National College of Ireland

Predictive Analytics for Credit Risk Assessment in Peer-to-Peer Lending Platforms.

MSc Research Project
Programme Name Financial Technology

Femi Benjamin Obadimu
Student ID: 22244336

School of Computing
National College of Ireland

Supervisor: Brian Byrne

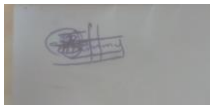
National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Femi Benjamin Obadimu
Student ID:	22244336
Programme:	MSc Financial Technology… **Year:** 2024
Module:	Practicum
Supervisor:	Brian Byrne
Submission Due Date:	12 August 2024
Project Title:	Predictive Analytics for Credit Risk Assessment in Peer to Peer Lending Platform
Word Count:	5995 **Page Count** 15

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: …12 August 2024

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):	

Predictive Analytics for Credit Risk Assessment in Peer-to-Peer Lending Platforms.

Femi Benjamin Obadimu
Student ID: 22244336

Abstract

Over the years, the continuous development of Fintech has accelerated the rapid growth of the P2P online lending market by presenting a unique platform for investment and effectively matching borrowers directly with lenders on the social lending platform. However, these online lending platforms have experienced a very high rate of default over the years which can lead to a potential reduction in the number of investors on the platform. This research study is carried out with the sole aim of indicating crucial variables to be considered during credit risk assessment by investors to make informed decisions and effectively reduce loan defaults and increase return on investment. This will be achieved by utilizing the data on the Bondora peer-to-peer lending platform which operates across Europe with its headquarters in Estonia. To achieve this result, three machine learning models will be utilized to rank these features according to the order of importance and build a robust credit scoring model. The machine learning used to rank the variables on the Bondora lending platform and build a robust credit risk assessment model includes logistic regression, XG boost, and Random forest.

Keywords: Peer-to-peer lending, Credit Risk Assessment, Bondora, Machine learning

1 Introduction

The exposure to World Wide Web services where tasks can be executed in a few seconds has changed the landscape of the finance industry, especially in terms of loans and loan applications thereby making online P2P loans play a vital role in financial consumption. Peer-to-peer lending is an end-to-end social lending platform that operates solely with information technology and the Internet. The P2P lending system is an effective way to enhance capital allocation and reduce ambiguity in the borrowing process. However, the borrowers and lenders in a social lending platform do not have physical contact with each other which tends to make the borrowing platform very risky because the lending risk cannot be measured accurately. This has led to the introduction of loan/credit risk assessment with the sole aim of ensuring adequate and proper operation of the peer-to-peer lending system and a healthy financial economy in the long run.

In this study, different ensemble techniques will be used to build a credit score model and generate a reliable result. Ensemble learning stands out as a powerful technique in machine learning offering a robust approach in improving model performance and predictive accuracy. Combining the strengths of multiple individual models, ensemble methods can outperform any single model making them a valuable machine learning toolkit. Ensemble techniques are effective because they reduce overfitting and improve generalization leading to a more robust model. An efficient credit system is very crucial environment as this will help to further reduce the probability of fraud and enhance continuous positive development of credit.

In credit assessment, there is an important decision to be made in terms of loans disbursement to intending borrowers and the accuracy of these decisions is based on the use of different algorithms instead of human intelligence (Ding 2018).

Furthermore, dataset and algorithms play an indispensable role in enhancing the performance of a credit risk assessment. The data used for this assessment (credit risk) usually have huge data anomalies and the features with less importance in this data eventually lead to reduced accuracy when data is being classified. Hassani et al (2020). The process of adequately choosing vital/important features (features selection) is needed for achieving predictive accuracy and further boosting accuracy. This technique is used in the predictive accuracy and further boosts accuracy. This technique is used in the processing stage of data classification.

In this study, Different machine learning models will be utilized to carry out a comparative analysis and further enhance the scalability and accuracy of the algorithms that will be used in assessing credit risks in peer-to-peer lending. The efficiency of these algorithms is assessed by using the peer-to-peer lending data set downloaded from the Bondora p2p platform.

Research question: What are the key factors for assessing credit risk on the Bondora Peer-to-Peer lending platform?

Research objective: To develop and validate an advanced predictive analytics model for credit risk assessment on the Bondora peer-to-peer platform aiming to enhance accuracy with a critical focus on creditworthiness evaluations and contribute to the overall risk management strategies within the peer-to-peer lending industry. This study will focus on integrating diverse data sources such as applicant profiles and loan characteristics to create a robust predictive model that effectively identifies and evaluates credit risks.

This objective outlines the specific focus on predictive analytics, credit risk assessment, and the incorporation of various data sources within the context of peer-to-peer lending platforms. It provides a clear direction for the research, emphasizing the goal of improving the credit evaluation process for better risk management in P2P lending.

2 Related Work

2.1. **Evolution and Growth/ Background**

P2P lending platform has changed how corporate individuals and private businesses access loans to finance their business. This platform was created to revolutionize finance and include technology by making funds easily available to potential borrowers, and excluding traditional financial institutions during the disbursement process.

The P2P platform began to grow rapidly between 2005-2009 across Europe and the United States with online lending platforms like Zoppa, Prosper, and Lending Club. These social lending platforms utilized technology to successfully assess and manage risks involved in borrowing funds from customers, and link both borrowers and lenders. P2P platforms also created a platform for lenders to earn while they borrow through the application of interest rate =s attached to disbursed loans. This made the platform to be well accepted in across Europe and America. In the last ten years, Peer to Peer lending platform has continued to grow across different continents including Africa and Asia, especially in countries like China and Kenya.

However, apart from offering personal loan services to customers, Online peer-to-peer platforms also offer other services like business loans and real estate financing. The inclusion of these loan services has made online peer-to-peer lending platforms more flexible for investors to further diversify their lending portfolios. Borrowers can also apply for various loans based on the customer's loan profile and financial statement (Bhanushali 2024).

2.2. Traditional lending and Peer to Peer lending.

P2P was created with the sole intention of eliminating financial intermediaries between borrowers and lenders.P2P is also referred to as marketplace lending. Both terms can be used to interchangeably but the difference between marketplace lending and P2P is that marketplace lending involves institutional investors participating in the lending market while participants in P2P are basically individuals and medium corporate businesses. Marketplace lending can also be further divided into (i) Consumer Lending (ii) Business lending (iii) Property lending. Consumer lending covers the larger part of marketplace lending. That is most loans generally fall under consumer lending. Consumer lending is approved for different reasons and they include debt consolidation, credit card refinancing and property development. Business lending is used majorly by production companies, engineering companies, and Transportation companies. Property lending firms are involved in providing flexible finance products including residential mortgages and commercial mortgages. The first P2P lending platform to commence the business of borrowing and lending without intermediaries is Zopa, a P2P lending platform founded in the UK in 2005, and Prosper (2006) founded in the US. These two companies pioneered the P2P lending platform in Europe and America fostering a decentralized market lending platform where prospective

borrowers and lenders can execute business directly with each other without the presence of an intermediary. Wang et al (2022)

2.3. Regulation Frameworks governing P2P governing P2P lending in different regions.

The rules and laws guiding the growth and innovation in Fintech are usually different across countries and regions all over the world. These rules convey different ideologies and methods used in analysing different problems and emerging positive development in the Fintech Industry (Uwaona et al 2023). Despite various rules and regulations governing the fintech industry in different countries use regulatory frameworks that are dynamic, and easily adapted to by emerging financial technology companies as consumer rights are held in high regard (Okoye et al, 2024). Due to unrestricted activities of technologies across different borders of different countries, it is imperative to ensure harmonization and coordination of rules and regulations across different border countries. (Ivanova 2019).

In the last few years, Regulatory sandboxes have been developed to further enhance innovation and provide existing fintech firms with a regulated environment where they can test-run their products and services intended for final consumers. The regulatory sandbox will allow regulators in the fintech industry to gain deeper understanding of the product, analyze the risks associated with the product and develop a flexible regulation regarding the product without discouraging future innovations and endangering the safety of customers. Regulatory sandboxes are regulatory innovation hubs that are specifically designed for fintech companies to seek regulatory support, guidance, and cooperation with Fintech regulators. Regarding their product with the sole aim of ensuring total compliance with the industry's regulations. (Atadoga et al 2024). Furthermore, government and regulatory authorities of different countries are very important stakeholders in regard to ensuring a safe and compliant environment for fintech startups in terms of integrity, stability, and efficiency in the industry. The central bank of these countries also ensures proper adherence to payment and settlement laws, formation of the monetary policy, and partnering with other regulators in the Fintech industry to resolve developing issues in the Fintech industry (Ebiogbe et al 2023).

Furthermore, International financial regulatory organizations including World bank, International Monetary Fund (IMF) and Financial Stability Board (FSB) also render assistance in terms of research, guidance and policy formulation to countries that actively engage in financial technology for economic growth and development (Uwaoma et al, 2023). Regulatory synchronization among different countries across Europe and Africa where fintech is actively taking over would further help to facilitate an efficient fintech ecosystem across the globe.

Regulatory frameworks utilized in the fintech industry must ensure safety and protection of consumers is highly prioritized when these policies are being formulated because consumers are critical to the success of any fintech product. These consumers also determine the continuous existence of emerging fintech companies. Therefore the regulatory framework must include equal customer service (elimination of segregation among customers), swift

financial transaction, disclosure of hidden charges and reliability in periodic reports of these fintech companies. (Alekseenko, 2022).

2.4. Credit Risk Assessment in Peer-to-Peer Lending

Evaluating a potential borrower's creditworthiness accurately has become one of the major issues faced by lenders and peer-to-peer platforms because these lending platforms give out loans to their customers without adequate collateral just in case the loan becomes a bad debt and becomes written off. Lenders tend to charge a very high interest rate which will serve as insurance for the loan in the event of default by the borrower (Mild et al 2015).

Credit loans have formed a crucial part of the financial sector and calls have been made by investors for the industry to minimize risks involved in borrowing. However, Credit risk evaluation is a very important aspect to be taken into deep consideration before loans are disbursed to borrowers (Bekhet 2014).

Khashman (2011) further indicated that credit risk assessment in the Peer-to-peer lending platform is very important for adequate credit and financial risk management. The inability to accurately recognize risks can have a negative effect on the decisions made by lenders to disburse funds to borrowers thereby making the investment unprofitable for investors.

Despite the rate at which the P2P has been widely accepted especially in Europe, the high probability of investment failure cannot be overlooked, and this can be attributed to the unsecured loans and information asymmetry on the platform. The fact that both lenders and borrowers do not meet physically to exchange documents in regards to the loan applied for has contributed to the high risk involved in the online business lending business (Lin and Zeng 2017).

Borrowers information including demographic and financial information is used to further analyse and determine loan approval by lenders. Lenders use these two factors to determine the extent of risk they are willing to undertake on a P2P lending platform. In addition, strong social networking can also play an important role in choosing on investment that will yield a high return (Lin et al 2013). Furthermore, Nyberg (2019) stated that it is more advantageous to disburse loans that have little or no risks attached, and a very low debt-to-income rate.

The rapid development of the P2P platform over the years has made the platform a good substitute in terms of alternative microfinance. However, the credit risk attached to the P2P platforms has discouraged potential investors from investing funds in the platform. However, Emektear et al (2015) discussed the crucial importance of borrowers' characteristics on loan default and modelled the probability of borrowers defaulting their loans with the use of binary logistic regression. The relationship between the duration of the loan and the possibility of default was further tested with the use of cox proportional hazard model. Lin et al (2017) also analysed risk attributed to borrowers from a P2P platform in China building a logistic regression model to examine and analyse the credit risk in a P2P platform and the outcome indicated that borrowers with minimal loan default risks are young adult, working class, married individuals and well-educated loan applicants.

2.5. Credit Scoring.

Credit scoring is used by financial institutions to assess a borrower's probability of default loan repayment. Credit scoring is an analytical approach that determines a loan the solvency of a borrower.

These credit scoring systems have the high capacity to successfully process and analyse huge amounts of loan applications quickly thereby eliminating excess costs that would be incurred in the process. The continuous increase in credit defaults in the financial sector has made the use of credit scoring methods very essential for lenders to make informed decisions regarding loan disbursement to potential borrowers. (Ghatge and Halkarnikar 2013). Credit scoring models are constructed using linear methods including logistic regression. However, Artificial intelligence and Machine learning methods including Decision trees, Random forest, Artificial Neural Networks, and logistic regressions are used in building credit scoring models (Bekhet and Eletter 2014).

Studies over the years have revealed that the lending process has been severely affected by demographic information including age, gender, and various historical records. Duarte et al (2012) stated that individuals who look trusted and reliable can get their loan applications approved with very low interest rates. By classifying lenders under manual and automated bids, Zhang and Chen (2017) revealed that people who invest in the peer-to-peer lending market do not make independent decisions based on personal research but would rather follow the choice of other investors by choosing to fund similar loans chosen by other investors thereby leading to imbalance and high risk in the P2P market. Using the data on Ren Ren Dai, an Indian digital lending platform for SMEs seeking loan, Wang and Zhang (2019) propose that the use of video information is an efficient approach to build trust. And affect the lending attributes/ behaviour of borrowers who have very low credit ratings. Zhang and Wang (2018) also assessed how risks are being categorized in P2P platforms in China by using entropy revised model and correlation analysis.

However, different models are being used by banks and several financial institutions to evaluate credit default risk assessment and these models include logistic regression, probit regression, and Bayesian Networks. Maria Rafael (2013). Tsai and Yen (2014) decided to evaluate the most appropriate method that best categorizes ensembles and decide on the best ensemble method that can be used for credit scoring assessment. The outcome of this evaluation concluded that decision tree ensembles which contains 80 classifiers had the most accurate and reliable result using the boosting method. Furthermore, King and Ragsdale (2015) also carried out a study that revealed that the stacking esemble method always perform better when used to forecast or predict credit risk assessment because the outcome or results would be reliable rather using just one approach like decision trees, logistic regression and Naïve Bayes to forecast credit risks that using ensemble method is more effective and reliable.

2.6 Predictive Analytics: An overview

Predictive analytics has become a critical part of lending decisions in the peer-to-peer lending system due to the ever-growing customer base and continuous demand for loans across Europe. Predictive analytics consists of Artificial Intelligence, Machine learning, and statistical methods used in critically analyzing previous and current data of consumers with the sole aim of arriving at an essential conclusion regarding future occurrences. The essence of predictive analytics is its ability to analyze and process huge amounts of data and discover trends and anomalies that can not be easily detected by humans. This data-driven method gives an objective and quantifiable (in terms of numbers) decision making unlike traditional lending where decisions are made depending merely on human instincts. In credit risk management, models built with predictive analysis can assess borrower's transaction patterns, social media behaviour, and financial statement with the aim of assessing and analyzing a customer's creditworthiness.

Predictive analytics can also be used to uncover fraud in the finance sector. Banking utilize modern learning models to identify unusual patterns that can detect fraudulent activities. These models also have the capability of gaining an understanding of different types of fraudulent activities and modifying the algorithms to match with the current trends thereby ensuring standard accuracy and relevance over time (Andriosopoulos et al 2019).

The introduction of predictive analytics with credit risk assessment can be attributed to the continuous modification of technologies used for data processing and analytics. Predictive analytics uses machine learning, artificial intelligence, and statistical techniques to interpret and assess historical data and draw conclusions about future events. However, in the situation of social lending, and credit risk assessment, Predictive analytics uses huge data to predict the probability of intending borrowers, being insolvent after a loan has been granted. Furthermore, predictive analytics ensures the reliability balance and profitability of financial institutions by curbing the high percentage of loss experienced by financial institutions regarding credit risks (Lera et al 2019). Wang and Greiner (2011) stated that online peer-to-peer lending platform was widely accepted due to low transaction costs for both parties (Borrowers and Lenders) and microloans are easily facilitated. However, the social lending platform is still plagued with information asymmetry which hinders the productivity of the market. In a case where the interest rate associated with a loan is significantly high, the loan is considered a risky investment for the borrower. However, due to lenders' limited access to information regarding the borrower, a high-interest-rate loan may not seem appealing to them Yum and Chae (2012). Investors prefer to approve short-term loans because these loans have feasible payback periods. Furthermore, financial information provided by the borrower regarding credit history is very important to the loan request being approved. Freedman and Jin (2008) also state that more loans were approved on the Prosper social lending platform because the social lending company requested more information (financial) from the customers.

2.7 Data in Predictive Analytics for Credit Risk Assessment.

Every database kept by financial institutions has different variables for every customer's credit history. The higher the number of variables the financial institution needs to deal with in the database, the tougher it is to successfully categorize customers and decide the effect of these variables regarding a customer's probability to default. However, explanatory variables are usually used to build credit models (Marshall et al 2010).

Volk (2023) stated that financial institutions should gather non-financial information regarding the borrower to foster a close relationship with their customers. As this information can be used to evaluate the liquidity and creditworthiness of potential borrowers. Grunert et al (2005) further analysed different European financial institutions and discovered that the use of financial and non-financial variables in building credit risk models facilitates accurate forecasting instead of using the variables separately (Financial and non-financial).

---

3 Research Methodology

This section describes the utilization of different machine learning models used in building a credit risk scoring model and they include XG boost, Random Forest, and Logistic Regression. Confusion matrix, F1 score, recall and precision were also used to analyse the key features in the data set.

Logistic Regression: This is referred to as a supervised machine learning algorithm used to execute binary classifications. Linear regression models are used to define the relationship between dependent and independent variables. Logistic regression is used to distribute variables into categories (Yes or No, 1 or 0). The independent variable in this research is referred to as binary variable which is categorized as 0 (late) 1 (repaid).. in logistic regression, a logit transformation is used to to evaluate the probability of success divided by the probability of failure. Which is also referred to as log odds. The formula for this logistic function is shown below.

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_n X_n)}}$$

P(Y=1IX) is the probability that the dependent variable is Y given that the independent variable is X

e = base of the natural algorithm

$Bo = intercept\ term$

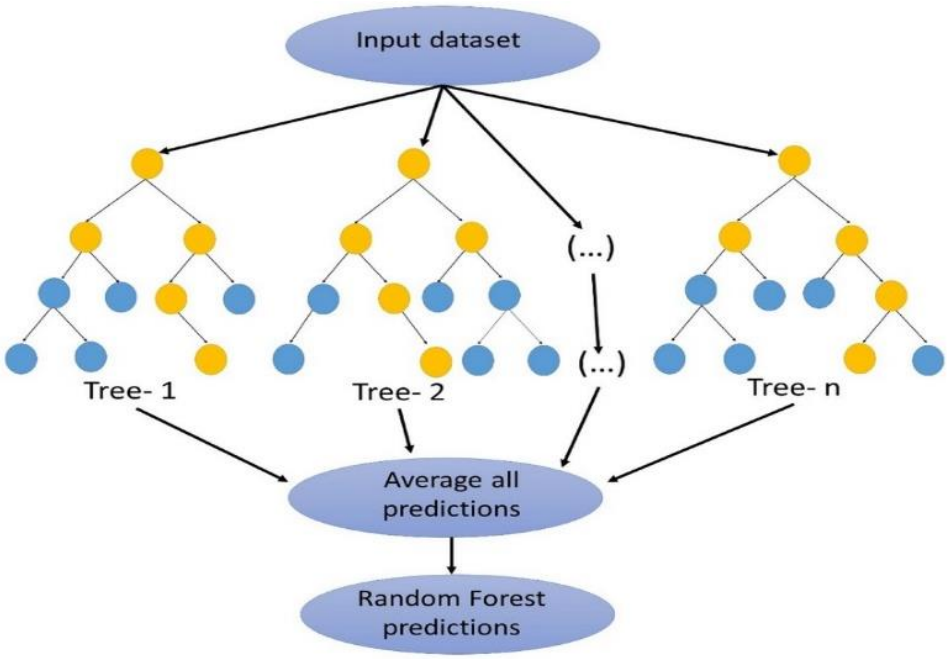
$B1$, B2, B3 are coefficients of the independent variables X1, X2 and Xn.

XG Boost: Extreme Gradient boosting is a machine learning algorithm classified under ensemble learning. This machine learning algorithm involves using decision trees as base learner. XG boost is widely used for predictive analytics because of its ability to analyze

complicated relationships between variables and its ability to handle overfitting in a data set. Overfitting occurs when a machine learning model can not execute other predictions form a different data set apart from the data set its has been trained to predict. XG Boost can handle large data sets and different data types including regression, ranking and classification. This machine learning algorithm is well known for efficient data processing, dealing with missing values, and carefully analysing variables of a data set.

Random Forest: This is referred to as a supervised classification algorithm which consist of ensemble decision trees. These trees can dictate rules at each node according to the trained data set. Random forest combines different results of different decision trees to reach a final conclusion. The random forest technique creates multiple models by using bagging selection training and random selection features.

In random forest when the nodes are being split, the model does not go for the most important feature in a subset but goes for the best features in a random subsets of features which helps to give a better result. Furthermore, in a random forest classifier, Random subsets of the features are the only subsets evaluated by the algorithm when nodes are separated/split.



Model Performance Measures.

Confusion Matrix, F1 score, recall, Precision were used to evaluate the performance of the machine learning used during this study.

Confusion Matrix: This matrix is used to measure the performance of a machine learning model like Logistic Regression and Random Forest. The matrix compares the actual target with predicted target with use of machine learning. It is used to visualize and summarise the performance of a classification algorithm. Confusion matrix displays the number of True positives, True Negatives, False positives and False negatives in a model's prediction.

F1 Score: F1 score is also known as harmonic mean is used to measure the accuracy of a model by combining the precision and the recall score. The F1 score is used to evaluate the balance between the precision and recall score of a prediction model.

Recall: This measures how a model can effectively identify relevant features in a data set. It compares predicted labels with actual labels. Recall identifies how accurately a machine learning model can identify true positives from positive samples in a data set.

Precision: This metric measures the quality of a positive prediction made by a model. It is the total number of true positives divided by total number of positive predictions.

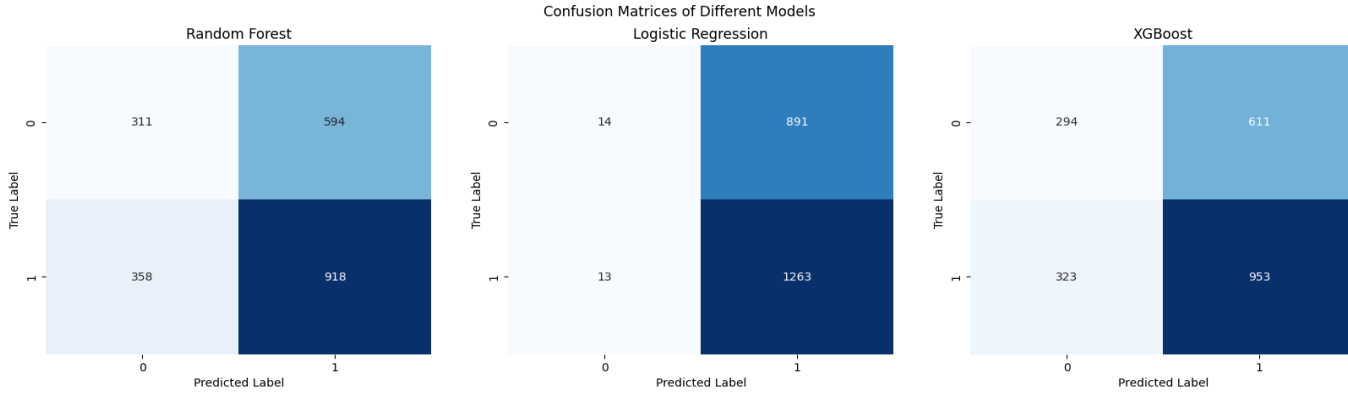
4 Design Specification

This section highlights the algorithm used during the course of this project. During this study, Google Colab hosts a Jupyter notebook service that allows writing and running of python codes and provides access to GPU and TPUs used to execute descriptive statistics and Exploratory Data Analysis (EDA). Python programming language has different libraries including numpy, pandas, matplotlib, and seaborn which was used during the course of this study. The google colab environment also supports machine learning algorithms like Random forest, Logistic Regression, and XG boost which were used analyse data during this study.

5 Implementation

This research was conducted using data downloaded from the Bondora lending peer to peer platform. The dataset was transformed into a CSV file in order to suit the Google colab environment and run successfully. After uploading the file into colab in a CSV format, python codes were developed and executed for further analysis. The chi-square test was used to determine how significant the variables in the data set are to the target variable (status) to ensure accurate P value of variables. Exploratory data analysis was also used to create a visual understanding of the variables and further highlight their significance to the target variable.

6 Evaluation



Logistic Regression.

True positive: Logistic Regression model predicted repaid loans as 1263, True Negative which identifies late loans as 14, false positive which is the total number of late loans wrongly predicted as repaid is 891, and false negative of 13 which indicates loans that were wrongly predicted as late rather than repaid. To further measure the performance of this model, the precision score which measures the number of positive predictions of a model calculated the precision of this model to be 59%. This means that the model accurately repaid loans on the Bondora platform at 59%. Furthermore, the recall of this model which measures how effective a model is in recognizing positive instances from actual positive samples in a dataset indicates that the logistic regression model was able to recognize 99% positive samples in the bondora data set which means 99% of the loans are categorized as repaid. Also F1 score which indicates a balance between recall and precision of models measure the balance of the machine learning model used to analyze repaid and late loans on the bondora lending platform as 73.6%. However, the overall accuracy of the model which measure how the model correctly predicts the loan status on the bondora p2p lending platform is 59% with a high recall rate of 99% further indicating that the model identifies a high percentage of repaid loans. The model correctly predicts a high number of repaid loan which is 1263 showing its effectiveness in its ability to recognize loans that are repaid.

Random forest.

Random forest model correctly predicted repaid loans as 918, True Negative, is the total number of loans predicted by the model to be late which is 311. The false positive according to the confusion matrix is 594. 594 late loans were predicted to be repaid by the random forest model and a total of 358 repaid loans were incorrectly categorized as late loans by the model. According to the confusion matrix of this model, The precision of the model was 60.7%. However, the model was able to predict 60.7% of loans as being repaid and effectively identify 71.9% of all repaid loans on the Bondora data set further indicating that 71.9% of loans has been repaid by borrowers. However, the F1 score which measures the

accuracy of both precision and recall of this model by ensuring a balance between recall and precision indicates that the outcome of both accuracy and precision of this model is 65.9% balanced. Furthermore, the accuracy of this model is measured at 56.4% which indicates that random forest model accurately predicts 56.4% of the loan status between 2018 and 2023 correctly.

XG Boost

According to the confusion matrix, XG boost model correctly predicted 953 loans as repaid (True Positive), True Negative which indicates total number of correctly predicted late loans on the Bondora platform as 294, false positives which indicates the number of loans identified as repaid but were paid late by the borrowers as 611, and total number of false negatives which indicates the number of repaid loans that were wrongly classified as late on the platform as predicted by XG boost was 323. However, the precision score of 0.609 (60.9%) indicates 60.9% of loans were accurately predicted as repaid, Also, the model identified 74.47% as the total number of actual positive cases which was correctly predicted (Recall). The F1 score indicates a positive balance of between both precision and accuracy (67.1%). However, the accuracy score of 57.2% indicates that the model correctly predicts the loan status 57.2% . indicate that the model was able to predict loan default and repayment on the bondora peer to peer platform accurately by 57.2%.

6.1 Discussion

Interpretation of Results/Discussion

The purpose of this study is to build predictive models that can analyse and forecast and assess credit risk in peer-to-peer lending platform . the data set used for the purpose of this study was downloaded from the Bondora peer to peer lending platform that operates in Estonia the data set contains 15821 rows and 47 columns. This data was analysed using python program via the Google Colab platform.

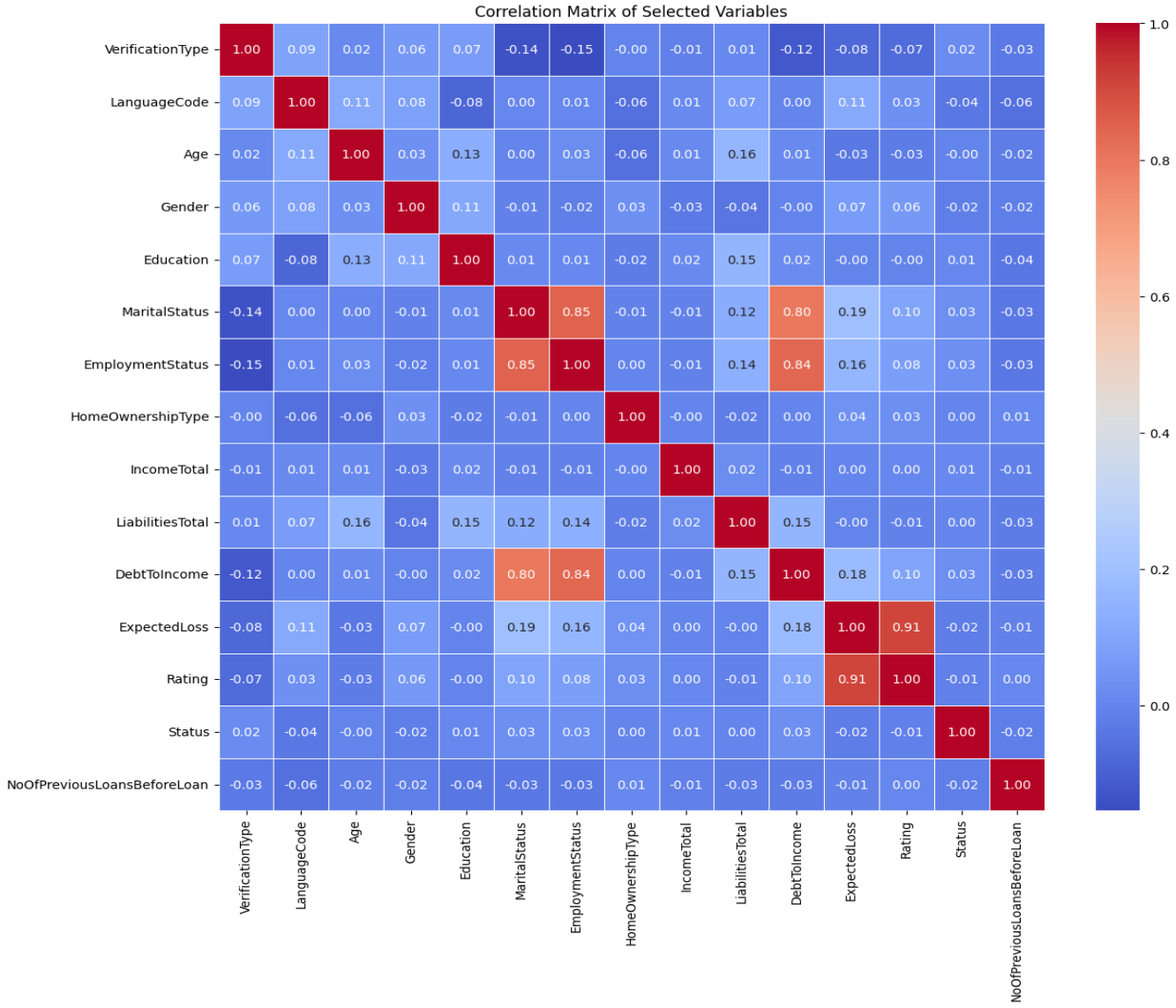
Feature Engineering

The chi-square test was used to ascertain the significant features in the Bondora loan data set having a strong correlation with the target variable (Status) using the chi-square test method which involves cleaning of the data and application of statistical tests to identify top variables based on their chi-square scores. The chi-square test is a tool used to examine the relationships between variables in a data set.

Feature		**Chi-Square Score**	**P value**
Principal Balance		14,190,890.00	0.0
PrincipalPayment made		7,076,571	0.0
Amount of previous loan		1899061	0.0
Interest and penalty		1,654,609	0.0
Previous repayment		197,292.5	0.0

As seen in the result above, variables including principal balance and principal payment made are strongly associated to the target variable. With a chi score of 14,190,890, P-value 0.0, and 7,7,076,571, P-value 0.0 respectively.
The chi-score, which is the statistical measure describes how important a variable is to the target variable (Status). Higher chi-score indicates a strong correlation to the target variable, while the P value helps to measure the statistical significance of the chi-square score. A p-value of O indicates the correlation of the ranked features to the target variable (Status). Other significant strongly associated with the target variable include amount of previous loans before the loan, planned interest post default, income total, principal, overdue by schedule, loan application started date and CreditscoreEemini.

Correlation Matrix



The correlation matrix above further highlights the relationship between variables in the loan data srt including marital status, employment and age. The correlation coefficients above exhibits the correlation between various variables. As seen above, the vaue of the correlation coefficient is between -1 to 1 where:
1- Indicates a positive linear correlation/ relationship.
-1 Indicates that there is a negative linear correlation between the variables.

0- indicates that there is no linear relationship/correlation between variables.
The relationship between the variables in the loan data set as shown in the correlation matrix above can further be explained below.
Marital Status and Employment: The correlation coefficient between marital status and employment in the matrix above shows 0.85 which can be further interpreted that there is positive correlation between marital status and employment status.
Debt to Income and Marital Status: The correlation above reads that there is a positive correlation (0.84) between Debt-to-income status and marital status of loan applicants. The debt to income and marital status of loan applicants are bound to affect the position of their loan application.
Expected loss and Rating: There is a correlation coefficient of 0.91 between the expected loss and rating of a loan applicant. If the rating of a loan applicant is very high. This means that the loan is expected to be repaid within the agreed time period. However, if the loan applicant has a very low credit score rating there is a high probability the loan be repaid on time and could end up as a bad loan. Credit rating on the Bondora platform is ranked from AA to HR where AA is the safest grade and HR is classified as the riskiest 'investment grade' rating.
The exhibition of a strong positive correlation between marital status and employment, Expected loss and rating, debt to income, and employment status amongst others can provide more predictive power for the target variable (Status) while variables that have a very weak correlation including language code and house ownership type, verification type and gender, have little or no significant linear correlation thereby making their impact on the target variable (Status) very weak.

7 Conclusion and Future Work

This research study focuses on crucial factors that influence the loan status of a borrower on a P2P lending platform. During this research, three credit scoring models were developed namely: XG Boost, Logistic regression, and Random forest. These predictive models were built with the sole aim of supporting investors' decision-making by answering the research question in this study and minimizing the risk of loan defaults on the P2P lending platform by using data from the Bondora P2P lending platform. The research study used loan applications between 2018 and 2023 on the Bondora lending platform which contains 15,821 variables including Income total, Education, gender, and principal balance. This study focused on loan status (Late and Repaid) which is the independent variable used in building the predictive model. During this study, it was further discovered that some variables have a positive linear correlation with each other and few others had a negative correlation. The variables with positive linear correlation include Debt to income and employment, expected loss and ratings, and marital status and employment, these variables have a high predictive power on the target variable while variables like language code and ownership have no impact on the target variable due to their weak correlation.

Predictive models used in this study ranked the variables in this data set according to their level of importance and how they influence the independent variable of this study (Status). Random Forest with an accuracy score of 56% indicates that variables including Number of previous loans, age, gender, education, number of previous loans, and expected loss influence the loan status (Late and Repaid) of borrowers on the Bondora lending platform while XG boost with an accuracy of 57% indicates that expected loss, gender, age, education and

number of previous loans influence the loan status of borrowers. However, the Logistic Regression model with the highest accuracy of 59% indicated that the probability of loan default variable in the data set is crucial when predicting credit risks on the Bondora peer-to-peer platform. The model also ranked other variables according to how they influence the loan status and they include expected loss, gender, age, education, verification type, use of loan and probability of default. As the model with the highest accuracy (59%). It can be concluded that investors must consider these factors when analyzing the credit risk and probability of default of a borrower as these variables provide more insights into the borrower's ability to repay the loan within or after the agreed time.

These findings will help investors make more informed business decisions regarding loan investment on the Bondora peer-to-peer lending platform. Investors on the Bondora peer-to-peer platform must consider a borrower's probability of default, marital status, verification type, use of loan and expected loss before making a lending decision.

For further studies, data can be collected from different years by utilizing both oversampling and undersampling technique to conclude on which model to use to support investment decisions on the Bondora lending platform before loan disbursement to customers thereby ensuring a high return on investment.

References

Alekseenko, A.P., 2022. Privacy, Data Protection, and Public Interest Considerations for Fintech. In *Global Perspectives in FinTech: Law, Finance and Technology* (pp. 25-49). Cham: Springer International Publishing.

Addy, W.A., Ugochukwu, C.E., Oyewole, A.T., Ofodile, O.C., Adeoye, O.B. and Okoye, C.C., 2024. Predictive analytics in credit risk management for banks: A comprehensive review. *GSC Advanced Research and Reviews*, *18*(2), pp.434-449.

Andriosopoulos, D., Doumpos, M., Pardalos, P.M. and Zopounidis, C., 2019. Computational approaches and data analytics in financial services: A literature review. *Journal of the Operational Research Society*, *70*(10), pp.1581-1599.

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

Cubiles-De-La-Vega, M.D., Blanco-Oliver, A., Pino-Mejías, R. and Lara-Rubio, J., 2013. Improving the management of microfinance institutions by using credit scoring models based on Statistical Learning techniques. *Expert systems with applications*, *40*(17), pp.6910-6917.

Duarte, J., Siegel, S. and Young, L., 2012. Trust and credit: The role of appearance in peer-to-peer lending. *The Review of Financial Studies*, *25*(8), pp.2455-2484.

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

Emekter, R., Tu, Y., Jirasakuldech, B. and Lu, M., 2015. Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. *Applied Economics*, *47*(1), pp.54-70.

Eboigbe, E.O., Farayola, O.A., Olatoye, F.O., Nnabugwu, O.C. and Daraojimba, C., 2023. Business intelligence transformation through AI and data analytics. *Engineering Science & Technology Journal*, *4*(5), pp.285-307.

Hassani, Z., Meybodi, M.A. and Hajihashemi, V., 2020. Credit risk assessment using learning algorithms for feature selection. Fuzzy Information and Engineering, 12(4), pp.529-544.

"Investor Bulletin: Accredited Investors," U.S. Securities and Exchange Commission, accessed May 28, 2024. Available at http://www.investor.gov/newsalerts/investor-bulletins/investor-bulletin-accredited-investors.
Igbinenikaro, E. and Adewusi, A.O., 2024. Financial law: policy frameworks for regulating innovations: ensuring consumer protection while fostering innovation. *Finance & Accounting Research Journal*, *6*(4), pp.515-530.
What is a logistic Regression available at https://www.ibm.com/topics/logistic-regression Accessed July 7 2024.
Jorge Newberry (2015): P2P Lending is not dead. Available at https://www.huffpost.com/entry/p2p-lending-is-not-dead_b_7028292 accessed 30 May 2024
Keval Bhanushali (2024): The evolution of peer-to-peer lending: A global journey. Available at https://1finance.co.in/blog/the-evolution-of-peer-to-peer-lending-a-global-journey/#:~:text=Origins%3A%20Pioneering%20the%20P2P%20Landscape&text=At%20a%20time%20when%20traditional,peers%20in%20need%20of%20financing. Accessed May 2024
King, M.A., Abrahams, A.S. and Ragsdale, C.T., 2015. Ensemble learning methods for pay-per-click campaign management. *Expert Systems with Applications*, *42*(10), pp.4818-4829.
A. Khashman, "Credit risk evaluation using neural networks: Emotional versus conventional models", *Applied Soft Computing*, vol. 11, no. 8, pp. 5477-5484, December 2011.
Klimowicz, A. and Spirzewski, K., 2021. Concept of peer-to-peer lending and application of machine learning in credit scoring. *Journal of Banking and Financial Economics*, (2 (16), pp.25-55.
Lenz, R., 2016. Peer-to-peer lending: Opportunities and risks. *European Journal of Risk Regulation*, *7*(4), pp.688-700.
Lin, X., Li, X. and Zheng, Z., 2017. Evaluating borrower's default risk in peer-to-peer lending: evidence from a lending platform in China. *Applied Economics*, *49*(35), pp.3538-3545.
Lee, T.S. and Chen, I.F., 2005. A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with applications*, *28*(4), pp.743-752.
Lera, I., Guerrero, C. and Juiz, C., 2019. YAFS: A simulator for IoT scenarios in fog computing. *IEEE Access*, *7*, pp.91745-91758.
Lin, M., Prabhala, N.R. and Viswanathan, S., 2013. Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending. *Management science*, *59*(1), pp.17-35.
Lin, X., Li, X. and Zheng, Z., 2017. Evaluating borrower's default risk in peer-to-peer lending: evidence from a lending platform in China. *Applied Economics*, *49*(35), pp.3538-3545.
Mild, A., Waitz, M. and Wöckl, J., 2015. How low can you go?—Overcoming the inability of lenders to set proper interest rates on unsecured peer-to-peer lending markets. *Journal of Business Research*, *68*(6), pp.1291-1305.
Nyberg, J., 2019. Exploring the factors affecting peer-to-peer lending performance with Self-Organizing Maps.
Pacces, A.M. and Heremans, D., 2012. Regulation of banking and financial markets. *Encyclopedia of Law and Economics: Regulation and Economics, 2nd Edition, AM Pacces and RJ Van den Bergh, eds., Elgar*.
Saimadhu Polamuri (2024) How the CART Alogrithm works. Available at https://dataaspirant.com/cart-algorithm/ accessed July 7 2024
Shaheen, S.K. and Elfakharany, E., 2018, October. Predictive analytics for loan default in banking sector using machine learning techniques. In *2018 28th International Conference on Computer Theory and Applications (ICCTA)* (pp. 66-71). IEEE.

Tsai, C.F., Hsu, Y.F. and Yen, D.C., 2014. A comparative study of classifier ensembles for bankruptcy prediction. *Applied Soft Computing*, *24*, pp.977-984.
Uwaoma, P.U., Eboigbe, E.O., Eyo-Udo, N.L., Daraojimba, D.O. and Kaggwa, S., 2023. Space commerce and its economic implications for the US: A review: Delving into the commercialization of space, its prospects, challenges, and potential impact on the US economy. *World Journal of Advanced Research and Reviews*, *20*(3), pp.952-965.
Okoye, C.C., Ofodile, O.C., Tula, S.T., Nifise, A.O.A., Falaiye, T., Ejairu, E. and Addy, W.A., 2024. Risk management in international supply chains: A review with USA and African Cases. *Magna Scientia Advanced Research and Reviews*, *10*(1), pp.256-264.
Ivanova, P., 2019. Cross-border regulation and fintech: are transnational cooperation agreements the right way to go?. *Uniform Law Review*, *24*(2), pp.367-395.
Volk, M., 2023. Do unrealised bank losses affect loan pricing?. *Available at SSRN 4504056*.
Wang, H., Yu, M. and Zhang, L., 2019. Seeing is important: the usefulness of video information in P2P. *Accounting & Finance*, *59*, pp.2073-2103.
Wu, C., Zhang, D. and Wang, Y., 2018. Evaluating the risk performance of online peer-to-peer lending platforms in China. *Journal of Risk Model Validation*, *12*(2), pp.63-87.
Wang, R., 2012. AdaBoost for feature selection, classification and its relation with SVM, a review. *Physics Procedia*, *25*, pp.800-807.
Wang, C., Chen, X., Jin, W. and Fan, X., 2022. Credit guarantee types for financing retailers through online peer-to-peer lending: Equilibrium and coordinating strategy. *European Journal of Operational Research*, *297*(1), pp.380-392.
Xu, P., Ding, Z. and Pan, M., 2018. A hybrid interpretable credit card users default prediction model based on RIPPER. *Concurrency and Computation: Practice and Experience*, *30*(23), p.e4445.
Zhang, K. and Chen, X., 2017. Herding in a P2P lending market: Rational inference OR irrational trust?. *Electronic Commerce Research and Applications*, *23*, pp.45-53.
Feng, G. and Buyya, R. (2016). Maximum revenue-oriented resource allocation in cloud, *IJGUC* 7(1): 12–21.
Gomes, D. G., Calheiros, R. N. and Tolosana-Calasanz, R. (2015). Introduction to the special issue on cloud computing: Recent developments and challenging issues, *Computers & Electrical Engineering* 42: 31–32.