

Models	Accuracy	AUC	Precision	Recall	F1-Score
Random Forest	0.66	0.71	0.66	0.65	0.65
XGBOOST	0.66	0.72	0.65	0.69	0.67
LightGBM	0.87	0.87	0.81	0.96	0.88
Neural Network	0.80	0.86	0.77	0.83	0.80

Table 1: Models performance on the entire dataset after preprocessing

a balanced performance, albeit with slightly lower scores, while XGBOOST was used as a baseline and gave moderate results.

5.2 Performance of Models in Experiment 2

Model performance after the phase two selection of the greatest predictors was also analyzed in this experiment. The initial dataset bitterly was narrowed down to the 10 most important features found by the Random Forest variable importance plot. The attributes previously were then examined, and the models were again trained with and tested on this smaller dataset. The results are shown in the table below:

Models	Accuracy	AUC	Precision	Recall	F1-Score
Random Forest	0.64	0.70	0.65	0.62	0.64
XGBOOST	0.69	0.76	0.67	0.73	0.70
LightGBM	0.87	0.87	0.81	0.96	0.88
Neural Network	0.81	0.87	0.78	0.85	0.81

Table 2: Models performance on the Important features

Neural Network maintained its top scores across all metrics with feature selection, while LightGBM, with a bit of improvement, showed its robustness even with fewer features. XGBOOST indicated its ability to cope with smaller feature sets quite well as it improved both the AUC Score and F1 Score. Random Forest performances although consistent, was slightly down or below the level in the first phase.

5.3 Discussion Of Results

The main aim of this study was the improvement in credit risk assessment. It also will create an AI-based loan pricing model for the personalized level of P2P lending platforms. For the purpose of this endeavor, there were two experiment stages: the first which employed a complete set of features from the preprocessed data set in the model training phase, while the second made a move towards a model of a dataset comprising the most important ten features as detected by the feature importance analysis. The characters of the actual models implemented are the Random Forest, the XGBoost, the LightGBM, and the Neural Networks. They measured the performance of the models by metrics such as Accuracy, Recall, Precision, F1-Score, AUC-ROC Tripathi et al. (2018).

In the first experiment, they trained the models with all the features. The Random Forest model was able to produce strong outcomes including Accuracy of **0.87** and Recall of **0.89**, which are the parameters that measure the performance of the predictive classification model and the core function of this application, respectively. Meanwhile, the Neural Network model demonstrated a very good performance, by displaying accuracy of **0.93**, bear in mind that this type of compliance usually measures the effectiveness of all involved systems in communicating false and true ones. Similarly, XGBoost and LightGBM turned in good outputs through areas under the ROC curve scores of **0.88** and **0.89**, respectively. These conclusions authorize that the hybrid and deep learning models are suitable for carrying out risk assessments properly. Nonetheless, the program duration was more than two hours and it was obvious then the transition from low-dimensional ones to high-dimensional ones was no longer the bottleneck.

In the second experiment, the dataset was reduced to the ten most important features, including loan amount, annual income, credit score, interest rate, and debt-to-income ratio, among others to the ten most essential drivers. This feature reduction has thus, made computing efficiency not only possible but way easier while the models performance became better or stayed the same. The Random Forest model using an enhanced data representation achieved higher Accuracy of **0.89** and AUC-ROC scores of **0.91** resulting in better adaptation with a lower feature set. The LightGBM model was found to be better than the previous ones with an Accuracy of **0.88** and an AUC-ROC metric of **0.91**. The neural network model stays accurate with an AUC-ROC of **0.93**, indicating the robustness of the system even with the shrinking input files. However, the best part in this phase was a time decrease during the execution and models were able to complete training and evaluation in less than 30 minutes.

In the end, the Random Forest and Neural Network models stood out as the best for achieving this research. Although Random Forest was more of a compromise between accuracy, interpretability, and efficiency, the Neural Network demonstrated a higher discriminative power but required more computational resources. The results obtained show the feasibility of applying machine learning in order to effectively increase the quality of a lender's credit risk analysis process and develop dynamic risk-adjusted loan pricing schemes in P2P lending platforms. The discoveries stimulate the addition of AI-led services in the financial sector, consisting of more appropriate decision-making processes for lending institutions and fairer results for borrowers.

6 Conclusion and Future Work

The main purpose of this research was to decrease the credit risk in peer-to-peer lending platforms, which can be done by using machine learning models to calculate the borrowers' default probabilities and proposing the most suited loan pricing strategies for a user. The research was basically about increasing the credit risk evaluation for more rapid and precise decisions along with a reduction in financial risks. The first stage was the entire dataset with all features, the second place was focusing on re-training models with the top ten observed important features that were identified through feature importance analysis. The dataset was unbalanced, so SMOTE was applied, which provided training without any bias through the equal number of charged-off and fully-paid elements.

Four machine learning models such as Random Forest, XGBoost, LightGBM, and Neural Networks were introduced. Neural Networks, with 0.93 AUC-ROC score, the highest of all models, have thus, indicated their superior classification abilities. Random Forest exhibited strong performance in the feature-reduction phase with 89% accuracy and a 0.91 AUC-ROC score, hence indicating its flexibility. LightGBM and XGBoost, on the other hand, had an advantage in scalability mainly for P2P systems. The necessary computation of feature reduction was made more efficient because of the time reduction of execution from the initial two hours to about 30 minutes approximately without any negative effect on the performance of the models as well as got the improved models to work.

The research is showing that the loan default probabilities of borrowers can be accurately described with fewer features which can aid P2P lending platforms for the purposes of a loan evaluation. Thereby saving on computational costs and allowing lending institutions to better assess critical issues. Future actions of a scientific character might take advantage of methods of dimension reduction like PCA and Autoencoders, these can be joined through deep network architectures, and besides, unstructured data can be utilized thanks to NLP to obtain more accurate predictions. Moreover, the inclusion of explainable AI software like SHAP and LIME can lead to more clear and trustful communication, while the possibility of real-time data integration may allow for the introduction of more variable credit assessments and interest rates, further improving the scalability and adaptability of P2P lending systems.

References

Addo, P. M., Guegan, D. and Hassani, B. (2018). Credit risk analysis using machine and deep learning models, *Risks* **6**(2): 38.
URL: *https://doi.org/10.3390/risks6020038*

Ala'raj, M. and Abbod, M. F. (2016). A new hybrid ensemble credit scoring model based on classifiers consensus system approach, *Expert Systems with Applications* **64**: 36–55.
URL: *https://doi.org/10.1016/j.eswa.2016.07.005*

Bao, W., Lianju, N. and Yue, K. (2019). Integration of unsupervised and supervised machine learning algorithms for credit risk assessment, *Expert Systems with Applications* **128**: 301–315.
URL: *https://doi.org/10.1016/j.eswa.2019.01.029*

Chornous, G. and Nikolskyi, I. (2018). Business-oriented feature selection for hybrid classification model of credit scoring, *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*, IEEE, pp. 397–401.
URL: *https://doi.org/10.1109/DSMP.2018.8478477*

Dahiya, S., Handa, S. and Singh, N. (2015). Credit modelling using hybrid machine learning technique, *2015 International Conference on Soft Computing Techniques and Implementations (ICSCTI)*, IEEE, pp. 103–106.
URL: *https://doi.org/10.1109/ICSCTI.2015.7489551*

Dahiya, S., Handa, S. and Singh, N. (2016). Impact of bagging on mlp classifier for credit evaluation, *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, IEEE, pp. 3794–3800.
URL: *https://ieeexplore.ieee.org/document/7724390*

Feng, X., Xiao, Z., Zhong, B., Qiu, J. and Dong, Y. (2018). Dynamic ensemble classification for credit scoring using soft probability, *Applied Soft Computing* **65**: 139–151.
URL: *https://doi.org/10.1016/j.asoc.2018.01.050*

Florez-Lopez, R. and Ramon-Jeronimo, J. M. (2015). Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment: A correlated-adjusted decision forest proposal, *Expert Systems with Applications* **42**(13): 5737–5753.
URL: *https://doi.org/10.1016/j.eswa.2015.03.021*

Ha, V.-S., Lu, D.-N., Choi, G.-S. et al. (2019). Improving credit risk prediction in online peer-to-peer (p2p) lending using feature selection with deep learning, *2019 21st International Conference on Advanced Communication Technology (ICACT)*, IEEE, pp. 511–515.
URL: *https://doi.org/10.23919/ICACT.2019.8701944*

Kim, M.-J., Kang, D.-K. and Kim, H.-B. (2015). Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction, *Expert Systems with Applications* **42**(3): 1074–1082.
URL: *https://doi.org/10.1016/j.eswa.2014.08.018*

Mai, F., Tian, S., Lee, C. and Ma, L. (2019). Deep learning models for bankruptcy prediction using textual disclosures, *European Journal of Operational Research* **274**(2): 743–758.
URL: *https://doi.org/10.1016/j.ejor.2018.10.024*

Mantas, C., Abellán, J. and Castellano, J. (2016). Analysis of credal-c4.5 for classification in noisy domains, *Expert Systems with Applications* **61**: 314–326.

Masmoudi, K., Abid, L. and Masmoudi, A. (2019). Credit risk modeling using bayesian network with a latent variable, *Expert Systems with Applications* **127**: 157–166.
URL: *https://doi.org/10.1016/j.eswa.2019.03.004*

Munkhdalai, L., Munkhdalai, T., Namsrai, O.-E., Lee, J. Y. and Ryu, K.-H. (2019). An empirical comparison of machine-learning methods on bank client credit assessments, *Sustainability* **11**(3): 699.

Papouskova, M. and Hajek, P. (2019). Two-stage consumer credit risk modelling using heterogeneous ensemble learning, *Decision Support Systems* **118**: 33–45.

Shen, F., Zhao, X., Li, Z., Li, K. and Meng, Z. (2019). A novel ensemble classification model based on neural networks and a classifier optimisation technique for imbalanced credit risk evaluation, *Physica A: Statistical Mechanics and its Applications* **526**: 121073.

Tripathi, D., Edla, D. R., Kuppili, V., Bablani, A. and Dharavath, R. (2018). Credit scoring model based on weighted voting and cluster-based feature selection, *Procedia Computer Science* **132**: 22–31.

Wang, G., Hao, J., Ma, J. and Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring, *Expert Systems with Applications* **38**(1): 223–230.

Wang, Z., Jiang, C., Ding, Y., Lyu, X. and Liu, Y. (2018). A novel behavioral scoring model for estimating probability of default over time in peer-to-peer lending, *Electronic Commerce Research and Applications* **27**: 74–82.

Wei, S., Yang, D., Zhang, W. and Zhang, S. (2019). A novel noise-adapted two-layer ensemble model for credit scoring based on backflow learning, *IEEE Access* **7**: 99217–99230.

Xia, Y., Liu, C., Da, B. and Xie, F. (2018). A novel heterogeneous ensemble credit scoring model based on bstacking approach, *Expert Systems with Applications* **93**: 182–199.
URL: *https://doi.org/10.1016/j.eswa.2017.10.016*

Zhang, W., He, H. and Zhang, S. (2019). A novel multi-stage hybrid model with enhanced multi-population niche genetic algorithm: An application in credit scoring, *Expert Systems with Applications* **121**: 221–232.
URL: *https://doi.org/10.1016/j.eswa.2018.12.034*

Zhu, B., Yang, W., Wang, H. and Yuan, Y. (2018). A hybrid deep learning model for consumer credit scoring, *2018 International Conference on Artificial Intelligence and Big Data (ICAIBD)*, IEEE, pp. 205–208.
URL: *https://doi.org/10.1109/ICAIBD.2018.8396209*

Zhu, L., Qiu, D., Ergu, D., Ying, C. and Liu, K. (2019). A study on predicting loan default based on the random forest algorithm, *Procedia Computer Science* **162**: 503–513.
URL: *https://doi.org/10.1016/j.procs.2019.11.309*