

University of Bahrain

College of Science



**Department of
Postgraduate Programs**

**A Comparative Study of Machine Learning
Models for Default Groups Clustering and Credit
Scoring Prediction for Peer-to-Peer Lending**

**Thesis submitted in partial fulfilment of the requirements for the
degree of M.Sc. In Big Data Science and Analytics**

**Prepared by
Fatema Mohamed Almehaiza
ID No: 20080330**

**Supervised by
Prof. Nabil Mahmood Hewahi**
Professor
University of Bahrain

**Kingdom of Bahrain
September 2023**



Thesis Discussion Committee Decision

Student Name: Fatema Mohamed Almehaiza	Student ID: 20080330
Department: Choose an item.	College: Science
Program: M.Sc. in Big Data Science and Analytics	
Thesis Title: Comparative Study of Machine Learning Models for Default Groups Clustering and Credit Scoring Prediction for Peer-to-Peer Lending	
Supervisor Name: Prof. Nabil Hewahi	Student GPA: 3.45

The Discussion Committee met on 28-Sep-23 In S41-1002. Upon discussing the student whose details are stated above in public, the discussion committee recommends the following:

The student is awarded a grade Choose an item. in the thesis subject.

- ☐ Without making any changes to the thesis.
- ☒ After making the modifications attached with this decision.

Discussion Committee Members Approval:

Name

1. Prof. Nabil Hewahi

(Advisor)

2. Dr. Riadh Ksantini

(Internal Examiner)

3. **Prof. Boubakeur Boufarma**

(External Examiner)

Signature

Nabil Hewahi

P.J.

B.B.

Abstract

With the intervention of financial technology (FinTech), peer-to-peer (P2P) lending platforms have emerged as a disruptive force in the financial sector, directly connecting borrowers with lenders without the need for traditional intermediaries or collateral. P2P lending platforms heavily depend on the accuracy of creditworthiness predictions for borrowers through assessing the credit risk involved in lending. Machine Learning (ML) models provide the opportunity to improve credit scoring predictions and boost the overall efficiency and sustainability of P2P platforms.

This thesis presents a comprehensive comparative study of machine learning models for credit scoring prediction in the field of P2P lending. Six prominent ML algorithms are explored: Logistic Regression (LR), Random Forest (RF), Naive Bayes (NB), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), and a hybrid model that combines the strengths of multiple algorithms using P2P lending dataset from "LendingClub" for the period (2015-2018) accepted loan applications dataset that is publicly available. The study aims to assess the predictive performance of these models and identify the most suitable approach for credit scoring on P2P lending platforms. The research methodology includes exploratory analysis and data preprocessing techniques, including various balancing techniques. Several performance metrics were used, such as accuracy, precision, recall, and F1-score. The findings of the research show that the SVM provides the most accurate performance with F1-score of 99.1% outperforming individual models and hybrid model (combining LR+SVM+XGBoost). The result of this research is in line with previous studies.

Dedication

This thesis is dedicated to my parents, family and friends who have been always the biggest source of support and encouragement to always proceed my goals, especially my elder sister; Shaikha Almehaiza, who has always provided me with support and pushed me to exceed my expectations. In addition, I would like to express my gratitude to my managers for their continuous encouragement and guidance in accelerating my learning and growing in my career.

Thank you for supporting me during this journey.

Acknowledgements

I would like to extend a special acknowledgment to Bahrain Network (BNET), University of Bahrain, and Liverpool John Moores University for the opportunity to join the Big Data Science and Analytics master's Program.

Also, I would like to acknowledge my thesis supervisor Professor Nabil Mahmood Hewahi for his support and invaluable guidance on the thesis development.

Dr. Sawsan Hilal, Dr. Hesham Al-Ammal, Dr. Ahmed Zeki, and Dr. Abdulla AlQaddoumi for their efforts throughout the program.

Table of Contents

Defense Committee’s Approval	A
Abstract.....	B
Dedication.....	C
Acknowledgements.....	D
Table of Contents.....	E
List of Tables	H
List of Figures.....	J
List of Abbreviations	L
Chapter One Introduction	2
1.1. Overview.....	2
1.2. Research Problem	4
1.3. Research Questions.....	5
1.4. Research Objectives.....	5
1.5. Research Significance.....	6
1.6. Thesis Scope and Limitation	8
1.7. Thesis Structure	8
Chapter Two Background and Literature Review	10
2.1. Background.....	10
2.1.1. Introduction to Peer-to-Peer Lending.....	10
2.1.2. Credit Risk	12
2.1.3. Credit Scoring.....	13
2.2. Introduction to Machine Learning.....	14

2.2.1. Machine Learning for Credit Scoring.....	17
2.3. Literature Review	18
2.3.1. Studies used Different P2P Lending Datasets	18
2.3.2. Studies using LendingClub Dataset.....	21
Chapter Three Methods and Procedures	30
3.1. Research Methodology	30
3.2. Data Collection and Procedures	33
3.2.1. Data Collection	34
3.2.1.1. Data Pre-Processing.....	36
3.3. Research Models and performance evaluation metrics	40
3.3.1. Logistic Regression	40
3.3.2. Random Forest.....	41
3.3.3. Naïve Bayes	44
3.3.4. Support Vector Machine.....	44
3.3.5. Extreme Gradient Boosting	45
3.3.6. Ensemble Method (Hybrid Model)	48
3.3.7. Performance Evaluation	49
3.4. Research Tool	53
Chapter Four Results and Discussion.....	55
4.1. Experiment.....	55
4.1.1. Data Preparation	55
4.1.1.1. Exploratory Analysis	55
4.1.1.2. Data Pre-Processing.....	62
4.2. Results.....	66

4.2.1. Logistic Regression	67
4.2.2. Random Forest.....	70
4.2.3. Naïve Bayes	73
4.2.4. Support Vector Machine.....	76
4.2.5. Extreme Gradient Boosting Machine	79
4.2.6. Hybrid Model.....	82
4.3. Discussion.....	87
4.3.1. Comparison Analysis on Data Balancing Methods.....	87
4.3.2. Comparison Analysis on Computational Time	95
4.3.3. Performance Comparison Analysis	96
Chapter Five Conclusion and Future Work	101
5.1. Conclusion	101
5.2. Future Work.....	103
References.....	104
Appendices.....	111
الملخص	118

List of Tables

Table (1): Summary of Literature Review - Different P2P Lending Datasets	20
Table (2): Summary of Literature Review - LendingClub P2P Lending Datasets.....	25
Table (3): Confusion Matrix.....	49
Table (4): Loan Grade against Average Interest Rate	61
Table (5): Specificity, Precision, Recall, Accuracy, and F1-score result of LR-No Balancing.....	67
Table (6): Specificity, Precision, Recall, Accuracy, and F1-score result of LR- Various Balancing Methods.....	70
Table (7): Specificity, Precision, Recall, Accuracy, and F1-score result of RF-No Balancing.....	71
Table (8): Specificity, Precision, Recall, Accuracy, and F1-score result of RF- Various Balancing Methods.....	73
Table (9): Specificity, Precision, Recall, Accuracy, and F1-score result of NB-No Balancing	74
Table (10): Specificity, Precision, Recall, Accuracy, and F1-score result of NB- Various Balancing Methods.	76
Table (11): Specificity, Precision, Recall, Accuracy, and F1-score result of SVM-No Balancing	77
Table (12): Specificity, Precision, Recall, Accuracy, and F1-score result of SVM- Various Balancing Methods.	79
Table (13): Specificity, Precision, Recall, Accuracy, and F1-score result of XGBoost -No Balancing.....	80

Table (14): Specificity, Precision, Recall, Accuracy, and F1-score result of XGBoost- Various Balancing Methods.....	82
Table (15): Specificity, Precision, Recall, Accuracy, and F1-score result of Hybrid Model -No Balancing.....	84
Table (16): Specificity, Precision, Recall, Accuracy, and F1-score result of Hybrid Model- Various Balancing Methods.....	86
Table (17): F1-score result of Different Models - Various Balancing Methods.....	88
Table (18): Experiment Results Ranking.	89
Table (19): Methods' Computation Time.....	95

List of Figures

Fig. 1: Crisp-Dm Methodology (Wirth & Hipp, 2000).....	30
Fig. 2: Overview of the CRISP-DM tasks and their outputs.....	32
Fig. 5: Research Experiment's Design	34
Fig. 6: Number of Loans Issued by Year.....	35
Fig. 16: Research Dataset Pre-Processing.	36
Fig. 3: Random Forest.	43
Fig. 4: Common Ensemble Architecture.	48
Fig. 7: Dataset Distribution.....	56
Fig. 8: Grade against Loan Status.....	57
Fig. 9: Interest Rate against Loan Status	57
Fig. 10 Loan Amount against Loan Status	58
Fig. 11: Purpose of Loan	58
Fig. 12: Debt-to-Income ratio	59
Fig. 13: Open Credit Lines Analysis	59
Fig. 14: Employment Title against Loan Grade	60
Fig. 15: Interest Rate against Loan Grade	61
Fig. 17: Correlation Heatmap.	64
Fig. 18: Principal Component Analysis (PCA).	65
Fig. 19: Logistic Regression Confusion Matrix – No Balancing.	67
Fig. 20: Logistic Regression Confusion Matrix – Various Balancing Methods.....	69
Fig. 21: Random Forest Confusion Matrix – No Balancing.	71
Fig. 22: Random Forest Confusion Matrix – Various Balancing Methods.....	72
Fig. 23: Naïve Bayes Confusion Matrix – No Balancing.....	74
Fig. 24: Naïve Bayes Confusion Matrix – Various Balancing Methods.....	75

Fig. 25: Support Vector Machine Confusion Matrix – No Balancing.	77
Fig. 26: Support Vector Machine Confusion Matrix – Various Balancing Methods.	78
Fig. 27: XGBoost Confusion Matrix – No Balancing.	80
Fig. 28: XGBoost Confusion Matrix – Various Balancing Methods.	81
Fig. 29: Hybrid Model Confusion Matrix – No Balancing.	83
Fig. 30: Hybrid Model Confusion Matrix – Various Balancing Methods.	85
Fig. 31: Logistic Regression Performance Comparative Balancing Methods.	91
Fig. 32: Random Forest Performance Comparative Balancing Methods.	91
Fig. 33: Naïve Bayes Performance Comparative Balancing Methods.	92
Fig. 34: Support Vector Machine Performance Comparative Balancing Methods.	93
Fig. 35: XGBoost Performance Comparative Balancing Methods.	93
Fig. 36: Hybrid Model Performance Comparative Balancing Methods.	94
Fig. 37: Performance Comparative - F1-score Various Balancing Methods.	94
Fig. 38: An Overview Comparison on Results including Results of Previous Studies.	98
Fig. 39: Experiment Results Summary.	98

List of Abbreviations

Abbreviation	Explanation
P2P	Peer to Peer
SMEs	Small and medium enterprises
ML	Machine Learning
FICO	Fair Isaac Corporation
LGD	Loss given Default
AI	Artificial Intelligence
DT	Decision Tree
RF	Random Forest
NB	Naïve Bayes
GB	Gradient Boosting
SVM	Support Vector Machines
CART	Classification And Regression Trees
LR	Logistic Regression
XGBoost	Extreme Gradient Boosting
GBM	Gradient Boosting Machines
BPNN	Back Propagation Neural Network
CRISP-DM	Cross-Industry Standard Process for Data Mining
MV	Majority Voting
DTI	Debt to Income
PCA	Principal component analysis
SMOTE	Synthetic Minority Oversampling Technique
RUS	Random Under-sampling
ROS	Random Over-sampling

Chapter One

Introduction

1.1. Overview

1.2. Research Problem

1.3. Research Questions

1.4. Research Objectives

1.5. Research Significance

1.6. Thesis Scope and Limitation

1.7. Thesis Structure

Chapter One

Introduction

1.1. Overview

With the wild spread of innovation technological advances in financial sector, FinTech has been introduced; the integration of technology and finance that improve the financial services. One of FinTech invention is crowdfunding also called as “Peer to Peer (P2P) lending”. The peer-to-peer lending is an online platform that connects the lenders and borrowers (business or individual) with a corresponding interest rate that determined based on borrowers’ information without face-to-face interaction or financial institutions intervention (CFI Education Inc., 2023)

P2P lending has rapidly grown worldwide as it has challenged the traditional lending’s implication and long approval cycle. According to Market Research (The Business Research Company , 2022), P2P lending market was valued at \$106.09 billion in 2022 and is expected to reach \$302.49 billion by 2026.

P2P lending has gained popularity for many reasons, one of which is that both parties -lenders and borrowers- benefits from this transactions, where with lenders receive higher return and borrowers receive their loan quicker and more easily than the traditional way (Ko, Lin, Do, & Huang, 2022). In addition, due to the bank’s cost associated with assessing borrowers’ risk of personal loans or small and medium enterprises (SMEs), resulting in inefficient process, rejections, and high interest rates. P2P lending has addressed these issues for individuals and SMEs; especially the ones struggle to obtain loans, offering them with access to capital with

competitive interest rates through more efficient process compared to banks (Saputra, Faturohman, & Kaderi Wiryono, 2021). P2P lending also balances and decreases the investors risk through distributing their investments to different borrowers (Ko, Lin, Do, & Huang, 2022).

On the other hand, P2P lending is vulnerable to high credit risk or default risk; the potential loss of investors due to borrowers not able to meet their commitment or deadlines to repay their loan (Tao & Chang, 2019), due to information asymmetry and lack of adequate risk management. Furthermore, the limitation of regulators guidelines or legal framework to the P2P lending as it is a new disruptive business model and it will require acceptance and time to develop new regulations that reduce the credit risk (Lenz, 2016). To bridge this gap, P2P lending platforms must implement a credit risk management system; manage uncertainties through conducting risk assessment, which can be achieved through credit scoring to determine whether to accept or reject the loan application (Saputra, Faturohman, & Kaderi Wiryono, 2021). Consequently, P2P lending platform deploy software to perform the credit assessment through big data analytics to assess large dataset of borrowers' data, including economic data, personal information, and other information obtained from various internet sources, in order to estimate loan amount, interest rates, and probability that borrowers will default (Lenz, 2016).

As a results, many researchers have developed and applied different Machine Learning (ML) models and algorithms to analyse P2P lending data. Aiming at producing the optimal model for credit scoring models with high accuracy to predict the loan application will be good or bad credit, non-performing loan. It is worth mentioning, based on different studies, comparing the average of non-performing loans of traditional banks to P2P platforms the figures are slightly higher on P2P platforms. Which means that with further enhancement in the ML algorithms the P2P platforms

performance might outperform the expert-knowledge assessment (Lenz, 2016).

1.2. Research Problem

The P2P lending sector has significantly transformed the way individuals obtain loans funding and invest their financial resources. However, P2P lending platforms are vulnerable to high credit risk of defaults, which exposed investors to high losses due to borrowers' failure to repay their loans. Further, it is a significant challenge for P2P platforms to accurately assess borrowers' credit risk in an effective and efficient manner in order to determine whether to accept or reject the loan application. In addition, due to its nature, P2P lending suffers from lack of collaterals and credit histories of borrowers, which lead to increase in uncertainty in predicting borrowers' behaviour.

ML models have been employed to conduct credit risk assessment and credit scoring based on historical data to enhance the accuracy of predicting the borrowers' status, default or not. Noting that the prediction is very critical and must be done with high accuracy, because any error would result in significant investment losses. Additionally P2P lending datasets often suffer the issue of class imbalance, wherein the number of default cases is significantly lower than non-default cases. Therefore, it is critical to develop models that can handle this issue and effectively predict defaults is essential.

This research aims to measure the ability for ML models to be effectively leveraged to enhance credit scoring prediction for P2P lending platforms, considering the unique characteristics of P2P lending data and the need for accurate risk assessment. By comparing different ML models

and a hybrid model to identify the ultimate model with the best performance.

By constructing a highly accurate model that aims to reduce default rates and maximize investor return, P2P lending platforms can enhance their long-term viability and facilitate more informed loan acceptance or rejection decisions.

1.3. Research Questions

The study main question is what is the best credit scoring algorithm that will ensure high accuracy and reduce errors. There are already several comparison studies of classification models, however, it still suffers from limitation of data origin, or handling data problems. Hence, this research focus on analysing credit scoring prediction in P2P lending aiming at answering the following research questions:

- What is the performance of traditional ML models compared to each other's and what is the optimal ML model for credit scoring prediction for P2P lending?
- Can a hybrid model that combines the strengths of multiple ML algorithms outperform individual models in terms of credit scoring prediction accuracy for P2P lending?
- Can a data pre-processing techniques and sample modifying techniques significantly improve classification performance?

1.4. Research Objectives

The objective of this study is to:

- To understand the factors of default, and non-default groups of P2P lending

- To compare different machine learning models performance on credit scoring prediction for P2P lending
- To compare different balancing techniques performance on credit scoring prediction for P2P lending
- To develop a hybrid model that leverages the strengths of multiple algorithms to enhance credit scoring prediction accuracy.
- To evaluate the performance of different machine learning algorithms to build a model that predict the credit scoring.

1.5. Research Significance

The prediction of credit score holds significant importance for P2P lending platforms since the primary objective of these platforms is to make informed decisions regarding the approval of loans to borrowers with a high probability of repayment.

The primary source of revenue for the platform is contingent upon the growth in the number of investors, and the volume of loans disbursed, which is mainly dependent on the investor confidence. Investor confidence relay on the assurance that their investments and funds will be allocated to borrowers with a lower risk of default.

Accurate credit scoring algorithms allow investors a comprehensive understanding of the risk entailed in their investments. When investors hold confidence in the platform's ability to appropriately evaluate the creditworthiness of borrowers, their probability of engaging in participation increases, thus resulting in enhanced availability of funding for borrowers.

Therefore, the accurate classification prediction of loan applications holds major significance, as even marginal improvements in the predictive results may contribute to the development of a strong business framework,

reduce loan default rates, and minimize losses for the platform, while increasing returns for investors.

Despite the existence of numerous studies on credit scoring for P2P lending platforms, the accuracy rate, to the best of the researcher's knowledge, is not at its highest. Furthermore, a considerable number of research have employed various machine learning (ML) models. However, only a limited number of studies have suggested the use of hybrid models to improve performance. Noting that the previously conducted hybrid models are a combination of Deep Learning (DL) and ML models. Consequently, is a noticeable gap in the application of a combination of ML algorithms for accurately determining the appropriate default category of a loan applicant. In addition, it is worth noting that despite the widespread utilization of traditional machine learning (ML) models such as Decision tree, Logistic Regression, Linear Discriminant Analysis, Support Vector Machine, Naïve Bayes, Random Forest, Extreme Gradient Boosting, and others in several studies, the research findings are contradictory.

Therefore, the main objective of this study is to evaluate and compare the effectiveness of several algorithms and a hybrid model in classifying data, specifically using a publicly accessible dataset on P2P lending. The aim is to achieve a significant impact to improve the efficiency of loan application performance, consequently facilitating the sustainable growth of peer-to-peer (P2P) lending. The understanding that credit scoring models play a critical role in evaluating the probability of borrower default is widely recognized. Through the process of refining models, P2P platforms can enhance their ability to accurately identify borrowers with a high risk of default. The outcomes of this study have the potential to advance the understanding of credit scoring methodologies and enhance the operational efficiency of P2P lending platform.

1.6. Thesis Scope and Limitation

This thesis aims to provide a comparative study of selected five popular ML classifiers; LR, RF, SVM, NB, XGBoost, and a hybrid model for credit scoring prediction for P2P lending platforms. The experiment will be executed using a publicly available loan dataset extracted from LendingClub for the period 2015-2018, particularly the accepted loan applications dataset. Specialized techniques will be used for data preprocessing, which are feature selection (including PCA), feature engineering, and dimensionality reduction, dataset partition, and various balancing techniques including oversampling (ROS), under sampling (RUS), and Synthetic Minority Over-sampling Technique (SMOTE). The ML models performance will be measure using performance metrics, such as accuracy, precision, recall, and F1-score.

1.7. Thesis Structure

The flow of this study is organized as follows: Chapter two describes P2P lending mechanism and credit scoring concept, and literature review. Chapter three discusses the methods and procedures including the study's methodology, study's tools, and description of the dataset, collection, and procedures. Chapter four include the results and discussion, and conclude with chapter five with the study's conclusion, and future work.

Chapter Two

Background and Literature Review

2.1. Background

2.2 Introduction to Machine Learning

2.3. Literature Review

Chapter Two

Background and Literature Review

2.1. Background

This chapter provides an overview on the study's field, introduction to ML algorithms with a particular interest in credit scoring, and literature review discussing previous studies conducted in the same field.

2.1.1. Introduction to Peer-to-Peer Lending

Peer-to-Peer Lending is a method of borrowing and lending money through an online platform that connects individuals or businesses seeking loans with investors willing to lend money. P2P lending has different types; for individual personal loans, study loans, mortgage, etc. or for SMEs to finance their businesses (Klaft, Online Peer-to-Peer Lending: A Lenders' Perspective).

P2P lending work through straightforward operating model where a potential borrower fills an online application that includes the loan amount, the purpose of the loan, and other relevant information on the selected lending platform. Then the platform assesses the application to identify the risk level, set the credit score and appropriate interest rate for the applicant. Once the applicant agrees on the platform's assessment; including the credit score and assigned interest rate, the platform publish the offer with defined period for the investors to proposes different loan options for the applicant. Once the total investments' offers matches the required loan amount, the applicant received the amount and investors received a credit claim which documents the borrower's legal commitment to repay the loan with interest rate over set period (usually monthly basis) (CFI, 2022).

The platform will be responsible of managing the entire process including credit scoring, loan payment, collections, and interest distribution. In addition, in case of defaults the platform act as a debt collector on behalf of the lenders (Lenz, 2016).

Starting from 2005 ‘Zopa’ was launched as the first lending platform in United Kingdom (Europe), followed by many other platforms that have been created year over year all around the world. Some of the biggest platforms were Lending Club, Prosper, Axiafunder, mintos, Debitum, etc (P2P Market Data, 2023). The P2P lending business has experienced significant growth and global expansion since its inception. Currently, there are more than 251 online P2P lending platforms registered with P2PMarketData, operating in various countries worldwide. (P2P Market Data).

One of the largest P2P lending platforms in the world is Lending-Club, helping more than 4 million people with over 15 billion dollar originated loans. The platform was launched in 2007 in the United States, providing unsecured personal loans and business loans ranging between \$1,000 up to \$40,000 with terms averaging three years (LendingClub, 2023). Lending Club generates revenue by charging service fees to both borrowers and lenders. Nevertheless, Lending-Club has made its data available to the public, sparking a flurry of scientific interest in the model's analysis. The data used for this study will be the Lending-Club’s data.

The platform uses a proprietary algorithm to assess the creditworthiness of the borrowers, with taking into consideration the borrower’s general information, employment background, credit score, income and debt details and other factors. Borrowers with higher credit scores they are likely to get lower interest rates on the requested loan (LendingClub, 2023).

Nevertheless, the responsibility for completing the application form on P2P lending platforms rests with the users. This situation presents a significant obstacle in achieving information symmetry between borrowers and investors. Certain individuals had a propensity to conceal accurate information in order to secure loans, or utilized deceptive information to obtain loans, resulting in investments being made for borrowers with a higher probability of default and subsequently amplifying the magnitude of losses incurred. Hence, the evaluation of credit risk assumes a significant role in upholding the stability of financial systems (Klafft, 2008).

2.1.2. Credit Risk

Credit risk has imposed a limitation in the financial sector especially in the lending activities, it considered the most significant risk that is difficult to manage. Credit risk can be defined as; ‘the potential that a contractual party will fail to meet its obligations in accordance with the agreed terms’ (Brown & Moles, 2016). A financial loss could be sustained by investors (creditors) as a result of the possibility that a borrower who received a loan would not be able to fulfil their duty to repay the money they were given on time. This increases the risk of a borrower defaulting due to failing to meet their obligation (Spuchl’áková, Valášková, & Adamko, 2015). Defaults occur for many reasons, some of the common reasons are borrowers in a weak financial state, facing a bankruptcy, a heavy debt burden, difficulties in maintaining steady source of income, or deficiency in a loan agreement causing a failure to uphold a borrower on the loan terms (Gestel & Baesens, 2009). As a result, the higher the credit risk of the borrower lead to higher interest rate.

The number of defaulted loans rises as a result of increased lending activities and credit loans. Financial institutions and lending platforms

must therefore make a clear distinction between "good" and "bad" applicants before granted the loan (Brown & Moles, 2016).

As a result, credit risk evaluation is key to manage default risk and evaluate the borrower application to differentiate between good and bad application before the loan request. Credit risk evaluation includes gathering, analysing, and evaluating various variables in order to take the decision to accept the loan application or not. It is the process of classifying applicant patterns and predicting if the applicant is likely to default using various analytical techniques (Leo, Sharma, & Maddulety, 2019).

There are numerous methods and tools to assess the credit risk and borrower's creditworthiness, including credit score, financial history, employment information, financial statements, and other related information. A collateral is frequently used to reduce credit risk.

2.1.3. Credit Scoring

Credit scoring is a popular technique in the financial sector and was one of the first credit risk management techniques developed (Leung, Cheong, & Cheong, 2008). Credit scoring is the process of evaluating the creditworthiness of a borrower through examining the borrowers' credit history and other financial information to assigning a numerical score that predicts the probability of the borrower repaying their debts on time. The model classifies the applicants into two groups; a "good credit" group who are more likely to repay their debt, and "bad credit" group who are more likely to default (West, 2000).

The credit scoring models take into account number of factors can be classified in four main groups: financial indicators, demographic indicators, employment indicators, and behavioural indicators (Mpofu & Mukosera, 2014). Number of parameters will be identified under each

group and each of these parameters will be given a certain weight in the model and the total score is calculated based on these weights.

The use of credit scoring has become extremely important in the financial sector and lending transactions, as it has developed a way for lenders to evaluate the lending's risk level of borrower quickly and objectively. It offers several advantages that have improved credit risk management. To start with, credit scoring has increased the objectivity of the loan approval process by ensuring that all applicants are evaluated according to the same standards, irrespective of race, gender, or any other legally banned traits (Mester). In addition, credit scoring is extremely scalable with huge as it uses a statistical methodology that can process large number of applicants' attributes simultaneously on large volume of applications (Ince & Aktan). Furthermore, the automation of the process allows quicker and more consistent loan evaluation, lower cost, and reduced risk associated with bad debt, which results in better decisions for accepting or rejecting the applications with less time and efforts expended (Mester, 1997). Last, credit scoring facilitates in the entire lending process by helping to determine the amount of the loan and an appropriate interest rate for the loan applicant (Ince & Aktan).

Hence credit scores are used by P2P lending platforms to determine whether to approve the borrower's loan request or not and define the interest rate (Leung, Cheong, & Cheong, 2008).

2.2. Introduction to Machine Learning

Machine Learning (ML) concept was introduced in 1959 by Arthur Samuel to enable machines the ability to continuously learn, adapt, and improve their performance without any human involvement. ML is a set of

algorithms that are designed to analyse and automate various complex problems of pattern recognition in massive databases (Zhang, 2010). ML algorithms operate dynamically and automatically adapt to changes in the data from experience using statistics and mathematical optimization to find meaningful insights and make predictions based on that learning (Guegan & Hassani, 2018).

ML algorithms can be classified in three types which are supervised learning which aims to predict the outcomes using the giving set of variables, unsupervised learning which focus on detecting and recognizing relationships, and patterns within data, and reinforcement learning that includes positive or negative enforced learning.

1. Supervised Learning: one of the most commonly used ML techniques. The technique entails training the model on a labeled data in order to predict new or unknown data. Supervised learning depends on a dataset with a set of features/variables and corresponding targets values (training data) to learn a specific function in order to accurately predict the target value for the new data (Russell & Norvig). The aim of supervised learning is to build a predictive model that can make good generalization to new unseen data. The algorithm works to identify patterns in the data to be used to provide accurate predictions. The algorithm also iteratively modifies its parameters to reduce the discrepancy between its predictions and the actual values in the training dataset.

There are two types of supervised learning: classification and regression. Classification is used to predict a categorical output variable, such as binary label of multi class label, while regression used to predict a continuous output variable, such as numeric value (Russell & Norvig).

2. Unsupervised Learning: this ML techniques builds a model that is trained with unlabelled dataset without specific output. Instead, the algorithm is developed to identify patterns and relationships in the data independently, with no specific goal. In this kind of ML technique, the algorithm simply uses the input data only to identify patterns and structures among the dataset's features. In particular, the algorithm can find clusters with similar data points, find patters that are connected with various variables, or detect outliers or inconsistencies within the dataset. There are many types of unsupervised learning, some of which include clustering, dimensionality reduction, generative models, outliers' detection, etc. The most popular method is clustering, where the algorithm discovers clusters or groups with related data points in the dataset (Russell & Norvig).

3. Reinforcement Learning: an approach for dealing with decision-making in dynamic and unpredictable environment. The learning occurs when an agent learns how to behave in a given scenarios based on the feedback received. The agent attempts to maximize or minimize the feedback that comes in a form of rewards or penalties. This algorithm incorporates an agent learning in an interactive environment based on trial-and-error. The aim of this learning is to enable the agent to learn policies that increase rewards on the long-term. This tool is used to build decision-making intelligent agents that are utilized for developing robotics, recommendation engines and gaming (Russell & Norvig).

Supervised learning is the most applicable techniques used for credit scoring models, where classification is used to predict loan status whether to “accept” or “reject” the loan application.

2.2.1. Machine Learning for Credit Scoring

Since the 2000s up until the present, ML has gained popularity as one of the strong tools for credit scoring. The ML algorithms are used to analyse large amount of data to defined patterns that can be utilized to predict a borrower’s creditworthiness. In addition, ML has outperformed the traditional financial methos, especially with the advance statistical techniques, ML models may spot correlations and trends in data that human analysts may not notice immediately. Furthermore, traditional financial approaches usually base a credit score on limited set of variables such as payment history and credit background. However, ML algorithms can analyse a considerably wider range of data, including social media and online purchasing patterns. Additionally, ML models have the ability to learn and adapt over time allows to modify the weighting of various variables to improve its predictive accuracy which will ultimately reduce the risk of default (Kumar, Sharma, & Mahd, 2021). ML models have become a significant instrument in the domain of P2P lending for credit scoring due to its advantageous characteristics, enabling the provision of credit scores that are more modified, adaptable, and accurate. P2P lending platforms possess the capability to proficiently discern debtors with a high-risk profile, hence enabling them to either decline loan applications or establish elevated interest rates.

Various algorithms, such as statistical, machine learning, artificial intelligence, and other models, are currently being employed in the development of credit scoring systems. In earlier times, credit scoring was

formulated through the utilization of traditional statistical models such as logistic regression and linear regression. These models were favoured for their straightforwardness and user-friendly nature. Nevertheless, due to the ongoing progress in credit scoring models and the pursuit of attaining optimal precision, wherein even a minor enhancement in accuracy yields substantial profits, increasingly intricate machine learning (ML) and artificial intelligence (AI) models are being employed for constructing credit scoring models. Several models commonly used in machine learning include Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB), Gradient Boosting (GB), Support Vector Machines (SVM), among others. Numerous studies have been undertaken to conduct comparative analysis in order to evaluate the performance of various models and ascertain the superior credit scoring model.

2.3. Literature Review

This section is divided into two parts. The first part is a review focused on previous studies on ML algorithms for credit scoring using different P2P lending datasets. While the second part is a review of previous studies that used LendingClub data.

2.3.1. Studies used Different P2P Lending Datasets

Credit risk management and credit scoring of P2P lending platform have become a hot research topic in the finance sector's ML applications and Artificial Intelligence (AI). Many studies have been conducted comparing classification algorithms for credit scoring on different data. Meanwhile, many P2P lending platforms have made their data available for researchers to improve the prediction accuracy of credit scoring models.

David West (West, 2000), has developed a study to investigate the accuracy of neural network (NN) models for credit scoring challenging the parametric and non-parametric models. The study involved five neural network models: multilayer perceptron, mixture-of-experts, radial basis function, learning vector quantization, and fuzzy adaptive resonance, two parametric models: linear discriminant analysis and logistic regression, using several non-parametric methods, such as K-Nearest Neighbor (KNN), kernel density, and decision trees (CART) experimented on two datasets; German credit scoring data of 700 instances and 24 attributes. As a results, logistic regression has outperformed the models with the lowest error rate.

Another study (He & Nawata, 2019) was conducted to predict the default risk of borrowers using ML models, which are logistic regression, random forest, neural network (multi-layer perceptron) and naïve bayes. The study was conducted on data from one of the largest P2P platforms in China, Renrendai with a dataset of 48,120 records. The researchers conclude by stating that all models have performed well with an accuracy above 86%, however naïve bayes has outperformed others by 97% accuracy.

In a subsequent study, (Padimi, Telu, & Ningombam, 2022) explored several ML algorithms to predict P2P lending loans default using data from European P2P Lending platform, Bondora. The dataset was extracted for the period 2009 with 220,906 records and 112 attributes. The study used logistic regression, decision tree, k-nearest neighbor, random forest, and naïve bayes to predict whether a loan would default or not. The simulation result shows that random forest accuracy outperformed the other supervised algorithms.

In Khatir et al. (2022) the researchers aimed to develop a comprehensive study to determine the optimal combination for credit

scoring including best ML classifier, feature selection techniques, and sampling methods – oversampling. Comparison analyses were conducted on several ML models: logistic regression, decision tree, k-nearest neighbor, neural networks, random forest, and gaussian naïve bayes. Similar results were observed in the (Padimi, Telu, & Ningombam, 2022) study, which concluded that random forest was the best model.

Table (1): Summary of Literature Review - Different P2P Lending Datasets

Author	Dataset	Models	Accuracy	Technique
David West (West, 2000)	German credit scoring data	Multi-layer perceptron	73.28%	-
		mixture-of-experts	75.66%	
		radial basis function	74.60%	
		Learning vector quantization	68.37%	
		fuzzy adaptive resonance	57.23%	
		Linear Discriminant Analysis	72.60%	
		Logistic Regression	76.30%	
		K Nearest Neighbour	67.60%	
		kernel density	69.20%	
		Classification & Regression Tree	69.56%	
Huixi Hea & Kazumitsu Nawata (He & Nawata, 2019)	Renrendai - China	Logistic Regression	86.40%	-
		Random Forest	95.10%	
		Multi-layer perceptron	86.74%	
		Naïve Bayes	96.74%	
Padimi et al. (Padimi, Telu, & Ningombam, 2022)	Bondora - European	Logistic Regression	92.73%	Under-sampling
		Naïve Bayes	89.33%	
		Random Forest	94.60%	
		K Nearest Neighbour	90.51%	
		Decision tree	94.29%	
Khatir et al. (Khatir, Almustfa, & Bee, 2022)	German retail credit dataset	Random Forest	85.40%	Oversampling
		Decision tree	81.10%	
		Gaussian Naive Bayes	69.10%	
		K Nearest Neighbour	68.00%	
		Neural Network	76.30%	
		Logistic Regression	68.86%	

Table 1 summaries the literature review of studies conducted on different P2P lending data, other than LendingClub. Overall, traditional ML models such as RF, NB, and LR have proven its effectively in the credit

scoring prediction field. Nevertheless, DL has demonstrated significant efficacy in a range of fields, including image identification, natural language processing, and speech recognition. However, numerous studies have demonstrated that its performance in credit scoring prediction is low.

2.3.2. Studies using LendingClub Dataset

This part will explore studies that used LendingClub data to conduct research on comparing different ML algorithms for credit scoring. Many studies have focused on comparing modern ML models for credit scoring for P2P lending platforms in comparison with logistic regression.

To start with, Teply and Polena (2020) explored to find the best classification model for P2P lending to build a credit scoring. Comparing ten different classification algorithms in their study including linear algorithms; logistic regression and linear discriminant analysis, number of non-linear algorithms, support vector machine, artificial neural network, k-nearest neighbor, naïve bayes, bayesian network, and few rule-based classifiers; classification and regression tree, random forest, and Linear support vector machine. The researchers have used big dataset of 212,252 records with 23 different variables. The study's findings shows that logistic regression outperformed other algorithms with an accuracy of 79%.

Jiaying Sun conducted another study in (2022) aimed to develop a predictive model for default based on LendingClub dataset of 201,350 cleaned datasets with 75 variables. In comparison with the logistic regression model, the researcher used Artificial Neural Networks (ANN). As a result, logistic regression has outperformed the ANN with superior discriminating abilities for predicting credit defaults. The LR has achieved an accuracy of 75%, whereas the ANN has achieved an accuracy of 66%.

To further challenged the random forest model when it comes to predict credit scoring, Yijie Fu (2017) proposed a new method of combining random forests and neural networks to predict the borrowers' classification, default or not. The researcher experiment includes comparing linear function, single random forest, single neural network, combination of random forest and neural network, and combination opposite neural network then random forest on dataset of 132,000 instances. The hybrid model, which combines neural network and random forest techniques, has greater accuracy performance compared to other methods, with an accuracy rate of 73.5%. In contrast, the single random forest model achieves the second greatest accuracy, albeit slightly lower at 73.3%.

Chen et al. (2021) employed different ML models including random forest, neural networks, and logistic regression to predict the default risk of P2P lending. The research dataset size includes 269,668 records. The researchers have introduced the issue of imbalanced data and applied sampling techniques to resolve it. Random Under-Sampling (RUS) technique was utilized. As a result, the random forest shows best performance among the three models.

A recent research study was conducted on the P2P lending platform, with the aim of predicting loan default using LendingClub dataset of 115,000 records and 102 attributes (Zhu, Qiu, Ergu, Ying, & Liu, 2019). The study conducted a comparison of several machine learning models, including decision trees (DT), logistic regression (LR), support vector machines (SVM), and random forests (RF). Moreover, the researchers have employed a range of techniques for the pre-processing of data, such as the utilization of balancing techniques including oversampling, data cleansing, and dimension reduction. The random forest methodology exhibits

superior performance in comparison to the other three algorithms employed in model construction, with an accuracy rate of 98%.

Another study by Yiheng Li and Weidong Chen under the titled “A Comparative Performance Assessment of Ensemble Learning for Credit Scoring” was conducted in (2020). The study compares the performance assessment of ensemble algorithms, including random forest, AdaBoost, XGBoost, LightGBM, and stacking, in the construction of a credit scoring model utilizing traditional ML models. Empirical evidence demonstrates that ensemble learning often outperforms individual learners. Moreover, it can be observed that the random forest algorithm exhibits superior performance across five measures with an accuracy of 81%, while XGBoost and LightGBM algorithms emerge as strong contenders in terms of performance (Li & Chen, 2020).

Nevertheless, Yiping Guo’s research suggested that neural network-based algorithms can outperform other traditional algorithms due to its ability to generalize and achieve higher accuracy. Accordingly, the researcher has conducted comprehensive experiment to compare between BP neural network and logistic regression for P2P lending platforms’ credit risk assessment. He discovered that the BP neural network-based algorithm performed better with an accuracy of 93.3% (Guo, 2019).

Next study conducted by Qian and Hu (2019) applied different ML algorithms for prediction of the default risk on P2P lending platforms. The researchers have conducted an empirical study on predicting the default risk using feedforward neural networks, logistic regression, and least squares support vector algorithms utilizing data from Lending Club, some of these algorithms were not commonly used in prior studies. The study demonstrates that the least squares support vector machine was more accurate compared to the other two methods achieving an accuracy of 92%.

Byanjankar et al. (2021) used LendingClub, 2015 data to predict the default probabilities of P2P loans as researchers seek to construct more accurate credit scoring models. The researchers experimented logistic regression, random forest, and XGBoost algorithms to identify the optimal model for the predictions. As a consequence, the proposed method, XGBoost has offer better performance than the other models achieving an accuracy rate of 74.2%.

In a recent study (Ko, Lin, Do, & Huang, 2022), researchers explored deeper P2P lending default prediction using statistical models and AI with the aim of improving credit scoring model accuracy. They employed a variety of ML models, including logistic regression, linear discriminant analysis, and bayesian classifier, as well as AI models including random forest, artificial neural network, decision tree, light gradient boosting Machine (LightGBM), and Convolutional Neural Network (CNN). According to the empirical study LightGBM has the best performance with the highest accuracy of 68.6%.

Chang et al. (2022) proposed number of ML and artificial neural network algorithms to build credit scoring models for P2P lendingClub platform including, logistic regression, decision tree, support vector machine, random forest, XGBoost, light gradient boosting machine and 2-layer neural networks. The dataset was extracted includes 282,763 records and 151 attributes. The experiment revealed that XGBoost has the optimal model for P2P credit scoring with an accuracy of 88%.

Last but not least, a study conducted by Qionglin Shan and Mikael Nilsson (2018) examines credit risk analysis and evaluates the classification performances of various algorithms which are logistic regression, support vector machine, decision tree, multilayer perception, probabilistic neural network, and deep learning. The study utilizes a substantial dataset from P2P lending, consisting of one million

observations, however after data cleansing 86,285 instances were examined with 101 attributes. The researchers' findings indicate that Support Vector Machines (SVM) demonstrate the highest level of accuracy, achieving a performance rate of 97%. Furthermore, the study has successfully addressed the ongoing controversies regarding the superior performer between logistic regression (LR) and multilayer perceptron (MLP). The findings of the study indicate that LR has exhibited slightly greater performance compared to MLP, despite their identical levels of performance. Finally, the study also affirms that when comparing ML and DL models, DL had the worst performance relative to the other models (Shan & Nilsson, 2018).

Table (2): Summary of Literature Review - LendingClub P2P Lending Datasets

Author	Dataset	Models	Accuracy	Technique
Teply and Polena (2020)	2009-2013	Logistic Regression	79.13%	-
		Linear Discriminant Analysis	79.04%	
		Support Vector Machine	78.18%	
		Artificial Neural Network	79.05%	
		K Nearest Neighbour	76.55%	
		Naïve Bayes	78.36%	
		Bayesian Classifier	78.78%	
		Classification and regression tree	76.59%	
		Random Forest	78.83%	
		Linear support vector machine	78.87%	
Jiaying Sun (2022)	2007-2015	Artificial Neural Network	66.00%	-
		Logistic Regression	75.00%	
Yijie Fu (2017)	2015	Linear function	65.60%	-
		Random Forest	73.30%	
		Neural Network	67.80%	
		Random Forest + Neural Network	72.00%	
		Neural Network + Random Forest	73.50%	
Chen et al. (2021)	2007-2015	Random Forest	63.93%	Under-sampling
		Neural Network	63.56%	
		Logistic Regression	63.24%	
Zhua et al. (Zhu, Qiu, Ergu, Ying, & Liu, 2019)	2019	Random Forest	98.00%	SMOTE
		Decision tree	95.00%	
		Support Vector Machine	75.00%	
		Logistic Regression	73.00%	

Yiheng Li and Weidong Chen (2020)	2018	Random Forest	81.05%	-
		AdaBoost	78.61%	
		Extreme Gradient Boosting	79.47%	
		Light Gradient Boosting Machine	79.47%	
		Stacking	79.19%	
		Neural Network	79.14%	
		Decision tree	79.00%	
		Logistic Regression	79.15%	
		Naïve Bayes	68.28%	
		Support Vector Machine	79.19%	
Yiping Guo's (Guo, 2019)	2015-2019	BP neural network-based	93.30%	-
		Logistic Regression	89.36%	
Qian and Hu (2019)	-	Logistic Regression	65.90%	-
		Bat-Optimization Algorithm of Feedforward neural network	91.45%	
		least squares support vector machine	92.17%	
Byanjankar et al. (2021)	2015	Logistic Regression	73.74%	-
		Random Forest	73.25%	
		Extreme Gradient Boosting	74.23%	
Chang et al. (2022)	2013-2014	Logistic Regression	86.81%	Under-sampling
		Support Vector Machine	87.38%	
		Decision tree	87.23%	
		Random Forest	87.58%	
		Extreme Gradient Boosting	87.90%	
		Light Gradient Boosting Machine	87.58%	
		Neural Network	87.06%	
Qionglin Shan and Mikael Nilsson (Shan & Nilsson, 2018)	2015-2018	Logistic Regression	93.74%	-
		Support Vector Machine	97.00%	
		Decision tree	95.64%	
		Multi-layer perceptron	93.62%	
		Neural Network	92.96%	
		Deep Learning	90.07%	

The literature has illustrated different results for credit scoring model, Table 2 summaries all study's findings. For years, LR model has consistently maintained its position as the predominant and widely utilized statistical model within the financial sector, particularly for the lending activities. As a results of many comparative studies, logistic regression has emerged as the industry standard for credit scoring model and used as a

benchmark for other models. The superior performance of LR over other models has been demonstrated by West (2000), Teply and Polena (2020), and Jiaying Sun (2022), irrespective of the dataset, volume, or parameters.

On the other hand, it has been demonstrated in several studies that ensemble models exhibit superior performance compared to generalized linear models. Random forest has demonstrated higher performance compared to other models in various studies, including (Fu, 2017), (Chen, Leu, Huang, Wang, & Takada, 2021), (Padimi, Telu, & Ningombam, 2022), and (Khatir, Almustfa, & Bee, 2022) . As a result, it has gained popularity as a preferred model for multiple P2P lending systems.

Still, other ML models such as Naïve Bayes (He & Nawata, 2019), Support Vector Machine (Shan & Nilsson, 2018), and others have gained credibility and are being utilized more frequently in practical applications.

Furthermore, there has been a growing fascination among academics in recent years with the remarkable efficacy of gradient boosting models (in particular XGBoost) in enhancing the precision of credit scoring for the purpose of predicting loan defaults, as evidenced by the findings of the study conducted by Chang et al. (2022) study.

Even though Yijie Fu (2017), and Yiping Guo (2019), are proposing neural network-based algorithms for P2P lending platforms' credit risk assessment. Many of the studies shows that deep learning models have failed in competing with other ML models for credit scoring. In fact, based on Shan and Nilsson's (2018) study, deep learning has performed the worst among other techniques used, which were logistic regression, decision tree, support vector machine, probabilistic neural network, and multilayer perception.

This research significantly diverges from previous studies in credit scoring prediction from P2P lending platforms due to its distinct focus on exploring a wide array of ML models, including traditional algorithms,

advanced preprocessing techniques, a hybrid approach, and handling imbalanced data techniques.

To begin with, unlike many previous studies that were focused on comparing the performance of traditional ML models, this research proposes a hybrid model. This model combines the strengths of multiple algorithms that were chosen based on the top three performing models from the selected traditional ML models. The aim is to optimize their complementing characteristics of variables in order to potentially achieve improved predictive accuracy.

Moreover, this study applies a collection of comprehensive preprocessing techniques to improve the quality and relevance of the dataset. Through utilizing feature engineering and data transformation techniques tailored for P2P lending data, the research aims to improve the input quality, thus potentially leading to better-performing models.

Next, by recognizing the inherent imbalance in credit scoring datasets, this research differentiates itself by addressing this challenge. Various balancing techniques for managing imbalanced data will be used, such as oversampling, under sampling, and SMOTE, highlight this research's commitment to tackling real-world data issues.

In essence, by designing the hybrid model, preprocessing, and balancing techniques specifically for the intricacies of P2P lending data, this study aligns its methodologies with the context at hand, departing from more general approaches used in other studies.

Chapter Three

Methods and Procedures

3.1. Research Methodology

3.2. Research Tools

3.3. Data Collection and Procedures

Chapter Three

Methods and Procedures

3.1. Research Methodology

This chapter provides an overview of the methodology employed in this research. For this research, Cross-Industry Standard Process for data mining (CRISP-DM) was conducted. CRISP-DM provides a standard process framework model for data mining projects, yet it can be adapted for research planning. It provides a structured method to manage projects through the following 6 cyclical phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment Fig.1 (Wirth & Hipp, 2000). CRISP-DM can be executed in a non-strict manner, researchers can move between phases. Each phase is a new learning experience that reveals an essential information throughout the process, and each phase is corelated to the others.

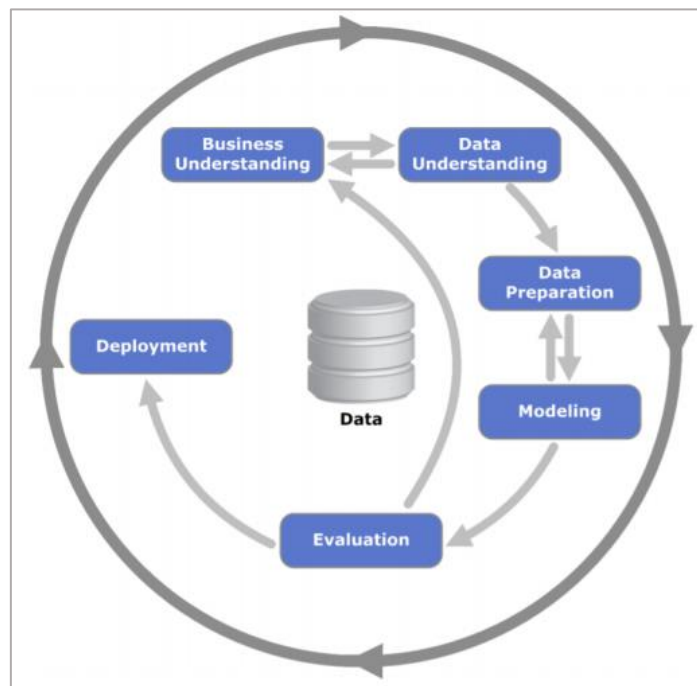


Fig. 1: Crisp-Dm Methodology (Wirth & Hipp, 2000)

The initial phase is business understanding, which entails understanding the research objectives and the problem that the researchers want to solve. In this phase, the researcher identified the research objectives and questions, and set the research expected outcomes. In addition, this phase includes conducting a comprehensive literature review analysis of previous studies that have been conducted on the same subject. This will allow to examine the work and experiments of other researchers. Next, the data understanding phase which focuses on understanding the data that will serve the research objectives. This phase consists of collecting the data from trusted sources, defining and examining the data, assessing the data and their relationships, and verifying data quality. In order to select the features for the chosen models, this phase identifies the key characteristics and explore initial patterns within the data. After that is data preparation phase, in which the researcher selects the dataset, cleans the data, then executes features engineering, and transform data format as necessary to fit the models and satisfy the research objectives. The fourth phase is modeling; employing the ML algorithms to solve the business problem and satisfy the research objectives. The modeling phase includes selecting the ML models that will be used for the experiment, design the model's test environment by splitting the data into training set and testing set, fit the models using the selected dataset that was prepared previously, and last assessing the models' performance through set of success evaluation metrics and identifying the best performer for the research objective. Last but not least, evaluation phase that includes reviewing the research process and evaluating the results. In order to determine whether to deploy or modify the data or models. Last phase is deployment, releases the model for use in real environment. This phase involves documenting the procedure, monitoring the results and quality of the model, and

developing the final report, Fig. 2 demonstrates an overview of CRISP-DM tasks and outputs (Wirth & Hipp, 2000).

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives Background Business Objectives Business Success Criteria Assess Situation Inventory of Resources Requirements, Assumptions, and Constraints Risks and Contingencies Terminology Costs and Benefits Determine Data Mining Goals Data Mining Goals Data Mining Success Criteria Produce Project Plan Project Plan Initial Assessment of Tools and Techniques	Collect Initial Data Initial Data Collection Report Describe Data Data Description Report Explore Data Data Exploration Report Verify Data Quality Data Quality Report	<i>Data Set</i> Data Set Description Select Data Rationale for Inclusion / Exclusion Clean Data Data Cleaning Report Construct Data Derived Attributes Generated Records Integrate Data Merged Data Format Data Reformatted Data	Select Modeling Technique Modeling Technique Modeling Assumptions Generate Test Design Test Design Build Model Parameter Settings Models Model Description Assess Model Model Assessment Revised Parameter Settings	Evaluate Results Assessment of Data Mining Results w.r.t. Business Success Criteria Approved Models Review Process Review of Process Determine Next Steps List of Possible Actions Decision	Plan Deployment Deployment Plan Plan Monitoring and Maintenance Monitoring and Maintenance Plan Produce Final Report Final Report Final Presentation Review Project Experience Documentation

Fig. 2: Overview of the CRISP-DM tasks and their outputs.

To achieve the research objective of predicting P2P lending credit scoring defaults through ML models, a number of models were selected for several reasons. After conducting the literature review on previous studies (illustrated in chapter two), the selection of the research methodology was on the most popular and preferred ML algorithms found in previous studies for the same research topic. To begin with, Many researchers identified logistic regression as the state-of-art benchmark for credit scoring models, as demonstrated by West (2000), Teply and Polena (2020), and Jiaying Sun (2022). On the other hand, recently random forest and extreme gradient boosting machine (XGBoost) has outperformed logistic regression and other models in predicting default risk on P2P Lending different datasets,

as stated by (Fu, 2017), Byanjankar et al. (2021), Chen et al. (2021), Padimi et al. (2022), (Khatir, Almustfa, & Bee, 2022), (Chang, Yang, Tsaih, & Lin, 2022), and (Schmitt, 2022). In addition, random forest and XGBoost were two of the most popular ML models employed for data science competitions (Goldbloom, 2015). Based on these reasons, logistic regression, random forest, naïve bayes, support vector machine and XGBoost were employed for experiment framework of this research. In addition to a hybrid model.

3.2. Data Collection and Procedures

To achieve the research objectives, a number of procedures have been considered to conduct this research experiment. This section includes a demonstration of the conducted experiment's design and procedures that used to prepare the data to execute the experiment. In addition to an exploratory data analysis and visualisation that have been conducted to understand the data.

Fig. 5 describes the experiment design that have been set to conduct this research. It illustrates the procedures that will be followed starting from collecting and selecting the data up to the research's comparative analysis.

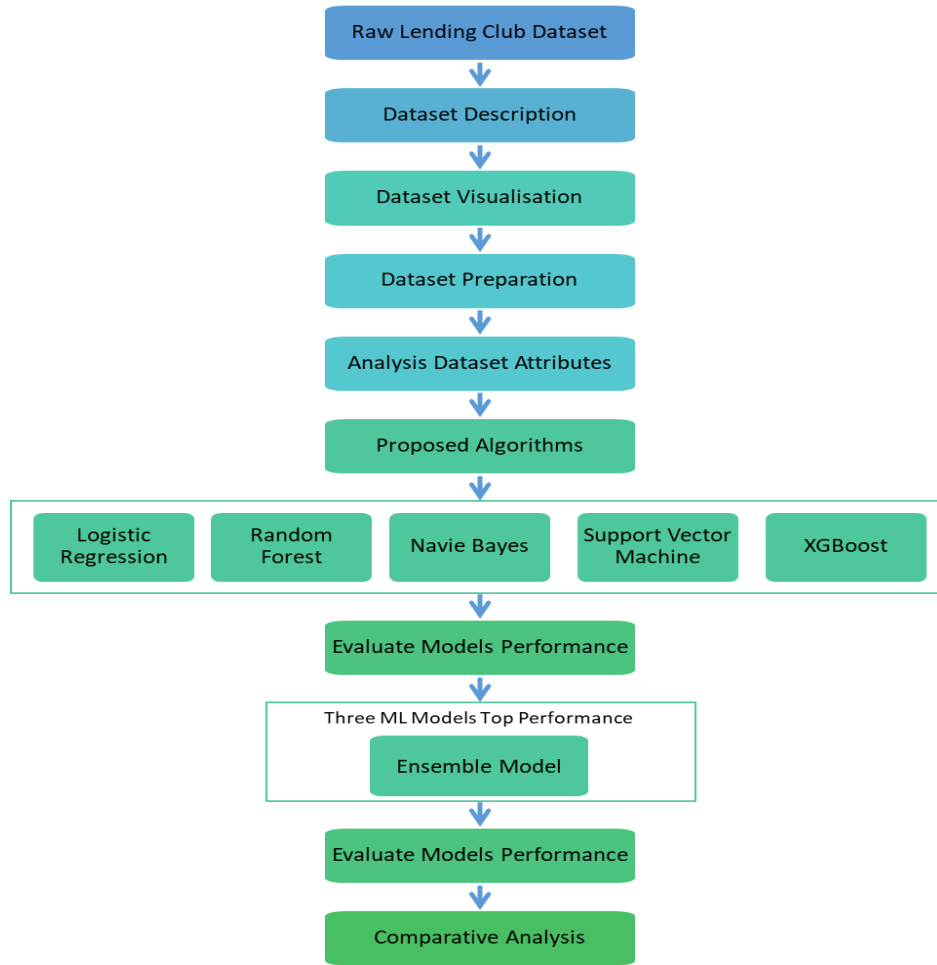


Fig. 5: Research Experiment's Design

3.2.1. Data Collection

The dataset belongs to the P2P lending platform “Lending Club” established in the United States, which contained real-world loans data. The data obtained from the data science community website “Kaggle”. The dataset is available at <https://www.kaggle.com/> (Kaggle, 2023). The data used covered the accepted loans from years 2015-2018, which contained 1,048,444 observations and 151 variables (all variables description can be found in Appendix A). The variables contained a combination of numerical and categorical variables, in which 112 are numerical variables and 39 are categorical. Due to the study’s objectives only accepted loans data will be considered for this study. Additionally, the loan status was additionally

identified as a dependent variable based on the selected dataset, while all other variables considered independent variables. Fig. 6 shows the number of loans issued per year per status.

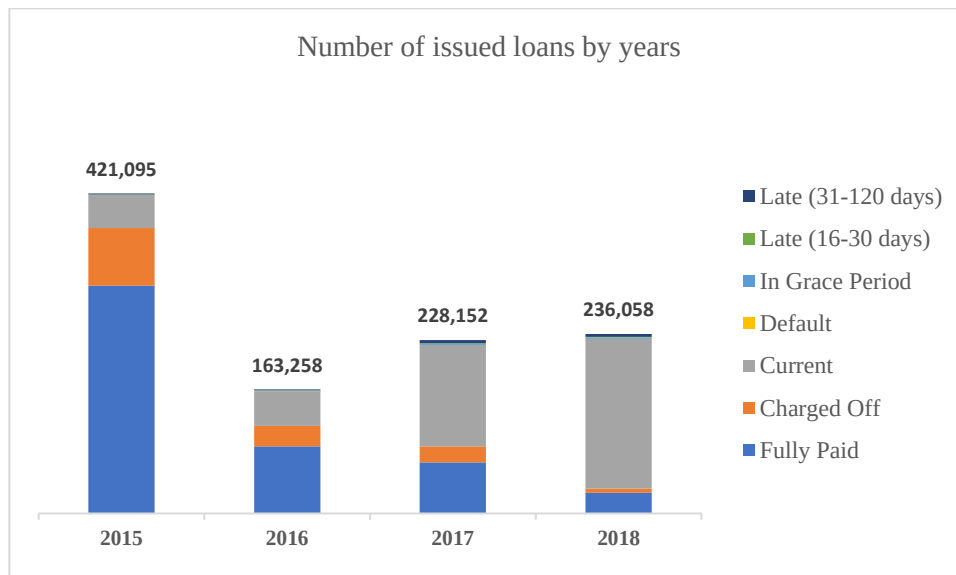


Fig. 6: Number of Loans Issued by Year.

Furthermore, as shown in Fig. 6, the loan status is categorised into seven levels, which are Fully Paid, Current, In Grace Period, Late (16-30 days), Late (31-120 days), Charged Off, and Default. Fully Paid represent loans which have been fully repaid, current for active loans with all due payment are made on time, in grace period means that the loan is past due however it is within the 15-days grace period, Late (16-30 days) represents loans that have due payment late for 16-30 days, late (31-120 days) represents loans that have due payment late for 31-120 days, default status occur on loans that have due payment last for more than 120 days, and charged off represents loans that are no longer expected to be paid off (LendingClub, 2023). For this research two loan statuses were selected which are Fully Paid and Charged Off, also a random sample of 100,000 instance will be consider for the experiment due to server limitation.

3.2.1.1. Data Pre-Processing

In order to construct the selected models in this research many techniques were used to prepare the data. It is critical to apply data cleaning, features selection, feature engineering, data reduction and other techniques to fit the selected models. Fig. 16 demonstrates the data preprocessing steps applied in this study.

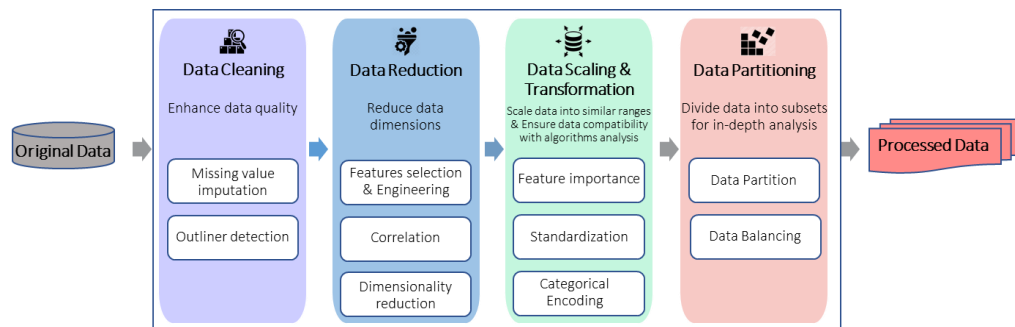


Fig. 16: Research Dataset Pre-Processing.

3.2.1.1.1. Data Cleaning

The purpose of performing some data cleaning techniques is to improve the quality and usability of the dataset. Data cleansing is a vital step in the data analysis process that enhances the reliability and accuracy of the results. The data cleaning techniques applied for this research, including missing value treatment, and outlier treatment.

3.2.1.1.2. Data Reduction

To ensure scalability and ease of interpretation, we applied several data reduction techniques, including Feature selection, Feature Engineering, Feature Standardization, Correlation, and Dimensionality Reduction.

Principal component analysis (PCA) was applied to determine the optimal number of components using the cumulative variance of the components. PCA is one of the methods that enables to explain the variance-covariance structure of a set of p-variables by using few linear combinations of these components (Richardson, 2009). It aims to lower the

dimension of the dataset while maximizing the variance of the linear combination of these variables. PCA can be applied to data that lack observational groups and variable fragmentation.

3.2.1.1.3. Data Partitioning

- **Data Split**

In order to measure the performance of the models, it is required to split the original dataset into two, one dataset will be training dataset and the other will be testing dataset.

- **Data Balancing**

One of the common problems for data analysts is class imbalance, where the distribution of the classes is not equal which can lead to biased analytical models and inaccurate prediction. It occurs on imbalance datasets with a majority of negative classes and a minority are positive classes, or vice versa. Most classification algorithms assume a balanced distribution of the classes, which maximizes classification accuracy by the most common class. Thus, models are constructed with minimum number of classification errors, regardless of the significance of misclassifying samples within each class. Although such models might be accurate, they are considered as ineffective when applied to problems in the real world.

As mentioned previously in exploratory analysis section, in this research the selected dataset (LendingClub data) has class imbalance ratio of 3.7, with “fully paid” loans representing the majority, and “charged off” loans being the minority class.

Researchers have conducted numerous studies for class imbalance problem in different fields of expertise. However, there is no absolute solution to the problem, as various classification problems with different data and classification task respond differently to any imbalance techniques. Data resampling was proposed as a solution applied on the data

level to achieve an equal proportions class, the solution includes under-sampling the majority class, oversampling the minority class, data generating synthetic sampling techniques, and others.

In addition, selection of performance matrices that are more suitable for imbalanced datasets, such as confusion matrix, precision, recall, and F1-score.

For this study, three widely used data balancing techniques (resampling) were applied to balance the datasets by modifying the distribution of the classes. There are three types of resampling methods. The first is under-sampling, which causes the larger class to shrink to the size of the smaller class. Secondly oversampling is used to increase the size of the smaller class to match that of the larger class. Last Synthetic Minority Oversampling Technique (SMOTE) generates synthetic instances of the minority class to balance the class distribution in a dataset.

1. Random Under Sampling

Random Under-sampling (RUS) involves eliminating instances from the majority class to create a balanced dataset. This is achieved by randomly selecting and removing instances from the majority class until the desired balance is reached. This method is beneficial in situations where the dataset is very large, and reducing the number of instances can lead to significant time and resource savings during training and analysis.

However, RUS can also lead to loss of valuable information and might not be suitable for all cases. This method is most effective when applied to datasets with less significant class imbalance and when the redundancy of the majority class instances is not important (Chen, Leu, Huang, Wang, & Takada, 2021).

2. Random Over Sampling

Random Over-sampling (ROS) is another technique used to balance class distribution in datasets. It achieves this balance by generating duplicates of instances from the minority class and adding them to the dataset. However, the randomness of this technique can lead to overfitting since it merely copies instances from the original dataset.

This method is best suited for situations where the dataset is small, and the analysis cannot afford to lose important data. While ROS is easy to implement and understand, it may not be ideal for all use cases due to its overfitting risks.

3. Synthetic Minority Oversampling Technique

Synthetic Minority Oversampling Technique (SMOTE) is a data balancing method that generates synthetic instances of the minority class to balance the class distribution in a dataset. By applying this technique, analysts can create a more evenly distributed data set to train their models. SMOTE works by selecting samples of the minority class and then generating new instances that are similar, but not identical, to the selected samples by interpolating feature values.

This technique is beneficial in situations where the minority class in a dataset has far less representation than the majority class. By oversampling the minority class, the predictive model can better recognize patterns and characteristics of the underrepresented class. SMOTE has proven effective in a wide range of applications, including fraud detection, medical diagnosis, and credit scoring.

3.3. Research Models and performance evaluation metrics

The first part will describe the ML models selected for this research which are, LR, RF, NB, SVM, XGBoost and hybrid mode. The second part define the performance evaluation metrics used to evaluate the ML models' performance.

3.3.1. Logistic Regression

Logistic regression (LR) is one of the most commonly utilized algorithms for credit scoring activities, due to its simplicity of implementation, interpretability, and ability to predict probabilities directly. Although LR can works with binary, multinomial, or ordinal, LR is immensely popular used for binary classification problems, which in this research predict default and non-default.

LR is one of the statistical models under the supervised learning and can be defined as 'A probability estimator that provides a hint of an object being in a tribe than another' (James, Witten, Hastie, & Tibshirani, 2017). In order to predict if the borrower will default or not, LR works through estimating the probability $p(y = 1|x_i)$ at a certain threshold of a borrower i will default given number of predictors x_i . It estimates the probability of a binary outcomes by using the 'logit function' to predicts the logarithm of the probability of a specific class as a linear combination of the input variables. By given set of input variable $x_1, x_2, x_3, x_4, \dots, x_n$, the logit function can be represented as:

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \dots \dots \beta_n x_n \quad \text{Equation (1)}$$

Where p is the probability of the specific class of interest (default or non-default), and α , and $\beta_1 \beta_2 \dots \beta_n$ are the regression coefficients

(parameters). The regression coefficients are derived by using the ‘maximum likelihood’ estimation method. To obtain the results the logistic regression function transforms the results to be ranged between 0 and 1 which represent the probability (default or non-default), this logistic function called ‘sigmoid’ which is used to obtain the probability of the specific class of interest, represented as:

$$P = \frac{1}{1 + \exp - (\alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)} \quad \text{Equation (2)}$$

LR was employed for this research as it has proven its effectiveness toward developing credit-scoring models. For the use of this research the probability of default can be represented in the following mathematical formula:

$$p = P(\text{loan default status} = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)}} \quad \text{Equation (3)}$$

As β_0 is the linear regression intercept, $x_i (i = 1, \dots, n)$ are independent variables, and $\beta_i (i = 1, \dots, n)$ the corresponding coefficient estimated in the model. Thus, the output is a binary variable with the value equal to 1 mean the loan status is default (charges off) and 0 mean the loan is non- default (fully paid).

3.3.2. Random Forest

RF considered one of the ensembles learning methods, which attempt to construct more than one decision tree through set of hypotheses and combine them (Ali, Khan, Ahmad, & Maqsood, 2012). This learning technique has various advantages such as, it can transform weak learners into vital learners. In addition, RF is a tree-based model that overcome the limitations of decision trees algorithms. Decision trees is a sort of supervised learning that predict using a structured tree-like method based on top-down technique. It begins by identifying an initial node as the ‘root’

as a start, and then develop a sequence of recursive branching of succeeding decision nodes until it reaches the terminal ‘leaf’ node, which provides the prediction output. Each decision node has two branches based on the feature value of the data and this process is repeated until it reaches the leaf nodes.

Although the DT is capable of handling large datasets with both quantitative and qualitative variables, it suffers from overfitting. Overfitting happens when a model learns training data too well resulting in poor performance. Accordingly, RF algorithm solves the overfitting issue in DT. RF features two well-known techniques which are boosting and bagging. Boosting has capacity to minimize the generalisation error and prevent overfitting, whilst bagging has the ability to reduce bias and variance (James, Witten, Hastie, & Tibshirani, 2017).

The RF model predicts by aggregating number of decision trees. It employs the bagging technique where each node is constructed using a bootstrapped training sample from the dataset, resulting in randomness and increase in trees diversity. Moreover, the trees are trained on a set of random samples of a predictor chosen as split candidates from the full set of predictors. The bagging technique then trains multiple decision trees with a random subset of input predictors on different training data subsets. As a result, the trees are combined with unweighted voting to provide a new prediction, figure 3.

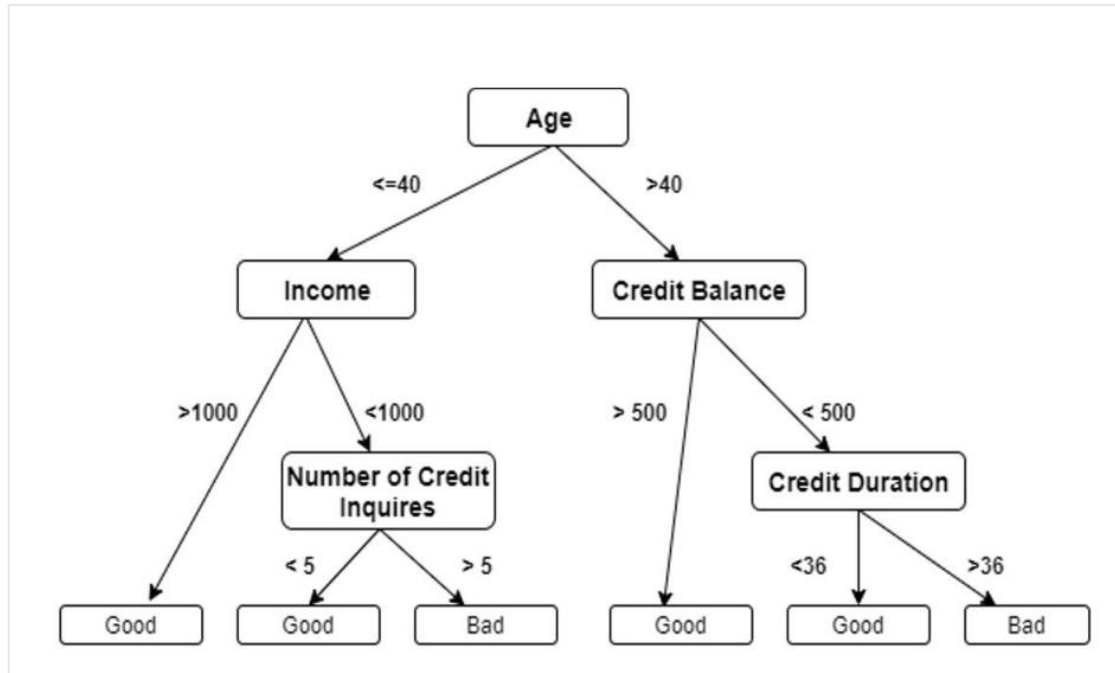


Fig. 3: Random Forest.

Further, during conducting a RF model, the number of trees (n) are defined by the researcher, as well as the number of random features used in the tree split. RF training process are as follow (Ali, Khan, Ahmad, & Maqsood, 2012):

1. Determine (n) number of trees bootstrap from the dataset.
2. In each bootstrap sample a decision tree is built, and in each decision tree, a set of random samples of a predictor (independent) selected as best split from the full predictors.
3. By aggregating the predictions from all the trees contracted, a prediction is made for a new data.

In general, the RF algorithm possesses several advantages that are well-suited for the chosen dataset in this research study. To enumerate some of these advantages, it is worth noting that RF demonstrates efficient performance while operating on large datasets, while also exhibiting exceptional accuracy in comparison to other algorithms. Furthermore, the RF model is capable of adapting to faults present in an unbalanced dataset's class population.

3.3.3. Naïve Bayes

Naïve Bayes (NB) is a popular non-linear algorithm used for classification problems that was developed as an improvement to the Bayesian method. NB is known for its scalability and efficiency in competing with other complex models. The term ‘Naïve’ refer to the model that assumes all proposed attributes are mutually independent of one another and generates conditional probabilities for combinations of attributes and the target predictor (International Business Machines Corporation). NB independence assumption on the conditional probability can be calculated with the following mathematical formula:

$$p(x|y) = \prod_{j=1}^q p(x_j|y) \quad \text{Equation (4)}$$

Where $p(x_j|y)$ is the conditional probability of individual input variables.

3.3.4. Support Vector Machine

Support Vector Machine (SVM) is a non-linear classifier that connects linear and non-linear classifiers by using ML theory to maximize predicted accuracy without overfitted the data. In numerous practical applications, SVM have demonstrated the ability to achieve superior accuracy in long-term prediction tasks when compared to alternative computing methodologies. SVM algorithm relies on the principle of decision plans that define the bounds for making decision. Using basic linear function (a hyper-plane in higher dimensions), SVM classifies different target classes, where input data are mapped into the higher dimensional feature space to segregate data by a hyper-plane (Deris, Zain, & Sallehuddin, 2011). The optimal hyper-plane obtains the largest separation margins between the two classes (Wendler & Gröttrup, 2021). The data transformation usually done through a kernel function which

avoid overfitting the data, there are different kernel functions. Kernel function can be defined as.

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle, \quad \text{Equation (5)}$$

Where $\Phi: X \rightarrow H$ is a projection from feature space into high-dimensional feature space.

A hyper-plane is then used to separate the data in the high-dimensional feature space. The optimal hyper-plane has the greatest separation margins between the two classes. Solving a quadratic optimization problem with constraints yields the greatest separation. The classification decision function with hyper-plane as a parameter can be stated mathematically as follows:

$$f(x) = \text{sing}(\langle \omega, \Phi(x) \rangle + b), \quad \text{Equation (6)}$$

Where ω is a solution of quadratic optimization and b is a constant.

The linear (SVM-L) kernel function was used for this study. The Support Vector Machine with linear kernel function (SVM-L) is the simplest kernel function, having the following expression:

$$k(x, x') = \langle x, x' \rangle \quad \text{Equation (7)}$$

SVM has been used for credit risk evaluation due to its impressive performance, effectiveness, and flexibility. According to an experiment conducted by (Shan & Nilsson, 2018), SVM recorded the highest accuracy rate of 97% among five other models, which is significant when compared to other studies.

3.3.5. Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) has recently amazed scientists with its performance and efficiency, outperforming existing supervised learning models. It has been widely utilized for binary classification

problems (extensively employed for credit risk prediction) and ranking problems. It has been recognised as the gold mine for Kaggle competitions due to its state-of-the-art results on any problem.

XGBoost is a type of Gradient Boosting Machines models classified under decision tree-based ensemble model. While random forest algorithm implemented through bagging to provide a prediction from a collection of outputs of number of trees, gradient boosting algorithm executed by combining bagging and boosting. It incorporates boosting that combines weak learners (decision trees) with only one split, on a sequential basis (happened in parallel), so that each new tree to corrects the errors of the prior one. By using the loss function, each tree's gets performance is evaluated before constructing a new tree that has lower loss than the first tree, thus boosting the prediction outcomes. This process is repeated until the desired level of accuracy is achieved.

The XGboost algorithm, which is based on the GTB algorithm, employs the regular function to control model over-fitting. The target function is as follows:

$$\begin{aligned} \min \{\text{obj}\} &= \min \{L(F_m(X), Y) + \Omega(f) + C\} \quad \text{Equation (8)} \\ &= \min \{\sum_i l(y_i, y'_i) + \sum_i \Omega(f_t) + C\} \end{aligned}$$

Where:

- X = as per data input, all variables in use for prediction;
- Y = the outcome variables must be predicted;
- y_i = the true worth of the I^{th} observation; the variable's final true worth must be predicted;
- y'_i = the predicted value of the I^{th} observation;
- $F_m(X)$ = after m iterations, m weak classification techniques with a fixed weight were incorporated into the final predictable model;

- $\iota(F_m(X), Y)$ = a function of the difference between the actual truth value and the predicted value of the I^{th} observation;
- $L(F_m(X), Y)$ = the error function; the discrepancy between the final predicted variables and the real ones after m iterations;
- $\Omega(f)$ = the final prediction model's regularization function after all iterations;
- $\Omega(f_t)$ = the regularization function in the t -wheel iterative process;
- Obj = the function is objective; the product of the loss and supplemental functions; each and every ML iterative process is a method that maximizes or minimizes the objective function. It must be ascertained if it is maximized or minimized based on real problem.

These formulas could be used to define the frequent term of the algorithm.

XGBoost is well-known for its scalability which enables it to manage a wide range of data types, relationships, hyperparameters, and distributions, in addition to its execution speed (Chen & Guestrin, 2016). In addition, XGBoost offers regularization which allows to control the overfitting issue. Furthermore, XGBoost provides feature importance, allowing for a better understanding of which variables are most significant in developing predictions.

These benefits complement credit scoring benefits since P2P lending platforms require a scalable algorithm that is fast and effective with high performance while making decisions on loan applications.

However, unlike gradient boosting, XGBoost uses maximum depth parameter specified and performs backwards pruning, which enhances computing productivity. During XGBoost training, the algorithm executes the tree-building process using parallelization technique.

3.3.6. Ensemble Method (Hybrid Model)

Ensemble Learning was proposed in 1965 by Nilsson for classification tasks. Ensemble learning concept evolve around training multiple learners as ensemble members then combining their predictions into a single output, Figure. 4. Leading to a better performance and accuracy of classification than ordinary learning member (Zhou, 2012). Instead of constructing one learner from training data like ordinary learning models, ensemble technique (Hybrid Model) works to construct multi classifiers and aggregate them to achieve higher performance. Hence, the error and deviation of a single classifier is compensated by the other members of ensemble on classification problem. Since 1990, ensembles techniques have emerged as a significant paradigm for learning, where they have proven its effectiveness and outstanding performance in numerous studies (Yang, 2017).

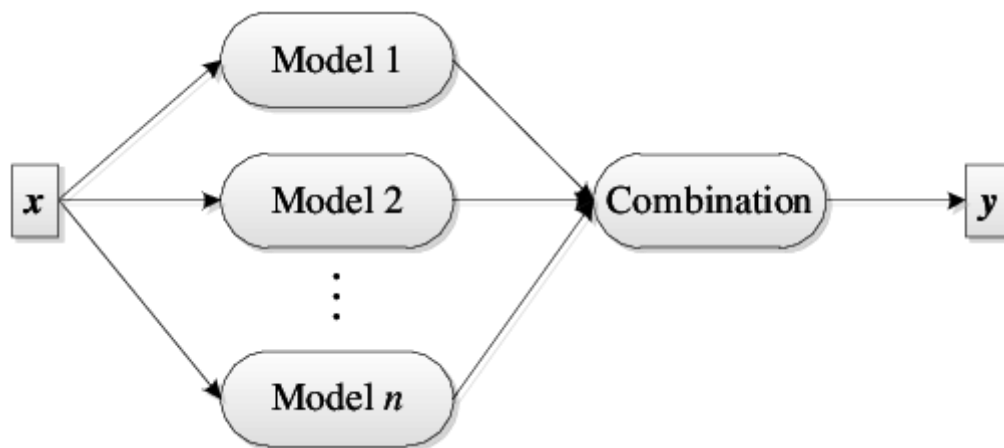


Fig. 4: Common Ensemble Architecture.

3.3.6.1.1. Majority Voting

Majority Voting (MV) is a simple form of ensemble learning method for classification tasks. It combines the predictions of multiple models where each model's output is taken as a 'vote', then a single output (decision) is formed based on majority voting. There are two techniques of

majority voting: hard voting and soft voting. Hard voting involves summing all the predictions for each class label and predicting the class label (output) with the most votes. While soft voting involves summing the predicted probabilities for each class label and predicting the class label (output) with the highest probability (Baker, Mahmood, & Shaker, 2022). The Majority Voting scheme with MAJORITY_VOTING_RULE allows for the evaluation of multiple models and improves overall performance.

3.3.7. Performance Evaluation

This sub-section illuminates the performance metrics that will be used to evaluate the performance of the selected models. As this research handle a binary classification issue the output will be value equal to 1 mean the default loan (Charged Off) and 0 mean the loan is non- default (Fully Paid). This research proposed list of performance metrics that suits the selected data and research's objectives.

3.3.7.1.1. Confusion Matrix

Confusion matrix is a structure table that visualizes the comparison of the classification between the actual classes versus predictions. It is considered as a standard measure for evaluating the performance of classification models (Bekkar, Djemaa, & Alitouche, 2013). It provides a summary of a classification model's prediction by using a contingency table, for binary classification it provides a 2 by 2 matrix with the two classes (positive class and negative class). Since credit scoring is a binary classification problem in which loan applications are classified as 'loan default' or 'fully paid', the confusion matrix can be represented as shown in Table 3.

Table (3): Confusion Matrix

		Predicted Class	
		Positive 1	Negative 0
Actual Class	Positive (Default Loan) 1	True Positive (TP)	False Negative (FN)
	Negative (Non-Default) 0	Fales Positive (FP)	True Negative (TN)

The classification framework assigned ‘default loan’ as positive class since it represents a more significant class to accurately predict due to the higher cost of misclassifying. If the model classifies loan default as non-default (fully paid) there is a high risk of losing investment. While ‘non-default’ assigned as negative class has lower risk if the model misclassified. In Table 3, the model categorises each instance of loan application to one of two classes: loan default or non-default. As a results, the instances can be mapped to one of four possible outcomes.

- True Positive (TP) represent the model correctly predicted instances as a loan default to be default.
- False Negative (FN) represent the model incorrectly predicted instances loan default as non-default.
- True negative (TN) represent the model correctly predicted instances of non-default to be non-default (fully paid).
- False Positive (FP) instances represent the model incorrectly predicted instances non-default as loan default.

Confusion matrix provides an overview of the performance of the classification models to identify the model's strength, weaknesses, and errors made.

3.3.7.1.2. Accuracy

Accuracy metrics considered one of the simplest measures for performance evaluation and it is commonly used for classification problems, due to its ease of application and interpretation. Concisely, it indicates the model's accuracy level in correctly predicting the class label, through calculating the number of correctly predicted instances by dividing the correct prediction (sum of TP and TN) by the total number of predictions and multiplying the result by 100. Accuracy can be represented in a mathematical form as the Percentage of Correctly Classified (PCC) instance over the total number of instances. While the opposite of accuracy represents the error rate which is the incorrect classification.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{Equation (9)}$$

Nevertheless, accuracy treats all classes equally regardless of the number of correct predictions for each class; this may cause a performance's limitation in case of imbalanced data, as it may provide misleading results. Therefore, various evaluation metrics were considered for this research, provided in subsection 3.0.7.

3.3.7.1.3. Precision and Recall

Precision can be defined as the measure of correctness, meaning the proportion of accurately labeled positive instances out of all instances labeled positive by the classifier, thus, precision is the measure that quantifies the accuracy of the labeled positive predictions (Bekkar, Djemaa, & Alitouche, 2013).

$$Precision = \frac{TP}{TP+FP} \quad \text{Equation (10)}$$

Whereas Recall is the proportion of positively labeled instances that were correctly predicted by the model. Recall aids in discovering missing positive predictions, which reflect the model's ability to predict the positive class (Bekkar, Djemaa, & Alitouche, 2013).

$$Recall = \frac{TP}{TP+FN} \quad \text{Equation (11)}$$

These two measurements are inversely related to one another. Also, they are not sensitive to imbalance data.

Specificity which measures the probability of true negative; the probability of the correctly prediction the negative label, has lower risk of the outcomes (predicting non-default to be non-default), however it will support the decision making to accept the loan application.

$$Specificity = \frac{TN}{TN+FP} \quad \text{Equation (12)}$$

3.3.7.1.4. F1-score

F1 score or F-score is a combination of two metrics; Precision and Recall and it calculated by the harmonic mean of these two. It is necessary to have both Precision and Recall are relatively high in order to have high F-score, because the harmonic mean tends to be closer to the smaller of the two numbers. As a results if one of the two is significantly low the F1 score will be low.

$$F1\ Score = \frac{2x(Precision \times Recall)}{Precision+Recall} \quad \text{Equation (13)}$$

The value of F1 range between 0 and 1, where the higher the value indicates higher classification performance of the model. Meaning, when the F-score value is the minimum (1), this indicates that the model failed to correctly predict the positive class. While if the F-score value the

maximum (1), that mean the model has perfectly predicted without any misclassification or errors.

In order to achieve the study purpose effectively, it is necessary to properly select a set of suitable ML algorithms and performance metrics. Knowing that most ML algorithms are built aimed at maximizing overall prediction accuracy and the common metrics align with this objective. For the purpose of this research, appropriate performance metrics need to be selected to suit the unbalanced binary classification issue. Metrics that highlight the higher relative importance of minority class predictions or the identification of loan defaults. Hence, the evaluation metrics employed primarily include the F1-score, accuracy, precision, recall, and confusion matrix.

3.4. Research Tool

This research experiment employing number of ML models to predict the credit scoring defaults for P2P lending using Python in Google Colab. According to Kaggle State of Data Science and Machine Learning 2022 report Python has ranked as the most common program used for data scientists in the last five years (Kaggle, 2022).

Chapter Four

Results and Discussion

4.1. Experiment

4.2. Results

4.3. Discussion

Chapter Four

Results and Discussion

4.1. Experiment

4.1.1. Data Preparation

Data preparation sub-section is divided into two parts. The first part covers an exploratory analysis and data visualization to understand the distribution of values across various features and to determine the relationships between response and predictor variables. And the second part focuses on data pre-processing where it covered the executed data cleaning, features selection, feature engineering, data reduction and other techniques to fit the selected models. In addition to correlation coefficient matrix of individual LendingClub variables. Last, a comprehensive review on data balancing and its various sampling techniques.

4.1.1.1. Exploratory Analysis

In order to understand the data and variables, exploratory analysis was conducted. Data visualization was performed to understand the distribution of values across various features and to determine the relationships between response and predictor variables. This sub-section is structured to include relationship between response and predictor variables, univariate predictors analysis, and multivariate predictors analysis.

The study explored the relationship between the loan status and predictor variables. The loan status was set as a response variable; where it focusses on two status which are fully paid and charged off. As a results,

total loans considered for this research were 87,891 loans for time frame from 2015-2018. Among these loans, 17,603 loans (20%) are default and 70,288 loan (80%) are fully paid, as shown in Fig. 7. This illustrate a clear imbalanced dataset which will be discussed in the next section.

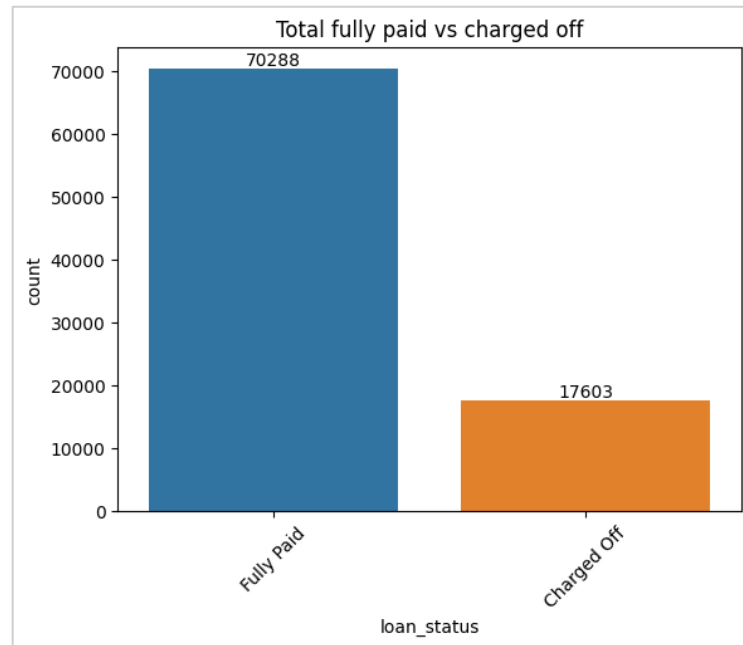


Fig. 7: Dataset Distribution

To begin with, Loan Grade; is a credit grade classification assigned by Lending Club from A to G where ‘A’ represents the less risky borrowers and as grades go higher the borrower’s riskiness increase. By exploring the relationship between loan grade and loan status (Fig. 8), most loans are funded to grades A-D which are grades with less risk, while grades E-G has a smaller number of loans due to its riskiness. Furthermore, analysing the percentage of defaulted per grade, grade ‘A’ loans has only 3.4% of defaulted loans (charged off), as opposed to Grade ‘G’ loans that has 38.2% of defaulted loans. Loans that are graded between A-D have a better repayment chance, while loans graded between E-G are more likely to for loan delinquency.

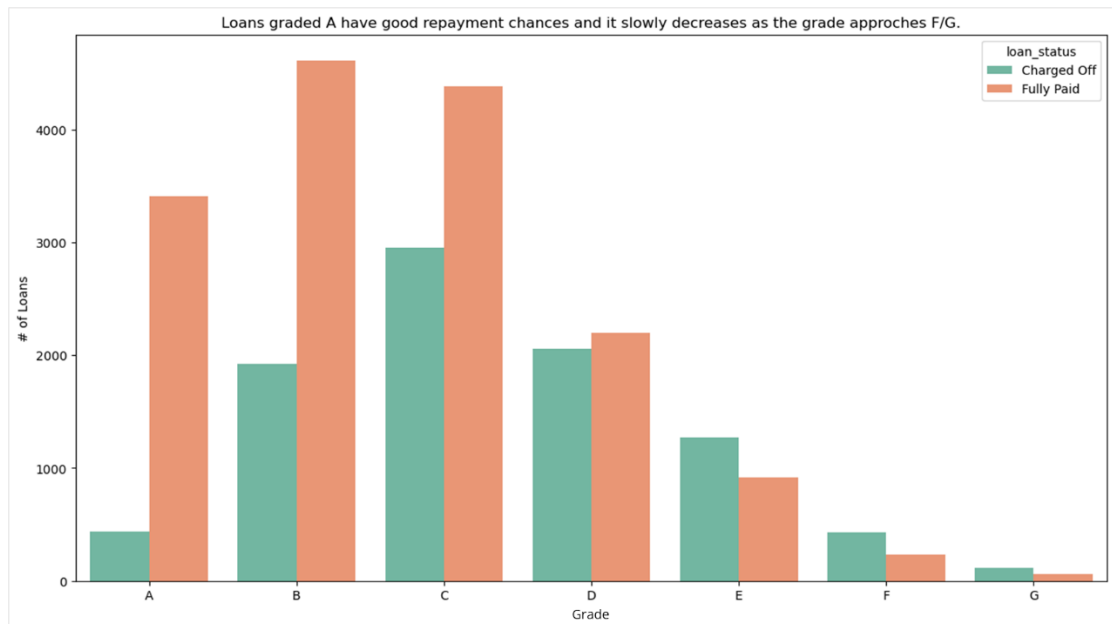


Fig. 8: Grade against Loan Status

Another relationship was explored between loan status and interest rate where loans with higher interest rates are more likely to default or be charged off, Fig. 9.

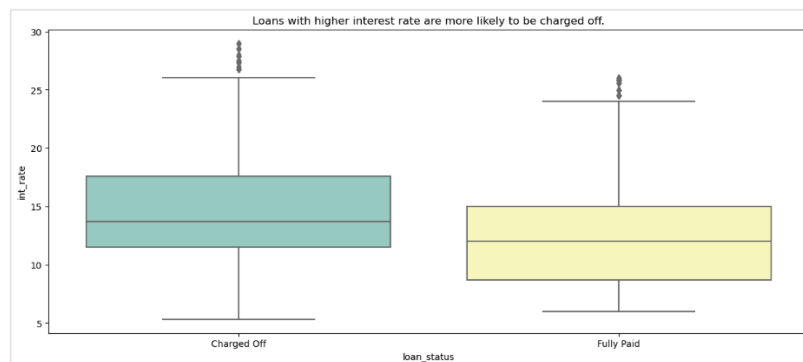


Fig. 9: Interest Rate against Loan Status

Consequently, same relationship occurs between loan status and loan amount where loans with higher loan amounts tends to be default or charged off, Fig. 10.

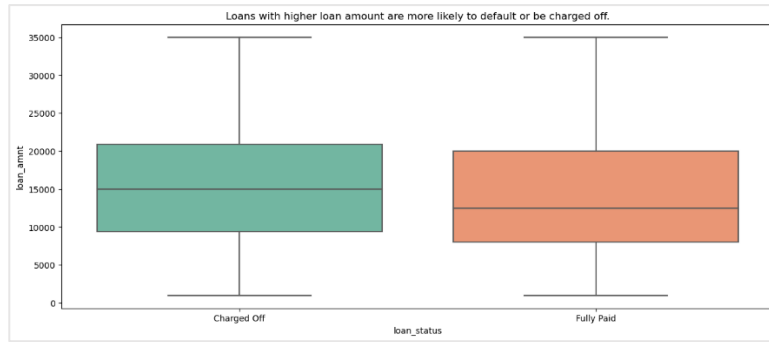


Fig. 10 Loan Amount against Loan Status

Univariate predictors analysis was conducted on features such as purpose, debt-to-income ratio, and number of open credit lines.

Starting with exploring the loan's purpose, which is understanding the reasons behind borrowers' loan request. As illustrated in Figure 11, the majority of the loans are for the reason of debt consolidation, which means taking a new loan to pay off other liabilities and consumer debts such as credit card bills and other loans. The second highest reason is credit card re-financing, followed by home improvement and other.

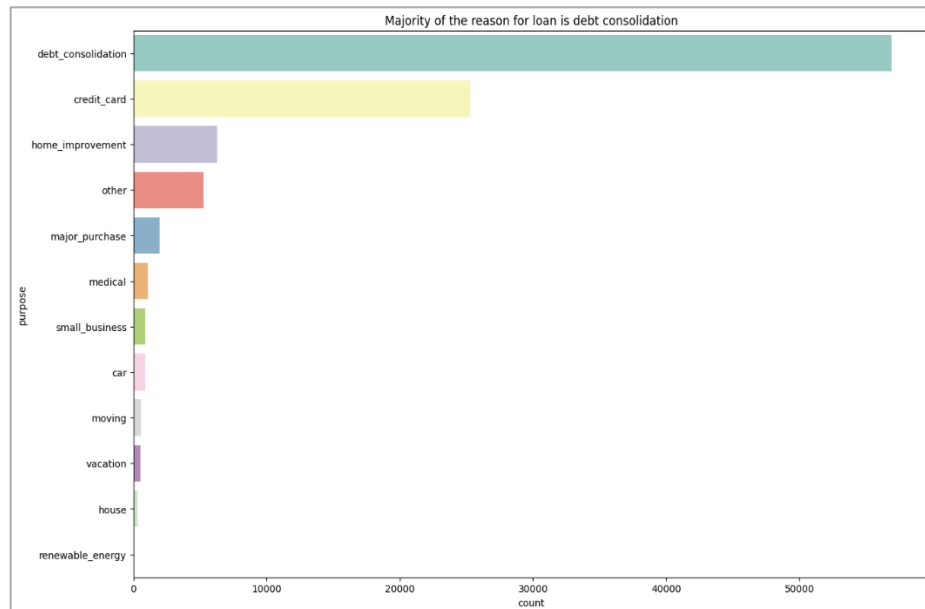


Fig. 11: Purpose of Loan

Moreover, an analysis was conducted on debt-to-income (DTI); which is a financial metrics used to determine individual's financial health by comparing their total debt payment and their total income. DTI is commonly used for financial/lending sector since it gives lenders insight

on borrower's ability to manage and repay debt on time which help them in evaluating loan applications and identifying any potential risks. In this research dataset, the DTI ratio has many outliers, very right skewed, illustrated in Fig. 12. However, it shows that DTI ratio was concentrated predominantly in the range of 0 to 40%.

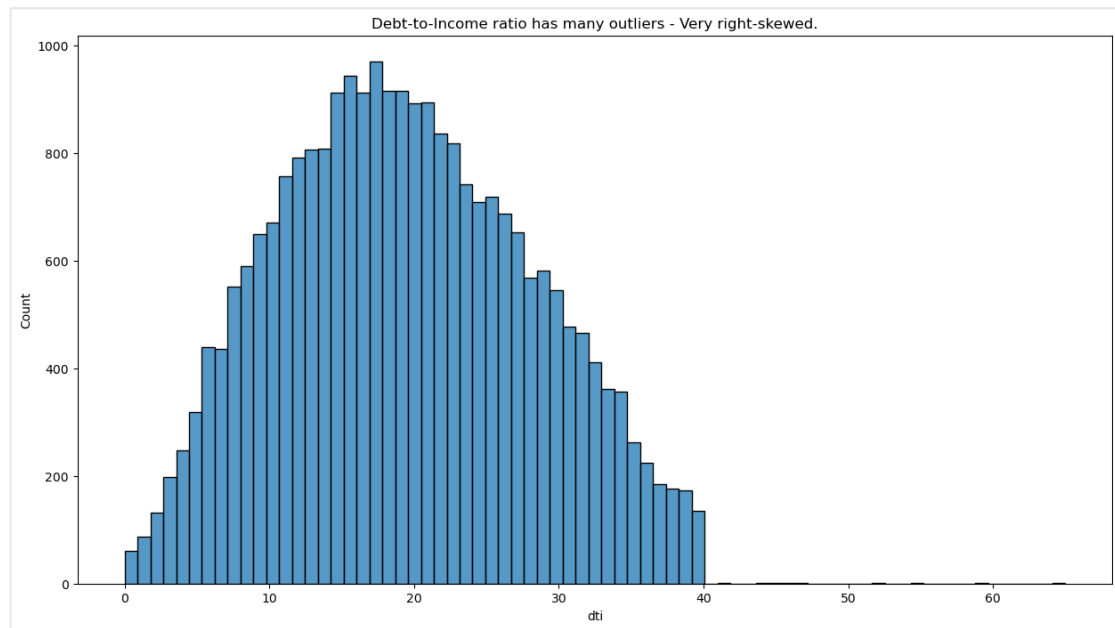


Fig. 12: Debt-to-Income ratio

Furthermore, based on Fig. 13 the number of open credits it was notice that the number of open credit lines in the borrower's credit file seems to be right-skewed.

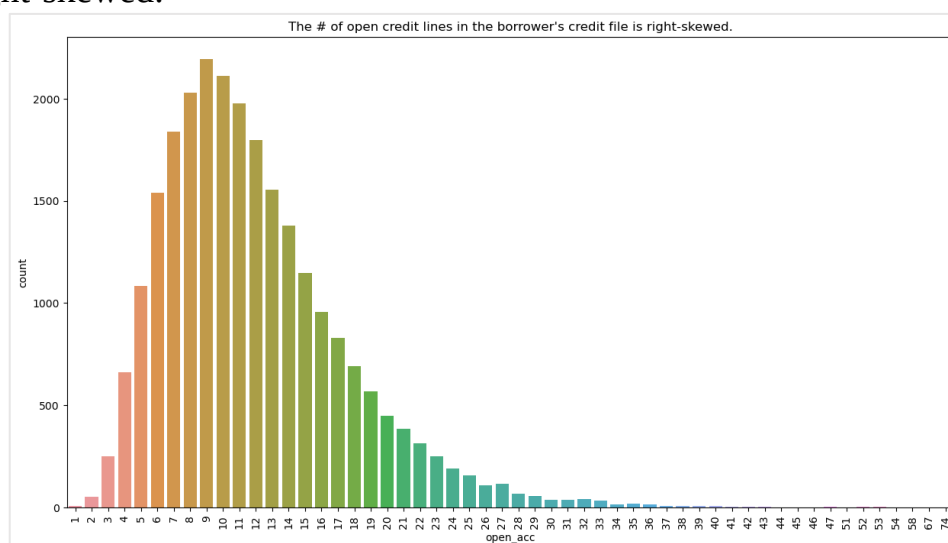


Fig. 13: Open Credit Lines Analysis

Nevertheless, multivariate predictors analysis was conducted to explore the relationships between different predictors. To begin with, a comparison between loan grade and professionals (identify in employment title) were conducted using a heatmap to identify the relationship between the two variables. As illuminated in Fig. 14, professionals such as attorneys, vice presidents, directors, engineers, and store managers tends to have the highest instances of Grade A category loans, while administrator, accountant, analyst, consultant, sales manager have the highest-grade in B category loans, and so on. However, it was discovered that most of the employment title score the lowest in grade F and G, which in line with the observation that the grades of E-G have the highest risk were most borrowers does not have employment titles.

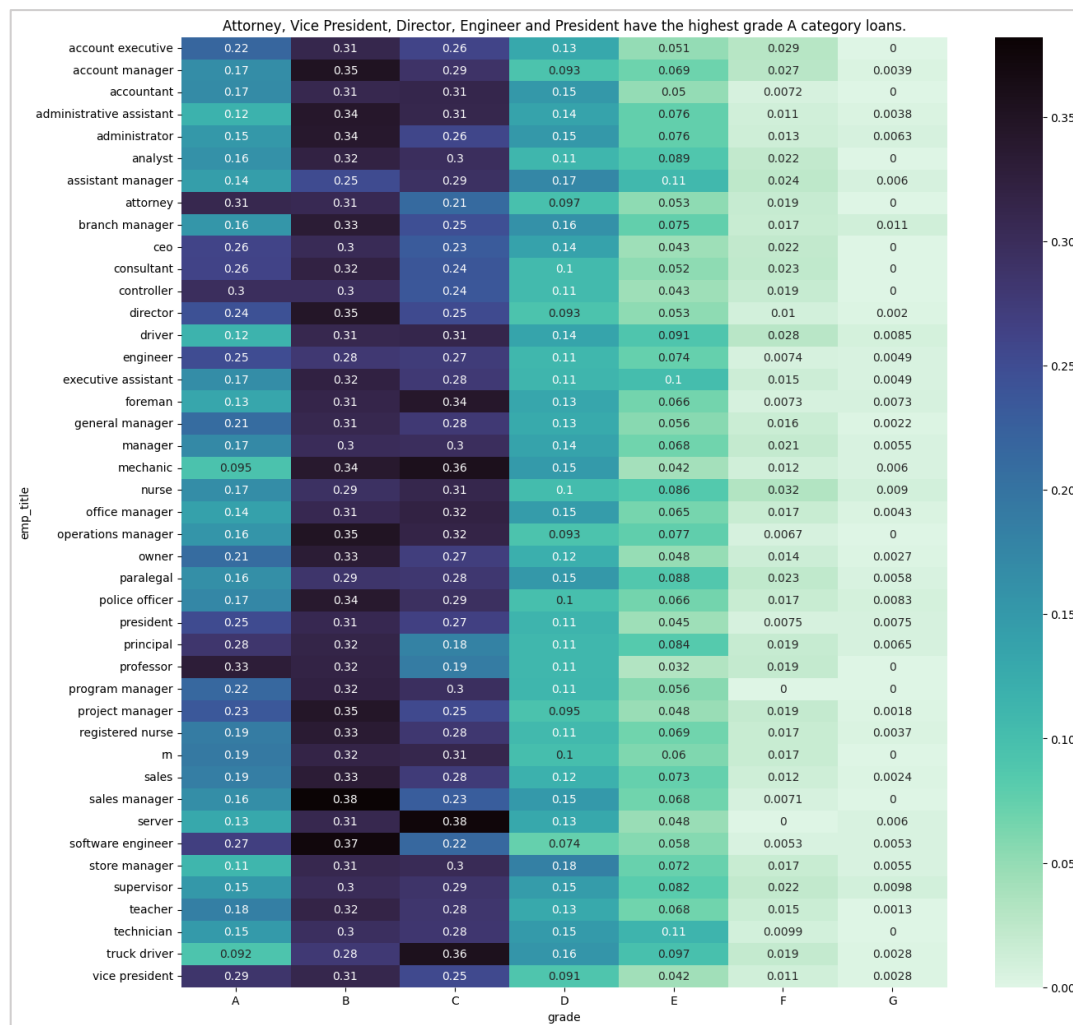


Fig. 14: Employment Title against Loan Grade

Table (4): Loan Grade against Average Interest Rate

Loan Grade	Average Interest Rate
A	6.9
B	10.3
C	13.9
D	17.9
E	21.4
F	25.6
G	29.0

As a consequence, an increase in the interest rate was observed where the interest rate varies drastically by the grade of the loan, as indicated in Table 4. This is reflecting that lower grade are associated with higher risk, which required higher rate of return.

In addition to the increase in the interest rates as grade goes from A to G, also applications type indicates the level of risk, as individual applications have lower interest rates than joint applications and the gap increases as it goes from grade A to grade G, Fig. 15.

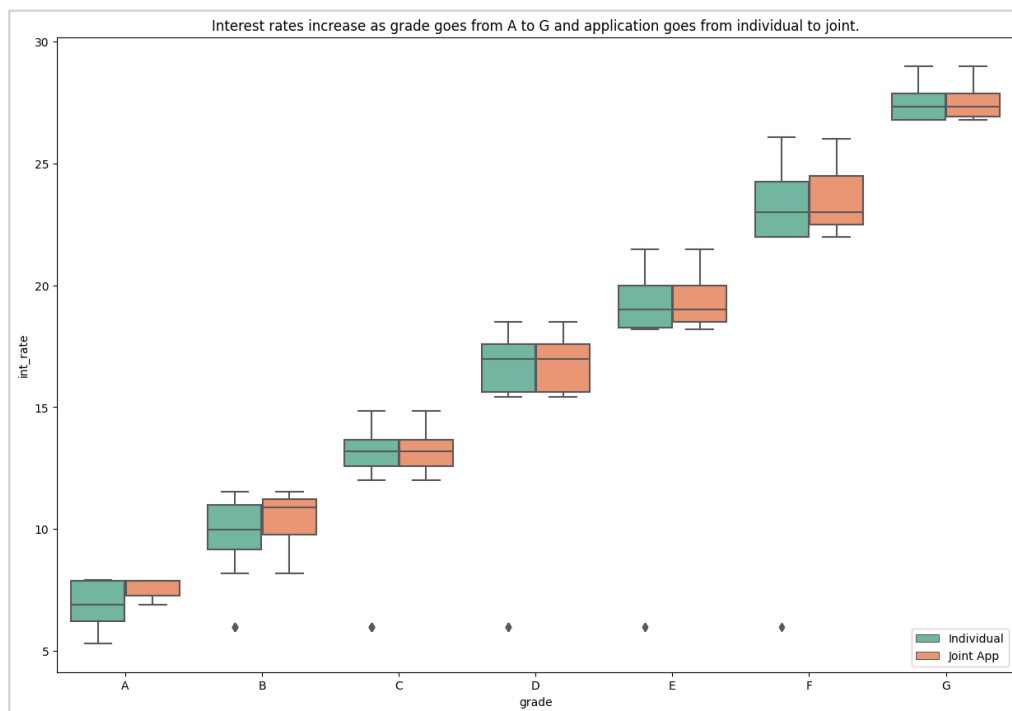


Fig. 15: Interest Rate against Loan Grade

The key findings highlight the importance of loan grading, interest rates, and other factors in predicting the likelihood of successful loan repayment. This information can be helpful for both lenders and borrowers in making informed decisions in the loan application and approval process.

4.1.1.2. Data Pre-Processing

4.1.1.2.1. Data Cleaning

The purpose of performing some data cleaning techniques is to improve the quality and usability of the dataset. Data cleansing is a vital step in the data analysis process that enhances the reliability and accuracy of the results. The data cleaning techniques applied for this research, including missing value treatment, and outlier treatment.

- **Missing Value imputation**

To maintain data integrity and accuracy, issue of missing values was addressed by identifying and removing features containing more than 30% missing values, which has excluded 58 variables. This was essential to ensure a proper and reliable analysis further down the line.

- **Outlier Detection**

The presence of outliers can distort the overall dataset and lead to incorrect conclusions. Box-plot representations were applied to illustrate the distribution of values within individual features, which subsequently facilitated the identification of outliers. Accordingly, the outliers exceeding the 95th percentile and falling below the 5th percentile were removed to smoothen the distribution and minimize its impact on subsequent analysis.

4.1.1.2.2. Data Reduction

- **Feature selection**

After examining the dataset, redundant features were removed, such as funded_amt and funded_amnt_inv, since both features represented by

the loan amount feature. Additionally, the "Employment Title" feature was eliminated, as it contained 37,529 unique categorical values and did not contribute meaningfully to predicting loan status. Moreover, principal component analysis (PCA) was applied to determine the optimal number of components.

▪ **Feature Engineering**

Several features were modified to improve their usefulness in the analysis, which are:

- emp_length (Employment Length) was converted into a numerical feature; between 0 and 10 where 0 means less than one year and 10 means ten or more years.
- fico_range_low and fico_range_high were merged into a single variable named "fico_range". Similarly, last_fico_range_low and last_fico_range_high were combined.

Following these changes, any features that appeared irrelevant for predicting loan status were removed, this has resulted in a dataset with 60 features and a target variable, in which 50 numerical and 10 categorical features.

▪ **Feature Standardization**

Standardized numerical features were separately from categorical data. Numerical features were standardized using the StandardScaler class from the Sci-Kit Learn library. Categorical data was standardized by implementing one-hot encoding techniques and creating dummy variables for existing features. The standardized datasets were then combined into a single data frame. As a results total number of features after standardizing was 119.

▪ **Correlation**

A correlation matrix was generated on the dataset and filtered features with a correlation coefficient of more than 0.9. This allows to eliminate

highly correlated features, which could affect the accuracy of the analysis. Visualize the correlation matrix using a heatmap matrix to visualize the correlation matrix, it can help in identifying pairs of features that are highly correlated, and accordingly can be eliminated, Fig. 17.

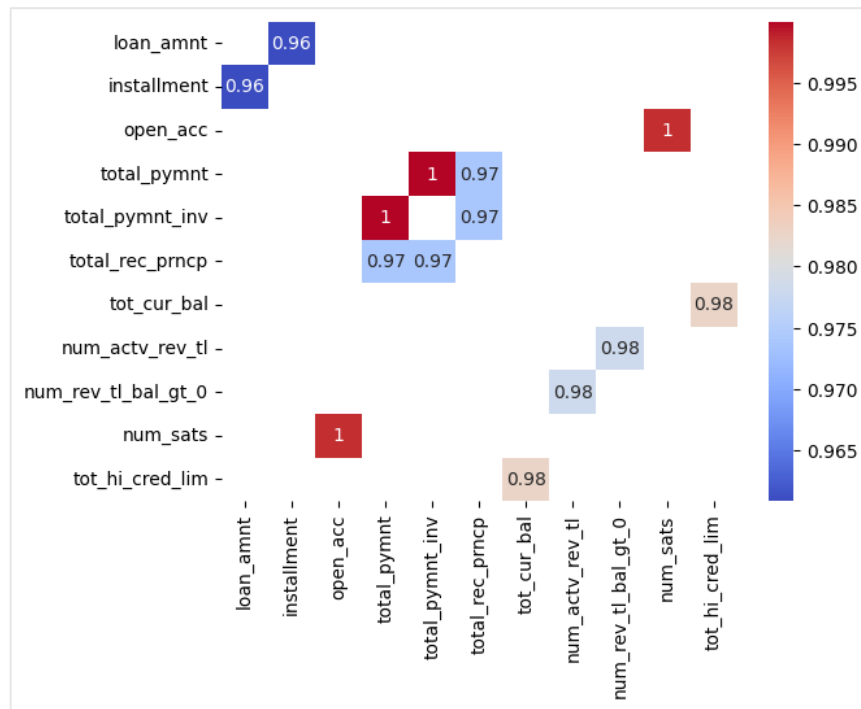


Fig. 17: Correlation Heatmap.

The features which are more than 90% correlated with other columns were removed, leading to a total number of features of 112.

▪ Dimensionality Reduction

By applying PCA in this study, initially 100 principal components were created and plotted the cumulative sum of explained variance. The optimal number of components was established 39, by selecting those that explained approximately 95% of the variance in the data. Finally, the optimal principal components were constructed, Fig. 18.

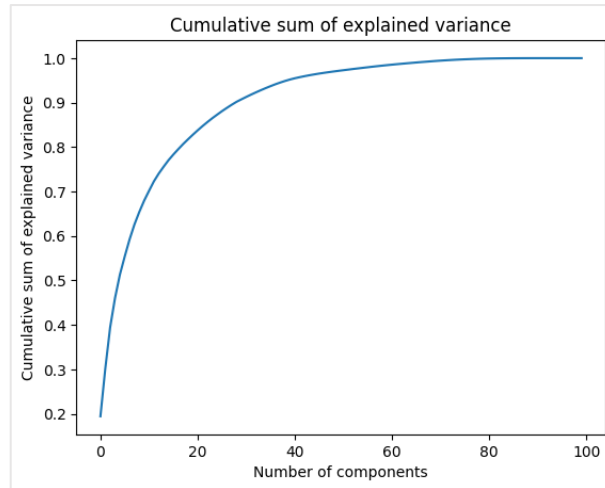


Fig. 18: Principal Component Analysis (PCA).

4.1.1.2.3. Data Scaling and Transformation

- **Feature Importance**

Loading scores of each feature were obtained which represents the significance of the feature in explaining the variance in the data. By plotting the importance of each feature, it enables to visualize the positive and negative influence of each feature on the variance in the data. The final dataset included these principal components as the features.

4.1.1.2.4. Data Partitioning

- **Data Split**

By having two datasets, the models can be trained based on the training set, and then evaluates their performance using the testing dataset. The common split ratio of the original dataset is between 70% and 80% for the training data and remaining 30% or 20% for testing. In this research, the split ratio used was 80% train data, and 20% test data.

4.2. Results

In the current competitive financial market, efficiency and effectiveness are key advantages that make online P2P lending a challenging competitor to traditional banks in the digital age. Therefore, identifying the most accurate model for assessing the risk associated with loan applications is critical for minimizing financial losses and bad debt. P2P Lending platforms strategies aim to limit credit risk by using different classification algorithms to analyse customer data and identify links between borrower's characteristics and attributes, and loan status (default or non-default). Therefore, accurately distinguishing between good and bad loan applicants is a crucial element of credit scoring using ML approaches, which are gaining popularity due to their improved accuracy rates.

Hence, this research aims at identifying the optimal model that can be used to predict defaulters for P2P lending through conducting various experiments to compare the performance and accuracy of five popular ML classifiers which are LR, RF, SVM, NB, XGBoost, and a hybrid model using selected publicly available loan dataset extracted from LendingClub.

As explained in the previous chapter, the experiment started with analysing the dataset, and thorough data pre-processing that includes data cleansing, selecting appropriate features, scaling, transforming some features, and partitioning to prepare the dataset to be used for the selected models.

The following section presents the results and findings of five ML algorithms and the hybrid model applied to a loan default dataset.

4.2.1. Logistic Regression

ML algorithms are heavily employed in the financial and lending sector, particularly to tackle classification challenges. LR is a commonly utilized algorithm for credit scoring, as it deals with binary classification problems. Since this research deals with binary classification issue, LR classifier has been used on the loan default dataset.

Starting the modeling with training the LR classifiers on the loan default dataset. The first results obtained from LR was on full dataset where no balancing methods were considered. The model's performance was presented through confusion matrix, shown in Fig. 19.

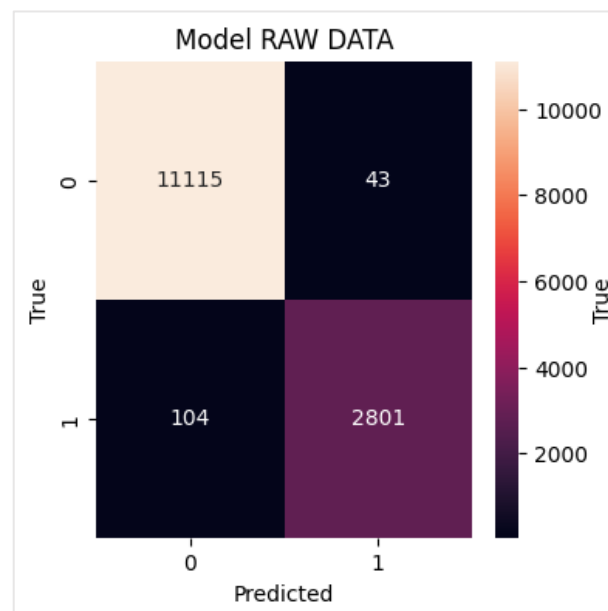


Fig. 19: Logistic Regression Confusion Matrix – No Balancing.

While the diagonal values are predicted accurately, the non-diagonal elements are inaccurate. In the confusion matrix, the Y axis represents correct or accurate labels are placed, while the X axis represents the predicated labels. Actual forecasts are 11,115 and 2,801, whereas faulty predictions are 43 and 104. As mentioned previously, a confusion matrix is composed of four cases: False Negatives of 104, False Positives of 43, True Negatives of 11,115, and True Positives of 2,801.

Table (5): Specificity, Precision, Recall, Accuracy, and F1-score result of LR-No Balancing

Balancing Method	Specificity	Precision	Recall	Accuracy	F1-score
No Balancing	0.9961	0.9895	0.9895	0.9895	0.9895

As shown in Table 5, the validation accuracy of LR with no balancing method is 98.95%. This means that the percentage of accurately predicted instances is 0.9895. In this experiment, class 0 represents non-default or Fully Paid, and class 1 represents the default or Charged Off, where class 0 has 11,158 instances considered, and class 1 has 2,905 instances. Overall, the model performs a specificity of 0.996, a precision of 0.989, a recall of 0.989, an accuracy of 0.989 and the result of F1-score of 0.989.

Moreover, as there is an imbalance issue in the dataset, the experiment was extended to include training the model using various balancing methods which are SMOTE, ROS, and RUS. The experiment includes training the LR model on different balanced datasets, performing grid search using K-Fold cross-validation and a predetermined set of hyperparameters to find the best combination of parameters for each model. Followed by training the models on the best set of parameters obtained from the grid search and then assessing the models' performance using various evaluation metrics. Fig. 20 shows confusion matrices were performed to evaluate performance of LR models using various data balancing techniques.

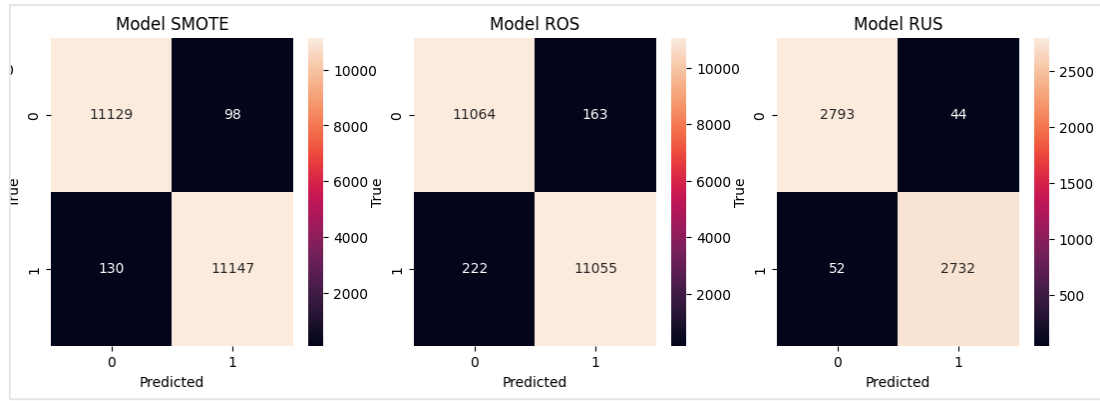


Fig. 20: Logistic Regression Confusion Matrix – Various Balancing Methods.

While performing the LR model using SMOTE balancing technique, the actual forecasts were 11,129 and 11,147, whereas faulty predictions were 98 and 130. The confusion matrix's four cases were False Negatives of 130, False Positives of 98, True Negatives of 11,129, and True Positives of 11,147. For the SMOTE case, class 0 (non-default or Fully Paid) had 11,227 instances, and class 1 (default or Charged Off) had 11,277 instances.

For LR model using ROS balancing method, the actual forecasts were 11,064 and 11,055, whereas faulty predictions are 163 and 222. The confusion matrix's four cases were False Negatives of 222, False Positives of 163, True Negatives of 11,064, and True Positives of 11,055. For this training (ROS), class 0 (non-default or Fully Paid) had 11,227 instances, and class 1 (default or Charged Off) had 11,277 instances.

The last balancing technique used for LR model was RUS where the actual forecasts were 2,793 and 2,732, whereas faulty predictions are 44 and 52. The confusion matrix's four cases were False Negatives of 52, False Positives of 44, True Negatives of 2,793, and True Positives of 2,732. The dataset used has class 0 (non-default or Fully Paid) of 2,837 instances, and class 1 (default or Charged Off) of 2,784 instances.

Table (6): Specificity, Precision, Recall, Accuracy, and F1-score result of LR- Various Balancing Methods.

Balancing Method	Specificity	Precision	Recall	Accuracy	F1-score
No Balancing	0.99614	0.98952	0.98954	0.98954	0.9895
SMOTE	0.99127	0.98987	0.98986	0.98986	0.9898
ROS	0.98548	0.98290	0.98289	0.98289	0.9828
RUS	0.98449	0.98292	0.98292	0.98292	0.9829

Table 6 shows a performance comparison for the LR models using no balancing, SMOTE, ROS, and RUS. The models' F1-scores were compared to determine the best performing model. A higher F1-score indicates superior performance in identifying positive cases (recall) and avoiding false positives (precision). Upon analysing the F1-score performance of the LR models trained using different data balancing techniques, SMOTE was found to yield the highest F1-score (0.9898). This indicates that for LR model, SMOTE technique is the most effective data balancing method among those evaluated in this study.

4.2.2. Random Forest

RF considered to be the second popular algorithm used for credit scoring, it has shown a significant performance in number of studies on credit scoring. For several reasons, this algorithm was chosen for the experiment. It efficiently handles large datasets and outperforms other algorithms in accuracy. Also, it can adapt with errors in datasets with unbalanced class populations.

Following the same steps done previously, RF models were trained on a training dataset. Starting with conducting the first RF model run on

full data with no balancing method, Fig. 21 present the results on the RF model through confusion matrix.

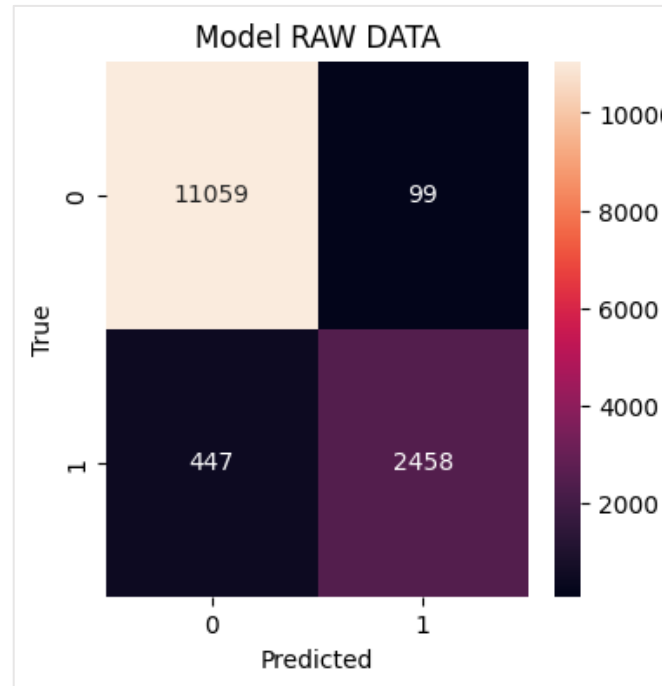


Fig. 21: Random Forest Confusion Matrix – No Balancing.

As illustrated in Fig. 21, the RF model performing using full data has an actual forecast of 11,059 and 2,458, whereas faulty predictions are 99 and 447. The confusion matrix four cases performed as False Negatives of 447, False Positives of 99, True Negatives of 11,059, and True Positives of 2,458.

Table (7): Specificity, Precision, Recall, Accuracy, and F1-score result of RF-No Balancing

Balancing Method	Specificity	Precision	Recall	Accuracy	F1-score
No Balancing	0.99112	0.96117	0.96117	0.96117	0.96023

As shown in Table 7, the validation accuracy of RF with no balancing method is 96%. This means that the percentage of accurately predicted instances is 0.961. Overall, the model performs a specificity of 0.991, a precision of 0.961, a recall of 0.961, an accuracy of 0.961 and the result of F1-score of 0.960.

Nevertheless, the experiment also includes training the RF model using the selected balancing methods which are SMOTE, ROS, and RUS. After training the model, performing grid search using K-Fold cross-validation, train the model on the best combination of parameters, and assessing the models' performance using various evaluation metrics. Fig. 22 shows confusion matrices were performed to evaluate performance of RF models using various data balancing techniques.

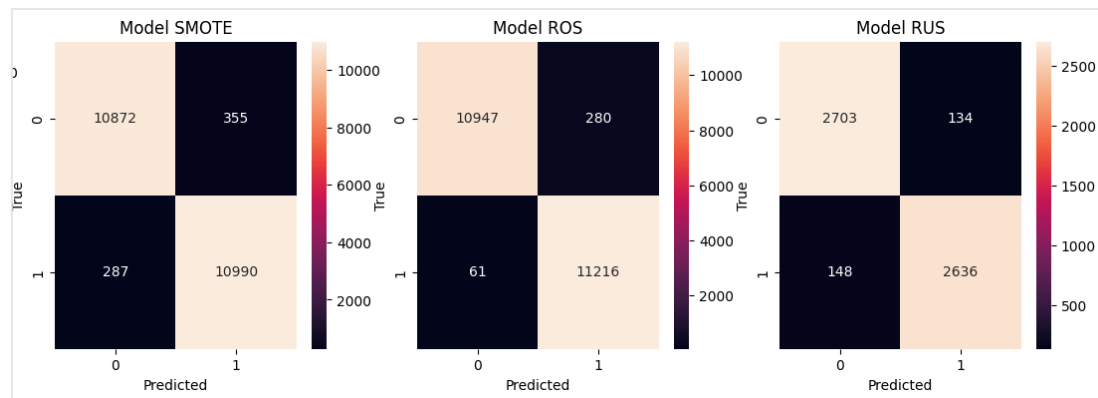


Fig. 22: Random Forest Confusion Matrix – Various Balancing Methods.

Fig. 22 the first confusion matrix on the left illustrates the performing of RF models using SMOTE balancing, it results with actual predictions of 10,872 and 10,990, where inaccurate predictions were 287 and 355. The four cases of confusion matrix were False Negatives of 287, False Positives of 355, True Negatives of 10,872, and True Positives of 10,990.

Next, RF model using ROS balancing method, the actual predictions were 10,947 and 11,216, while inaccurate predictions are 61 and 280. The four cases of confusion matrix were False Negatives of 61, False Positives of 280, True Negatives of 10,947, and True Positives of 11,216.

The last balancing technique used for RF model was RUS where the actual predictions were 2,703 and 2,636, whereas inaccurate predictions are 148 and 134. The four cases of confusion matrix were False Negatives of 148, False Positives of 134, True Negatives of 2,703, and True Positives of 2,636.

Table (8): Specificity, Precision, Recall, Accuracy, and F1-score result of RF- Various Balancing Methods.

Balancing Method	Specificity	Precision	Recall	Accuracy	F1-score
No Balancing	0.99112	0.96117	0.96117	0.96117	0.96023
SMOTE	0.96837	0.97148	0.97147	0.97147	0.97147
ROS	0.97506	0.98503	0.98484	0.98484	0.98484
RUS	0.95276	0.94983	0.94983	0.94983	0.94982

Table 8 shows a performance comparison for the RF models using no balancing, SMOTE, ROS, and RUS. The models' F1-scores were compared to determine the best performing model. After evaluating the F1-score performance of RF models trained with various data balancing techniques, it was discovered that ROS produced the highest F1-score 0.984. This suggests that ROS is the most effective data balancing method for RF models among those assessed in this study.

4.2.3. Naïve Bayes

NB was considered for this research as a popular non-linear algorithm used for classification problems. Since NB is known for its scalability and efficiency in competing with other complex models it was applied to predict the credit scoring to assess the loan applicants. In Huixi Hea and Kazumitsu Nawata study “Application of Machine Learning Algorithms in Predicting the Borrower’s Default Risk in Online Peer-to-Peer Lending” (He & Nawata, 2019), NB has outperformed other models with 96% accuracy. In additional, NB can perform reasonably well with imbalanced data though techniques such as oversampling or under sampling.

NB models were trained on a training dataset using cross-validation (K-fold). Firstly, NB model performance with full data with no balancing method reflected in confusion matrix presented in Fig. 23.

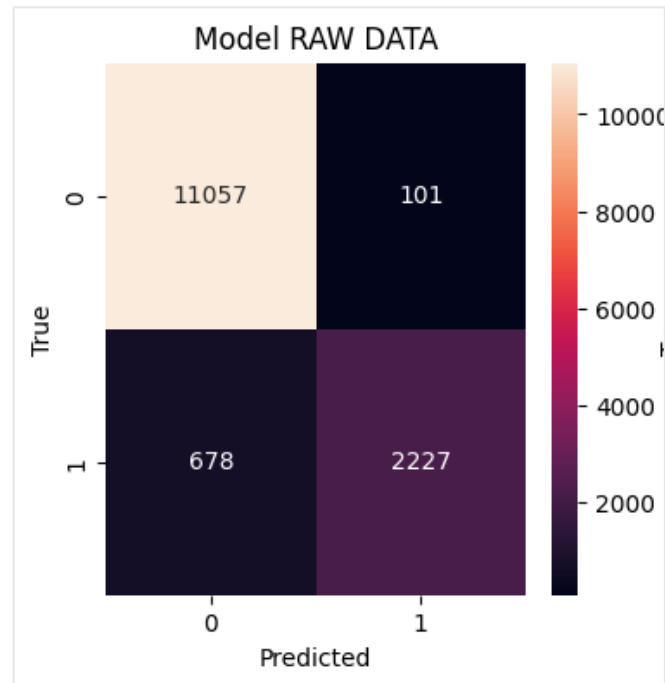


Fig. 23: Naïve Bayes Confusion Matrix – No Balancing.

The NB actual predictions were 11,057 and 2,227, while inaccurate predictions were 678 and 101. The confusion matrix four cases performed as False Negatives of 678, False Positives of 101, True Negatives of 11,057, and True Positives of 2,227.

Table (9): Specificity, Precision, Recall, Accuracy, and F1-score result of NB-No Balancing

Balancing Method	Specificity	Precision	Recall	Accuracy	F1-score
No Balancing	0.99094	0.94519	0.94460	0.94460	0.94225

As shown in Table 9, NB with no balancing technique performed an accuracy of 94%. This means that the percentage of accurately predicted instances is 0.94. The overall performance specificity of 0.99, a precision of 0.94, a recall of 0.94, an accuracy of 0.94 and the result of F1-score of 0.94.

Nonetheless, the experiment also includes training the NB model using the selected balancing methods which are SMOTE, ROS, and RUS. After applying all the required steps to executing the models. The models' performance was assessed using various evaluation metrics. Fig. 24 presents confusion matrices were performed to evaluate performance of NB models using various data balancing techniques.

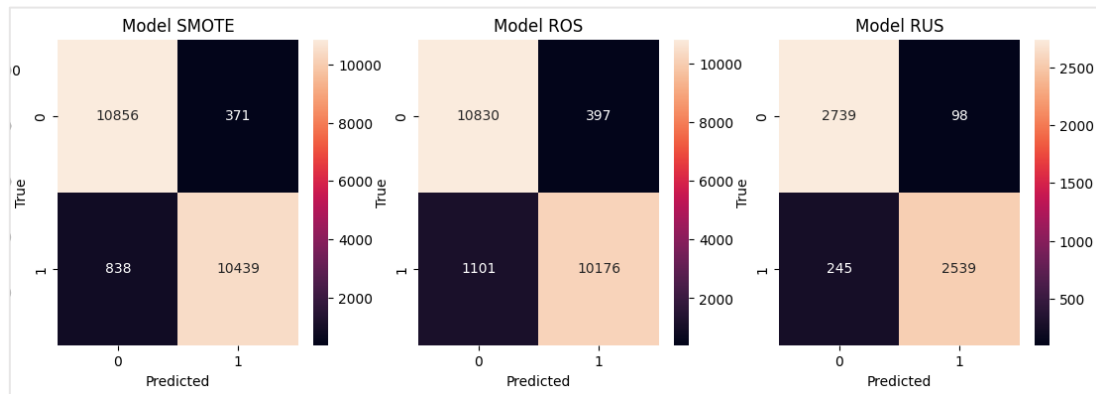


Fig. 24: Naïve Bayes Confusion Matrix – Various Balancing Methods.

Through analysing the confusion matrix of NB using SMOTE balancing technique, the results show actual predictions of 10,856 and 10,439, against 371 and 838 in accurate predictions. Further analysis on the four cases of False Negatives of 838, False Positives of 371, True Negatives of 10,856, and True Positives of 10,439.

Another analysis on NB confusion matrix using ROS balancing techniques, where it is resulted with 10,830 and 10,176 accurate predictions and 397 and 1,101 inaccurate predictions. Reflected in the four confusion matrix cases of False Negatives of 1,101, False Positives of 397, True Negatives of 10,830, and True Positives of 10,176.

Last analysis on NB confusion matrix using RUS balancing as shown in Fig. 24 in right, the actual forecasts were 2,739 and 2,539, while the incorrect forecasts were 245 and 98. The four cases of confusion matrix were False Negatives of 245, False Positives of 98, True Negatives of 2,739, and True Positives of 2,539.

Table (10): Specificity, Precision, Recall, Accuracy, and F1-score result of NB- Various Balancing Methods.

Balancing Method	Specificity	Precision	Recall	Accuracy	F1-score
No Balancing	0.99094	0.94519	0.94460	0.94460	0.94225
SMOTE	0.96695	0.94705	0.94627	0.94627	0.94625
ROS	0.96463	0.93514	0.93343	0.93343	0.93337
RUS	0.96545	0.94015	0.93897	0.93897	0.93892

Table 10 presents NB models performance using various balancing techniques through different measurements. Focusing on F1-score to determine the best balancing technique to deploy NB model, SMOTE has yield the highest F1-score of 0.946. Which conclude that the best deployment of the NB model using loan default dataset is using SMOTE balancing technique.

4.2.4. Support Vector Machine

The SVM algorithm has been proven effective in assessing credit risk, particularly for classification issues. It is considered as a powerful ML algorithm in evaluating the creditworthiness of borrowers. The selected dataset has large number of features, where SVM performs effectively in high dimensional spaces and can handle complex relationships between variables. In addition, SVM can handle imbalanced data via using different data balancing techniques. The first part of SVM experiment was done through using the full dataset without any balancing method, Fig. 25 demonstrates RF performance using confusion matrix.

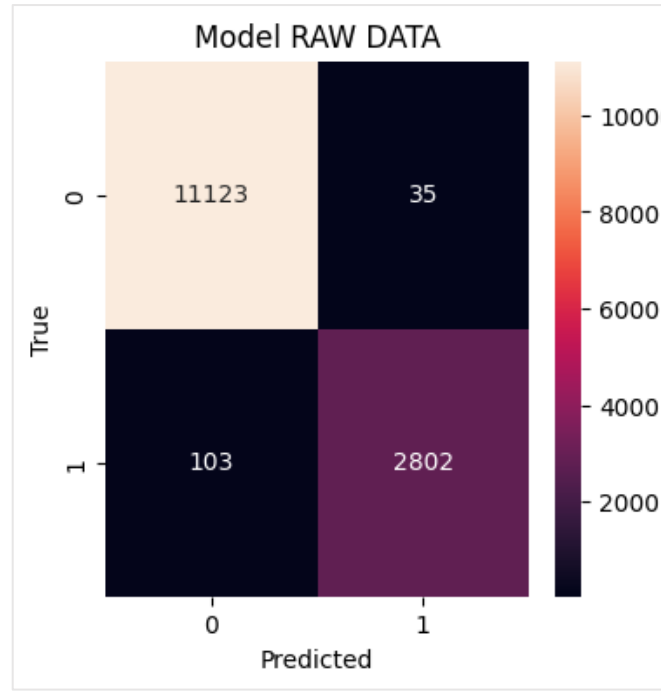


Fig. 25: Support Vector Machine Confusion Matrix – No Balancing.

The SVM accurate predictions were 11,123 and 2,802, while inaccurate predictions were 35 and 103. The confusion matrix four cases performed as False Negatives of 103, False Positives of 35, True Negatives of 11,123, and True Positives of 2,802.

Table (11): Specificity, Precision, Recall, Accuracy, and F1-score result of SVM-No Balancing

Balancing Method	Specificity	Precision	Recall	Accuracy	F1-score
No Balancing	0.99686	0.99017	0.99018	0.99018	0.99014

By Evaluating the performance of SVM using the appropriate metrics, the model performed an accuracy of 0.99, which means that the percentage of accurately predicted instance were 0.99. Other metrics measured as well such as specificity of 0.996, a precision of 0.99, a recall of 0.99, an accuracy of 0.99 and the result of F1-score of 0.99.

On the other hand, like other algorithms the SVM model was trained using the selected balancing methods which are SMOTE, ROS, and RUS.

Once all executing steps were done, the models' performance was compared through confusion matrix as shown in Fig. 26.

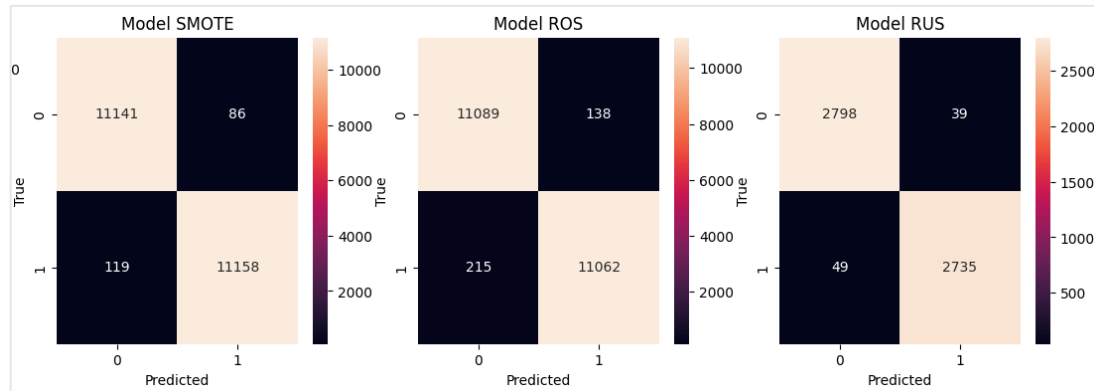


Fig. 26: Support Vector Machine Confusion Matrix – Various Balancing Methods.

As illustrated in Fig. 26 the first confusion matrix illustrates the performing of SVM models using SMOTE balancing, it results with actual predictions of 11,141 and 11,158, where inaccurate predictions were 119 and 86. The four cases of confusion matrix were False Negatives of 119, False Positives of 86, True Negatives of 11,141, and True Positives of 11,158.

Next, SVM model using ROS balancing method, the actual predictions were 11,089 and 11,062, while inaccurate predictions are 138 and 215. The four cases of confusion matrix were False Negatives of 215, False Positives of 138, True Negatives of 11,089, and True Positives of 11,062.

The last balancing technique used for SVM model was RUS where the actual predictions were 2,798 and 2,735, whereas inaccurate predictions are 49 and 39. The four cases of confusion matrix were False Negatives of 49, False Positives of 39, True Negatives of 2,798, and True Positives of 2,735.

Table (12): Specificity, Precision, Recall, Accuracy, and F1-score result of SVM- Various Balancing Methods.

Balancing Method	Specificity	Precision	Recall	Accuracy	F1-score
No Balancing	0.99686	0.99017	0.99018	0.99018	0.99014
SMOTE	0.99233	0.99089	0.99089	0.99089	0.99089
ROS	0.98770	0.98433	0.98431	0.98431	0.98431
RUS	0.98625	0.98435	0.98434	0.98434	0.98434

As reflected in Table 12, by comparing the performance of SVM models using no balancing, SMOTE, ROS, and RUS, particularly the F1-score. It was found that SVM model using SMOTE technique has outperformed the others with the highest F1-score of 0.9908.

4.2.5. Extreme Gradient Boosting Machine

XGBoost has recently become popular and effective algorithm used for credit scoring due to its excellent performance among other ML applications. It has high predictive power as it can capture complex and non-linear relationships between features and the response, which is considered valuable for credit scoring.

The XGBoost experiment started using the full dataset with no balancing technique applied on the dataset. Accordingly, Fig. 27 demonstrates the XGBoost performance using confusion matrix.

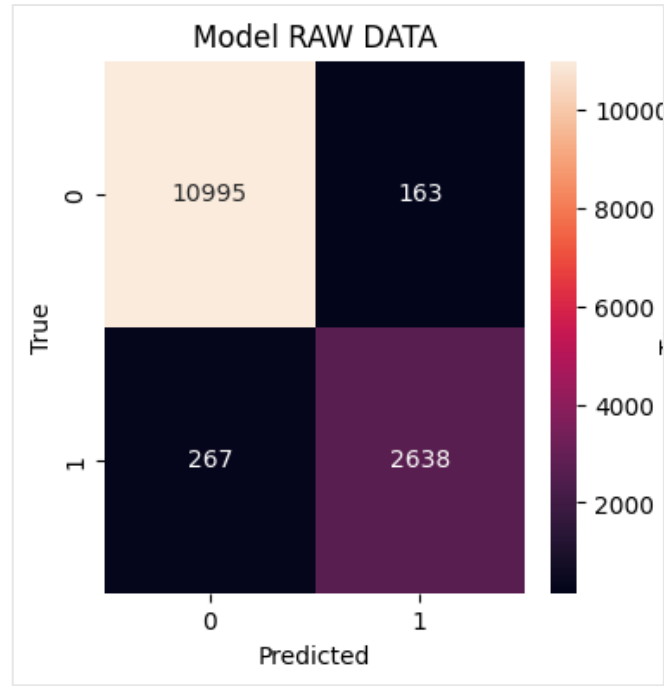


Fig. 27: XGBoost Confusion Matrix – No Balancing.

The XGBoost actual predictions were 10,995 and 2,638, while inaccurate predictions were 163 and 267. The confusion matrix of four cases performed as False Negatives of 267, False Positives of 163, True Negatives of 10,995, and True Positives of 2,638.

Table (13): Specificity, Precision, Recall, Accuracy, and F1-score result of XGBoost -No Balancing

Balancing Method	Specificity	Precision	Recall	Accuracy	F1-score
No Balancing	0.985391	0.96916	0.96942	0.96942	0.969215

As shown in Table 13, XGBoost with the imbalanced data has performed an accuracy of 0.96, which mean that the accurately predicted instance where 0.96. Other metrics performance were specificity of 0.985, a precision of 0.969, a recall of 0.969, an accuracy of 0.969 and the result of F1-score of 0.969.

In opposition, further training was conducted on XGBoost model using other selected balancing methods which are SMOTE, ROS, and

RUS. An assessment of the model's performance has been conducted through confusion matrix, shown in Fig. 28.

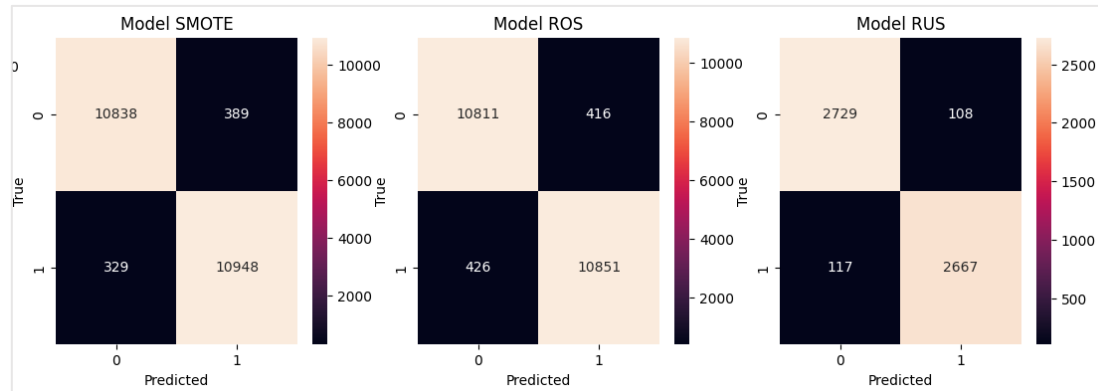


Fig. 28: XGBoost Confusion Matrix – Various Balancing Methods.

As shown in Fig. 28 the XGBoost model with SMOTE balancing technique has resulted with 10,838 and 10,948 actual predictions, in addition to 329 and 389 inaccurate predictions. Also, the four cases of confusion matrix were False Negatives of 329, False Positives of 389, True Negatives of 10,838, and True Positives of 10,948.

Consequently, XGBoost model using ROS balancing method has an actual prediction of 10,811 and 10,851, while inaccurate predictions are 416 and 426. The four cases of confusion matrix were False Negatives of 426, False Positives of 416, True Negatives of 10,811, and True Positives of 10,851.

Lastly, using RUS balancing technique with XGBoost model, the results have 2,729 and 2,667 actual predictions, whereas inaccurate predictions were 108 and 117. The four cases of confusion matrix were False Negatives of 117, False Positives of 108, True Negatives of 2,729, and True Positives of 2,667.

Table (14): Specificity, Precision, Recall, Accuracy, and F1-score result of XGBoost- Various Balancing Methods.

Balancing Method	Specificity	Precision	Recall	Accuracy	F1-score
No Balancing	0.98539	0.96916	0.96942	0.96942	0.96921
SMOTE	0.96535	0.96810	0.96809	0.96809	0.96809
ROS	0.96294	0.96258	0.96258	0.96258	0.96258
RUS	0.96193	0.95997	0.95997	0.95997	0.95997

After evaluating the F1-score performance of XGBoost models trained using different datasets with different balancing techniques, it was found that the XGBoost performed the best with the full dataset with no balancing method, it has yielded the highest F1-score of 0.969.

4.2.6. Hybrid Model

As explained in chapter 3, a hybrid model can be highly effective for credit scoring by combining the predictions of multiple traditional models, leveraging their strengths, while mitigating their weaknesses. The goal of this methodology is to enhance the performance and generalization capabilities of the overall system. This diversity in learning helps the hybrid model to cover a wide range of perspectives, thus ensuring a comprehensive and detailed understanding of the underlying patterns and relationships in the data.

In this experiment, considering the five ML models executed earlier, which are LR, RF, NB, SVM, and XGBoost. A combination of the three top performance models were considered for the hybrid model. As a results, the hybrid model combined LR, SVM, and XGboost.

Furthermore, majority voting technique was employed in hybrid models. The method takes predictions from each individual model, assigning a "vote" for each possible class. The class with the majority votes is considered the final output. This approach allows the hybrid model to benefit from the strengths of each model, ensuring that the final decision is made based on a comprehensive understanding of the data. Moreover, Voting Classifier helps in reducing the impact of any single weak classifier, thus offering a more robust and accurate model.

The modeling started through combining the models with their best parameters on the loan default training dataset that was defined previously, followed by identifying the voting classifier. To begin with, the first results were obtained from the hybrid model on full dataset where no balancing methods were considered. The model's performance was presented through confusion matrix, shown in Fig. 29.

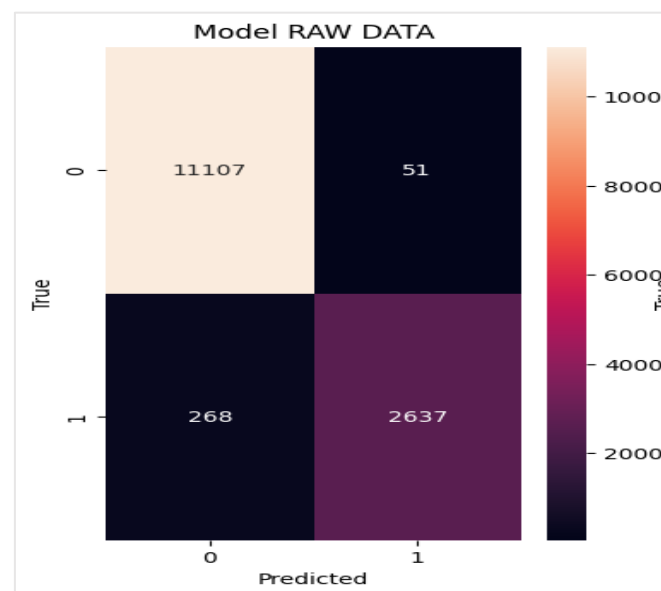


Fig. 29: Hybrid Model Confusion Matrix – No Balancing.

Based on the confusion matrix results, the hybrid model performed actual predictions of 11,107 and 2,637, while inaccurate predictions are 51 and 268. The confusion matrix four cases performed as False Negatives of

268, False Positives of 51, True Negatives of 11,107, and True Positives of 2,637.

Table (15): Specificity, Precision, Recall, Accuracy, and F1-score result of Hybrid Model -No Balancing

Balancing Method	Specificity	Precision	Recall	Accuracy	F1-score
No Balancing	0.99542	0.9773871	0.977316	0.97731	0.97698

As presented in Table 15, hybrid model with no balancing performed an accuracy of 0.977. This means that the percentage of accurately predicted instances is 0.97. The overall performance specificity of 0.995, a precision of 0.977, a recall of 0.977, an accuracy of 0.977 and the result of F1-score of 0.976.

Nonetheless, the experiment also includes training the hybrid model using the selected balancing methods which are SMOTE, ROS, and RUS. Applying all the required steps to executing the models In addition, the models were trained on a training dataset using cross-validation to reduce the risk of overfitting; K-Fold cross validation was performed which involves splitting the training dataset into k equal-sized subsets called folds, then training the models k times, each time using (k-1) folds for training and the remaining fold for validation. The models' performance was assessed using various evaluation metrics. Fig. 30 presents confusion matrices were performed to evaluate performance of hybrid models using various data balancing techniques.

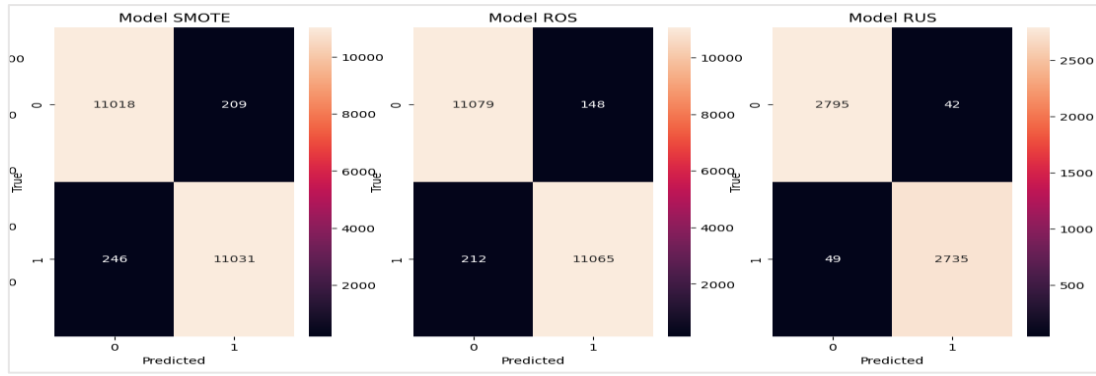


Fig. 30: Hybrid Model Confusion Matrix – Various Balancing Methods.

As shown in Fig. 30 the first confusion matrix illustrates the performing of hybrid models using SMOTE balancing, it results with actual predictions of 11,018 and 11,031, where inaccurate predictions were 209 and 246. The four cases of confusion matrix were False Negatives of 246, False Positives of 209, True Negatives of 11,018, and True Positives of 11,031.

Next, hybrid model using ROS balancing method, the actual predictions were 11,079 and 11,065, while inaccurate predictions are 148 and 212. The four cases of confusion matrix were False Negatives of 212, False Positives of 148, True Negatives of 11,079, and True Positives of 11,065.

Last, hybrid model using RUS has resulted with actual predictions were 2,795 and 2,735, whereas inaccurate predictions are 42 and 49. The four cases of confusion matrix were False Negatives of 49, False Positives of 42, True Negatives of 2,795, and True Positives of 2,735.

Table (16): Specificity, Precision, Recall, Accuracy, and F1-score result of Hybrid Model- Various Balancing Methods.

Balancing Method	Specificity	Precision	Recall	Accuracy	F1-score
No Balancing	0.99542	0.9773871	0.977316	0.97731	0.97698
SMOTE	0.98138	0.979786	0.979781	0.97978	0.97978
ROS	0.98681	0.984018	0.984002	0.98400	0.98400
RUS	0.98519	0.983813	0.9838107	0.98381	0.98381

As shown in Table 16, a performance comparison for the hybrid models using various balancing technique. The models' F1-scores were compared to determine the best performing model. Upon analysing the F1-score performance of the hybrid models trained using different data balancing techniques, ROS was found to yield the highest F1-score of 0.984.

4.3. Discussion

This section will provide and analyse the findings derived from the conducted experiment. The portion is divided into two separate parts. The first part evaluates the outcomes achieved through the utilization of different balancing techniques and their influence on the models. The next part involves a comparative analysis of the outcomes obtained from the five machine learning models, together with an evaluation of their respective performance in comparison to the hybrid model.

4.3.1. Comparison Analysis on Data Balancing Methods

As mentioned in chapter two, a comprehensive literature review was conducted on previous studies covering ML algorithms for credit scoring using LendingClub data. It was found that limited studies have used balancing techniques in balancing the selected data before deploying ML models, yet most of the studies have not discussed the data imbalance in their findings. Accordingly, part of this research aimed to compare the ML models on LendingClub data using different balancing techniques, such as SMOTE, ROS, and RUS, since these techniques has proven its effectiveness through the years. Also, to identify the most appropriate data balancing technique that can enhance the performance of the ML model that developed to predict the credit scoring based on LendingClub data.

Whereas this sub-section will evaluate and compare the performance of different data balancing techniques for training the selected ML models, against training the same model without data balancing.

Table (17): F1-score result of Different Models - Various Balancing Methods.

ML Model	No Balancing	SMOTE	ROS	RUS
LR	0.9895	0.9898	0.9828	0.9829
RF	0.9602	0.9714	0.9848	0.9498
NB	0.9422	0.9462	0.9333	0.9389
SVM	0.9901	0.9908	0.9843	0.9843
XGBoost	0.9692	0.9680	0.9625	0.9599
Hybrid	0.9769	0.9798	0.9840	0.9838

In order to compare the effectiveness of various data balancing techniques, and assess their impact on the ML models performance, a comparative analysis focused on F1-score was presented in Table 17. Through analysing the performance of the individual ML models trained which are LR, RF, NB, SVM, and hybrid (LR+SVM+XGBoost), excluding XGBoost model, using different data balancing techniques. SVM was found to yield the highest F1-score of 0.9908 using SMOTE. By comparing the three balancing techniques, SMOTE performed better in three models out of 5 models. Followed by ROS where it has improved the performance of RF and hybrid model. Generally, this indicates that SMOTE balancing technique can provide better performance when individual ML models are used for credit score prediction for P2P lending.

Table (18): Experiment Results Ranking.

No.	ML Method	Accuracy	F1-score	Balancing Method
1	SVM	0.99089	0.99089	SMOTE
2	LR	0.98986	0.9898	SMOTE
3	RF	0.98484	0.98484	ROS
4	Hybrid (LR+SVM+ XGBoost)	0.98400	0.98400	ROS
5	XGBoost	0.96942	0.96921	No Balancing
6	NB	0.94627	0.94625	SMOTE

The results of the selected ML models where ranked based on their best performance in the conducted experiment, presented in Table 18. The SVM model employing the SMOTE approach has been ranked as the top performer among the other models. Subsequently, LR was employed, demonstrating a notable improvement in performance by the utilization of the SMOTE. Both RF and Hybrid (LR+SVM+XGBoost) models exhibited similar performance, with both achieving their highest accuracy when employing the ROS technique. XGBoost, known for its capability to adapt to imbalanced data, achieved a 96.9% accuracy rate without employing any balancing techniques, thus securing the 5th position. XGboost is an ensemble model, which has a certain property that make them less sensitive to class imbalances compared to other individual ML models. Lastly, the NB model has the lowest performance compared to other models, achieving an accuracy rate of 94.6%.

Nevertheless, NB model may not be the best choice for predictive modeling tasks, including P2P lending prediction, due to its assumption of independence where NB assumes that the features used in the classification are conditionally independent, given the class label. In P2P lending prediction, this assumption may not hold because the features often have

complex relationships. In addition, NB has limited model expressiveness compared to more complex algorithms. It may not capture intricate patterns and interactions within the data, which are important in P2P lending prediction.

On the other hand, despite the initial high expectations surrounding the hybrid model and its anticipated superior performance. It is commonly understood that hybrid models frequently leverage the inherent strengths of different models in order to effectively address class imbalances. The hybrid model utilized in this study incorporates the combination of LR, SVM, and XGBoost algorithms. The results of this experiment indicate that the hybrid model did not demonstrate competitive performance when compared to individual traditional algorithms such as SVM and LR.

In summary, since the SVM model yield the highest F1-score among other models, SVM with SMOTE balancing is the recommended method for predicting credit scoring for the P2P lending.

By further analysing the performance of balancing techniques and their impact on the ML models. The analysis will commence by evaluating the efficacy of the balancing methods on the logistic regression (LR) model. As illustrated in Figure 31, the application of balancing techniques has a negligible effect on the performance of the LR model. The disparity between the utilization of the SMOTE and the absence of any balancing technique is negligible, with SMOTE revealing better results by a margin of 0.03%.

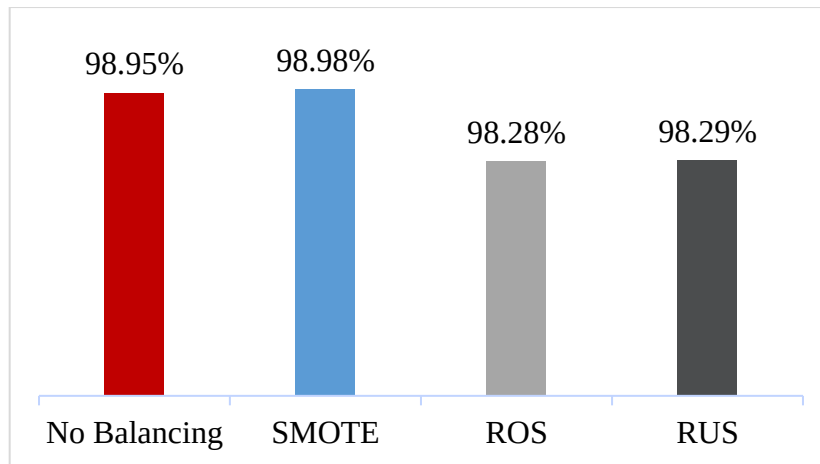


Fig. 31: Logistic Regression Performance Comparative Balancing Methods.

Next, the performance of the Random Forest (RF) model will be assessed by a comparative analysis of different balancing methods. The results of the experiment indicate that the implementation of balancing techniques has had a positive impact on the performance of RF models, with the exception of the RUS technique. However, by doing a comparative analysis of the performance of the RF model utilizing ROS technique compared to the performance of the RF model trained on the whole dataset without any technique applied. The performance has experienced a significant impact, resulting in an increase in accuracy from 96% to 98%, as demonstrated in Fig. 32.

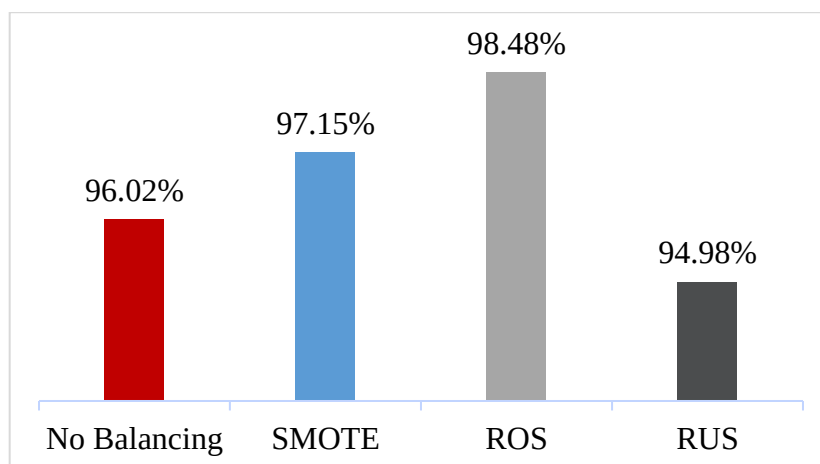


Fig. 32: Random Forest Performance Comparative Balancing Methods.

Moreover, as shown in Figure 33, the performance of the Naïve Bayes (NB) classifier was observed to be marginally affected when various balancing approaches were applied. According to the experimental results, the application of the SMOTE technique compared to others resulted in a slight enhancement of the NB model. Specifically, the SMOTE technique led to a 0.4% improvement in the performance of the NB model compared to the use of the imbalanced dataset.

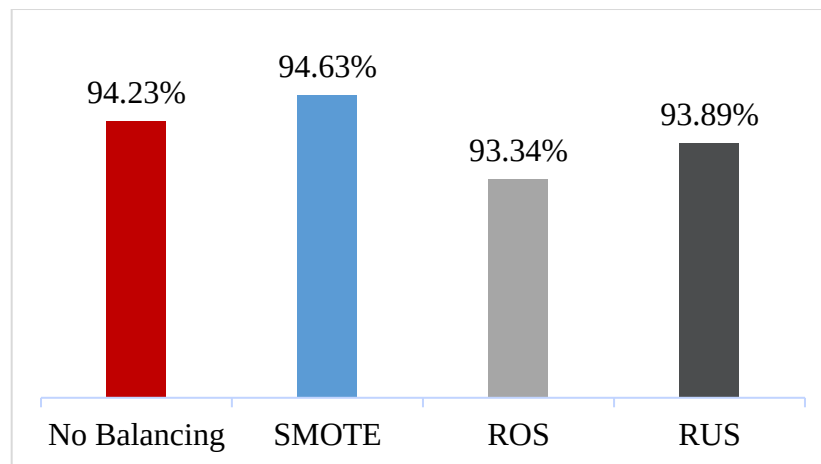


Fig. 33: Naïve Bayes Performance Comparative Balancing Methods.

On the other hand, the support vector machine (SVM) has exhibited a slightly favourable impact when diverse balancing techniques were employed, particularly with the utilization of SMOTE. The application of the SMOTE technique has yielded a slight enhancement in the accuracy of SVM model. The utilization of the SMOTE resulted in a significant enhancement in the performance of the SVM algorithm, achieving a remarkable accuracy rate of 99.09%. On the other hand, the SVM achieved a classification accuracy of 99.01% without utilizing any form of balancing technique, showed in Fig. 34.

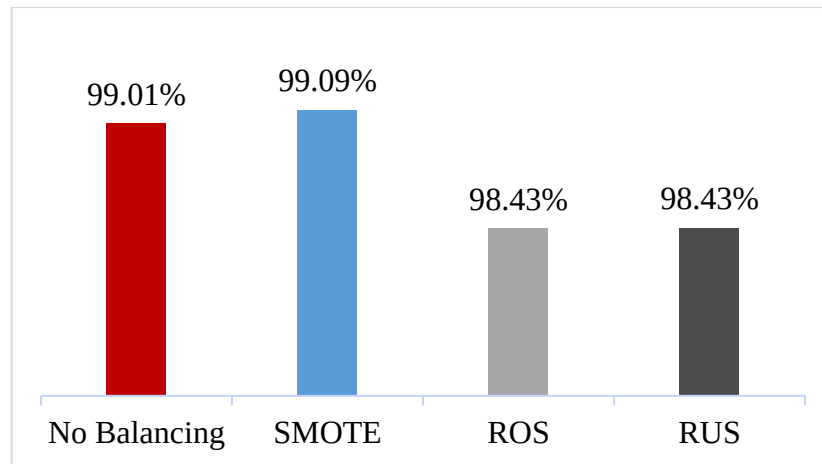


Fig. 34: Support Vector Machine Performance Comparative Balancing Methods.

Last but not least, it is worth noting that XGBoost achieved the maximum level of accuracy in the absence of any balancing approaches. The findings presented in, Figure 35, indicate that XGBoost has adaptability when dealing with imbalanced datasets.

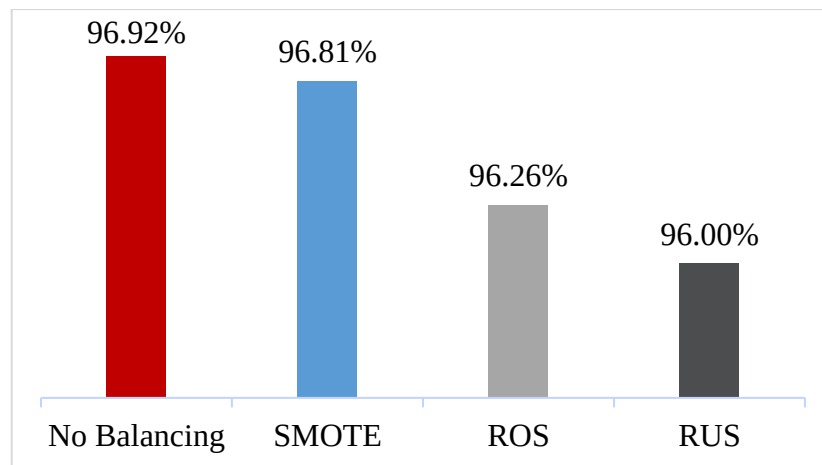


Fig. 35: XGBoost Performance Comparative Balancing Methods.

Finally, the hybrid model exhibited superior performance when ROS technique employed. Although, it is well-established that hybrid models frequently leverage the inherent capabilities of different models in order to effectively address class disparities, yet the performance of the hybrid model has improved using ROS. The hybrid model employed in this study

combines the algorithms of LR, SVM, and XGboost. Fig. 36, presents the performance results of hybrid model using various balancing techniques.

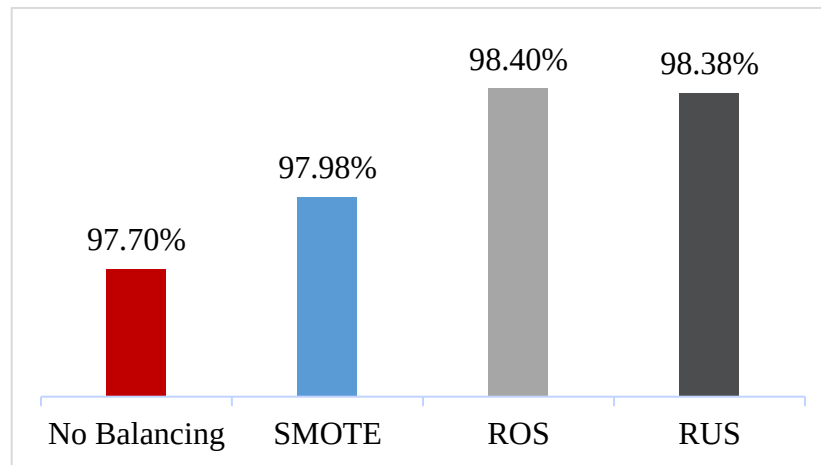


Fig. 36: Hybrid Model Performance Comparative Balancing Methods.

In general, the utilization of balancing approaches has a notable influence on the performance of the RF and NB models, while it has a modest positive effect on the LR, SVM, and hybrid models. The most effective method for credit scoring prediction, out of all the models examined, was found to be the SVM model with the utilization of SMOTE technique, as shown in Fig. 37. This experiment indicates that individual ML models generally achieve optimal performance when employing balancing approaches, with the exception of ensemble algorithms like XGBoost.

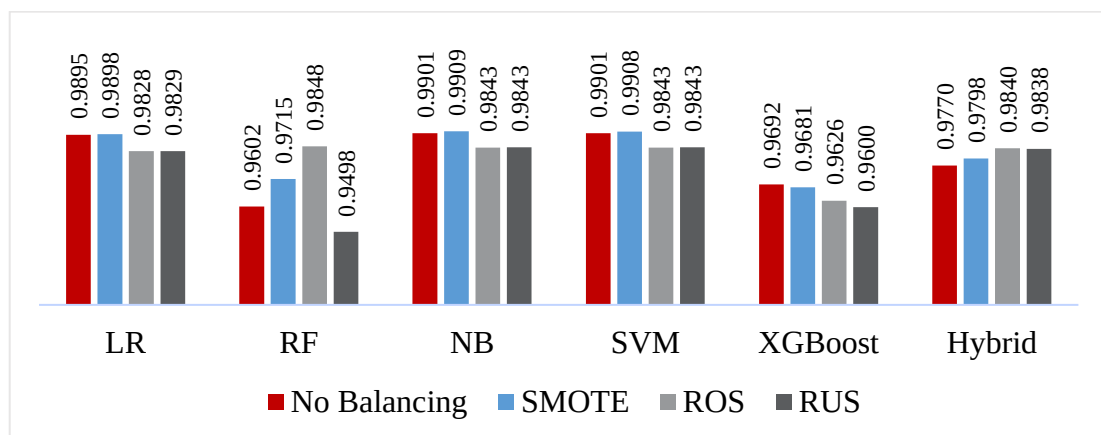


Fig. 37: Performance Comparative - F1-score Various Balancing Methods.

4.3.2. Comparison Analysis on Computational Time

While comparing one classification method to another that differs in functionality, by assessing their computational time complexity. This refers to how the time required to execute an algorithm scale with the size of the input data and relevant parameters. The computing time consumed was measured for all balancing scenarios, as shown in Table (19).

Table (19): Methods' Computation Time.

ML Method	No Balance	SMOTE	ROS	RUS
LR	1.82s	1.74s	1.314s	0.618s
RF	28.999s	45.894s	41.379s	9.17s
NB	0.346s	0.682s	0.425s	0.382s
SVM	143.051s	823.263s	503.863s	20.696s
XGBoost	3.209s	5.588s	6.079s	1.481s
Ensemble	187.521s	828.119s	558.226s	22.389s

NB was the most computationally efficient, knowing that NB is a straightforward and simple probabilistic model. It makes a strong assumption of feature independence, which simplifies probability calculation. The computational complexity of prediction is reduced as a result of this simplicity. The second most efficient computational time is LR. LR solve the linear optimization problem by assuming a linear relationship between the input features and the log-odds of the target variable. SVM, on the other hand, had a long computation time ranging from 20.696s to 823.263s. In SVM, the computational time grows significantly as the size of the dataset increases. The model aims to find the optimal hyperplane that maximizes the margin between classes, this can lead to intensive computation. Furthermore, SVM requires storing the entire dataset in memory, especially when utilizing the scikit-learn library,

resulting in time-consuming calculation. While RF and XGBoost are often more efficient than traditional decision trees and can handle larger datasets effectively.

Furthermore, the computational time can be influenced by different balancing techniques for imbalanced datasets. In the scenario of no balancing, the computational time is fast, yet the performance of the model may be poor. Additionally, because the RUS balancing technique removes some instances from the majority class, the data size is reduced, resulting in a reduction in computational time. Nonetheless, SMOTE and ROS tend to have the most expensive computing time due to the increase in the dataset size.

4.3.3. Performance Comparison Analysis

Through a comparative analysis of the performance of various machine learning models, namely Logistic Regression (LR), Random Forest (RF), Naive Bayes (NB), Support Vector Machines (SVM), XGBoost, and hybrid combination of LR, SVM, XGBoost, utilizing different data balancing techniques, it was found that SVM demonstrated the highest level of accuracy in classification tasks. In particular, SVM achieved an F1-score of 0.9908, indicating its superior performance in accurately classifying the data. This finding aligns with the outcomes published by Qionglin Shan and Mikael Nilsson in their study conducted in (2018). According to their research, it has been found that the performance of SVM outperformed that of other competing techniques in many applications. However, despite the fact that Shan and Nilsson's study utilized the same dataset employed in this experiment. The results of this experiment have demonstrated superior performance in SVM compared to their study. The study conducted by the researchers yielded a

level of accuracy of 97%, whereas the present experiment achieved a higher accuracy rate of 99.5%.

Additionally, LR exhibited a commendable F1-score of 0.989, while XGBoost achieved a lower F1-score of 0.969. This finding demonstrates a potential link with the outcomes obtained by Jiaying Sun (2022) and Teply and Polena (2020). Both studies have demonstrated that the performance of LR outperformed that of other methods offered in their respective investigations. Even though it is generally recognized that LR is the most accurate model among traditional algorithms, it fails to compete with SVM in this research.

It worth mentioning that RF model has performed its best using ROS balancing techniques with a F1-score of 0.984. However, it fails behind in competing with SVM. This is opposite to the studies done by Chen, et al. (2021), and Yiheng Li and Weidong Chen (2020) where RF has best performance for default predication using the same LendingClub data.

However, the study conducted by Zhu, et al. (2019) yielded comparable findings to the present research. Their study demonstrated that the Random Forest (RF) algorithm exhibited superior performance compared to other algorithms, achieving an accuracy rate of 98.00%. Notably, this was accomplished by the implementation of SMOTE balancing techniques (Zhu, Qiu, Ergu, Ying, & Liu, 2019). However, the present study has yielded comparable findings, with the RF model attaining an accuracy rate of 98.48% through the use of the oversampling technique. It is important to note, however, that the comparison between the two studies is subject to variation due to the utilization of different datasets (different period).

Furthermore, this finding is contrary to the research conducted by Fiji Fu (2017), who also compared the hybrid model to traditional models, where the researcher obtained the highest performance accuracy from the

hybrid model, combining Neural Network + Random Forest compared to other models in their study. However, this study indicate that single ML models have outperformed the hybrid model conducted in this experiment.

Yet, it is important to recognize that the performance observed in the current experiment was significantly better compared to the outcomes reported in Fu's study, The comparison between the two research is open to fluctuation due to the exploitation of various datasets, namely from different periods. Figure 38, present overview comparison between this experiment results in comparison with other previous studies.

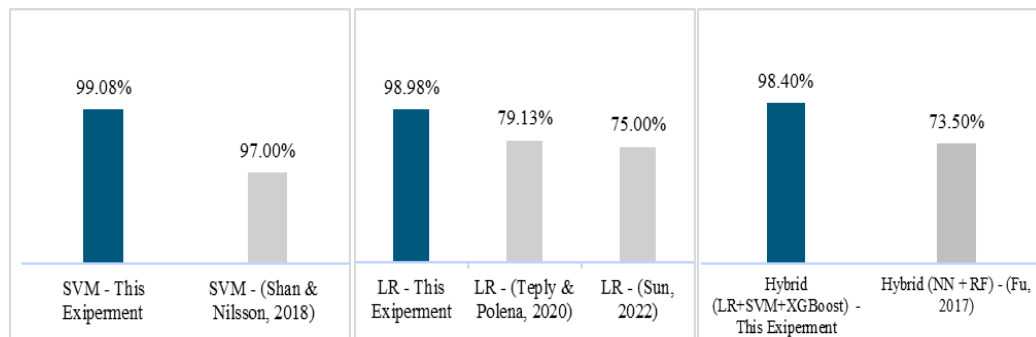


Fig. 38: An Overview Comparison on Results including Results of Previous Studies.

Overall, the findings show that Support vector machine (SVM) with SMOTE balancing techniques provides the most accurate performance followed by Logistic Regression (LR), Random Forest (RF), hybrid combination of LR+SVM+XGBoost, XGBoost and Navie Bayes (NB), as shown in Fig. 39.

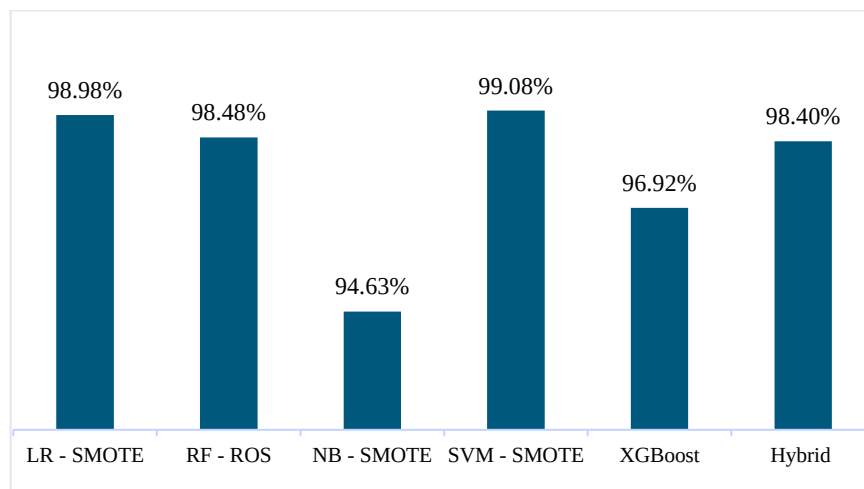


Fig. 39: Experiment Results Summary.

The main contribution of this research is the reapplication of different ML techniques, a hybrid model, and comprehensive data preprocessing using P2P lending dataset from "LendingClub". The findings from SVM model are consist with previous studies.

The performance of the credit scoring models has been greatly enhanced as a result of this research, which is in contrast to the other findings reported in chapter two. Thus, due to the extensive data pre-processing, feature selection, features engineering, and other pre-processing techniques, the procedures undertaken in preparing the data have substantially improved the performance of predicting defaulters or Non-defaulters for P2P lending. This is because of the comprehensive nature of the data pre-processing.

Chapter Five

Conclusion and Future Work

5.1. Conclusion

5.2 Future Work

Chapter Five

Conclusion and Future Work

5.1. Conclusion

The study embarks on a comprehensive exploration of machine learning models for credit scoring prediction in the realm of P2P lending platforms. The main objective was to improve the accuracy and efficacy of credit scoring prediction to classify defaulter and non-defaulter loan applications, which will help investors in making better decision. Through evaluating the selected ML algorithms proposed for this research which are Logistic Regression (LR), Random Forest (RF), Naive Bayes (NB), Support Vector Machine (SVM), XGBoost, and a hybrid model using P2P lending dataset from "LendingClub" publicly available.

The findings of this study emphasize the significance of accurate credit scoring in P2P lending, specifically in light of the distinctive attributes of the platform, including limited credit history of borrowers and varying economic conditions. By employing extensive preprocessing techniques, conducting feature engineering, and implementing dimensionality reduction, the dataset was refined to capture essential borrower attributes and loan features, thus creating a strong foundation for model training and evaluation. In addition to utilize various balancing techniques. The evaluation of the performance was conducted by diverse range of performance metrics, which are accuracy, precision, recall, and F1-score.

The results have reveals that the balancing techniques have exhibits notable improvements on the individual ML models performance, except

for NB. Among the three balancing techniques, SMOTE was the most suitable techniques that have improved the performance of the majority of the individual models. As a results, the utilization of balancing techniques was deemed necessary in implementing a ML model for credit scoring predication, knowing that the SVM model achieved the highest performance of 99.1% using SMOTE balancing techniques. These results are aligned with previous studies conducted by (Shan & Nilsson, 2018) obtaining highest accuracy with SVM achieving 97%, however, the result of this study has outperformed their study.

By conducting a comparative analysis of the models' performance, the results indicate that the SVM model with SMOTE balancing techniques exhibits the highest level of accuracy, achieving a rate of 99.1%. This is closely followed by the LR model, which achieves an accuracy of 98.9%. While the RF model achieves an accuracy of 98.5%, which is slightly better than the performance of hybrid model, which achieved an accuracy of 98.4%. The XGBoost model demonstrates a lower accuracy of 96.9%, and the NB model exhibits the lowest accuracy among the models, with a rate of 94.6%. This finding is consistent with the previous literature.

These findings have implications for a range of stakeholders in the P2P lending ecosystem. Platform operators have the opportunity to utilize the findings to improve their credit risk assessment mechanisms, hence promoting a more secure lending environment.

In the rapidly evolving landscape of P2P lending, this study contributes to the ongoing dialogue on credit scoring methodology by offering comprehensive and comparative analysis of ML models. Although the SVM model demonstrated significant potential, it is crucial to recognize that the effectiveness of such models may differ depending on the datasets used, economic circumstances, and regulatory frameworks.

5.2. Future Work

Future research can be delved deeper into refining hybrid models and exploring ensemble techniques to further optimize predictive accuracy and interpretability.

This study highlights the potential of ML models in improving credit scoring prediction for P2P lending, emphasizing its transformative nature. By adopting innovative practices while maintaining a commitment to transparency, the P2P lending sector has the potential to further develop as a more inclusive, secure, and efficient platform for borrowing and lending, therefore transforming the contemporary financial landscape.

Furthermore, the research findings can be employed to apply ML techniques to a different dataset that is tailored to the specific behaviours and characteristics of Bahrain and GCC citizens.

Knowing that since the Central Bank of Bahrain (CBB) has introduced the Conventional Financing Based Crowdfunding Regulation, followed by the introduction of Shariah-compliant Financing Crowdfunding Platform (P2B) two years later (Central Bank of Bahrain, 2023). The financial-based crowdfunding market in Bahrain is growing, today its limited to SMEs and real estate projects based in Bahrain, however, it is expected to offer for individual borrowers soon. Nowadays, there are number of P2P lending platforms in Bahrain that are in different development stages, such as Suyula, Amal Invest, Raseed, and Tasleefa (Crowdfunding in Bahrain, 2022). The findings of this research can contribute to the grow of the FinTech industry in Bahrain.

References

- Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random Forests and Decision Trees. *International Journal of Computer Science Issues*, 272-278.
- Baker, M. R., Mahmood, Z. N., & Shaker, E. H. (2022). Ensemble Learning with Supervised Machine Learning Models to Predict Credit Card Fraud Transactions. *Revue d'Intelligence Artificielle*, 36(4), 509-518. doi:<https://doi.org/10.18280/ria.360401>
- Bekkar, M., Djemaa, H. K., & Alitouche, T. A. (2013). Evaluation Measures for Models Assessment over Imbalanced Data Sets. *Journal of Information Engineering and Applications*, 3, 27-38.
- Brown, K., & Moles, P. (2016). *Credit Risk Management*. Great Britain: Edinburgh Business School.
- Byanjankar, A., Heikkilä, M., & Mezei, J. (2015). Predicting Credit Risk in Peer-to-Peer Lending: A Neural Network Approach. *2015 IEEE Symposium Series on Computational Intelligence*, 719-725.
- Byanjankar, A., Mezei, J., & Heikkilä, M. (2021). Data-driven optimization of peer-to-peer lending portfolios based on the expected value framework. *Intelligent Systems in Accounting, Finance and Management*, 119-129.
- Central Bank of Bahrain. (2023). *Central Bank of Bahrain*. Retrieved from Central Bank of Bahrain: <https://www.cbb.gov.bh/fintech/>
- CFI. (2022, December 12). *Peer-to-Peer Lending*. Retrieved from CFI: <https://corporatefinanceinstitute.com/resources/wealth-management/peer-to-peer-lending/>
- CFI Education Inc. (2023, May 23). *Peer-to-Peer Lending*. Retrieved from Corporate Finance Institute:

<https://corporatefinanceinstitute.com/resources/wealth-management/peer-to-peer-lending/>

- Chang, A.-H., Yang, L.-K., Tsaih, R.-H., & Lin, S.-K. (2022). Machine learning and artificial neural networks to construct P2P lending credit-scoring model: A case using LendingClub data. *Quantitative Finance and Economics*, 6(2), 303-325. doi:10.3934/QFE.2022013
- Chen, T. (n.d.). *Story and Lessons Behind the Evolution of XGBoost*. Retrieved from nttrungmt-wiki: <https://sites.google.com/site/nttrungmtwiki/home/it/data-science---python/xgboost/story-and-lessons-behind-the-evolution-of-xgboost>
- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. doi:<https://doi.org/10.48550/arXiv.1603.02754>
- Chen, Y.-R., Leu, J.-S., Huang, S.-A., Wang, J.-T., & Takada, J.-I. (2021). Predicting Default Risk on Peer-to-Peer Lending Imbalanced Datasets. *IEEE Access*, 73103-73109.
- Deris, A. M., Zain, A. M., & Sallehuddin, R. (2011). Overview of Support Vector Machine in Modeling Machining Performances. *Procedia Engineering*, 308-312.
- Fu, Y. (2017). Combination of Random Forests and Neural Networks in Social Lending. *Journal of Financial Risk Management*, 6, 418-426. doi:<https://doi.org/10.4236/jfrm.2017.64030>
- Gao, B., & Balyan, V. (2022). Construction of a financial default risk prediction model based on the LightGBM algorithm. *ournal of Intelligent Systems*, 767-779. doi:<https://doi.org/10.1515/jisys-2022-0036>

- Gestel, T. V., & Baesens, B. (2009). *Credit Risk Management: Basic Concepts: Financial Risk Components, Rating Analysis, Models, Economic and Regulatory*. New York: Oxford University Press Inc.
- Goldbloom, A. (2015). *What algorithms are most successful on Kaggle?* Retrieved from Kaggle: <https://www.kaggle.com/code/antgoldbloom/what-algorithms-are-most-successful-on-kaggle/notebook>
- Guegan, D., & Hassani, B. (2018). Regulatory learning: How to supervise machine learning models? An application to credit scoring. *The Journal of Finance and Data Science*, 157-171.
- Guo, Y. (2019). Credit Risk Assessment of P2P Lending Platform towards Big Data based on BP Neural Network. *J. Vis. Communcation & Image Representation*. doi:<https://doi.org/10.1016/j.jvcir.2019.102730>
- He, H., & Nawata, K. (2019). The Application of Machine Learning Algorithms in Predicting the Borrower's Default Risk in Online Peer-to-Peer Lending. *International Conference on Education, Management and Information Technology*, 89-95.
- IBM SPSS Modeler 18.3 In-Database Mining Guide. (International Business Machines Corporation). New York: International Business Machines Corporation.
- Ince, H., & Aktan, B. (2009). A Comparison of Data Mining Techniques for Credit Scoring in Banking: a Managerial Perspective. *Journal of Business Economics and Management*, 233-240.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning*. New York: Springer.
- Kaggle. (2022). *Kaggle's 2022 State of Data Science and Machine Learning survey*. Kaggle. Retrieved from <https://www.kaggle.com/kaggle-survey-2022>

- Kaggle. (2023, February 02). *Kaggle*. Retrieved from Kaggle: <https://www.kaggle.com/>
- Khatir, H. A., Almustfa, A., & Bee, M. (2022). Machine Learning Models and Data-Balancing Techniques for Credit Scoring: What Is the Best Combination? *Risks*. doi:<https://doi.org/10.3390/>
- Klaftt, M. (2008). Online Peer-to-Peer Lending: A Lenders' Perspective. *SSRN Electronic Journal*, NA.
- Klaftt, M. (2008). Online Peer-to-Peer Lending: A Lenders' Perspective. *SSRN Electronic Journal*.
- Ko, P.-C., Lin, P.-C., Do, H.-T., & Huang, Y.-F. (2022). P2P Lending Default Prediction Based on AI and Statistical Models. *entropy*, 24-801.
- Kumar, A., Sharma, S., & Mahd, M. (2021). Machine Learning (ML) Technologies for Digital Credit Scoring in Rural Finance: A Literature Review. *risks*(9). doi:<https://doi.org/10.3390/risks9110192>
- LendingClub. (2023). *Lending Club*. Retrieved from Lending Club: <https://www.lendingclub.com/company/about-us>
- LendingClub. (2023, Jun 08). *What Do the Different Note Statuses Mean?* Retrieved from LendingClub: <https://www.lendingclub.com/help/investing-faq/what-do-the-different-note-statuses-mean>
- Lenz, R. (2016). Peer-to-Peer Lending: Opportunities and Risks. *European Journal of Risk Regulation*, 7(4), 688-700.
- Leo, M., Sharma, S., & Maddulety, K. (2019). Machine Learning in Banking Risk Management:. *Risks*, 1-22.
- Leung, K., Cheong, F., & Cheong, C. (2008). Consumer Credit Scoring using an Artificial Immune System. *IEEE Congress on Evolutionary Computation*, 3377-3384.

- Li , Y., & Chen, W. (2020). A Comparative Performance Assessment of Ensemble Learning for Credit Scoring. *Mathematics*, 1-19. doi:10.3390/math8101756
- Mester, L. J. (1997). What's the Point of Credit Scoring? *Business Review*, 3-16.
- Mpofu, T. P., & Mukosera, M. (2014). Credit Scoring Techniques: A Survey. *International Journal of Science and Research (IJSR)* , 165-168.
- Nguyen, T.-T., Khuat, T.-S., Ngo, T.-T.-T., Nguyen, H.-N., & Tran, M.-D. (2019). Improve Risk Prediction in Online Lending (P2P) Using Feature Selection and Deep Learning. *IJCSNS International Journal of Computer Science and Network Security*, 19(11), 216-222.
- P2P Market Data. (2023, June 6). *The History of Peer-to-Peer Lending*. Retrieved from P2P Market Data: <https://p2pmarketdata.com/articles/p2p-lending-history/>
- Padimi, V., Telu, V. S., & Ningombam, D. D. (2022). Applying Machine Learning Techniques To Maximize The Performance of Loan Default Prediction. *Journal of Neutrosophic and Fuzzy Systems (JNFS)*, 2, 44-56. doi:10.54216/JNFS.020204
- Perzhanovskiy, N. (2022, April 20). *Crowdfunding in Bahrain*. Retrieved from LenderKit: <https://lenderkit.com/blog/bahrain-crowdfunding-business/>
- Qian, M., & Hu, F. (2019). An Empirical Study on Prediction of the Default Risk on P2P Lending Platform. *IOP Conference Series: Materials Science and Engineering*, 490(6). doi:10.1088/1757-899X/490/6/062048
- Richardson, M. (2009). *"Principal component analysis"*. Retrieved from "Principal component analysis": <http://aurora.troja.mff.cuni.cz/nemec/idl/09bonus/pca.pdf>

- Russell, S. J., & Norvig, P. (2009). *Artificial Intelligence: A Modern Approach* (Third Edition ed.). Essex : Pearson Education Limited.
- Saputra, O., Faturhman, T., & Kaderi Wiryono, S. (2021, March). Social Media Data to Improve Credit Scoring Accuracy with A Data Mining Approach Based on Support Vector Machine: Case Study of An Online Peer to Peer Lending in Indonesia. *International Journal of Accounting, Finance and Business (IJAFB)*, 6(32), 1-14.
- Schmitt, M. (2022). Deep Learning vs. Gradient Boosting: Benchmarking state-of-the-art machine learning algorithms for credit scoring. *ArXiv*. doi:10.48550/arXiv.2205.10535
- Shan, Q., & Nilsson, M. (2018). *Credit risk analysis with machine learning techniques in peer-to-peer lending market*. Stockholm: Stockholm Business School.
- Spuchľáková, E., Valášková, K., & Adamko, P. (2015). The Credit Risk and its Measurement, Hedging and Monitoring. *Procedia Economics and Finance*, 675-681.
- Sun, J. (2022). Data Mining Techniques to Predict Default in LendingClub. *Journal of Economics, Business and Management*, 10, 60-64. doi:10.18178/joebm.2022.10.1.674
- Tao, W., & Chang, D. (2019). Credit Risk Assessment of P2P Lending Borrowers based on SVM. *Advances in Economics, Business and Management Research*, 80, 182-190.
- Teplý, P., & Polena, M. (2020). Best classification algorithms in peer-to-peer lending. *The North American Journal of Economics and Finance*, 51, NA. doi:https://doi.org/10.1016/j.najef.2019.01.001
- The Business Research Company . (2022). *Peer-to-Peer (P2P) Lending Global Market Report 2022: Ukraine-Russia War Impact* . Lyon: Reportlinker .

- Wendler, T., & Gröttrup, S. (2021). *Data Mining with SPSS Modeler*. Cham: Springer.
- West, D. (2000). Neural Network Credit Scoring Models. *Computers & Operations Research*, 1131-1152.
- Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. *Practical application of knowledge discovery and data mining*, 29-40.
- Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining.
- Yang, Y. (2017). *Temporal Data Mining via Unsupervised Ensemble Learning*. Elsevier. doi:Temporal Data Mining via Unsupervised Ensemble Learning
- Zhang, Y. (2010). *New Advances in Machine Learning*. Rijeka: InTech.
- Zhou, Z.-H. (2012). *Ensemble Methods: Foundations and Algorithms* (1st ed.). New York: Chapman and Hall (CRC). doi:<https://doi.org/10.1201/b12207>
- Zhu, L., Qiu, D., Ergu, D., Ying, C., & Liu, K. (2019). A study on predicting loan default based on the random forest algorithm. *Procedia Computer Science*, 503-513. doi:10.1016/j.procs.2019.12.017
- Zhua, L., Qiu, D., Ergu, D., Ying, C., & Liu, K. (2019). A study on predicting loan default based on the random forest algorithm. *Procedia Computer Science*, 503-513. doi:10.1016/j.procs.2019.12.017

Appendices

Appendix A – Features Dictionary

Features	Description	Type	possible values
id	A unique LC assigned ID for the loan listing.	Categorical	-
member_id	A unique LC assigned Id for the borrower member.	Categorical	-
loan_amnt	The listed amount of the loan applied for by the borrower, or adjusted by the credit department.	Numeric	(500 - 40,000)
funded_amnt	The total amount committed to that loan at that point in time.	Numeric	(1,000 - 40,000)
funded_amnt_inv	The total amount committed by investors for that loan at that point in time.	Numeric	(775 - 40,000)
term	The number of payments on the loan. Values are in months.	Categorical	36 months, 60 months
int_rate	Interest Rate on the loan	Numeric	(5.32 - 30.99)
installment	The monthly payment owed by the borrower if the loan originates.	Numeric	(21.6 - 1451.12)
grade	LC assigned loan grade	Categorical	A - G
sub_grade	LC assigned loan subgrade	Categorical	A1 - G5
emp_title	The job title supplied by the Borrower when applying for the loan.	Categorical	0
emp_length	Employment length in years.	Categorical	between 0 and 10 where means less than one year and 10 means ten or more years.
home_ownership	The home ownership status provided by the borrower during registration .	Categorical	Rent, Own, Mortgage, Other
annual_inc	The self-reported annual income provided by the borrower during registration.	Numeric	2,500 - 7,000,000
verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified	Categorical	Verified, Not verified, Source was verified
issue_d	The month which the loan was funded	Categorical	0
loan_status	Current status of the loan	Categorical	Fully paid, Charged off
pymnt_plan	Indicates if a payment plan has been put in place for the loan	Categorical	n, y
url	URL for the LC page with listing data.	Categorical	-
desc	Loan description provided by the borrower	Categorical	-
purpose	A category provided by the borrower for the loan request.	Categorical	education, medical, wedding, vacation, small business, etc.
title	The loan title provided by the borrower	Categorical	0
zip_code	The first 3 numbers of the zip code provided by the borrower in the loan application.	Categorical	-
addr_state	The state provided by the borrower in the loan application	Categorical	Shortname of the US States

dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.	Numeric	0 - 9999
delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years	Numeric	0 - 39
earliest_cr_line	The month the borrower's earliest reported credit line was opened	Categorical	1994 - 2013
fico_range_low	The lower boundary range the borrower's FICO at loan origination belongs to.	Numeric	660 - 845
fico_range_high	The upper boundary range the borrower's FICO at loan origination belongs to.	Numeric	664 - 850
inq_last_6mths	The number of inquiries in past 6 months.	Numeric	0 - 6
mths_since_last_delinq	The number of months since the borrower's last delinquency.	Numeric	0 - 226
mths_since_last_record	The number of months since the last public record.	Numeric	0 - 123
open_acc	The number of open credit lines in the borrower's credit file.	Numeric	0
pub_rec	Number of derogatory public records	Numeric	0
revol_bal	Total credit revolving balance	Numeric	0
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.	Numeric	0
total_acc	The total number of credit lines currently in the borrower's credit file	Numeric	0
initial_list_status	The initial listing status of the loan.	Categorical	W - F
out_prncp	Remaining outstanding principal for total amount funded	Numeric	0
out_prncp_inv	Remaining outstanding principal for portion of total amount funded by investors	Numeric	0
total_pymnt	Payments received to date for total amount funded	Numeric	0 - 44772.7
total_pymnt_inv	Payments received to date for portion of total amount funded by investors	Numeric	0 - 44772.7
total_rec_prncp	Principal received to date	Numeric	0 - 33843
total_rec_int	Interest received to date	Numeric	0 - 22157
total_rec_late_fee	Late fees received to date	Numeric	0 - 1188.8
recoveries	post charge off gross recovery	Numeric	0 - 25355
collection_recover_y_fee	post charge off collection fee	Numeric	0 - 4563.9
last_pymnt_d	Last month payment was received	Categorical	dd-mm-yy
last_pymnt_amnt	Last total payment amount received	Numeric	0 - 21996
next_pymnt_d	Next scheduled payment date	Categorical	dd-mm-yy
last_credit_pull_d	The most recent month LC pulled credit for this loan	Categorical	dd-mm-yy
last_fico_range_high	The upper boundary range the borrower's last FICO pulled belongs to.	Numeric	(0 - 850)
last_fico_range_low	The lower boundary range the borrower's last FICO pulled belongs to.	Numeric	(0 - 845)

collections_12_mths_ex_med	Number of collections in 12 months excluding medical collections	Numeric	0 - 14
mths_since_last_major_derog	Months since most recent 90-day or worse rating	Numeric	0 - 226
policy_code	publicly available policy_code=1 new products not publicly available policy_code=2	Numeric	1, 2
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers	Categorical	Individual, joint, direct pay
annual_inc_joint	The combined self-reported annual income provided by the co-borrowers during registration	Numeric	11000 - 1837000
dti_joint	A ratio calculated using the co-borrowers' total monthly payments on the total debt obligations, excluding mortgages and the requested LC loan, divided by the co-borrowers' combined self-reported monthly income	Numeric	0 - 69
verified_status_joint	Indicates if the co-borrowers' joint income was verified by LC, not verified, or if the income source was verified	Categorical	Verified, Not verified, Source was verified
acc_now_delinq	The number of accounts on which the borrower is now delinquent.	Numeric	0 - 14
tot_coll_amt	Total collection amounts ever owed	Numeric	0 - 6214661
tot_cur_bal	Total current balance of all accounts	Numeric	0 - 1159895
open_acc_6m	Number of open trades in last 6 months	Numeric	0 - 18
open_act_il	Number of currently active installment trades	Numeric	0 - 53
open_il_12m	Number of installment accounts opened in past 12 months	Numeric	0 - 25
open_il_24m	Number of installment accounts opened in past 24 months	Numeric	0 - 51
mths_since_rcnt_il	Months since most recent installment accounts opened	Numeric	0 - 446
total_bal_il	Total current balance of all installment accounts	Numeric	0 - 585583
il_util	Ratio of total current balance to high credit/credit limit on all install acct	Numeric	0 - 464
open_rv_12m	Number of revolving trades opened in past 12 months	Numeric	0 - 26
open_rv_24m	Number of revolving trades opened in past 24 months	Numeric	0 - 53
max_bal_bc	Maximum current balance owed on all revolving accounts	Numeric	0 - 79754
all_util	Balance to credit limit on all trades	Numeric	0 - 175
total_rev_hi_lim	Total revolving high credit/credit limit	Numeric	0 - 1629000
inq_fi	Number of personal finance inquiries	Numeric	0 - 32
total_cu_tl	Number of finance trades	Numeric	0 - 68
inq_last_12m	Number of credit inquiries in past 12 months	Numeric	0 - 49
acc_open_past_24mths	Number of trades opened in past 24 months.	Numeric	0 - 64
avg_cur_bal	Average current balance of all accounts	Numeric	0 - 188442
bc_open_to_buy	Total open to buy on revolving bankcards.	Numeric	0 - 138182

bc_util	Ratio of total current balance to high credit/credit limit for all bankcard accounts.	Numeric	0 - 318
chargeoff_within_12_mths	Number of charge-offs within 12 months	Numeric	-
delinq_amnt	The past-due amount owed for the accounts on which the borrower is now delinquent.	Numeric	-
mo_sin_old_il_acct	Months since oldest bank installment account opened	Numeric	-
mo_sin_old_rev_tl_op	Months since oldest revolving account opened	Numeric	-
mo_sin_rcnt_rev_tl_op	Months since most recent revolving account opened	Numeric	-
mo_sin_rcnt_tl	Months since most recent account opened	Numeric	-
mort_acc	Number of mortgage accounts.	Numeric	-
mths_since_recent_bc	Months since most recent bankcard account opened.	Numeric	-
mths_since_recent_bc_dlq	Months since most recent bankcard delinquency	Numeric	-
mths_since_recent_inq	Months since most recent inquiry.	Numeric	-
mths_since_recent_revol_delinq	Months since most recent revolving delinquency.	Numeric	-
num_accts_ever_120_pd	Number of accounts ever 120 or more days past due	Numeric	-
num_actv_bc_tl	Number of currently active bankcard accounts	Numeric	-
num_actv_rev_tl	Number of currently active revolving trades	Numeric	-
num_bc_sats	Number of satisfactory bankcard accounts	Numeric	-
num_bc_tl	Number of bankcard accounts	Numeric	-
num_il_tl	Number of installment accounts	Numeric	-
num_op_rev_tl	Number of open revolving accounts	Numeric	-
num_rev_accts	Number of revolving accounts	Numeric	-
num_rev_tl_bal_gt_0	Number of revolving trades with balance >0	Numeric	-
num_sats	Number of satisfactory accounts	Numeric	-
num_tl_120dpd_2m	Number of accounts currently 120 days past due (updated in past 2 months)	Numeric	-
num_tl_30dpd	Number of accounts currently 30 days past due (updated in past 2 months)	Numeric	-
num_tl_90g_dpd_24m	Number of accounts 90 or more days past due in last 24 months	Numeric	-
num_tl_op_past_12m	Number of accounts opened in past 12 months	Numeric	-
pct_tl_nvr_dlq	Percent of trades never delinquent	Numeric	-
percent_bc_gt_75	Percentage of all bankcard accounts > 75% of limit.	Numeric	-
pub_rec_bankruptcies	Number of public record bankruptcies	Numeric	-
tax_liens	Number of tax liens	Numeric	-
tot_hi_cred_lim	Total high credit/credit limit	Numeric	-
total_bal_ex_mort	Total credit balance excluding mortgage	Numeric	-
total_bc_limit	Total bankcard high credit/credit limit	Numeric	-

total_il_high_credit_limit	Total installment high credit/credit limit	Numeric	-
revol_bal_joint	Sum of revolving credit balance of the co-borrowers, net of duplicate balances	Numeric	-
sec_app_fico_range_low	FICO range (high) for the secondary applicant	Numeric	-
sec_app_fico_range_high	FICO range (low) for the secondary applicant	Numeric	-
sec_app_earliest_cr_line	Earliest credit line at time of application for the secondary applicant	Categorical	-
sec_app_inq_last_6mths	Credit inquiries in the last 6 months at time of application for the secondary applicant	Numeric	-
sec_app_mort_acc	Number of mortgage accounts at time of application for the secondary applicant	Numeric	-
sec_app_open_acc	Number of open trades at time of application for the secondary applicant	Numeric	-
sec_app_revol_util	Ratio of total current balance to high credit/credit limit for all revolving accounts	Numeric	-
sec_app_open_act_il	Number of currently active installment trades at time of application for the secondary applicant	Numeric	-
sec_app_num_rev_accts	Number of revolving accounts at time of application for the secondary applicant	Numeric	-
sec_app_chargeoff_within_12_mths	Number of charge-offs within last 12 months at time of application for the secondary applicant	Numeric	-
sec_app_collections_12_mths_ex_med	Number of collections within last 12 months excluding medical collections at time of application for the secondary applicant	Numeric	-
sec_app_mths_since_last_major_derog	Months since most recent 90-day or worse rating at time of application for the secondary applicant	Numeric	-
hardship_flag	Flags whether or not the borrower is on a hardship plan	Categorical	N
hardship_type	Describes the hardship plan offering	Categorical	-
hardship_reason	Describes the reason the hardship plan was offered	Categorical	Unemployment, divorce family death, excessive obligation, income curtailment, medical, natural disaster, reduced hours.
hardship_status	Describes if the hardship plan is active, pending, canceled, completed, or broken	Categorical	Broken, Completed
deferral_term	Amount of months that the borrower is expected to pay less than the contractual monthly payment amount due to a hardship plan	Numeric	-
hardship_amount	The interest payment that the borrower has committed to make each month while they are on a hardship plan	Numeric	0.63 - 943
hardship_start_date	The start date of the hardship plan period	Categorical	dd-mm-yy
hardship_end_date	The end date of the hardship plan period	Categorical	dd-mm-yy

payment_plan_start_date	The day the first hardship plan payment is due. For example, if a borrower has a hardship plan period of 3 months, the start date is the start of the three-month period in which the borrower is allowed to make interest-only payments.	Categorical	dd-mm-yy
hardship_length	The number of months the borrower will make smaller payments than normally obligated due to a hardship plan	Numeric	-
hardship_dpd	Account days past due as of the hardship plan start date	Numeric	0 - 32
hardship_loan_status	Loan Status as of the hardship plan start date	Categorical	Current, In Grace Period Issued, Late (16-30 days Late (31-120 days),
orig_projected_additional_accrued_interest	The original projected additional interest amount that will accrue for the given hardship payment plan as of the Hardship Start Date. This field will be null if the borrower has broken their hardship payment plan.	Numeric	1 -- 2267
hardship_payoff_balance_amount	The payoff balance amount as of the hardship plan start date	Numeric	55- 39542
hardship_last_payment_amount	The last payment amount as of the hardship plan start date	Numeric	0 - 1407
disbursement_method	The method by which the borrower receives their loan. Possible values are: CASH, DIRECT_PAY	Categorical	Cash, DirectPay
debt_settlement_flag	Flags whether or not the borrower, who has charged-off, is working with a debt-settlement company.	Categorical	N, Y
debt_settlement_flag_date	The most recent date that the Debt_Settlement_Flag has been set	Categorical	dd-mm-yy
settlement_status	The status of the borrower's settlement plan. Possible values are: COMPLETE, ACTIVE, BROKEN, CANCELLED, DENIED, DRAFT	Categorical	Active, Broken, Comple
settlement_date	The date that the borrower agrees to the settlement plan	Categorical	dd-mm-yy
settlement_amount	The loan amount that the borrower has agreed to settle for	Numeric	107 - 27000
settlement_percentage	The settlement amount as a percentage of the payoff balance amount on the loan	Numeric	0 - 166
settlement_term	The number of months that the borrower will be on the settlement plan	Numeric	0 - 118

أدى تطبيق تقنية موازنة البيانات (SMOTE) إلى تحسن كبير في نماذج خوارزميات التعلم الآلي الفردية الأخرى. وبمقارنة نماذج تعلم الآلة الفردية الخمسة من أجل دقة التصنيف، أظهر آلة المتجهات الدعمة SVM أعلى مستوى من الدقة من بعد النموذج الهجين. في النهاية، تساهم هذه الدراسة في تطوير التنبؤ بالدرجات الائتمانية في صناعة الإقراض P2P من خلال تقديم تقييم منهجي لنماذج التعلم الآلي. تساعد النتائج مشغلي المنصات والمستثمرين والمقترضين على اتخاذ قرارات مستنيرة بشكل أفضل. وبالتالي، تساهم هذه النتائج في خلق بيئة أكثر أمانًا وفعالية لمعاملات الإقراض P2P.

الملخص

مع تطور التكنولوجيا المالية، التي تتم الإشارة إليها بالاختصار (FinTech) ، التي تسعى لمنافسة الأساليب التقليدية في تقديم الخدمات المالية. برزت منصات الإقراض من نظير إلى نظير (P2P) كقوة جديدة في القطاع المالي كاحدى منتجات التكنولوجيا المالية .

توفر منصات الإقراض خيارًا بديلاً أسهل وأسرع للاقتراض من خلال ربط المقترضين بالمقرضين مباشرة دون الحاجة إلى المؤسسات التقليدية أو ضمانات تقليدية. تعتمد منصات الإقراض P2P بشكل كبير على دقة تنبؤات الجدارة الائتمانية للمقترضين الجديرين من خلال تقييم مخاطر الائتمان التي ينطوي عليها الإقراض. تشير مخاطر الائتمان إلى احتمالية وفاء المقترضين بالتزاماتهم في سداد قروضهم، مما قد يؤدي بشكل مباشر إلى خسائر استثمارية كبيرة. ومن ثم فإن تقييم مخاطر الائتمان وتسجيل الائتمان يلعبان دورًا حاسمًا في التمييز بين المتعثرين وغير المتعثرين. توفر نماذج التعلم الآلي (ML) الفرصة لتحسين تنبؤات التصنيف الائتماني وتعزيز الكفاءة والاستدامة الشاملة لمنصات الإقراض P2P ، و تقليل مخاطر التخلف عن السداد.

تقدم هذه الأطروحة دراسة مقارنة شاملة لخوارزميات التعلم الآلي للتنبؤ بالدرجات الائتمانية في مجال منصات الإقراض من نظير إلى نظير. تم استخدام ستة خوارزميات بارزة لتعلم الآلي: الانحدار اللوجستي (LR)، والغابات العشوائية (RF)، والنيفي بايز (NB)، وآلة المتجهات الدعامة (SVM)، و تعزيز التدرج الشديد (XGBoost)، ونموذج هجين جديد يجمع بين نقاط قوة الخوارزميات المتعددة باستخدام مجموعة بيانات الإقراض P2P. تم الاستعانة على بيانات سابقة من إحدى منصات الإقراض وهي لينديق كلوب "LendingClub" مجموعة بيانات طلبات القروض المقبولة المتاحة للعامة لفترة 2015-2018. تهدف الدراسة إلى تقييم الأداء التنبؤي لهذه النماذج وتحديد النهج الأكثر ملاءمة لتسجيل الائتمان على منصات الإقراض P2P.

تتضمن منهجية البحث التحليل الاستكشافي وتقنيات المعالجة المسبقة للبيانات؛ بما في ذلك تنقية البيانات، واختيار الميزات، وهندسة الميزات، وتقليل الأبعاد، وكلها تهدف إلى تعزيز تمثيل ملفات تعريف المقترضين. بالإضافة إلى ذلك، يتم تقسيم مجموعة البيانات إلى مجموعات تدريب واختبار لتسهيل تقييم النموذج. علاوة على ذلك، ونظرًا لمجموعة البيانات غير المتوازنة، تم استخدام تقنيات موازنة مختلفة. تم إجراء تقييم الأداء من خلال النظر في مقاييس الأداء المختلفة، مثل الدقة والإحكام والاستدعاء ودرجة F1.

تظهر نتائج البحث إلى أن آلة المتجهات الدعامة (SVM) مع تطبيق تقنية موازنة البيانات (SMOTE) ، يوفر الأداء الأكثر دقة مع درجة F1 متفوقة على النماذج الفردية بنسبة 99%.

جامعة البحرين

قسم برامج الدراسات العليا



كلية العلوم

دراسة مقارنة خوارزميات التعلم الآلي على بيانات المتعثرين في
الدفع للتنبؤ بالتقييم الائتماني للإقراض النّظير بالنّظير عبر
منصات التمويل الرقمي

إطروحة مقدمة كجزء من متطلبات الحصول على درجة الماجستير
في علوم وتحليل البيانات الضخمة

إعداد

فاطمة محمد إبراهيم المهيزع

الرقم الجامعي: 20080330

إشراف

د. نبيل محمود الحويحي

أستاذ

جامعة البحرين

مملكة البحرين

سبتمبر 2023