Research Report

"ENHANCING LOAN ELIGIBILITY CHECKER USING MACHINE LEARNING ALGORITHM"

Written by NWADOCHELI LOUIS

Date: January 16, 2024.

ABSTRACT

The dissertation titled "Enhancing Loan Eligibility Checker Using Machine Learning Algorithm" investigates the integration of the XGBoost algorithm to augment the accuracy and efficiency of loan eligibility assessments. In response to the evolving landscape of financial technology, there is a growing need for sophisticated tools that leverage machine learning for precise decision-making in the lending domain. This research delves into the intricacies of XGBoost, a powerful gradient boosting algorithm, and explores its application in optimizing the loan eligibility checker. Through a meticulous analysis of the dataset, the study identifies and emphasizes the key features influencing loan approval decisions, employing XGBoost's inherent interpretability. Utilizing cross-validation techniques, the dissertation evaluates the algorithm's performance, aiming to provide borrowers and financial institutions with a robust and interpretable model for making informed lending decisions. The outcomes of this research contribute to the advancement of machine learning applications in the financial sector, offering an enhanced loan eligibility checker that not only improves predictive accuracy but also enhances transparency in decision-making, ultimately benefiting both lenders and borrowers in the financial ecosystem.

TABLE OF CONTENTS

Abstract………………………………………………………………………….……………2

Acknowledgements……………………………………………………….…………………6

Chapter 1 (Introduction………………………………….………………….….……………..7

Problem and rational…………………………………………………………….……………7

Project scope……………………………………………………………….………………..11

Research question……………………………………………………………………………...11

Aims & objectives……………………………………………………………………………...11

Project deliverables……………………………………………………….........................13

Project benefits………………………………………………………………………………...13

Chapter 2 (Literature review)………………………………………………….........................17

Background and significance of credit risk assessment…………………….........................17

Role of automated system in improving accuracy……………………………………………..18

Credit risk assessment and default ……………………………………………………………19

Key variables…………………………………………………………………………………..19

Role of data analytics in credit risk assessment ……………………………….......................22

Machine learning methods ………………………………………………….......................23

Data collection…………………………………………………………………………………24

Research gaps…………………………………………………………….……………………25

Chapter summary ……………………………………………………………………………...25

Chapter 3 (Research Methodology)…………………………………………………………26

Research philosophy……………………………………………………………………………...26

Research approach……………………………………………………………………………..29

Research strategy………………………………………………………………………………30

Research design……………………………………………………………..…………………31

Research technique and procedures: Data collection and analysis …………………………………32

Machine learning algorithms……………………………………………..……………………34

Why XGBoost …………………………………………………………………………………35

Strategies for model building …………………………………………………………………35

Software architecture………………………………………………………….………………36

Benefits of python over R ……………………………………………………………………..37

Software testing ………………………………………………………………………………..37

Ethics, risks, social and legal issues……………………………………….…………………...37

Chapter Summary ……………………………………………………………………………..39

Chapter 4 (Model Development)……………………………………..………………………40

Data pre-processing tools ……………………………………………………………………...41

Data cleaning…………………………………………………………..……………………...41

Data transformation……………………………………………..……………………………..45

Handling imbalance data……………………………………………………………………….47

Challenges posed by imbalanced datasets ……………………………………………………48

Techniques to address imbalanced datasets …………………………………………………48

Benefits of over-sampling over under-sampling techniques ……………………………...49

Types of sampling method……………………………………………………………………..49

Correlation matrix…………………………………………………………………………...51

Features selection……………………………………………………………………………51

Splitting dataset ……………………………………………………………………………..52

Model building..…………………………………..……………………………...........53

Chapter summary …………………………………………………………………………...56

Chapter 5 (Data Analysis & Testing)……………………………………………………..56

Data overview…………………………………………………………...........................56

Key features………………………………………………………………………………57

Pre-processing steps………………………………………………………………………57

Train-test split………………………………………………………………………….58

Data quality assurance ……………………………………………………………………59

Model training……………………………………………………………………………59

Model performance metrics…………………………………………….........................60

Key findings……………………………………………………………………………62

Interaction effects ………………………………………………………………………...67

Model validation………………………………………………………………………….68

Implication for decision making ……………………………………….…………………68

Recommendations ……………………………………………………………………….69

Models' comparison ……………………………………………………………………...69

User evaluation ………………………………………………………………………….70

Chapter summary …………………………………………………………………………...70

Chapter 6 (Conclusion, Limitations and Future work………………………………………..72

Conclusion ……………………………………………………………………………………72

Model limitation………………………………………………………………………………72

Future work……………………………………………………………………………………73

References……………………………………………………………..............................74

Appendix…………………………………………………………………………………...83

CHAPTER 1

INTRODUCTION

1.1 Problem and rationale

Credit risk assessment entails the prequalification criteria designed by lenders to assess an applicant's creditworthiness and ability to repay a loan before advancing credit (Sann, K. K. 2019).

Loan eligibility checker on the other hand, is a tool that helps individuals assess their chances of being approved for a loan based on their inherent credit risks before formally applying to the financial institution (Behn, M., et al. 2022). By using a loan eligibility checker, individuals can avoid the time-consuming process of filling out a complete loan application form only to find out later that they don't qualify. This saves time and effort for both the borrower and the lender.

Using the loan eligibility checker helps the banks to gain more customers as these customers know beforehand their chances of getting credit before approaching the financial institution. Knowing one’s chance of assessing credit gives confidence to the borrower. Most would-be borrowers fail to approach the bank as they are always sceptical of the bank’s position in terms of their pre-qualification criteria (Weder di Mauro, B., & Zettelmeyer, J. (2017).

It’s an obvious fact that loan default is a major risk facing financial institutions in developing and developed societies and reducing it will help increase the bottom line of the financial institutions (Taghizadeh-H, et al, (2020). Credit risk is a fundamental concept in finance and lending, and it is a concern for both lenders and investors who provide capital to borrowers. Hence, understanding and managing credit risk is essential in various financial activities, including banking, lending, investing in bonds or securities, and managing a credit portfolio (Yung-Chia C, et al, 2018).

This project is aiming to develop a machine learning model in place of traditional credit scoring methods or other existing models that will predict the level of credit risk inherent in every personal or small business credit application better in a financial institution. This knowledge will help the financial institutions take wiser decisions when extending credit to their customers. This study will also aid the loan applicants with self-assessment before approaching the bank for any financial support. This will serve as loan eligibility checker for both individual loan applicant to self-assess one’s credit worthiness and the financial institutions to assess their customers credit worthiness.

Knowing your eligibility in advance can boost your confidence when applying for a loan as it provides a clear understanding of your financial standing and the likelihood of approval.

Machine learning is a subset of artificial intelligence (AI) that focuses on the development of algorithms and statistical models that enable computer systems to learn and improve their performance on a specific task (Shinde, A., et all, 2022). In terms of credit assessment, adopting machine learning algorithms in decision making enhances the accuracy of such credit risk decisions. Machine learning models utilizes quantitative and qualitative data to make more accurate predictions regarding applicant’s risk assessment. Machine learning has the potential to significantly transform the way credit risk is determined and evaluated in the financial industry (Leo Sharma, et al, 2019). Traditional credit risk assessment methods, while effective, often rely mostly on historical quantitative data and some predefined rules, which may not fully capture the complexity and dynamic nature of credit risk (Bouteille. S, 2021).

Previous researchers have made significant contributions in addressing enhancing credit risk assessment. They have used diverse machine learning algorithms to address this all-important feature in lending. Researchers such as (Michael Bucker, et al. 2022) had researched about transparency, auditability, and interpretability of models in credit scoring. They stated that effective modelling should be transparent and interpretable to be able to address issues surrounding credit risk (Michael Bucker, et al, 2022). Also, other notable researchers (Shinde A, et al. 2022) discussed the use of Random Forest and Logistic Regression to assess credit risk of loan applicants and accuracy levels of 79% and 82% were achieved respectively. Additionally, this is backed up by (Yuelin Wang, et al. 2019) where they discovered that Random Forest predicted better than other machine learning algorithms tested as it gave them better output accuracies and efficiencies.

The output of these models by these researchers needs improvement in many areas including over-fitting issues, efficiency, and interpretability by adopting the following strategies:

- Regularization parameters - Regularization is a technique used to prevent overfitting in machine learning models. These techniques penalize certain aspects of the model's complexity, helping to prevent it from fitting the training data too closely and improving its ability to generalize to new, unseen data.
- Feature importance - Feature importance indicates the relative importance of each feature in contributing to the model's predictions. Feature importance is derived from the contribution of each feature in constructing decision trees during the boosting process. The contribution of

a feature is computed based on the number of times it is used to split the data across all the trees in the ensemble and the improvement it brings to the model's performance. This strategy helps to address model interpretability.

- Early stopping - Implementing early stopping halts training once the model's performance on a validation set ceases to improve. This helps prevent overfitting and reduces computation time. The primary goal of early stopping is to halt the training process when the model's performance on a validation dataset ceases to improve. This helps prevent overfitting by avoiding the point where the model starts fitting the noise in the training data, and it improves efficiency by stopping the training process once optimal performance is reached.

While traditional machine learning algorithms demonstrate success in specific scenarios, they may fail to attain satisfactory results due to their limitations in accounting for multiple solutions (Polikar 2012). To address the challenges mentioned earlier and enhance the effectiveness of susceptibility models, ensemble algorithms that incorporate diverse learning techniques have been introduced and are currently in use. Recently introduced as a robust ensemble learning model, the XGBoost algorithm has garnered significant attention and has been widely adopted by numerous academic communities (Wang et al. 2020). The reason XGBoost is frequently utilized across various disciplines lies in its ability to effectively predict the nature of numerous complex real-world problems (Kavzoglu, T., & Teke, A. (2022).

By implementing these strategies above, we can harness the power of XGBoost while addressing overfitting, improving model interpretability, and enhancing efficiency in this loan eligibility checker project.

Most of the research work as itemized earlier used low volume data set ranging from 200 to 600, however this research has a large dataset of up to 140,000 in volume. With a large dataset, models often have more information to learn from, which can lead to better generalization and performance. Working with larger datasets in modelling often enables the identification and understanding of nuances, small disjuncts, and other nonlinearities that may be challenging or impossible to discern from smaller datasets (Kelleher, J. D. (2019). One of the algorithms used by the researchers is Random Forest, however an alternative algorithm is XGBoost being a gradient boosting algorithm, builds trees sequentially, with each tree trying to correct the errors of the previous ones (Bentéjac C, et al. 2021). This often leads to better predictive performance compared to the independent tree construction in Random Forest. There is often lack of interpretability with Random Forest models.

This lack of interpretability can be a significant drawback, especially in situations where transparency and regulatory compliance are essential (Islam, S. et al. 2019). Random Forest models can be prone to overfitting when they are too complex or when the number of trees in the forest is too high. Overfitting can lead to poor generalization performance and unreliable credit risk predictions (Fratello, M., et al. 2018)

Credit is the largest risk faced by any financial institution whose main obligation is financial mediation (Greenbaum S. I., 2019). Continuous increase in the scale of financial transactions in banks and other financial institutions has necessitated the use of technology to manage the risks associated with lending activities (Van Gruening, et all, S. B. (2020). Bad lending decisions can trigger a credit crunch, where lenders become reluctant to extend credit, even to creditworthy borrowers, out of fear of further losses (Baradaran, M. 2020). Bad lending decisions can contribute to economic downturns, especially when they are part of a broader trend of irresponsible lending. For example, the subprime mortgage crisis in 2008 played a significant role in the global financial crisis. Borrowers who are unable to repay loans may experience financial hardship, including bankruptcy, foreclosure, and damaged credit histories. This can have long-lasting effects on their personal and financial well-being (Argys L. M., et al, 2020).

This research therefore intends to build a model using XGBoost machine learning algorithm to enhance the prediction of credit risk assessment of would-be borrowers. The project is adopting XGBoost machine learning algorithm because of it’s high accuracy level and ability to address overfitting, interpretability and efficiency concerns (Bentéjac C, et al. 2021), (Qin, C., et al. 2021). This can be a worthy solution to good lending decisions and build both borrowers and investors’ confidence.

This research therefore addresses the salient issues raised with other machine learning models by adopting XGBoost as the machine learning algorithm on a large dataset. XGBoost has the potential to address complex dataset, accuracy issues, interpretability and so many other model issues which will be suitable to address the research question as discussed above (Qin, C., et al. 2021).

This model will be used as an eligibility checker for corporate use by financial institutions or for personal use for an intending borrower. The result of this research work can streamline the lending process, improve customer satisfaction, reduce risk, and enhance operational efficiency, making it a valuable tool for both borrowers and lenders. Eligibility checkers often provide detailed information about the requirements for a particular loan. This can include minimum income criteria, credit score requirements, and other factors that influence loan approval. This information helps borrowers understand what is needed for approval.

1.2 Project Scope

This dissertation seeks to comprehensively enhance the loan eligibility checker using advanced machine learning algorithm, XGBoost, with a primary focus on improving accuracy, efficiency, interpretability, and fairness in lending practices. The research involves a thorough analysis of historical loan data, identification of relevant features, and the development of a predictive model capable of accurately assessing the eligibility of loan applicants. Additionally, the project addresses ethical considerations, including fairness and bias mitigation, to ensure that the enhanced loan eligibility checker aligns with ethical standards and promotes inclusivity in the lending process. The outcomes of this work aim to contribute to the optimization of loan appraisal systems for individuals and financial institutions, fostering a more transparent, efficient, and equitable financial landscape.

1.3 Research Question

How can ensemble machine learning algorithms such as XGBoost be optimized to enhance the accuracy of loan eligibility checker prediction models?

1.4 Aims

The goal of this dissertation is to revolutionize loan eligibility assessment by leveraging on advanced machine learning algorithms. Focusing on improving accuracy, efficiency, interpretability and fairness, the project’s aims are listed below:

- Conduct Precise Credit Risk Assessment: The main objective is to precisely evaluate the credit risk associated with borrowers through an analysis of their financial profile, repayment history, creditworthiness, and other pertinent factors. The system strives to offer a dependable and unbiased assessment of the borrower's capacity and willingness to fulfil debt obligations.
- Spot High-Risk Borrowers: The system aims to pinpoint borrowers with an increased probability of defaulting on loan obligations or displaying other credit-related issues. Identifying high-risk borrowers enables lenders to implement suitable risk mitigation strategies, including adjustments to loan terms, collateral requirements, or interest rates.

- To Enhance the Accuracy of Credit Risk Prediction: Develop and implement advanced models to enhance the precision and reliability of credit risk assessment, contributing to improved decision-making processes.
- To Contribute to the Advancement of Credit Risk Research: Contribute new knowledge to the field of credit risk assessment by advancing methodologies, exploring emerging trends, and providing insights that can inform future research and industry practices.
- To Optimize Credit Limit Determination: Develop strategies for optimizing credit limits to balance risk exposure and customer satisfaction, with a focus on minimizing default rates and maximizing profitability.

With the above aims in mind, this research work intends to build a loan checker that will automatically provide instant loan application status to the user by providing the level of inherent risk based on the risk assessment result. This output is then used for decision making by the organisation. The user might be the applicant or the financial institution who is processing loan for their client. The loan eligibility checker will output the decision based on the data provided using the machine learning algorithm that will run backend.

1.5 Objectives

Below are the objectives of this research work, detailing the processes and steps adopted for building a machine learning model for predicting loan eligibility of borrowers.

1. Literature review: Explore existing academic knowledge and findings around credit risk assessment, gain deeper understanding from previous work. Discover any gaps and it will become the focus of this research work.

2. Research design: Do a critical evaluation of the research methodology and philosophy surrounding this research work. Adopting a suitable research methodology to achieve the research goal. The methodology will form the logic of this research work.

3. Data preparation and future engineering: This project will source for data, prepare the data, and ensures the dataset is clean, well-structured, and ready for model training. It will also carry out feature engineering whereby new features are created, or existing ones are modified to enhance the model's ability to capture patterns and make accurate predictions.

4. Data analysis and python libraries: The project will utilize python libraries such as pandas for data analysis and manipulation, matplotlib and seaborn for data visualization. These will aid in identifying patterns from the dataset and presented on dashboards for easy visualization.
5. Model training and evaluation: Data will be split into training and testing dataset. This involves teaching the model to recognize patterns in the input data so that it can make accurate predictions or classifications. The model will then be tested to assess the accuracy of the prediction of the model.
6. Conclusion: Draw conclusion from the test result about the accuracy of the model and how it answers the research question. Document your findings and make recommendations for further studies.

The output of this research work is to deliver a model that can predict the credit risk inherent in a customer who wants to approach a financial institution for loan. A couple of machine learning algorithms will be explored in comparison with XGBoost machine learning algorithm which is believed to predict more accurately financial modeling tasks. The dataset will be clean and prepared for analysis using different modeling techniques. This research will compare some other machine algorithms with XGBoost for efficiency, accuracy, and output interpretability.

1.6 Project Deliverable

My deliverable is to build a model with a machine learning algorithm that enhances the accuracy of loan eligibility checker for borrowers and lenders. The enhanced loan eligibility checker will leverage advanced machine learning algorithms such as xgboost for improved accuracy, efficiency, interpretability, and fairness in assessing loan eligibility. This deliverable will include the developed predictive model, comprehensive documentation, and recommendations for usage.

1.7 Project Benefits

This research is aiming at leveraging the power of machine learning algorithms to accurately assess the credit worthiness of each borrower. The use of XGBoost machine learning algorithm will fill the gaps and lapses that other algorithms find it difficult to fill such as accuracy, efficiency, interpretability, and overfitting issues. Also, the project will fill additional gaps with using traditional credit risk assessments such as bias because of limitations of human involvement (Pandey. R, et al, 2022).

There is risk inherent in every application which must be determined or assessed before approving any loan request by the financial institution (Van Greuning, H., et al, 2020). Credit risk assessment is both beneficial to the credit institution and the borrower (Aduda, J., & Obondy, S. (2021). Credit risk assessment helps the applicant to know his credit worthiness. Knowing his credit worthiness will spur him to make a wise financial decision. Credit risk management by financial institutions remains a serious concern for players in this sector and controlling it becomes inevitable in their day-to-day business (Jeucken, M, et al, 2017). Effective credit risk management is crucial for maintaining the stability and soundness of financial institutions and markets. Credit risk management helps lenders make informed lending decisions, reduces the likelihood of financial crises, and protects the interests of investors and depositors (Vives, X. 2019).

The benefit of this application is to aid the financial institution who uses our application in decision making when they want to advance credit to their customers thereby reducing the rate of credit default and improving their balance sheet in the long run.

Furthermore, below are some of the reasons why loan eligibility assessment using machine learning is very crucial before extending loans to borrowers by the financial institutions:

- Risk Management: Lenders, whether banks, credit unions, or online lenders, face financial risks when they extend loans. The loan eligibility checker helps mitigate these risks by ensuring that borrowers are financially stable and likely to repay the loan on time. This risk management is essential to protect the lender's financial stability (Taghizadeh-H. F, et al, 2020).
- Responsible Lending: Lending institutions have a responsibility to ensure that they lend money to individuals and businesses who can afford to repay it without causing undue financial hardship (Kiviat, B, 2019). The loan eligibility checker helps lenders to adhere to responsible lending practices and avoid lending to individuals who are at high risk of default.
- Interest Rate Determination: Loan eligibility assessment can also influence the interest rate offered to borrowers. Lenders may offer more favourable terms, such as lower interest rates or longer repayment periods, to borrowers who are considered less risky based on their creditworthiness. Likewise, businesses and individuals perceived with higher risks though manageable tend to be offered higher interest rates (Chen, D., et al, 2017).
- Portfolio Management: Lenders use loan eligibility assessment to manage their loan portfolios effectively. They can diversify their lending by offering various types of loans to borrowers

with different risk profiles based on their business needs. This helps spread risk and maintain a balanced portfolio (Metawa, N., et al, 2017).

- Improved Predictive Accuracy: Machine learning algorithms can handle a broader range of data sources and complex data relationships, leading to more accurate predictions of creditworthiness. They can identify non-linear patterns and subtle indicators that may be missed by traditional credit scoring models (Bazarbash, M. (2019).
- Real-Time Decision-Making: Machine learning models can process and analyse data in real time, allowing for quick, automated lending decisions. This is crucial for industries like online lending, where speed and efficiency are essential. This automation can lead to more efficient and cost-effective lending operations (Gopal, S., et al, 2023).
- Reduced Human Bias: Traditional credit scoring methods can be influenced by human bias and subjective judgments. Machine learning models rely on data-driven patterns, potentially reducing bias in credit risk assessments and making lending decisions fairer. Human beings are prone to bias due to emotions which affect their decision making on approving loans including the amount and tenure (Lee, M. K. (2018).
- Regulatory Compliance: Machine learning models can help lenders comply with evolving regulatory requirements. They can be trained to incorporate specific rules, guidelines, and fairness considerations, which is particularly important in the financial industry (Barocas S., et al, 2017).

1.8 Achieving the research objectives.

The initial proposal to this research work can be viewed in the appendix section of this research work, appropriate approval was gotten before commencement of the research work and ethics form duly signed. The project flow is described in the figure below. Achieving the objectives starts from identifying the problem of credit risk assessment. Then continues to literature review, then to research methodology and data collection. The next stage is data preparation which involves the identification of influential features from the dataset, followed by data analysis and model development. Model development involves cleaning of dataset while leaving the influential features. The next is training the algorithm with some part of the dataset. Different machine learning algorithms will be developed and compared with XGBoost to and confirm if XGBoost has a better performance. Then testing phase with input from sample users and output will be analysed for conformity with accuracy as expected.

Finally, the research work will draw a conclusion on the level of success of the research work and scope of future research and development.

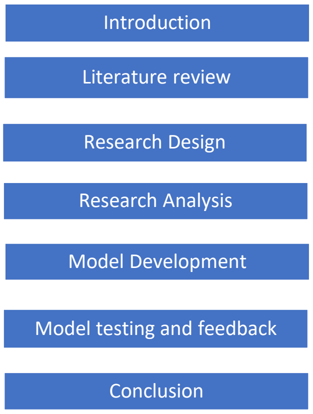


Figure 1.1- Project flow overview

Chapter summary

This project focuses on advancing credit risk assessment through a loan eligibility checker, employing the XGBoost machine learning algorithm. This tool aims to predict credit risk for individuals and small businesses, offering a more efficient alternative to traditional credit scoring methods. Leveraging a substantial dataset of up to 140,000 entries, the study addresses limitations in previous models by implementing strategies such as regularization, feature importance assessment, and early stopping. The research emphasizes the benefits, objectives, and scope, aiming to revolutionize credit risk prediction, improve decision-making processes, and contribute to fair lending practices. The final deliverable is an enhanced loan eligibility checker, accompanied by comprehensive documentation and usage recommendations, ultimately fostering transparency and efficiency in lending activities.

CHAPTER 2

LITERATURE REVIEW

This literature review begins by exploring the existing body of knowledge on loan eligibility assessment methods. It delves into traditional approaches employed in the financial industry, discussing their strengths and limitations. The review then transitions to the emergence of machine learning algorithms, with a focus on the application of XGBoost in predictive modelling for loan approval. Key studies addressing similar topics, advancements in machine learning for credit risk assessment, and the evolving landscape of financial technology are examined. This literature review aims to provide a comprehensive understanding of the current state of loan eligibility assessment and sets the stage for proposing enhancements through XGBoost in the subsequent dissertation chapters.

2.0.1 Background and significance of loan eligibility assessment

Existing knowledge on loan eligibility assessment methods encompasses a range of traditional approaches employed by financial institutions. These methods traditionally involve evaluating an applicant's creditworthiness based on criteria such as credit history, income, debt-to-income ratio, and employment status (Chen, N., et al. 2016). Credit scoring models, like FICO scores, have been widely utilized to quantify the risk associated with lending to an individual. Furthermore, traditional methods often rely on predefined rules and cutoffs to determine eligibility, leading to a somewhat rigid and rule-based decision-making process (Hohnen, P., et al. 2021). While these methods have been effective to a certain extent, they may not capture the full complexity of individual financial profiles and can be less adaptive to changing economic conditions or evolving borrower characteristics.

Credit risk assessment plays a pivotal role in the financial industry, influencing lending decisions, risk management, and the overall stability of financial institutions (Park, H., et al 2020). Over the years, researchers and practitioners have employed various methods and systems to evaluate the creditworthiness of borrowers (Ubarhande, P., et al., 2021). When financial institutions are granting credit facilities to individuals or corporate entities, there are lots of parameters being considered such as 5Cs of credit (character, capital, capacity, condition of credit, collateral) which enable them to investigate the client’s credit worthiness (Konovalova et al, 2016).

A lack of reliable credit risk assessment system has caused commercial banks and other financial institutions huge financial loss because of increased loan defaults (Okuthe Paul, et al., 2017). While assessing credit risk, there are also other variables such as cashflow, type of business, age of business, age of applicant, sex of applicant etc. that need to be considered to enable the credit institution to arrive at a good conclusion regarding the customer’s ability or inability to assess credit from the financial institution (Habtamu D., 2019).

The management of credit risk incorporates a comprehensive evaluation of both quantitative and qualitative measures, enhancing the decision-making process when extending credit (Konovalova et al, 2016).

Credit risk assessment plays a pivotal role in the financial industry, influencing lending decisions, risk management, and the overall stability of financial institutions (Park, H., et al 2020). Over the years, researchers and practitioners have employed various methods and systems to evaluate the creditworthiness of borrowers (Ubarhande, P., et al., 2021). When financial institutions are granting credit facilities to individuals or corporate entities, there are lots of parameters being considered such as 5Cs of credit (character, capital, capacity, condition of credit, collateral) which enable them to investigate the client’s credit worthiness (Konovalova et al, 2016).

A lack of reliable credit risk assessment system has caused commercial banks and other financial institutions huge financial loss because of increased loan defaults (Okuthe Paul, et al., 2017). While assessing credit risk, there are also other variables such as cashflow, type of business, age of business, age of applicant, sex of applicant etc. that need to be considered to enable the credit institution to arrive at a good conclusion regarding the customer’s ability or inability to assess credit from the financial institution (Habtamu D., 2019).

The management of credit risk encompasses a meticulous consideration of both quantitative and qualitative measures, aiming to improve decision-making when approving credit. (Konovalova et al, 2016).

2.0.2 Rise of automated credit risk assessment systems

Lending institutions encounter a challenge in deciding whom to lend money to, the loan amount, and the applicable interest rate (Khemakhem et al., 2018). The shift from manual systems to technology is deemed more dependable in predicting loan default. Presently, there is an increasing interest in creating automated applications utilizing diverse statistical and machine learning models to precisely forecast the inherent risk associated with borrowers (Shrawan Kumar Trivedi, 2020).

2.0.3 Importance of predicting personal loan defaults

The global credit crunch, originating in the United States in 2006, underscored the critical importance of credit risk management. It revealed that imprudent lending choices can result in substantial financial losses (Ken Brown, et al., 2016). Precision in predicting the likelihood of default in loans is pivotal for informed decision-making, contributing to enhanced financial stability for institutions. Every instance of extending credit creates an asset, prompting financial institutions to carefully assess the risk-reward ratio before disbursing funds (Lajis S.M, 2017).

2.1 Credit Risk Assessment and Default Prediction

Credit Risk Assessment and Default Prediction are fundamental processes in the financial industry, particularly in banking and lending (Doumpos, et al., 2019). These processes aim to evaluate the likelihood that a borrower will fail to meet their financial obligations, such as repaying a loan or fulfilling a credit agreement (Doumpos, et al., 2019). Credit risk assessment, often referred to as credit risk analysis, is the process of evaluating the creditworthiness of an individual, business, or entity seeking financial services, such as loans, credit cards, or mortgages (Bouteille, S., et al., 2021). The goal here is to determine the level of risk associated with lending money to the applicant. Default prediction is a specific aspect of credit risk assessment that focuses on forecasting the likelihood of a borrower failing to meet their financial obligations, i.e., defaulting on a loan or credit agreement. Default prediction involves predicting whether a borrower will miss payments, go delinquent, or ultimately default on their financial obligations (Cheng, D., et al., 2019).

2.1.1 Traditional credit risk assessment methods

Traditional credit risk assessment methods refer to the conventional approaches used by lenders and financial institutions to evaluate the creditworthiness of potential borrowers (Khemakhem et al., 2018). Historically, traditional credit risk assessment relied heavily on qualitative methods and expert judgment. Credit scores and credit reports from credit bureaus have been fundamental in assessing an individual's creditworthiness (Siddiqi, N. 2017). These scores are typically based on factors like payment history, credit utilization, length of credit history, new credit inquiries, and types of credit used (Doumpos, et al., 2019). Notable models include the FICO score in the United States and the credit scoring models of Equifax, Experian, and TransUnion. These models have undergone evolution and remain pivotal in consumer lending (Zakowska, A., 2023). Employed for numerous years, these methods rely on historical data, financial ratios, and pertinent details to gauge the risk level associated with lending to individuals or businesses. Emphasizing factors like cash flow analysis, age of the business or individual, credit report, collateral, employment history, and lender-provided information on the application form, these variables, while significant, fall short in comprehensively assessing an individual's credit risk. This limitation is evident in the high default rates experienced by major financial institutions (Khemakhem et al., 2018).

2.1.2 Statistical and Machine Learning Models

The limitations of traditional approaches have spurred interest in leveraging machine learning algorithms for loan eligibility assessment. This shift allows for a more dynamic and data-driven evaluation of borrowers, potentially improving accuracy and providing a more nuanced understanding of credit risk. The literature suggests that integrating advanced machine learning techniques can enhance the predictive power of loan eligibility models by considering a broader range of features and patterns in the data.

Statistical and machine learning models have gained prominence in credit risk assessment due to their ability to analyse large datasets and identify complex patterns (Leo, M., et al., 2019). These models include logistic regression, decision trees, random forests, and gradient boosting. They utilize a broader range of features, potentially including transaction history, behavioural data, and alternative data sources to enhance predictive accuracy. Research in this area has focused on model validation, overfitting prevention, and feature selection techniques to improve model performance (Ding, J., Tarokh, V., et al., 2018).

2.1.3 Credit Risk Assessment and Regulatory Compliance

The Basel Accords, particularly Basel II and Basel III, have had a significant impact on credit risk assessment practices (Baud, C., et al. 2017). These international banking regulations emphasize the importance of sound risk management and the need to align capital requirements with a bank's risk profile. Research in this area has explored the implications of these regulations and their influence on credit risk modelling and management.

2.1.4 Challenges in predicting loan defaults.

Predicting loan default is such a daunting task as lots of variables are involved and predicting accurately can be very challenging.

The predictive model this research work intends to build is expected to address the challenges by providing solution to bias, interpretability, efficiency, and accuracy as it leverages the power of advanced machine learning algorithm such as XGBoost gradient.

Some of the challenges that affect loan default prediction are listed below:

1.Shifting macroeconomic conditions pose a challenge in traditional loan default prediction, given the constant global economic changes. Relying on a traditional single-period approach becomes impractical due to these ongoing economic dynamics. Traditional quantitative analysis primarily focuses on financial statements, proving insufficient for accurate loan default prediction (Ivana Tomas Žiković, 2017).

2.Data quality plays a pivotal role in the accuracy of loan default prediction. Inadequate or incomplete data can introduce bias and inaccuracies into the predictions.

3.Limited credit history is a significant constraint, as decisions solely based on credit history may not be comprehensive. Some customers lack documented credit information, especially new borrowers with no established credit history, leading to flawed decisions (Hurley et al., 2016).

4.The continuously evolving lending landscape, driven by information technology and the introduction of new products and services, makes it challenging to predict defaults using traditional models (Hasan et al., 2020).

5.Interpretability presents a hurdle, as some outputs from highly predictive models require specific skill sets for interpretation, such as feature importance analysis, interpretability techniques, and data visualization skills (Chakraborty, S., et al. 2017).

5.External factors, such as natural events (global financial crises or disasters), can impact the outcomes of loans previously deemed low risk (Laframboise, M. N., 2012).

7.Balancing regulatory compliance with data utilization for predicting loan defaults can be challenging, requiring careful navigation to avoid regulatory guideline violations (Wachter, S. 2018).

2.1.5 Role of automated systems in improving accuracy.

Automated systems significantly contribute to enhancing the precision of loan default predictions (Alonso et al., 2020). These systems employ machine learning or statistical models, incorporating a broader set of input variables to forecast the likelihood of credit default. This results in a decline in non-performing loans within financial institutions, eliminating human errors and integrating non-quantitative data into credit risk modelling.

Credit risk assessment is a critical function in the financial industry, influencing lending decisions and risk management practices (Legesse B. M. 2017). The rise of machine learning and data analytics has transformed the way credit risk is assessed, leading to higher accuracy and efficiency in the process (Schmitt, M. (2020). This literature review examines key trends and developments in credit risk assessment, focusing on the application of XGBoost machine learning technique in comparison with Decision Tree, Random Forest and KNN.

This literature review discusses the dataset and the specific features of the dataset. It also discusses the different types of algorithms that will be compared with the base algorithm, XGBoost.

2.2 The role of data analytics in credit risk assessment

Data analytics plays a pivotal role in credit risk assessment, transforming the traditional methods of evaluating creditworthiness by leveraging advanced techniques to analyse vast datasets (Patel, K. 2023). Here are key aspects highlighting the significance of data analytics in credit risk assessment:

Credit Scoring Models: Data analytics facilitates the development of predictive models, such as credit scoring models, that assess the likelihood of a borrower defaulting on a loan. These models use historical data and various features to quantify credit risk (Xia, Y., et al. 2018).

Data-Driven Insights: Data analytics provides lenders with data-driven insights into the financial behaviours of borrowers (Patel, K. 2023). This information guides more informed decision-making by assessing the risk associated with lending to specific individuals or businesses.

Identifying Relevant Variables: Data analytics helps in identifying and engineering features that are most relevant to credit risk assessment (Bhatore, S., et al. 2020). This involves selecting key variables, transforming data, and creating new features that enhance the predictive power of models.

Tracking Financial Behaviours: Through data analytics, lenders can track and analyse the financial behaviours of borrowers over time. This includes monitoring payment history, debt utilization, and other financial activities that contribute to a comprehensive understanding of credit risk.

2.3 Machine learning models

Machine Learning (ML) is a subset of artificial intelligence (AI) that focuses on developing algorithms and statistical models that enable computer systems to improve their performance on a specific task (Tyagi, A. K., & Chahal, P. (2022). A machine learning model is an algorithm that is trained on data to make predictions, decisions, or automate actions (Alzubaidi, L., et al. 2023). The model learns patterns and relationships within provided data and can generalize that knowledge to make predictions on new, unseen data (Le, T., et al. 2021). In credit risk assessment, various machine learning algorithms such as logistic regression, decisions trees, K-nearest neighbours (KNN), Neural networks and gradient boosting are used to build models. In this section, we are looking at the research work done using these models to address credit risk assessment.

Using Random Forest, the researcher (Ghatasheh, N. (2014) utilizing the power of multiple trees in random forest model, created a risk assessment model using several variables for over 100 trees and got an accuracy of 78%. The researchers noticed there was a need to improve the performance of the model by enhancing the way of growing the decision trees by classification and better variable selection. The accuracy is not sufficient for a credit risk application. Pragya and other researchers (Pragya Pandey, et al. 2022) delved into building their model with K-NN (K-nearest neighbour) to determine the model effect in solving credit risk assessment problems. They used 30,000 datasets with 24 features each to decide if a credit request will be approved or rejected. The model achieved an accuracy of 81.2% which performed lower than Decision tree in their own analysis in terms of both accuracy of classification. Their result shows a low AUC score which measures the ability of a classification model to distinguish between positive and negative examples. Specifically, it represents the area under the ROC curve. Another set of researchers, (Kui Wang, et al. 2021) did a thorough work on using xgboost algorithm to evaluate credit risk. According to the researchers, XGBoost model is a tree-based gradient lifting integrated model with high efficiency and prediction accuracy. They stated that xgboost algorithm reduces the model's complexity, mitigating the occurrence of model overfitting. The model got an accuracy of 87.1% and area under curve (AUC) value of 0.94. The model used regularization to reduce over-fitting, however, they were silent on interpretability, feature selection using xgboost and class imbalance.

Previous research had significant contributions towards enhancing credit risk assessment, however there are still obvious gas that need to be filled to improve the enhancement of credit risk assessment. It is also quite obvious that xgboost gave a the best result from other researches which supports the reason why this research is delving more into the capabilities of xgboost by addressing some research gas that not filled in previous researches. Such gaps include, addressing class imbalance, over-fitting, interpretability, feature selection.

2.4 Data collection

Joseph D., et al. (2021) utilized Penn Machine learning benchmarks (PMLB) in sourcing for their dataset to build a standard machine learning benchmarking model for comparison of various models. PMLB represents one of the biggest sources of open dataset for use in machine learning (Joseph D., et al. 2021). To build a merchant credit risk detection system, Chih-Hisiung et al, (2019) collected over 56,000 datasets from private data source belonging to merchants of China Union pay (Chih-Hisiung et al, (2019). These sources are very relevant to this research work as one focused on private data source from a financial organisation while the other focused on data from public data collection source.

Similarly, Tushar S, et al. (2022) embarked on their research using data collected through survey. Survey was conducted on over 400 research papers and data was collected and used for their source code analysis using different algorithms. This research work is not adopting the use of survey for source of data because a more comprehensive dataset was gotten from an open data source and well certified by the owner. The dataset constitutes raw data from financial institution, large and suitable for this research work.

Tian Lu, et al. (2019) collected data from alternative data sources such as phone usage, online shopping records and social media presence. The data collected was used in their research work for finding the value of alternative source of data in credit risk prediction. In contrast, Miliūnaitė, L. (2023) utilized both alternative data and financial data from Lithuanian SMEs to carry out their financial risk assessment of SMEs. This approach would have been adopted by this research; however, this research work is not focusing on SMEs alone but both business and non-business credit risk with so many other variables. Also, embarking on survey will not be feasible because the project needs large dataset for better algorithm performance and accuracy. Gathering this level of dataset will be time consuming and very expensive which the budget and time for this research cannot support.

Meanwhile, Ni, D, et al. (2023) embarked on their research work with data obtained from about 441 companies for their financial data and other data from network activity such as visits to some popular websites to predict corporate credit risk. The challenge noticed here is the use of small dataset in combination with network activity of the intended borrowers to predict their credit worthiness. This research has robust dataset including both financial and non-financial data to carry out the credit risk analysis. From the various dataset collection methods, it’s obvious that the type and source of datasets matter much in terms of data integrity and privacy issues. This research understood this and got a set of data from reliable and approved source showcasing real and anonymized dataset suitable for this type of research work.

2.5 Research gaps

Though lots of work have been done in credit risk assessment model development, more research is still needed to explore the integration of a wider range of alternative data sources, such as social media activity, geolocation data, and non-traditional financial indicators, into credit risk models. This is crucial for assessing creditworthiness, particularly for individuals with limited credit histories (Müller, M. M. (2023). This research aims to explore methods for selecting a compact set of features that provide not only predictive power but also interpretability for regulatory compliance and decision-making transparency. Credit risk datasets are often imbalanced, with majority of borrowers being low-risk and a minority being high-risk (Abedin, M. Z., 2023). This research focuses on feature selection techniques specifically designed to handle imbalanced datasets and minimize bias in model predictions. Also, research gaps exist in over-fitting model efficiency which this research work intends to solve. Most credit risk assessment models primarily rely on structured data such as credit scores, financial statements, and payment history (Doumpos, M., et al, 2019). There's also research gap in effectively incorporating unstructured data sources such as text data from loan applications, social media activity, and news sentiment analysis into feature selection and risk assessment models (Pejić Bach, M., et al., 2019). This research will not focus on social media activity and news sentiments due to the time and budget available to the project.

Chapter Summary

In conclusion, the existing knowledge of machine learning models in credit risk assessment is vast and multifaceted. The integration of advanced algorithms, coupled with innovative data sources and thorough model evaluation, presents a promising path towards more accurate and efficient credit scoring systems. In line with papers reviewed in this work, a great milestone has been achieved in using machine learning algorithm to address credit risk assessment challenges. However, there still exists some research gaps in interpretability, accuracy, over-fitting, and feature selection issues. This research work will delve into improving on these research gaps.

CHAPTER 3

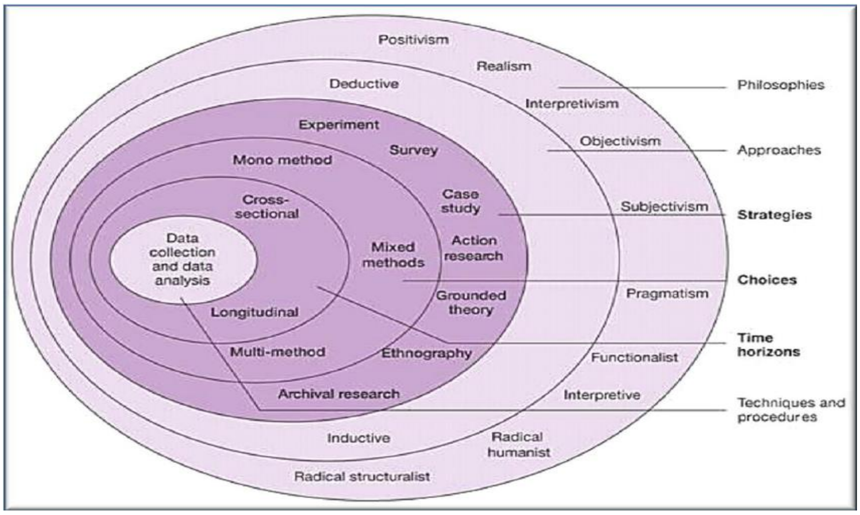
RESEARCH METHODOLOGY

This research work incorporated various research approaches to achieve our research question, aims and objectives as discussed below. The design detailed data collection, data analysis and interpretation of results.

The study utilized a suitable research methodology to achieve the research goal. The methodology will form the logic of this research work. The research methodology addressed the research goals, research question, the design approach, and the rationale behind selection of the design methodology. (Ian Sanders et al., 2022), (Esser, F., & Vliegenthart, R. (2017).

This research will use research onion framework to illustrate the layers and stages of our research methodology. This was developed by Saunders et al. (2019) to help researchers understand and plan the various components of their research process. In the context of credit risk assessment, the research onion will be a useful framework to guide our research methodology. Below is a diagrammatic representation of the research onion and as it relates to this project.

Figure 3.1 – Research onion



3.1 Research Philosophy – Positivism

Research philosophy refers to the set of beliefs, principles, and assumptions underlying a research study. It’s a guide on the researcher’s approach to designing, conducting, and interpreting research. The choice of research philosophy reflects the view of this research work about research methods, data collection, analysis, and the nature of knowledge (Kivunja, C., et al. 2017).

There are different kinds of research philosophy. Positivism, and Interpretivism and Pragmatism.

Positivism: This is a research philosophy that emphasizes the objective and empirical analysis of observable phenomena (Park, et al. 2020). In the context of this credit risk assessment project, a positivist approach focuses on using quantitative data and statistical analysis to understand and predict credit risk.,

Interpretivism: Interpretivism is a research paradigm and method that focuses on understanding and interpreting the subjective meanings individuals attach to their experiences, behaviours, and social phenomena (Alharahsheh, H. H., & Pius, A. 2020). Interpretivism research primarily employs qualitative research methods, such as interviews, observations, focus groups, and content analysis to achieve their aims (Khan, S. N. (2017). This research philosophy also has to do with behavioural and intent issues in the context of credit risk. Interpretivism acknowledges the subjective nature of creditworthiness. It recognizes that an individual's financial behaviour and creditworthiness may be influenced by personal circumstances, cultural factors, and individual experiences, which may not be fully captured by quantitative models alone. This is vital in this research as credit risk does not only consider capacity and capital but also character which is more of qualitative. In fact, character is the first “C” in 5Cs of credit. Interpretivism plays a vital role in credit risk assessment research like this one.

Pragmatism: This type of research philosophy combines some elements of other research philosophies such as positivism and interpretivism to achieve it’s outcome (Mitchell, A., & Education, A. E. 2018). It encourages the use of a variety of research methods, both quantitative and qualitative, as well as mixed methods approaches. This research will consider some qualitative data to be able to achieve a very reliable credit risk prediction. Pragmatism supports the use of multiple research methods based on their effectiveness in solving the research problem. In credit risk assessment, this may involve employing diverse techniques such as data analytics, machine learning, and behavioural studies to gain a comprehensive understanding of risk factors. This approach can be useful in this research as it involves data analysis and machine learning.

Positivism suits this research work most appropriately as it tends to focus on quantitative approach which is in line with the data and aims of this research. benefits of positivism over other research philosophies. Another benefit of positivism is that positivism aligns with the principles of the scientific method, emphasizing rigorous research design, hypothesis testing, and systematic inquiry. In credit assessment, where accuracy and reliability are crucial, a rigorous scientific approach can enhance the validity of the prediction models (Kankam, et al. 2019).

Therefore, based on our research objectives and the available data, this research work is adopting a positivism approach in the following ways:

Objective Data: This work is a positivist research work in credit risk assessment, and we rely on objective and quantifiable data. This typically involves using financial and economic data, credit scores, historical performance records, and other numerical measures to assess credit risk (Siddiqi. N. et al. 2017). The emphasis is on data that can be measured and verified, reducing the influence of personal biases.

Hypothesis Testing: Positivist researchers formulate hypotheses based on existing theories or knowledge and then test these hypotheses using statistical methods (“Park, Y et al. 2020). This research work, being an example of positivism is working on a presumption that XGBoost machine learning algorithm will enhance credit risk assessment better than other algorithm types. This hypothesis is based on the relationship between some specific financial data to predict the likelihood of loan default, and then use empirical data to confirm or refute this hypothesis.

Generalizability: Positivism seeks to produce findings that are generalizable to a broader population or context (“Park, Y et al. 2020). In credit risk assessment, this means that the research aims to produce insights and models that can be applied to a wider range of borrowers or financial institutions, not just the specific cases under study. This research aims to produce a model that can predict default in loan applications, it can be used by different stakeholders in the finance sector and the borrowers themselves to check their level of credit worthiness before approaching the Bank.

Deductive Reasoning: Positivist research often follows a deductive reasoning approach, where researchers start with a theory or a set of hypotheses and then collect and analyse data to test these theories (Casula. M, et al. 2021). For instance, you might start with a theory that certain financial variables are strongly correlated with credit risk and then gather data to confirm or disprove this theory. This research has gathered data of several features to be used to test the effectiveness of the machine learning algorithms.

Quantitative Analysis: Positivist researchers primarily use quantitative research methods and statistical analysis to interpret data. This can involve techniques such as regression analysis, machine learning, and mathematical modelling to identify patterns and relationships in the data. (Park, et al. 2020) This approach is what we are adopting in this research to achieve our research objectives.

Emphasis on Reliability and Validity: Positivist research places a strong emphasis on the reliability and validity of the research findings (Maxwell, J. A. (2017). This research work aims to ensure that the data collection and analysis methods are robust and that the results are replicable and dependable.

Reduction of Bias: Positivist researchers are concerned with reducing the influence of researcher bias and subjectivity (Levitt, et al. 2022). We employed standardized data collection procedures and aim for transparency in this research. The data source is verifiable and certified for use. The data is also anonymized to ensure there is no sense of bias.

3.2 Research Approach – Deductive

Research approach covers the plan to conduct a study and achieve its objectives. It encompasses the general methodological framework and the logic that guides the design, data collection, analysis, and interpretation of research. The research approach provides a roadmap for how the research will be conducted, and it shapes the way this research work will take it’s decisions throughout the entire research process (Rashid, et al. 2019).

There are two types of research approach namely deductive and inductive approaches.

Inductive approach: The inductive research approach is a method of reasoning in which the researcher moves from specific observations to broader generalizations or theories (Woo. SE, et al 2017). It involves generating theories or hypotheses based on the analysis of specific instances, phenomena or patterns. Inductive reasoning is often associated with qualitative research methods, and it is particularly useful when exploring new or complex topics where there are no available theories.

Deductive approach: A deductive research approach in a credit risk assessment project involves starting with a well-defined theory or set of hypotheses and then collecting and analysing data to test and confirm these hypotheses. This approach is characterized by a top-down process, where you begin with a general premise and seek to draw specific conclusions from it Azungah, T. (2018).

One of the benefits of deductive over other approaches is that it starts with a theoretical framework or existing knowledge and derive specific predictions. In credit risk assessment, a deductive approach allows researchers to test hypotheses based on established theories of creditworthiness, financial risk, and economic factors (Mantelaers, E., & Zoet, M. (2018). Also, deductive reasoning emphasizes identifying causal relationships between variables. In credit risk assessment, understanding the causal links between financial indicators, borrower characteristics, and credit outcomes is critical for accurate prediction models (Azungah, T. 2018).

Based on the above and the intended deliverables, this research work is adopting a deductive approach. Below are the steps we intend to follow:

Formulating a Theory or Hypotheses: Deductive research begins with the development of a theory or specific hypotheses related to credit risk assessment (Stol, K. J. et al. 2017). These hypotheses are typically derived from existing literature, financial theories, or expert opinions. In this research work, we hypothesize that some financial and non-financial data such as income, loan type, nature of business, tenure, sex, age, residence type, loan-to-value ratio etc are strongly correlated with credit risk.

Data Collection: In a deductive approach to credit risk assessment, you collect data that is relevant to the hypotheses Azungah, T. 2018). This data typically includes financial statements, credit histories, and other quantitative measures related to borrowers or credit applicants. The focus is on collecting objective and quantifiable data.

Data Analysis: With the collected data in hand, we will use statistical and analytical techniques to analyse it. This may involve various quantitative methods such as regression analysis, correlation analysis, and other statistical tools to test the relationships predicted by our hypotheses.

Testing Hypotheses: Our primary objective is to test each hypothesis we've formulated. For example, we hypothesized that some financial and non-financial data such as income, loan type, nature of business, tenure, sex, age, residence type, loan-to-value ratio etc are strongly correlated with credit risk, you would use statistical analysis to determine whether this relationship holds true in our dataset.

Drawing Conclusions: Based on the results of your data analysis, you draw conclusions about whether your hypotheses are supported or refuted. In a deductive approach, our aim is to draw specific conclusions based on the theoretical framework we started with.

Theory Validation: The final step involves summarizing our research and discussing the validity of our initial theory or hypotheses. If the research supports our hypotheses, it provides empirical evidence to validate the theory. If not, we may need to revise or refine the theory based on our findings (Sarfraz, M., 2018).

3.3 Research Strategy - Quantitative

The next layer involves selecting your research approach. It’s either you are using a quantitative approach, relying on numerical data and statistical methods, a qualitative approach, which involves more in-depth analysis of qualitative information like interviews and case studies or mixed method which involves both approaches (Tashakkori, A., et al. 2020).

Various research methods are commonly employed, including quantitative, qualitative, and mixed methods (Bryman, 2016). Quantitative Research involves collecting and analysing numerical data, using structured surveys, experiments, or observational methods alongside statistical techniques for interpretation (Suphat Sukamolson, 2007). Qualitative Research, on the other hand, delves into non-numerical data, such as opinions and experiences, using methods like interviews or content analysis to explore meanings and themes (Mills, K. A., 2019). Mixed methods combine both quantitative and qualitative approaches, collecting and critically evaluating both numerical and non-numerical data to achieve comprehensive insights (Ishtiaq, M., 2015). Our project has partly quantitative and partly qualitative which aligns with mixed method. However, the qualitative features will be transformed to numeric which is quantitative to enable machine learning algorithm to understand the data.

This research work is however adopting quantitative research methodology. The nature of the research work and the type of data encouraged the use of quantitative method. The data needed for credit analysis in this research context are numeric data and as such would require quantitative research approach.

This research work will utilize some notable tools, techniques, models, software architectures and development environments such as XGBoost machine learning algorithm, python programming language, python libraries, jupyter notebook to achieve the desired result.

3.4 Research Design – Longitudinal

Research design refers to the plan or blueprint that guides the conduct of a research study. It outlines the structure, process, and strategy that researchers use to collect, analyse, and interpret data to address their research questions or objectives. There are several types of research designs, each with its own strengths, weaknesses, and suitability for specific research purposes (Dawadi, S. et al. 2021)

Cross-sectional: Cross-sectional research design is a type of observational study that collects data from participants at a single point in time. This design is characterized by its focus on a snapshot view of a phenomenon, allowing researchers to gather information about variables of interest at a specific moment. Cross-sectional studies are often used in various fields, including epidemiology, sociology, psychology, and marketing. This type of research design does not align with our deliverable in this research work hence, it will not be adopted.

Longitudinal research design: Longitudinal research design is a study design that involves the repeated observation or measurement of the same individuals or groups over an extended period. Unlike cross-sectional studies that collect data at a single point in time, longitudinal studies track changes, developments, or trends over time (Dawadi, S. et al. 2021). This design is particularly useful for understanding the dynamics of individual and group behaviours, relationships, and outcomes.

Longitudinal research is more appropriate if you want to investigate how credit risk evolves over time, track changes in the creditworthiness of individuals or portfolios, and understand the factors contributing to these changes. It is particularly useful for modelling credit risk dynamics and predicting future credit events. Example: Following a cohort of borrowers over several years to assess how their credit scores and financial behaviours change and analysing the factors contributing to these changes. This research is employing longitudinal design approach as we intend to analyse and identify individual credit worthiness or risk inherent in individual applications to make an informed decision (Spector, P. E. 2019).

3.5 Research Techniques and Procedures: Data Collection and Analysis:

Data Collection: This layer involves selecting the specific methods and techniques you'll use to gather data related to credit risk assessment. Common methods include financial statement analysis, credit scoring models, historical data analysis, surveys, interviews, and data mining (Siddiqi, N. (2017). This research is using both financial analysis, credit scoring and historical data. We are not using surveys and interviews due to the volume and quality of data needed for the research coupled with the time and budget for the research work.

- **Data Source**: This research work collected dataset from a public data source. The dataset is certified and approved for use by the owner. The dataset was gotten from a credible source which involves actual credit data from customers of a financial institution. It includes both financial and non-financial dataset.
- **Data Acquisition**: The dataset was acquired legally as it's approved and certified by the owner for research use. The customer's details where anonymized for data protection reasons.
- **Data Quality Assurance:** Ensure the data you collect is of high quality. This includes checking for missing values, duplicates, and inconsistencies. Data quality assurance is crucial to the reliability of your analysis.

Data Analysis: This refers to the process of inspecting, cleaning, transforming, and modelling data to uncover valuable insights, draw conclusions, and support decision-making (Vassakis, K., et al. 2018). In credit risk assessment, this may involve statistical techniques like logistic regression, machine learning algorithms, or credit scoring models to assess the likelihood of default or late payment. Below are some steps this project will follow to analyse collected dataset for our model development:

- **Feature Selection Techniques:** Selecting relevant features is crucial for accurate predictions. Feature selection helps to eliminate irrelevant or redundant features that might negatively impact model performance or increase computational cost. The research will employ techniques like Recursive Feature Elimination (RFE) to remove less important features and tree-based models to provide feature importance scores (Weidong Guo et all., 2022).
- **Data cleaning:** The collected data needs to be processed to make it suitable for analysis. This research work employed some data cleaning tools such as Pandas, NumPy to handle missing values, and transforming the data into a structured format (A. Feeldersa et al, 2020).
- Addressing Imbalanced Data: Loan default datasets often exhibit imbalances, where non-default instances significantly outnumber default instances, leading to biased models. Relying solely on accuracy as an evaluation metric in imbalanced data scenarios can be misleading. Techniques like oversampling, under-sampling (random or informed selection), or employing algorithms designed for imbalanced data, such as cost-sensitive learning or adjusting decision thresholds, help partially balance the class distribution (Sotiris Kotsiantis et al., 2017).

- Validation and Cross-Validation: To evaluate the model's performance on unseen data and prevent overfitting, data is partitioned into training and testing sets or subjected to k-fold cross-validation. In k-fold cross-validation, the data is divided into k subsets, and the model is trained and evaluated k times, using each subset as a testing set once (Kaliappan et al., 2023).
- Model Interpretability Techniques: Various stakeholders, including regulators, borrowers, and internal teams, demand model interpretability. It aids in explaining model predictions by identifying each feature's contributions to the outcome, essential for transparent outcomes. Loan applicants have the right to know the machine learning prediction outcome used to decline their loan request, making interpreting the model outcome crucial. This article will explore one or two model interpretability techniques for machine learning outcomes (Andreas C. et al., 2022).

3.6 Machine Learning Algorithms

Credit risk assessment is a critical aspect of financial decision-making, playing a pivotal role in determining whether individuals or entities are deemed creditworthy and eligible for loans. As financial landscapes evolve and data availability expands, the integration of machine learning algorithms has emerged as a transformative force in enhancing the accuracy and efficiency of credit risk assessment processes (Boukherouaa, et al. 2021). Traditional methods, while effective, often rely on rigid rules and predefined criteria, potentially overlooking subtle patterns and dynamic factors that influence creditworthiness. Machine learning methods are widely used in credit risk assessment due to their ability to handle complex and large datasets (Lappas, P. Z., et al. 2021). Some popular machine learning algorithms we intend to use in the research work include:

Logistic regression: Logistic regression is a classification algorithm, not a regression one. It models the probability of an instance belonging to a particular class and is particularly well-suited for problems where the dependent variable is binary (two classes: 0 or 1) (Dosalwar, S., et al. 2021).

K-nearest neighbour (K-NN): K-Nearest Neighbors (K-NN) is a simple and intuitive machine learning algorithm used for both classification and regression tasks. K-NN is a type of instance-based, lazy learning algorithm, which means it makes predictions based on the similarity of data points in the training set to the data point you want to classify or predict (Song, Y., et al. 2017).

Extreme Gradient Boosting (XGBoost): XGBoost (Extreme Gradient Boosting) is a powerful and popular machine learning algorithm used for both regression and classification tasks. It is known for its efficiency, speed, and high predictive accuracy, making it a preferable choice for various data science and machine learning activities. XGBoost is an ensemble learning method that combines the predictions of multiple individual models (typically decision trees) to improve predictive performance (Kavzoglu, T., et al. 2022).

3.7 Algorithm selection

In selecting the algorithm to use, different factors were considered ranging from accuracy and predictive power, handling data imbalance, robustness to noisy data, feature selection, interpretability etc. In this project, we are using the XGBoost algorithm as to build our model for obvious reasons as mentioned above. We will also compare the algorithm with other ones to see their different levels of accuracy and interpretabilities. XGBoost is renowned for its high predictive accuracy. It excels in capturing complex relationships within the data, which is crucial in credit risk assessment where identifying subtle patterns and risk factors is essential (Somi, S., et al., 2023). It’s ensemble nature, combining multiple weak models, makes it powerful in making accurate predictions. Credit risk assessment datasets are often imbalanced, with most borrowers being low-risk and a minority being high-risk. XGBoost can handle imbalanced datasets effectively and avoid the problem of biased predictions in favour of the majority class (He, S., et al., 2021). XGBoost is robust to noisy data, which can be prevalent in credit risk assessment due to discrepancies in credit reports, data entry errors, or fraud. It's less prone to overfitting compared to some other algorithms (Simão, S. B. S., 2023). XGBoost provides ways to interpret model predictions, making it easier to understand why a particular decision was made. This is crucial in the credit risk assessment field, where transparency is often required for regulatory compliance. XGBoost is designed for parallel and distributed computing, which speeds up model training and prediction. This is valuable for handling large credit risk datasets efficiently (Qin, C., Zhang., et al., 2021).

3.8 Why Xgboost Algorithm

XGBoost (Extreme Gradient Boosting) is a popular machine learning algorithm that has proven to be effective in various domains, including credit risk assessment. XGBoost is an ensemble learning algorithm that combines the predictions of multiple weak learners (typically decision trees) to create a strong learner. The ensemble approach helps improve the model's predictive performance, generalization, and robustness (Ma, M., et al. 2021).

XGBoost has the potential to address some of the research gaps of overfitting, interpretability and efficiency raised in this research work. XGBoost provides a mechanism for assessing feature importance, allowing analysts and model users to understand the factors contributing most to the model's predictions (Ma, Y., et al. 2023). This transparency is crucial in credit risk assessment, where interpretability is often required for regulatory compliance and risk management. In credit risk assessment, where accuracy, interpretability, and the ability to handle complex patterns are crucial, XGBoost's combination of ensemble learning, regularization, and handling of imbalanced data make it a powerful and preferred choice for building predictive models (Wang, C., et al. 2020).

3.9 Stages of model building

Below are the sequential stages outlined by this research work for the development loan eligibility checker model for improved predictive performance.

Data collection: The research work will gather relevant data from approved data gathering data sources that will include historical loan data, credit scores, income information, and other important features needed for model building.

Data processing: The data will be cleaned and pre-processed, addressing missing values, outliers, and encoding categorical variables. We will also normalize or scale numerical features to bring them to a standard scale.

Exploratory Data Analysis (EDA): The research will conduct exploratory data analysis to gain insights into the distribution of key variables, identify patterns, and understand the characteristics of low-risk and high-risk borrowers.

Feature selection: The research will utilize the power of xgboost to select the most relevant and significant features (input variables) from the overall features to train a model. The goal is to enhance model performance, reduce complexity, and improve interpretability.

Feature Engineering: The research work will include creating new features or transform existing ones to capture relevant information about borrowers' financial behaviour. Derive features such as debt-to-income ratios, payment history if not included originally in the dataset.

Data Splitting: The dataset will be divided into training, and testing datasets. We will allocate most of the data for training, and a smaller dataset for testing of the model. Separating the dataset into training and testing prevents the model from gaining knowledge about the testing dataset during training, ensuring a fair evaluation (Raschka, S. (2018).

Hyperparameter Tuning: The next stage is fine tuning the model's hyperparameters using the validation set to optimize its predictive performance. This will be achieved by adjusting some parameters such as learning rates, tree depths, or regularization.

Model Evaluation: The model will be evaluated for it's performance on the test set using relevant metrics such as accuracy, precision, recall, F1-score, and AUC-ROC.

3.10 Software Architecture

This project has option of using Python or R as the programming language being popular data science languages in the research work. The choice between Python and R often depends on personal preference, the specific needs of a project, and the existing tools and ecosystems used by a data science team. Below is a brief of the benefits of Python over R.

3.11 Benefits of Python over R

Python has a rich ecosystem of libraries and frameworks that are widely used in data science. Popular libraries like NumPy, pandas, scikit-learn, TensorFlow, and PyTorch provide robust tools for data manipulation, machine learning, and deep learning. This project has much of visualizations and Python offers a variety of data visualization libraries, such as Matplotlib, Seaborn. These libraries provide flexibility in creating static and interactive visualizations for data exploration and communication. In terms tools and platform, many tools and platforms in data science, such as Jupyter Notebooks, have strong support for Python. Jupyter Notebooks are widely used for interactive data analysis and code sharing (Sarkar, D., et al. 2018).

Based on the above strengths, this project intends to use Python, python libraries such as pandas, numpy, matplotlib scikit-learn as design tools. Also, for planning and project implementation monitoring, agile project management methodology such as Scrum will be used.

3.12 Software Testing

Software testing constitutes a significant aspect of the software development lifecycle, aiming to identify bugs, defects, ensure software quality, and enhance the final product's reliability. This research employs diverse software testing methods, each tailored to address specific software aspects and reveal distinct defect types, including Unit testing, Integrated testing, functional testing, performance testing, security testing, useability testing, and acceptance testing (Jamil et al., 2018).

Ethics, Risks, Social, and Legal Issues:

This project utilizes real data from previous borrowers, encompassing both performing and non-performing loans. To ensure ethical considerations and mitigate associated risks, the following measures are implemented:

- **Privacy Violations:** Mitigation involves obtaining consent from data owners, preventing third-party access, and securely destroying data post-analysis (Filkins et al., 2016). Data minimization is applied, collecting only necessary information for credit risk assessment, avoiding irrelevant sensitive details.
- **Data Security:** Robust security measures, including encryption, secure protocols, and access controls, are implemented to safeguard data from unauthorized access during storage and transmission (Bandari, V., 2023). Adherence to industry best practices for data security is ensured.
- **Legal and Regulatory Compliance**: The project adheres to regulatory guidelines and relevant data protection regulations (e.g., GDPR, CCPA), ensuring the loan eligibility checker model aligns with legal standards for privacy protection (Tikkinen-Piri et al., 2018).
- **Informed Consent**: The project strictly employs datasets approved and certified for use by the owner, with explicit consent obtained. Stakeholders, including the model evaluator, are informed about the purposes, potential outcomes, and duration of data usage.
- **Data Misuse**: To prevent potential misuse, the research establishes explicit ethical guidelines, policies, and fair decision-making practices for using the credit risk assessment model's results (Kaur et al., 2021).
- **Anonymization and De-identification:** Personal loan data undergoes thorough anonymization or de-identification to minimize reidentification risks and protect individuals' privacy (Joo et al., 2018), involving the replacement or removal of direct identifiers.

Chapter Summary

The research methodology employs a comprehensive approach, utilizing the research onion framework and adopting a positivist research philosophy. Positivism aligns with the project's quantitative nature and aims for credit risk assessment. The deductive research approach facilitates hypothesis testing, generalizability, and a focus on objective data (Park, Y. S., et al. 2020). XGBoost is chosen due to its accuracy, interpretability, and robustness, addressing imbalanced datasets. The longitudinal research design tracks credit risk changes over time. Ethical considerations involve privacy protection, data security measures, legal compliance, and mitigating data misuse. A detailed software architecture plan favours Python for its extensive data science libraries, and rigorous software testing methods ensure quality. The stages of model building include data collection, processing, exploratory data analysis, feature selection, engineering, data splitting, hyperparameter tuning, and model evaluation.

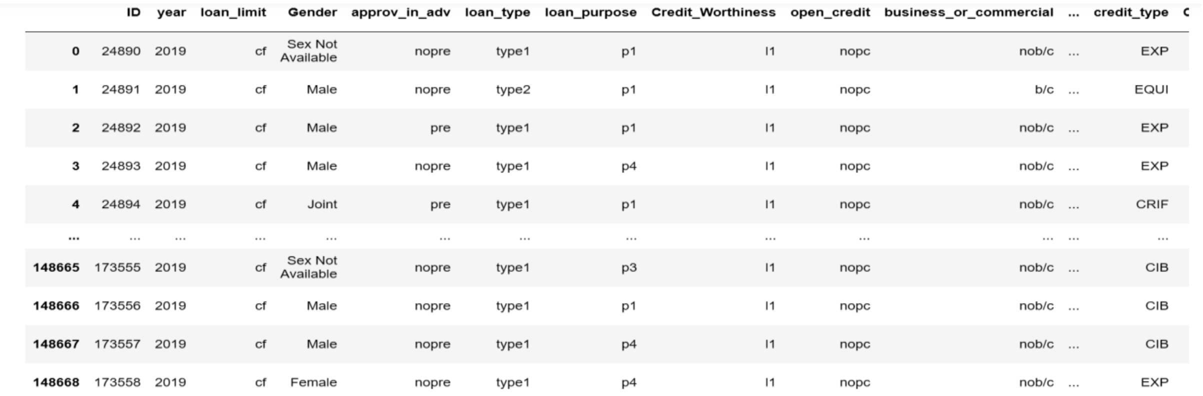
CHAPTER 4: DATA COLLECTION, TOOLS, TECHNIQUES AND MODELS

This chapter outlines the different factors considered and the stages involved in building the model ranging from data collection, transformation to model building.

4.1 Data collection method

The two main types of data collection are primary and secondary data collection. Primary datasets are often collected through experiments, surveys, observations, or other direct data collection methods Mazhar. S.A, et all. 2021). While secondary datasets include publicly available datasets, datasets obtained from previous studies, or data sourced from external databases.

This research work is using dataset collected from an open data source. The dataset was dully certified, approved for use, raw and not yet cleaned. The dataset used for this research had 34 features before data processing was initiated.



	ID	year	loan_limit	Gender	approv_in_adv	loan_type	loan_purpose	Credit_Worthiness	open_credit	business_or_commercial	...	credit_type	C
0	24890	2019	cf	Sex Not Available	nopre	type1	p1	l1	nopc	nob/c	...	EXP	
1	24891	2019	cf	Male	nopre	type2	p1	l1	nopc	b/c	...	EQUI	
2	24892	2019	cf	Male	pre	type1	p1	l1	nopc	nob/c	...	EXP	
3	24893	2019	cf	Male	nopre	type1	p4	l1	nopc	nob/c	...	EXP	
4	24894	2019	cf	Joint	pre	type1	p1	l1	nopc	nob/c	...	CRIF	
...	...	...	...	...	...	...	...	...	...	...	...	...	
148665	173555	2019	cf	Sex Not Available	nopre	type1	p3	l1	nopc	nob/c	...	CIB	
148666	173556	2019	cf	Male	nopre	type1	p1	l1	nopc	nob/c	...	CIB	
148667	173557	2019	cf	Male	nopre	type1	p4	l1	nopc	nob/c	...	CIB	
148668	173558	2019	cf	Female	nopre	type1	p4	l1	nopc	nob/c	...	EXP	

Figure 4.1 - Sample raw dataset

4.2 Data preprocessing tools

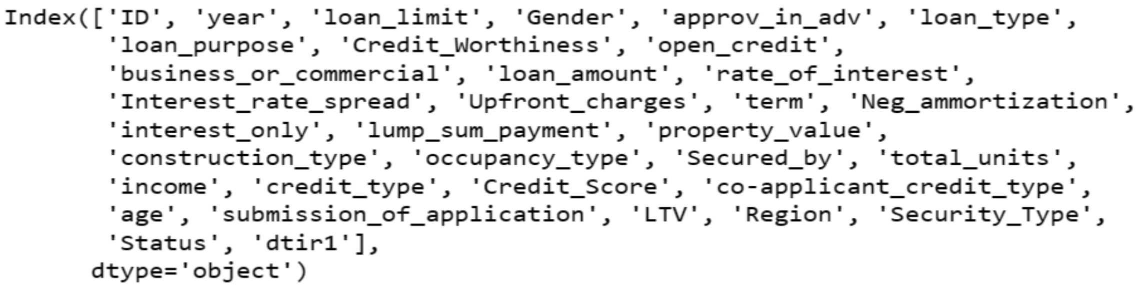
The dataset contains both quantitative and qualitative data types.

The dataset details the credit history, employment status, income, outstanding debts, cashflow, other financial indicators. The collected data needs to be processed to make it suitable for analysis. This research work employed some data cleaning tools such as Pandas, NumPy to handle missing values, and transforming the data into a structured format (A. Feeldersa et al, 2000).

Techniques for Feature Selection: The accurate prediction of outcomes relies on the careful selection of relevant features. Feature selection is instrumental in eliminating irrelevant or redundant features, preventing potential negative impacts on model performance, or increased computational costs. In this research, methods such as Recursive Feature Elimination (RFE) were applied to eliminate less important features, and tree-based models were utilized to generate feature importance scores (Weidong Guo et al., 2022).

4.3 Data cleaning

The data was cleaned to address issues like missing values, outliers, and noise. The issue of duplicate values was addressed and removed. Cleaning ensures that the data is of high quality and ready for analysis. Lots of columns that do not have any meaningful impact to credit risk were removed.

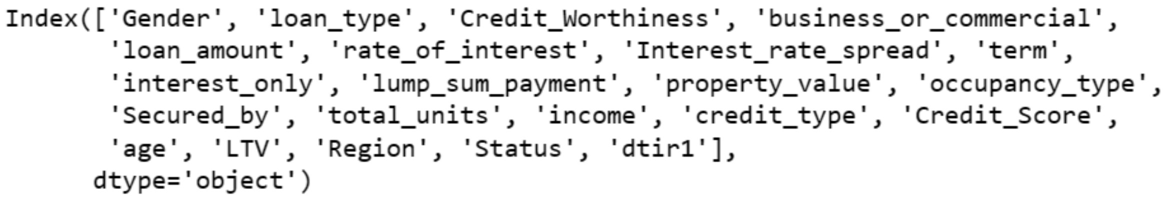


```
Index(['ID', 'year', 'loan_limit', 'Gender', 'approv_in_adv', 'loan_type',
       'loan_purpose', 'Credit_Worthiness', 'open_credit',
       'business_or_commercial', 'loan_amount', 'rate_of_interest',
       'Interest_rate_spread', 'Upfront_charges', 'term', 'Neg_ammortization',
       'interest_only', 'lump_sum_payment', 'property_value',
       'construction_type', 'occupancy_type', 'Secured_by', 'total_units',
       'income', 'credit_type', 'Credit_Score', 'co-applicant_credit_type',
       'age', 'submission_of_application', 'LTV', 'Region', 'Security_Type',
       'Status', 'dtir1'],
      dtype='object')
```

Figure 4.2 - All columns

The following columns were classified as not relevant and were therefore dropped from the dataset because of their limited impact on the outcome as discussed below:

ID, year, loan limit, approved in advance, loan_purpose, open credit, upfront_charges, co-applicant_credit_type, submission_of_application, security_type, construction_type, neg_amortisation.



```
Index(['Gender', 'loan_type', 'Credit_Worthiness', 'business_or_commercial',
       'loan_amount', 'rate_of_interest', 'Interest_rate_spread', 'term',
       'interest_only', 'lump_sum_payment', 'property_value', 'occupancy_type',
       'Secured_by', 'total_units', 'income', 'credit_type', 'Credit_Score',
       'age', 'LTV', 'Region', 'Status', 'dtir1'],
      dtype='object')
```

Figure 4.3 - New columns

Reasons for the removal of the listed features:

ID: The 'ID' in the dataset is meaningless as it does not have any contribution to risk assessment.

Year: The dataset has same year for all records; hence, it is not needed.

Loan limit: This supposed to be set after the credit risk assessment. It is not needed to assess the credit risk of an individual.

Approved in advance: Approving loans in advance does not have any effect on credit risk assessment.

Loan Purpose: Loan purpose, though important in approving loans, it does not have a direct relationship with the risk in the person or business. It only comes in when we are considering the type of loan to give to suit the purpose of the loan. So, it is not needed in this research work.

Open credit: The dataset does not have any open credit record. All the records are the same as 'nopc', meaning no open credit. It therefore does not have any impact on the outcome of the assessment since it's the same for all.

Upfront charges: This feature does not have any impact on risk assessment. Whether or not there is upfront charge, the credit risk remains the same.

Co-applicant credit type: Both applicants have the same application type as they are requesting for same credit. The feature has no effect on the dataset.

Submission of application: The manner application is submitted is inconsequential. Instant or not instant. It has no effect on the result of the credit risk assessment.

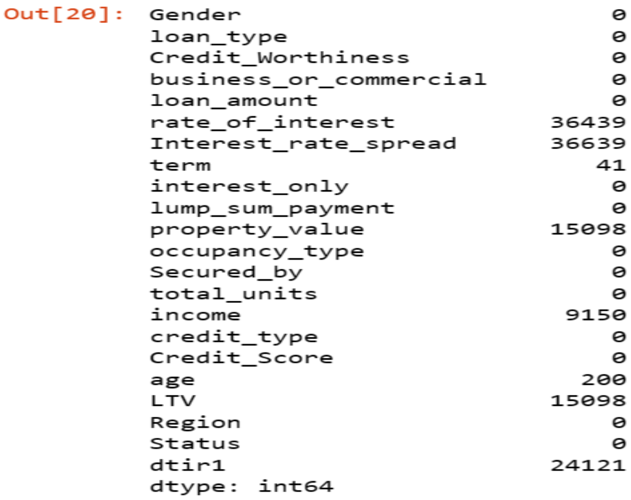
Security type: Security type is an import feature in credit risk assessment; however, the security type has already been listed as ‘secured_by’ option which is explicit. The ‘security_type’ feature in the dataset is meaningless, hence it was dropped.

Construction type: Construction type is not a relevant feature in credit risk, therefore it was dropped from the dataset.

In total, 12 features were removed from the dataset to give a proper dataset with features relevant to building a model for credit risk assessment.

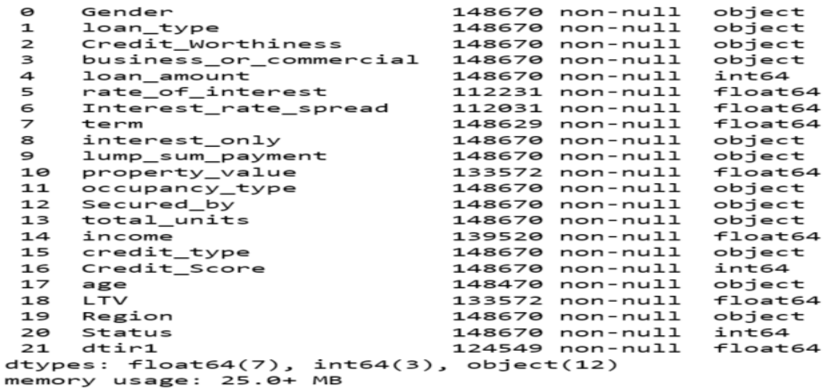
Negative amortization: There is no room for negative amortization as it means there is already a default which should not be a feature.

Looking at the dataset, there are lots of null values that need to be addressed. Considering the features, the mean of the values will to be taken and used to replace the null values for rate_of_interest, interet_rate_spread, property_value, term, LTV and dtir1.



```
Out[20]: Gender                           0
         loan_type                        0
         Credit_Worthiness                0
         business_or_commercial           0
         loan_amount                      0
         rate_of_interest             36439
         Interest_rate_spread         36639
         term                            41
         interest_only                    0
         lump_sum_payment                 0
         property_value               15098
         occupancy_type                   0
         Secured_by                       0
         total_units                      0
         income                        9150
         credit_type                      0
         Credit_Score                     0
         age                            200
         LTV                          15098
         Region                           0
         Status                           0
         dtir1                        24121
         dtype: int64
```

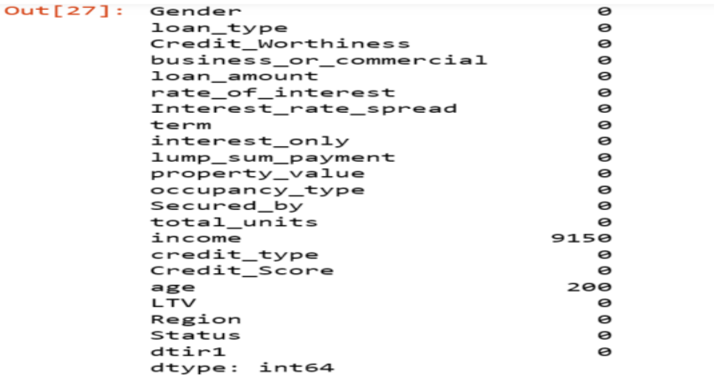
Figure 4.4 – Null values 1



```
 0    Gender                       148670 non-null   object
 1    loan_type                    148670 non-null   object
 2    Credit_Worthiness            148670 non-null   object
 3    business_or_commercial       148670 non-null   object
 4    loan_amount                  148670 non-null   int64
 5    rate_of_interest             112231 non-null   float64
 6    Interest_rate_spread         112031 non-null   float64
 7    term                         148629 non-null   float64
 8    interest_only                148670 non-null   object
 9    lump_sum_payment             148670 non-null   object
 10   property_value               133572 non-null   float64
 11   occupancy_type               148670 non-null   object
 12   Secured_by                   148670 non-null   object
 13   total_units                  148670 non-null   object
 14   income                       139520 non-null   float64
 15   credit_type                  148670 non-null   object
 16   Credit_Score                 148670 non-null   int64
 17   age                          148470 non-null   object
 18   LTV                          133572 non-null   float64
 19   Region                       148670 non-null   object
 20   Status                       148670 non-null   int64
 21   dtir1                        124549 non-null   float64
dtypes: float64(7), int64(3), object(12)
memory usage: 25.0+ MB
```

Figure 4.5 – Features info

The null values were replaced with features mean values to give the figure below.



```
Out[27]:  Gender                          0
          loan_type                       0
          Credit_Worthiness               0
          business_or_commercial          0
          loan_amount                     0
          rate_of_interest                0
          Interest_rate_spread            0
          term                            0
          interest_only                   0
          lump_sum_payment                0
          property_value                  0
          occupancy_type                  0
          Secured_by                      0
          total_units                     0
          income                       9150
          credit_type                     0
          Credit_Score                    0
          age                           200
          LTV                             0
          Region                          0
          Status                          0
          dtir1                           0
          dtype: int64
```

Figure 4.6 – Null values 2

The next is to drop the null values of ‘age’ since the null values are not much and the income with age has a wide variability as is shown in the figure below.

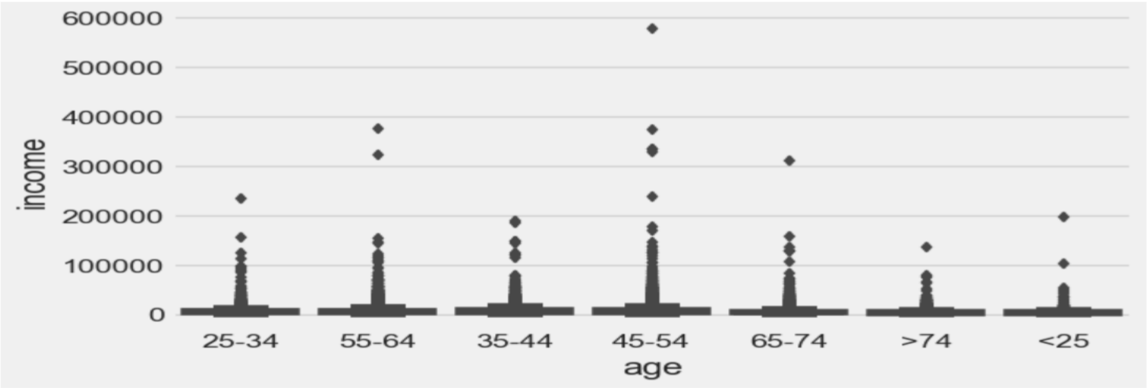
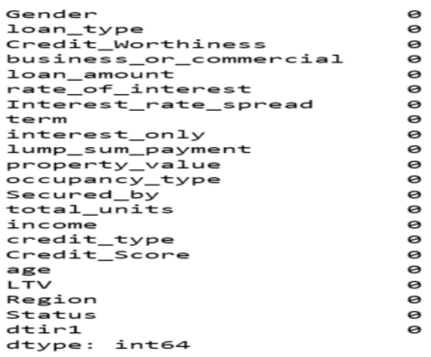


Figure 4.7 – Income to age

Due to this income variability with age, the null values in income feature will be dropped to ensure better model accuracy.



```
Gender                       0
loan_type                    0
Credit_Worthiness            0
business_or_commercial       0
loan_amount                  0
rate_of_interest             0
Interest_rate_spread         0
term                         0
interest_only                0
lump_sum_payment             0
property_value               0
occupancy_type               0
Secured_by                   0
total_units                  0
income                       0
credit_type                  0
Credit_Score                 0
age                          0
LTV                          0
Region                       0
Status                       0
dtir1                        0
dtype: int64
```

Figure 4.8 – No Null values

4.4 Data Transformation

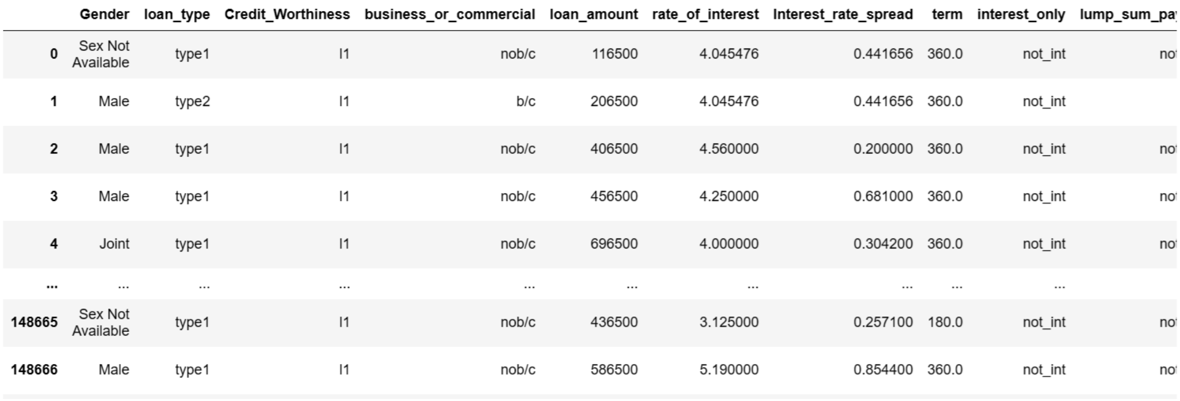
Data transformation is a crucial step in the machine learning (ML) process that involves modifying or converting the categorical input data to numeric representations to make it more suitable for model training. The goal is to enhance the quality and efficiency of the learning process. There are different types of data transformation such as label encoding, one-hot encoding or binary encoding (Cheraghi, Y., et al. 2021). This project used label encoding to achieve data transformation.

- Label encoding is a method of converting categorical data into numerical form. Each unique category is assigned a unique integer label.
- One-hot encoding is a technique for converting categorical variables into a binary matrix, where each category becomes a binary column.
- Binary encoding is a compromise between label encoding and one-hot encoding. It converts each integer label into binary code and represents it in columns.

This project will used label encoding to achieve data transformation for the following reasons:

- Label encoding is suitable when the categorical variable has an intrinsic ordinal relationship. For example, if a variable represents different credit ratings (e.g., low, medium, high), label encoding assigns numeric labels in ascending order, preserving the ordinal nature of the categories.

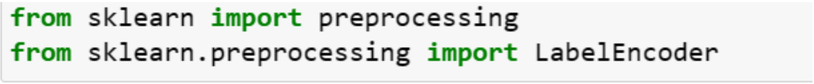
- Compared to one-hot encoding, label encoding decreases dimensionality, providing an advantage when faced with numerous unique categories in a categorical variable. This reduction in dimensionality can enhance the efficiency of model training, particularly with algorithms such as XGBoost.
- Label encoding can contribute to the interpretability of the model. The encoded numeric labels directly represent the order or hierarchy of categories, making it easier to interpret the significance of each category in the credit risk assessment.
- Being an ensemble learning algorithm rooted in decision trees, XGBoost leverages the innate capability of decision trees to handle ordinal data seamlessly. The use of label encoding, wherein integer labels are assigned based on category order, aligns effectively with the decision-making splits intrinsic to decision trees, rendering it a fitting choice for XGBoost.



	Gender	loan_type	Credit_Worthiness	business_or_commercial	loan_amount	rate_of_interest	Interest_rate_spread	term	interest_only	lump_sum_pa
0	Sex Not Available	type1	l1	nob/c	116500	4.045476	0.441656	360.0	not_int	no
1	Male	type2	l1	b/c	206500	4.045476	0.441656	360.0	not_int	
2	Male	type1	l1	nob/c	406500	4.560000	0.200000	360.0	not_int	no
3	Male	type1	l1	nob/c	456500	4.250000	0.681000	360.0	not_int	no
4	Joint	type1	l1	nob/c	696500	4.000000	0.304200	360.0	not_int	no
...	...	...	...	...	...	...	...	...	...	
148665	Sex Not Available	type1	l1	nob/c	436500	3.125000	0.257100	180.0	not_int	no
148666	Male	type1	l1	nob/c	586500	5.190000	0.854400	360.0	not_int	no

Figure 4.9 – untransformed data

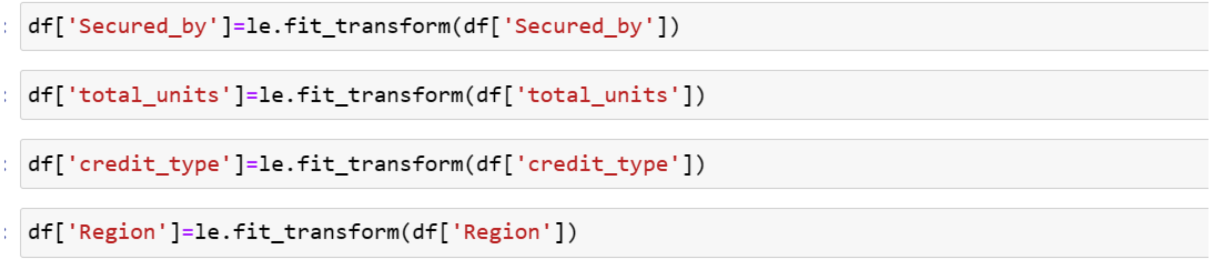
Applying transformation, we imported the LabelEncoder class from sklearn library.



```
from sklearn import preprocessing
from sklearn.preprocessing import LabelEncoder
```

Figure 4.10 - importing preprocessing.

The actual transformation code:



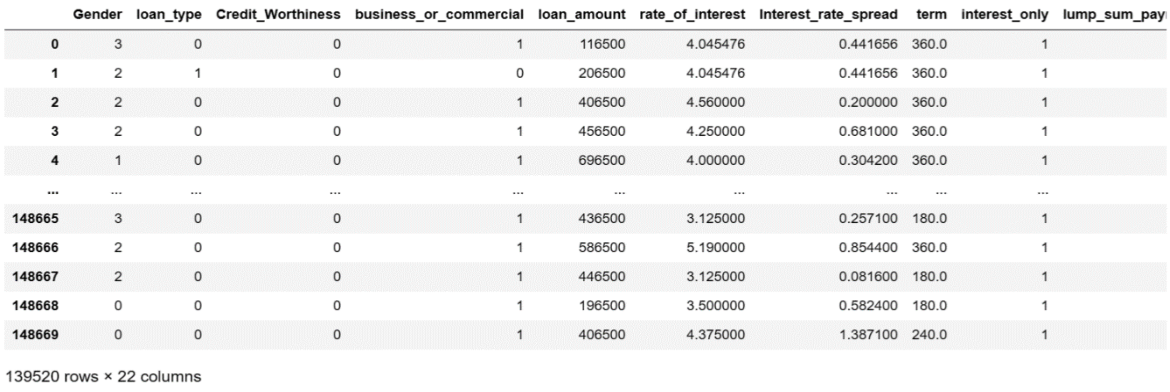
```
df['Secured_by']=le.fit_transform(df['Secured_by'])
```

```
df['total_units']=le.fit_transform(df['total_units'])
```

```
df['credit_type']=le.fit_transform(df['credit_type'])
```

```
df['Region']=le.fit_transform(df['Region'])
```

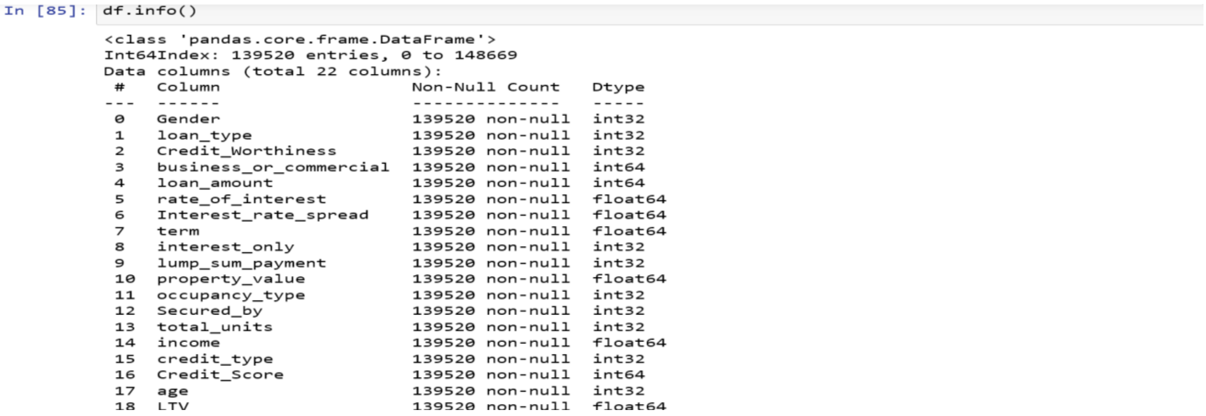
Figure 4.11 - transformation code



	Gender	loan_type	Credit_Worthiness	business_or_commercial	loan_amount	rate_of_interest	Interest_rate_spread	term	interest_only	lump_sum_pay
0	3	0	0	1	116500	4.045476	0.441656	360.0	1	
1	2	1	0	0	206500	4.045476	0.441656	360.0	1	
2	2	0	0	1	406500	4.560000	0.200000	360.0	1	
3	2	0	0	1	456500	4.250000	0.681000	360.0	1	
4	1	0	0	1	696500	4.000000	0.304200	360.0	1	
...	...	...	...	...	...	...	...	...	...	
148665	3	0	0	1	436500	3.125000	0.257100	180.0	1	
148666	2	0	0	1	586500	5.190000	0.854400	360.0	1	
148667	2	0	0	1	446500	3.125000	0.081600	180.0	1	
148668	0	0	0	1	196500	3.500000	0.582400	180.0	1	
148669	0	0	0	1	406500	4.375000	1.387100	240.0	1	

139520 rows × 22 columns

Figure 4.12 - transformed dataset.



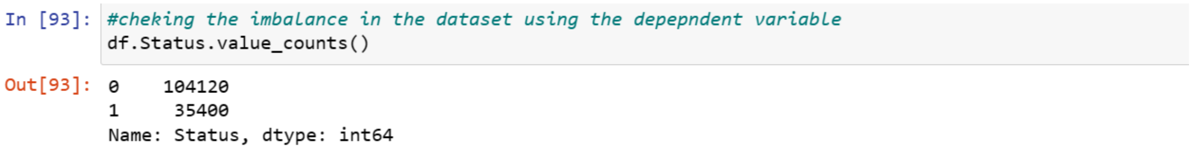
```
In [85]: df.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 139520 entries, 0 to 148669
Data columns (total 22 columns):
 #   Column                  Non-Null Count   Dtype
---  ------                  --------------   -----
 0   Gender                  139520 non-null  int32
 1   loan_type               139520 non-null  int32
 2   Credit_Worthiness       139520 non-null  int32
 3   business_or_commercial  139520 non-null  int64
 4   loan_amount             139520 non-null  int64
 5   rate_of_interest        139520 non-null  float64
 6   Interest_rate_spread    139520 non-null  float64
 7   term                    139520 non-null  float64
 8   interest_only           139520 non-null  int32
 9   lump_sum_payment        139520 non-null  int32
 10  property_value          139520 non-null  float64
 11  occupancy_type          139520 non-null  int32
 12  Secured_by              139520 non-null  int32
 13  total_units             139520 non-null  int32
 14  income                  139520 non-null  float64
 15  credit_type             139520 non-null  int32
 16  Credit_Score            139520 non-null  int64
 17  age                     139520 non-null  int32
 18  LTV                     139520 non-null  float64
```

Figure 4.13 - feature data types

4.5 Handling Imbalanced Dataset

An imbalanced dataset refers to a situation where the distribution of the target classes is not uniform. In other words, one or more classes in the dataset have significantly fewer instances compared to other classes. Class imbalance is a common issue, and it can impact the performance of machine learning models, particularly in classification tasks (Tahir, et al. 2019).

In this project scenario, where there are only two classes, there is an existing imbalanced dataset where the majority class is 75% and the minority class is 25%.



```
In [93]: #cheking the imbalance in the dataset using the depepndent variable
         df.Status.value_counts()

Out[93]: 0    104120
         1     35400
         Name: Status, dtype: int64
```

Figure 4.14 - class imbalance

4.6 Challenges posed by imbalanced datasets

Below are some challenges posed by having or using an imbalanced dataset.

Bias in Model Training: Machine learning models trained on imbalanced datasets may become biased toward the majority class, leading to poor generalization on minority class instances.

Poor Predictive Performance: The performance metrics, such as accuracy, may not be reliable indicators of a model's effectiveness on imbalanced datasets. A model that predicts the majority class for every instance may achieve high accuracy but perform poorly on the minority class.

Difficulty in Learning Minority Class Patterns: Models may struggle to identify patterns in the minority class due to the limited number of examples, leading to lower sensitivity or recall for that class (Abokadr, et al. 2023).

4.7 Techniques to address imbalanced datasets

Resampling: Oversampling involves increasing the number of instances in the minority class. While under-sampling involves reducing the number of instances of the majority class. The goal is to balance the class distribution, allowing the model to better learn patterns in the minority class. Below are two major types of resampling technique:

- **Over-sampling**: This involves increasing the number of instances in the minority class by duplicating or generating synthetic examples (e.g., using SMOTE - Synthetic Minority Over-sampling Technique).

- **Under-sampling**: This involves decreasing the number of instances in the majority class by randomly removing instances of the majority class (Taha, et al. 2021).

4.8 Benefits of over-sampling over under-sampling techniques

Retention of Information: Over-sampling involves increasing the number of instances of the minority class by duplicating or generating synthetic samples. This helps retain more information from the minority class compared to under-sampling, where instances from the majority class are removed. Retaining more information from the minority class can lead to a more representative and robust model.

Avoidance of Information Loss: Under-sampling involves removing instances from the majority class, potentially leading to information loss. This can be critical in scenarios where the majority class contains important patterns or variations that contribute to the overall understanding of the data. Over-sampling avoids this issue by preserving the majority class instances.

Mitigation of Bias: Under-sampling may introduce bias into the model, especially when the removed instances from the majority class contain crucial information. Over-sampling can help mitigate this bias by ensuring that all instances, both from the majority and minority classes, contribute to the learning process (Ahmed, S. et al. 2021).

Compatibility with Algorithms: Some machine learning algorithms may not perform well when trained on imbalanced datasets. Over-sampling can enhance the compatibility of these algorithms with imbalanced data, as it provides more instances for the minority class, helping the algorithm better learn its patterns (Shelke, et al. 2017).

Based on the above benefits, the project is using over-sampling to address class imbalance in the dataset.

4.9 Types of over-sampling methods

There are several over-sampling methods used to address class imbalance in machine learning. Each method has its own approach to generating synthetic instances for the minority class. Here are some common types of over-sampling methods:

Random Over-sampling: Random over-sampling method randomly duplicates instances from the minority class until a more balanced distributed of dataset is achieved. The disadvantage is that it can lead to over fitting.

SMOTE (Synthetic Minority Over-sampling Technique): Synthetic minority over-sampling technique (SMOTE) generates synthetic instances by interpolating between existing minority class instances. SMOTE introduces diversity and reduces the risk of overfitting.

ADASYN (Adaptive Synthetic Sampling): Like SMOTE, ADASYN focuses on generating synthetic instances for those minority class instances that are harder to learn by the model. However, it is computationally more expensive than SMOTE.

Because SMOTE handles class imbalance without over fitting and computationally inexpensive, this research work is adopting it (Shelke, et al. 2017).

Synthetic Minority Oversampling Technique – SMOTE

The Synthetic Minority Over-Sampling Technique (SMOTE) is a popular algorithm used to address the class imbalance problem in machine learning. It works by generating synthetic samples for the minority class, thereby increasing its representation in the dataset. SMOTE was introduced by (Chawla et al. in 2002) as a method to improve the learning of classifiers on imbalanced datasets.

Below is a brief process of how SMOTE works:

Selecting Instances: For each instance in the minority class, SMOTE selects k nearest neighbours from the same class. The value of k is a user-defined parameter.

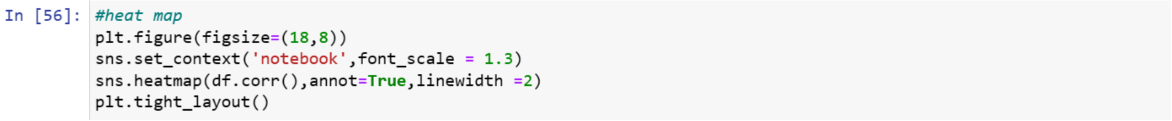
Generating Synthetic Samples: For each selected instance, SMOTE generates synthetic instances along the line segments connecting that instance to its k nearest neighbours. The number of synthetic instances to generate is also a user-defined parameter, often denoted as the "oversampling ratio."

Adding Synthetic Samples: The synthetic instances are added to the original dataset, effectively oversampling the minority class.

4.10 Correlation matrix

This project used heat map to visualize the correlation matrix of features. This helps identify pairs of features that are highly correlated with each other, most importantly in relationship to the dependent variable.

Heatmaps provide a visual summary of the correlation matrix, making it easier to identify patterns and relationships between variables.



```
In [56]: #heat map
         plt.figure(figsize=(18,8))
         sns.set_context('notebook',font_scale = 1.3)
         sns.heatmap(df.corr(),annot=True,linewidth =2)
         plt.tight_layout()
```

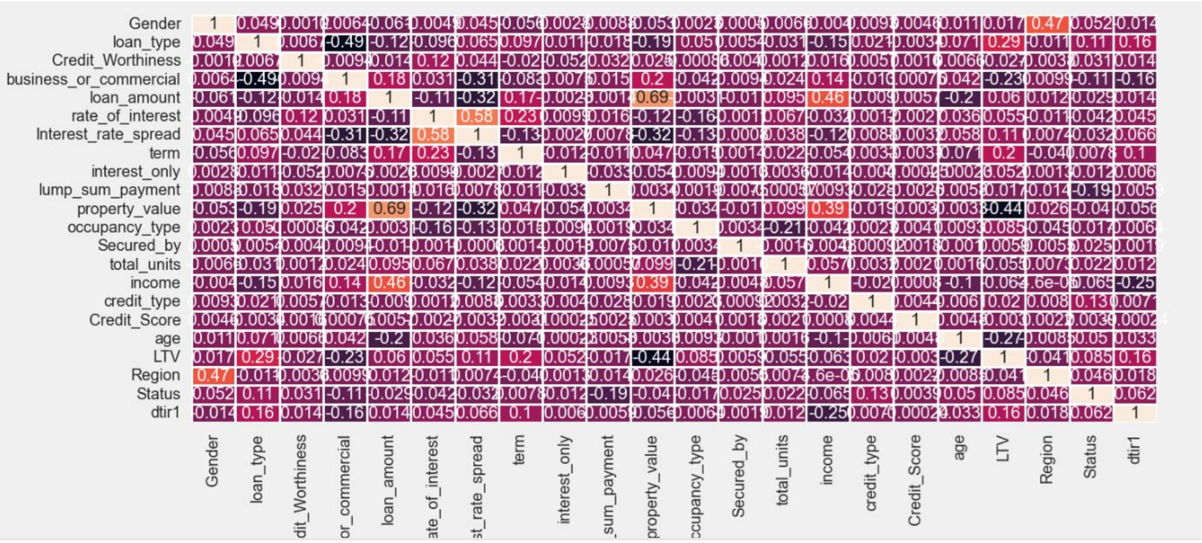


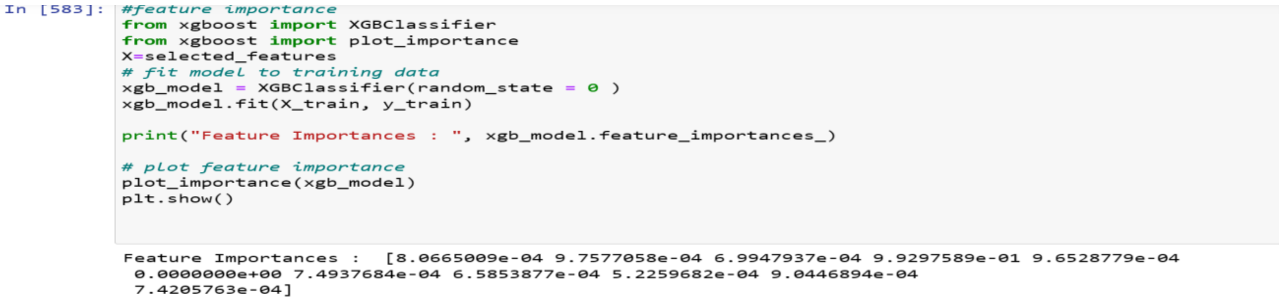
Figure 4.15 - Heat map

4.11 Feature Selection

Feature selection is a process in machine learning that involves choosing a subset of relevant features or variables from a larger set of features. The goal of feature selection is to improve the performance of a model by selecting the most informative and discriminative features while discarding irrelevant or redundant ones. This research work made use of Recursive Feature Elimination (RFE). This process of feature selection iteratively removes the least important features based on model performance (Li, J. et al. 2017). This was further confirmed by using the xgboost feature importance classification to generate the features with the highest weighted influence on the dependent variable which is ‘status’ in the dataset.

This is often a crucial step in the model development process to enhance model interpretability, reduce overfitting, and improve overall performance.

s



```
In [583]: #feature importance
          from xgboost import XGBClassifier
          from xgboost import plot_importance
          X=selected_features
          # fit model to training data
          xgb_model = XGBClassifier(random_state = 0 )
          xgb_model.fit(X_train, y_train)

          print("Feature Importances : ", xgb_model.feature_importances_)

          # plot feature importance
          plot_importance(xgb_model)
          plt.show()

          Feature Importances :  [8.0665009e-04 9.7577058e-04 6.9947937e-04 9.9297589e-01 9.6528779e-04
           0.0000000e+00 7.4937684e-04 6.5853877e-04 5.2259682e-04 9.0446894e-04
           7.4205763e-04]
```

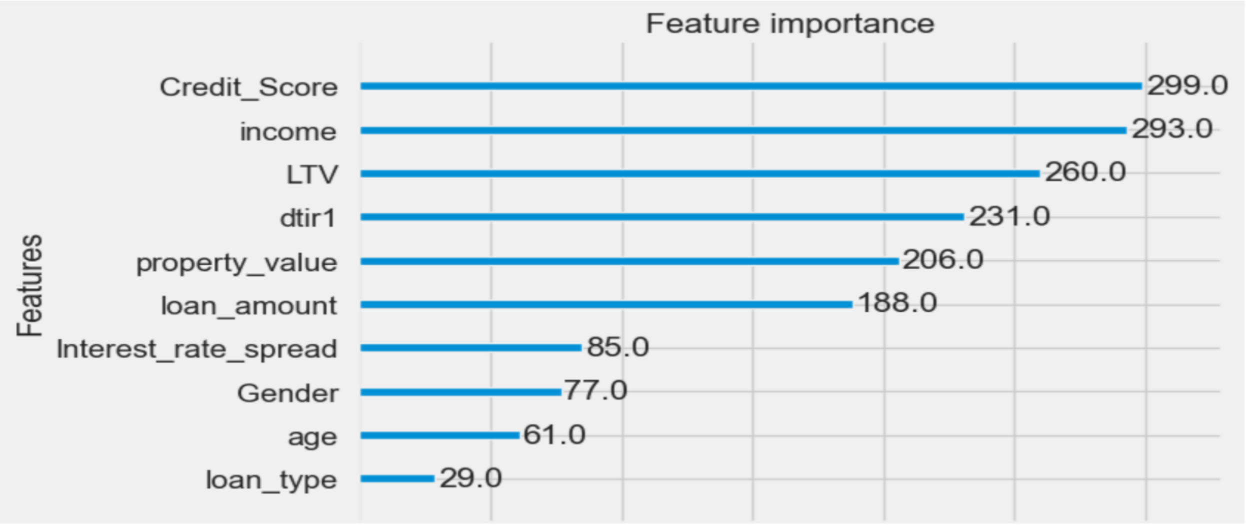


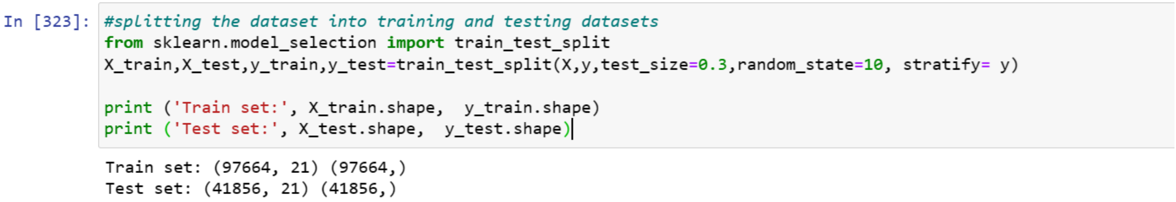
Figure 4.16 - important features

The above code gives us a list of the most important features that influence the dependent variable. This ensures better interpretability, reduce overfitting and improve the overall performance of our model.

4.12 Splitting Dataset

The cleaned dataset was split into two parts for training and testing. The training dataset was 70% of the total dataset while the testing dataset was 30% of the total dataset.

This research work is using SMOTE to address class imbalance in the dataset for better quality and accuracy of prediction.



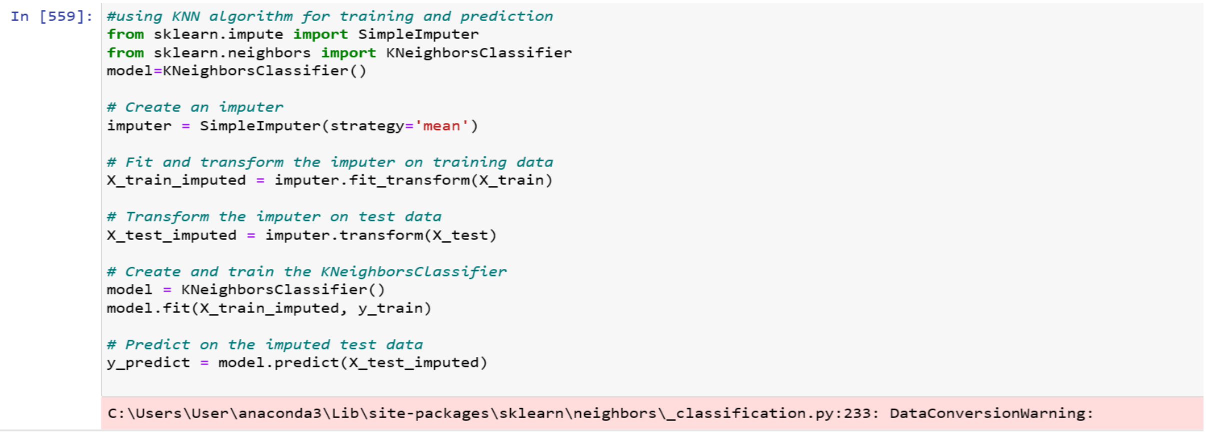
```
In [323]: #splitting the dataset into training and testing datasets
          from sklearn.model_selection import train_test_split
          X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.3,random_state=10, stratify= y)

          print ('Train set:', X_train.shape,  y_train.shape)
          print ('Test set:', X_test.shape,  y_test.shape)

          Train set: (97664, 21) (97664,)
          Test set: (41856, 21) (41856,)
```

Figure 4.17 - splitting dataset.

4.13 Model Building

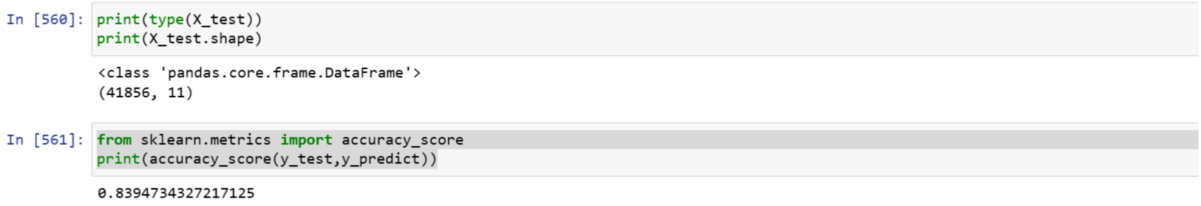
This project is researching on enhancing loan eligibility using xgboost algorithm. Hence few algorithms were considered, and their results compared to that of xgboost algorithm. The first algorithm considered for model development was K-NN.

Using K-NN algorithm to build a model, we got an accuracy of 83.9%.



```
In [559]: #using KNN algorithm for training and prediction
          from sklearn.impute import SimpleImputer
          from sklearn.neighbors import KNeighborsClassifier
          model=KNeighborsClassifier()

          # Create an imputer
          imputer = SimpleImputer(strategy='mean')

          # Fit and transform the imputer on training data
          X_train_imputed = imputer.fit_transform(X_train)

          # Transform the imputer on test data
          X_test_imputed = imputer.transform(X_test)

          # Create and train the KNeighborsClassifier
          model = KNeighborsClassifier()
          model.fit(X_train_imputed, y_train)

          # Predict on the imputed test data
          y_predict = model.predict(X_test_imputed)

          C:\Users\User\anaconda3\Lib\site-packages\sklearn\neighbors\_classification.py:233: DataConversionWarning:
```

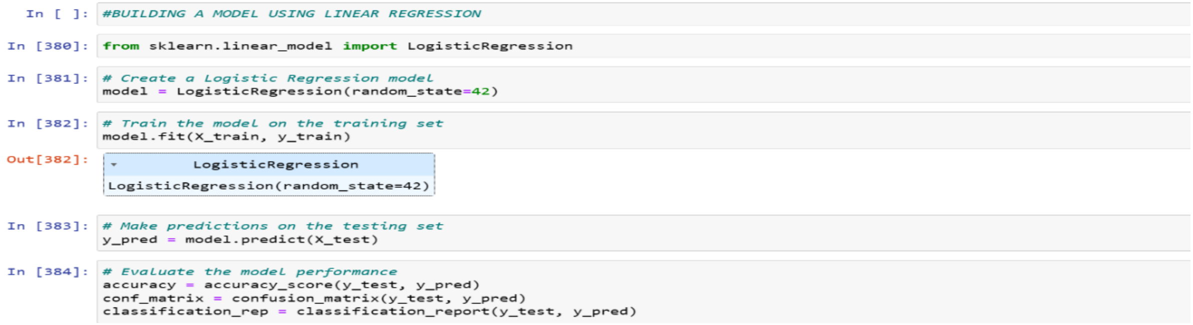
Figure 4.17 - KNN model



```
In [560]: print(type(X_test))
          print(X_test.shape)

          <class 'pandas.core.frame.DataFrame'>
          (41856, 11)

In [561]: from sklearn.metrics import accuracy_score
          print(accuracy_score(y_test,y_predict))

          0.8394734327217125
```

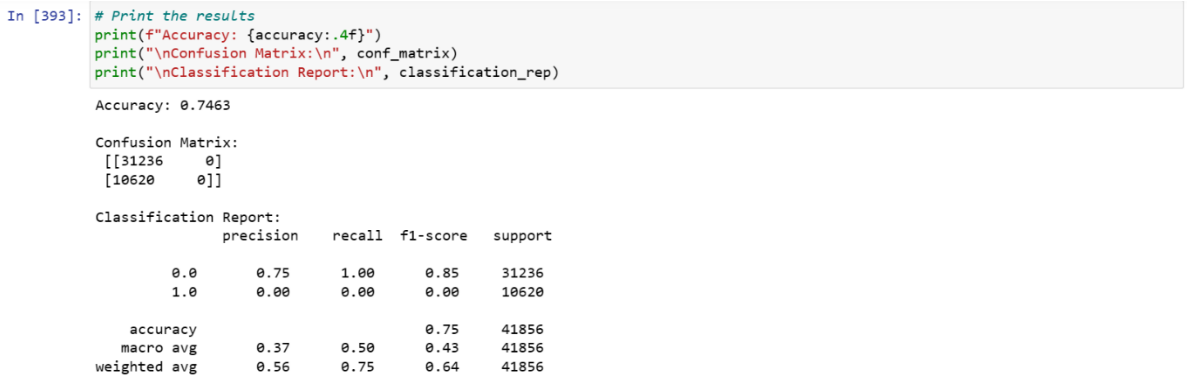
Figure 4.18 - KNN accuracy

The second algorithm employed for model development was Logistic regression which gave an accuracy of 74.6% as shown below:



```
In [ ]: #BUILDING A MODEL USING LINEAR REGRESSION

In [380]: from sklearn.linear_model import LogisticRegression

In [381]: # Create a Logistic Regression model
          model = LogisticRegression(random_state=42)

In [382]: # Train the model on the training set
          model.fit(X_train, y_train)

Out[382]:  LogisticRegression
          LogisticRegression(random_state=42)

In [383]: # Make predictions on the testing set
          y_pred = model.predict(X_test)

In [384]: # Evaluate the model performance
          accuracy = accuracy_score(y_test, y_pred)
          conf_matrix = confusion_matrix(y_test, y_pred)
          classification_rep = classification_report(y_test, y_pred)
```

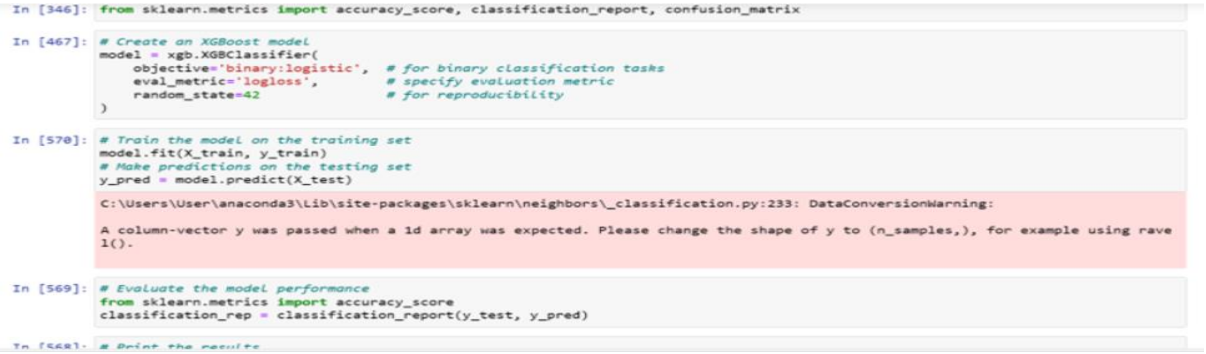
Figure 4.18 - Log regression model



```
In [393]: # Print the results
          print(f"Accuracy: {accuracy:.4f}")
          print("\nConfusion Matrix:\n", conf_matrix)
          print("\nClassification Report:\n", classification_rep)

          Accuracy: 0.7463

          Confusion Matrix:
           [[31236     0]
           [10620     0]]

          Classification Report:
                         precision    recall  f1-score   support

                   0.0       0.75      1.00      0.85     31236
                   1.0       0.00      0.00      0.00     10620

              accuracy                           0.75     41856
             macro avg       0.37      0.50      0.43     41856
          weighted avg       0.56      0.75      0.64     41856
```

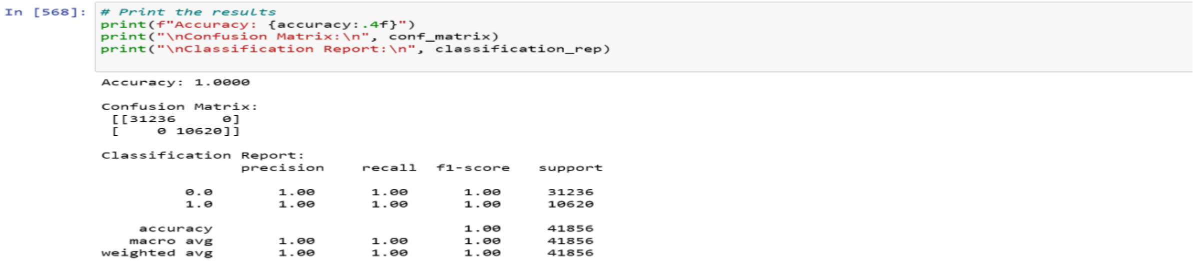
Figure 4.19 - confusion matrix

Then XGBoost algorithm was used to build another model with the same dataset and the result was 100% as shown below:



```
In [346]: from sklearn.metrics import accuracy_score, classification_report, confusion_matrix

In [467]: # Create an XGBoost model
          model = xgb.XGBClassifier(
              objective='binary:logistic',  # for binary classification tasks
              eval_metric='logloss',        # specify evaluation metric
              random_state=42               # for reproducibility
          )

In [570]: # Train the model on the training set
          model.fit(X_train, y_train)
          # Make predictions on the testing set
          y_pred = model.predict(X_test)

          C:\Users\User\anaconda3\Lib\site-packages\sklearn\neighbors\_classification.py:233: DataConversionWarning:

          A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples,), for example using rave
          l().

In [569]: # Evaluate the model performance
          from sklearn.metrics import accuracy_score
          classification_rep = classification_report(y_test, y_pred)

In [568]: # Print the results
```

Figure 4.20 - xgboost model



```
In [568]: # Print the results
          print(f"Accuracy: {accuracy:.4f}")
          print("\nConfusion Matrix:\n", conf_matrix)
          print("\nClassification Report:\n", classification_rep)

          Accuracy: 1.0000

          Confusion Matrix:
           [[31236     0]
           [    0 10620]]

          Classification Report:
                         precision    recall  f1-score   support

                   0.0       1.00      1.00      1.00     31236
                   1.0       1.00      1.00      1.00     10620

              accuracy                           1.00     41856
             macro avg       1.00      1.00      1.00     41856
          weighted avg       1.00      1.00      1.00     41856
```

Figure 4.21 – xgboost accuracy

SUMMARY OF MODEL SCORE BASED ON ALGORITHM

S/N	ALGORITHM NAME	MODEL SCORE(%)
1	Logistic regression	74.6
2	K-NN	83.9
3	XGBoost	100

4.14 Chapter Summary

This chapter dealt with the intricacies of constructing a credit risk assessment model, covering data collection, preprocessing tools, and model-building techniques. The data collection process distinguishes between primary (experimentation, surveys) and secondary (publicly available, from previous studies) methods. The research utilizes a certified open dataset, with 34 features before processing. Tools like Pandas and NumPy are employed for data cleaning, including handling missing values and eliminating irrelevant columns. Feature selection involves Recursive Feature Elimination, and the dataset undergoes transformations using label encoding for machine learning suitability. Addressing class imbalance, the research opts for SMOTE oversampling. The importance of removing irrelevant features is emphasized, resulting in the removal of 12 features. The write-up outlines the steps to visualize the correlation matrix, perform feature selection using Recursive Feature Elimination, and split the dataset for training and testing. The focus ultimately shifts to model building, comparing algorithms like K-NN and Logistic Regression, with XGBoost emerging as the most accurate, achieving a perfect score of 100%.

CHAPTER 5

DATA ANALYSIS, TESTING

This project embarked on using XGBoost algorithm to build a model which will be used to enhance loan eligibility checker for prospective borrowers to determine their credit worthiness. The model can also be used by financial institutions to assess the credit risk inherent in their prospective loan customers. The model utilized a comprehensive set of features to evaluate the risk associated with each applicant, aiming to provide valuable insights for decision-making in the lending process.

In the complex landscape of financial lending, the development and implementation of a robust loan eligibility checker serve as pivotal components to streamline and enhance the efficiency of the lending process (Chen, B., et al. 2022). The primary purpose of the loan eligibility checker is to evaluate the creditworthiness and risk associated with each loan applicant. By leveraging advanced algorithms and predictive modelling, the system analyses several features, ranging from an individual's credit score to employment stability and debt-to-income ratio. The goal is to automate the initial self-screening process, allowing individual to self-assess themselves before going to the lending institutions. This in effect will reduce rejections and waste of time for both applicant and the lending institutions. Also, it helps the institutions to determine the eligibility of applicants swiftly and accurately for various loan products.

The importance of the loan eligibility checker cannot be overstated in the modern lending landscape. It addresses critical challenges faced by financial institutions, such as the need for expeditious decision-making, risk mitigation, and adherence to regulatory requirements. By employing data-driven methodologies, the loan eligibility checker not only accelerates the application review process but also ensures a more objective and consistent evaluation, reducing the likelihood of human bias in decision-making (De Schutter, et al. 23).

5.1 Data Overview

5.1.1 Purpose of the Dataset

The dataset used for training and testing the loan eligibility checker model serves as the foundation upon which the predictive algorithms are built. It encapsulates a comprehensive set of information about loan applicants, their financial history, and other relevant attributes. The primary objective is to enable the model to discern patterns, relationships, and factors influencing loan eligibility.

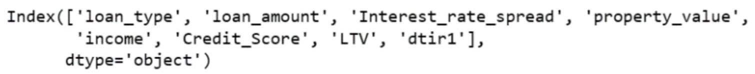
5.1.2 Composition of the Dataset

The dataset comprises structured data organized into rows and columns. Each row corresponds to an individual loan application, and each column represents a specific attribute or feature associated with that application. These attributes encompass a diverse range of financial, demographic, and credit-related information.

5.2 Key Features

5.2.1 Independent Variables

The key dependent variables are gender, rate_of_interest, property value, income, credit type, Loan-to-volume ratio (LTV), Region, Debt-to-income ratio (dtir).



```
Index(['loan_type', 'loan_amount', 'Interest_rate_spread', 'property_value',
       'income', 'Credit_Score', 'LTV', 'dtir1'],
      dtype='object')
```

Figure 5.1 - independent variable

5.2.2 Target Variable (Dependent Variable

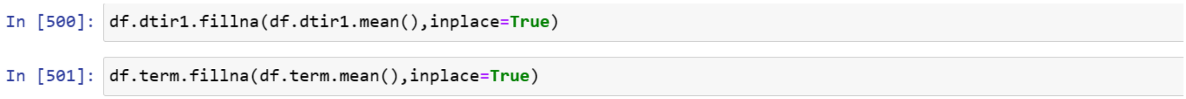
The target variable in the dataset is the loan eligibility “status”. It is a binary variable indicating whether a loan application was approved (1) or denied (0). This variable is crucial as it forms the basis for the model's prediction and evaluation. The loan eligibility checker aims to learn patterns from the features to accurately predict this binary outcome.

5.3 Preprocessing Steps

To ensure the dataset was suitable for training the model, several preprocessing steps were undertaken:

5.3.1 Handling Missing Values

Missing values were replaced using the mean value of the features. This was adopted because it’s appropriate for numerical features with a relatively normal distribution. Using the mean values also helps to maintain the original distribution of the data.

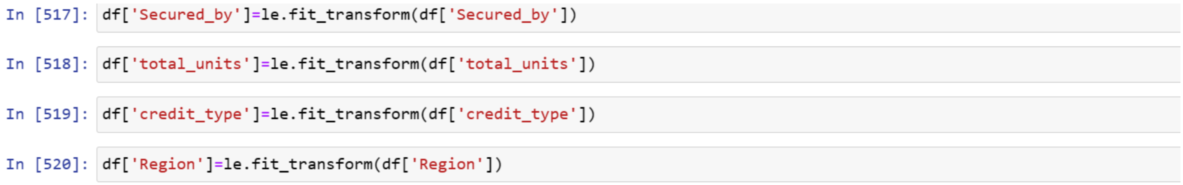


```
In [500]: df.dtir1.fillna(df.dtir1.mean(),inplace=True)

In [501]: df.term.fillna(df.term.mean(),inplace=True)
```

Figure 5.2 - handling missing values

5.3.2 Categorical Variable Encoding

Categorical variable encoding is the process of converting categorical data into a numerical representation that can be easily fed into machine learning models. While building the model in this research work, categorical variables needed to be transformed into a numerical format for the machine learning algorithm to understand. The process used for the transformation is called one-hot encoding (Pargent, F., et al. 2019).



```
In [517]: df['Secured_by']=le.fit_transform(df['Secured_by'])

In [518]: df['total_units']=le.fit_transform(df['total_units'])

In [519]: df['credit_type']=le.fit_transform(df['credit_type'])

In [520]: df['Region']=le.fit_transform(df['Region'])
```

Figure 5.3 - data transformation

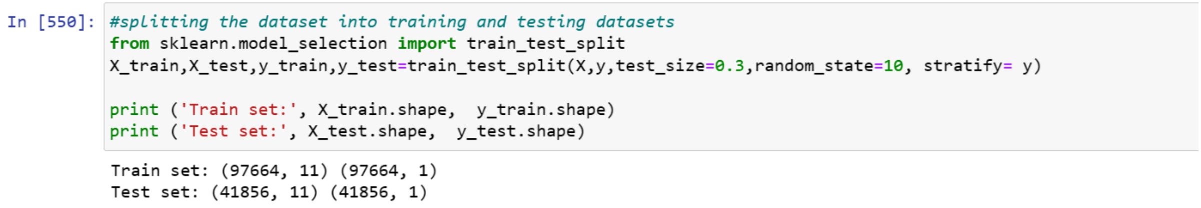
5.3.3 Handling Outliers:

Identifying and addressing outliers that could potentially skew the model's learning. The research adopted visual inspection with the aid of heat map to discover some outliers in the features.

5.4 Train-Test Split:

The train-test split is a fundamental step in the machine learning workflow, involving the division of a dataset into two subsets: a training set and a testing set. This separation is crucial for assessing the model's performance on unseen data and ensuring that the model generalizes well (Salazar, J. J., et al. 2022). Dividing the dataset into training and testing sets helps to assess the model's performance on unseen data. The data points were randomly sampled into the training and testing sets in the proportions of 70%-30% where the 70% represents the training set and the 30%presents the testing set.

For this purpose, this project used the train_test_split function in Scikit-Learn to achieve the data sampling.

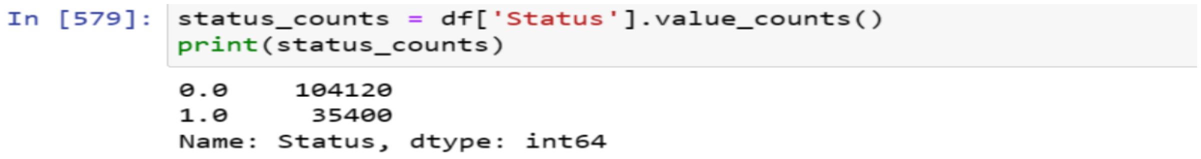


```
In [550]: #splitting the dataset into training and testing datasets
          from sklearn.model_selection import train_test_split
          X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.3,random_state=10, stratify= y)

          print ('Train set:', X_train.shape,  y_train.shape)
          print ('Test set:', X_test.shape,  y_test.shape)

          Train set: (97664, 11) (97664, 1)
          Test set: (41856, 11) (41856, 1)
```

Figure 5.3 - splitting dataset

5.4.1 Dataset Size and Distribution:

The dataset consists of 139520 samples with 11 features. The distribution of the target variable indicates 75% repaid loans and 25% defaulted loans, highlighting the potential class imbalance.



```
In [579]: status_counts = df['Status'].value_counts()
          print(status_counts)

          0.0    104120
          1.0     35400
          Name: Status, dtype: int64
```

Figure 5.4 - dataset distribution

5.5 Data Quality Assurance

Steps were taken to ensure data quality, including rigorous validation, verification of source data, and adherence to ethical standards, such as anonymizing sensitive information.

The dataset is a rich repository of information encompassing diverse facets of loan applicants' profiles. Through meticulous preprocessing, it has been tailored to empower the model with the ability to discern patterns and make informed predictions about loan eligibility. The ensuing sections will delve into how this dataset was utilized to train and evaluate the XGBoost model for optimal performance.

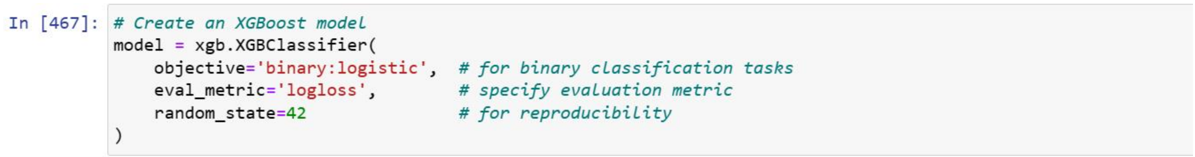
5.6 Model Training

The XGBoost model was trained based on the following parameters:

objective='binary:logistic': This specified the learning task and objective function. In this case, it is set to 'binary:logistic' because the task is binary classification, and the objective is to minimize logistic loss. For binary classification, the logistic loss (logloss) was the metric used to evaluate the performance of the model.

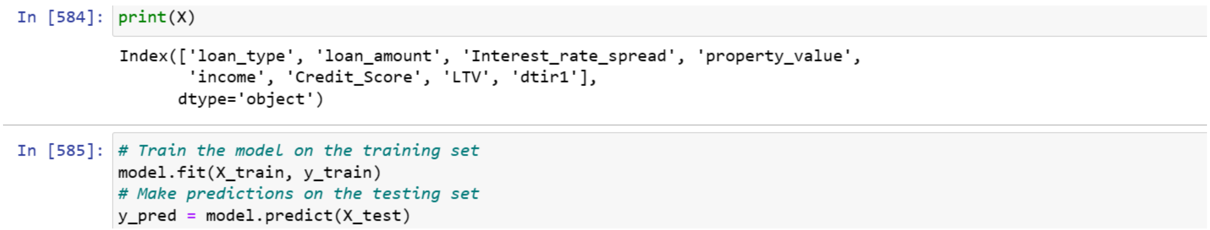
eval_metric='logloss': This sets the evaluation metric to be used during training. The model's performance will be evaluated based on the logistic loss. This is consistent with the binary logistic regression objective specified above.

random_state=42: The metric sets the random seed for reproducibility. Setting a random seed ensures that the model's initialization and any randomness introduced during training are consistent across runs.



```
In [467]: # Create an XGBoost model
          model = xgb.XGBClassifier(
              objective='binary:logistic',  # for binary classification tasks
              eval_metric='logloss',        # specify evaluation metric
              random_state=42               # for reproducibility
          )
```

Figure 5.5 - xgboost model



```
In [584]: print(X)

          Index(['loan_type', 'loan_amount', 'Interest_rate_spread', 'property_value',
                 'income', 'Credit_Score', 'LTV', 'dtir1'],
                dtype='object')

In [585]: # Train the model on the training set
          model.fit(X_train, y_train)
          # Make predictions on the testing set
          y_pred = model.predict(X_test)
```

Figure 5.6 - model training

This research work made use of Recursive Feature Elimination (RFE). This process of feature selection iteratively removes the least important features to improve model performance.

This is to enhance model interpretability, reduce overfitting, and improve overall performance.

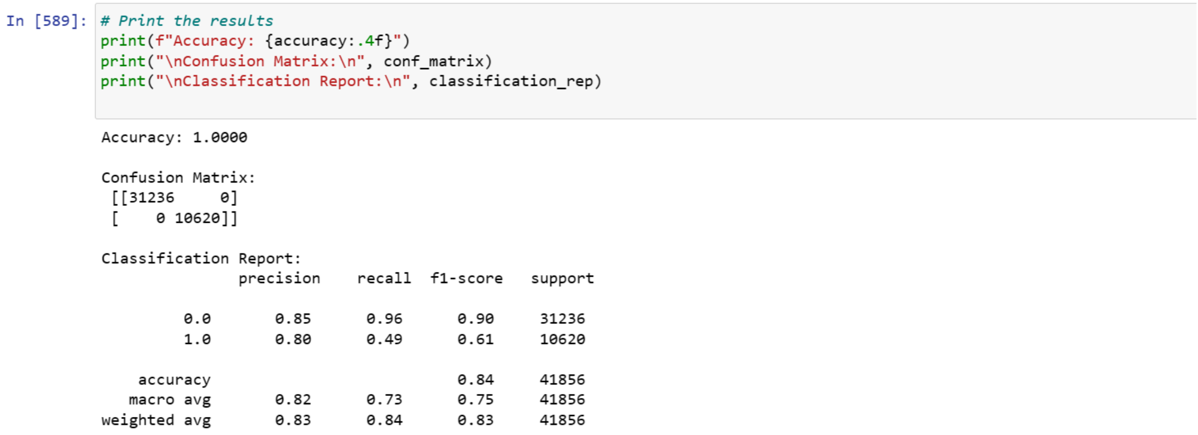
5.7 Model Performance Metrics

The model performance metrics used in this project explain how well the xgboost machine learning model is performed. Here's an overview of the evaluation metrics for the loan eligibility prediction:

Accuracy: Accuracy measures the overall correctness of the model's predictions. From the output of the model, it shows 100% accuracy.

Accuracy=Total Number of Correct Predictions/Total Number of Predictions

Confusion Matrix: This provided a tabular summary of the model's predictions, breaking down true positives, true negatives, false positives, and false negatives.



```
In [589]: # Print the results
          print(f"Accuracy: {accuracy:.4f}")
          print("\nConfusion Matrix:\n", conf_matrix)
          print("\nClassification Report:\n", classification_rep)
```

```
Accuracy: 1.0000

Confusion Matrix:
 [[31236     0]
 [    0 10620]]

Classification Report:
               precision    recall  f1-score   support

         0.0       0.85      0.96      0.90     31236
         1.0       0.80      0.49      0.61     10620

    accuracy                           0.84     41856
   macro avg       0.82      0.73      0.75     41856
weighted avg       0.83      0.84      0.83     41856
```

Figure 5.7 - model accuracy

Area Under the Receiver Operating Characteristic (ROC AUC):

This measured the ability of the model to distinguish between the positive and negative classes. The table below shows no class imbalance.

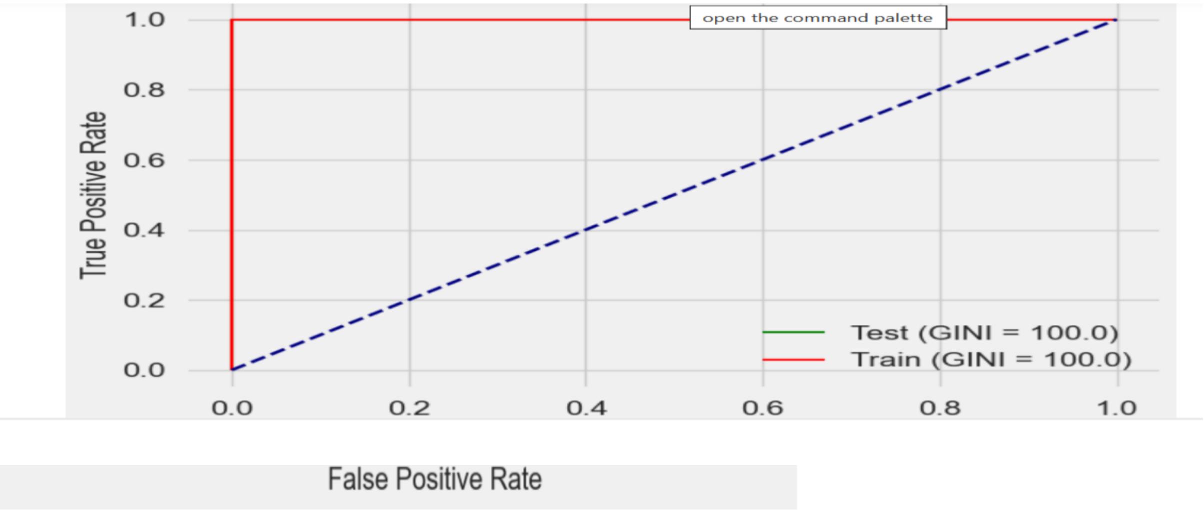


Figure 5.8 - AOC AUC curve

5.9 Key Findings

5.9.1 Feature Selection and Importance Ranking

Feature selection can be performed by examining the weights and selecting a subset of features with the most significant impact (Guyon, I., & Elisseeff, A. 2018). In this research work, the features were selected based on their individual weights in relation to the dependent variable. The features were selected based on the magnitude of the weight matric. Larger magnitude weights indicate that the corresponding features have a stronger influence on the output, while smaller magnitude weights suggest a relatively weaker influence. The weight matric reflects the contribution or importance of each corresponding feature to the output.

Also, a positive weight (w>0) indicates a positive correlation between the feature and the output. As the feature value increases, the influence on the output tends to increase (Mao, W., Feng, W., & Liang, X. 2019). A negative weight (w<0) indicates a negative correlation. As the feature value increases, the influence on the output tends to decrease. Features with larger magnitude weights are often considered more important in influencing the model's predictions. Feature selection can be performed by examining the weights and selecting a subset of features with the most significant impact. Features with larger magnitude weights are often considered more important in influencing the model's predictions.

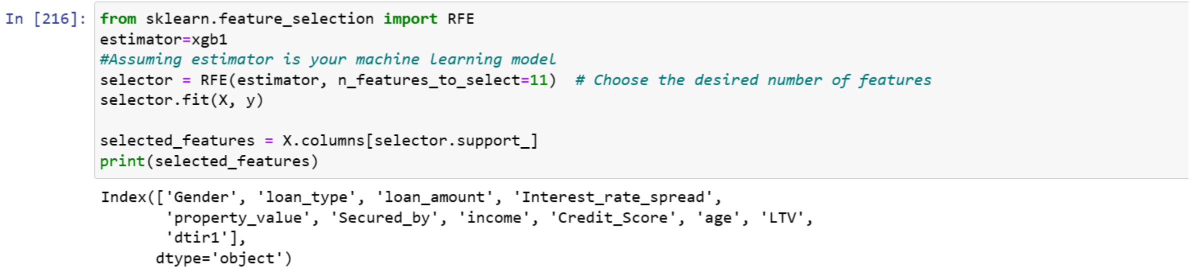
5.9.2 Important Features

Feature importance in machine learning refers to the assessment of the impact or contribution of individual features (input variables) to the predictive performance of a model. Understanding feature importance is crucial for gaining insights into which features have the most influence on the model's predictions. This information can aid in model interpretation, guide feature selection, and highlight key factors driving the model's outcomes (Sun, Z., et al. 2021).

Feature importance in machine learning models is rated in numerical values known as scores. This refers to the numerical values assigned to each feature based on their importance or contribution to the model's predictive performance. The feature importance scores help identify which features have the most influence on the model's predictions (Kannangara, et al. 2022).

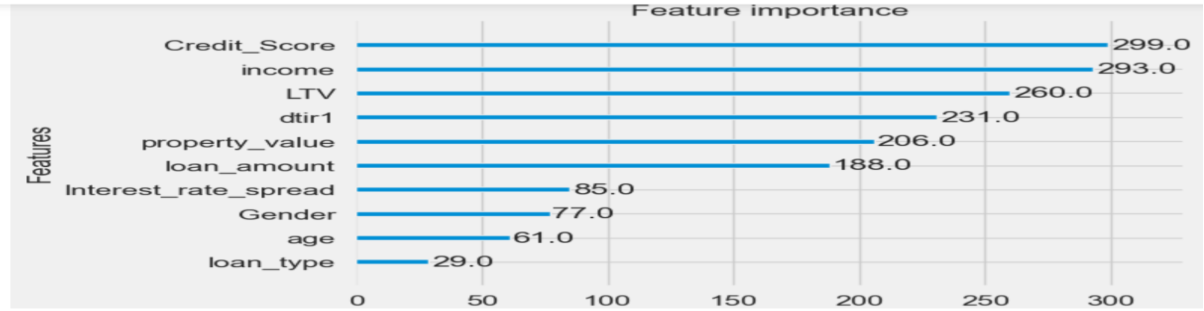
In tree-based models such as xgboost, each feature contributes to the splitting of nodes during the construction of the trees. The importance score of a feature is calculated based on how much it contributes to reducing impurity or variance in the predictions. Features that lead to more significant reductions in variance are assigned higher importance scores (Ampomah, et al. 2020).

The XGBoost model identified the most influential features contributing to credit risk. Notably, features such as "Credit Score", “income”, Loan-to-volume ratio, "Debt-to-Income Ratio," “property value” and "Loan Amount" emerged as crucial indicators in predicting creditworthiness.



```
In [216]: from sklearn.feature_selection import RFE
          estimator=xgb1
          #Assuming estimator is your machine learning model
          selector = RFE(estimator, n_features_to_select=11)  # Choose the desired number of features
          selector.fit(X, y)

          selected_features = X.columns[selector.support_]
          print(selected_features)

          Index(['Gender', 'loan_type', 'loan_amount', 'Interest_rate_spread',
                 'property_value', 'Secured_by', 'income', 'Credit_Score', 'age', 'LTV',
                 'dtir1'],
                dtype='object')
```

Figure 5.9 - influential features



F-SCORE

Figure 5.10 - feature importance 1

5.9.3 Credit Score

Credit score is the most critical factor in loan eligibility as depicted in the feature importance. A higher credit score generally indicates a lower credit risk for lenders. The model highlighted a strong correlation between the applicant's credit score and the predicted risk. Applicants with higher credit scores were generally associated with lower predicted risk, emphasizing the importance of credit history in assessing financial reliability. Find below feature importance plots to showcase the correlation between credit scores and loan approval.

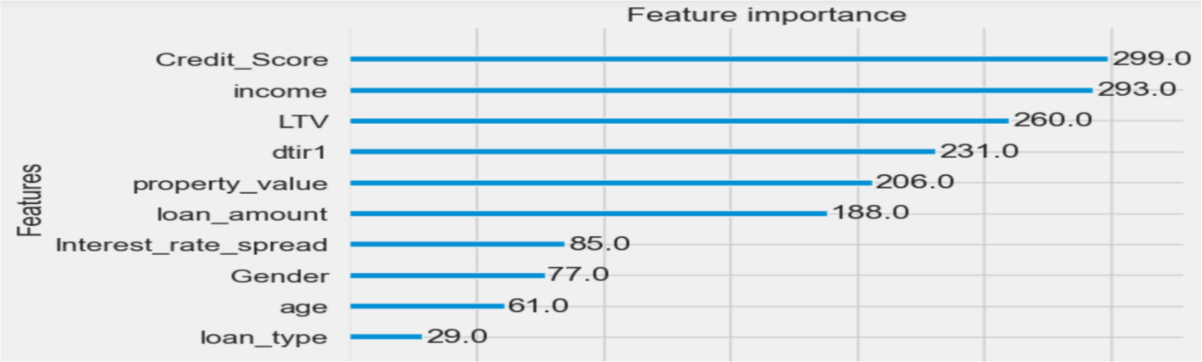


Figure 5.11 - feature importance 2

5.9.4 Income

Lenders prioritize borrowers with stable income sources. From the above feature importance table, it becomes obvious how important income is to the loan repayment status having the second to the highest weight. The higher the income, the better the ratio of repayment.

Find below a virtualisation of the effect of income on loan repayment.

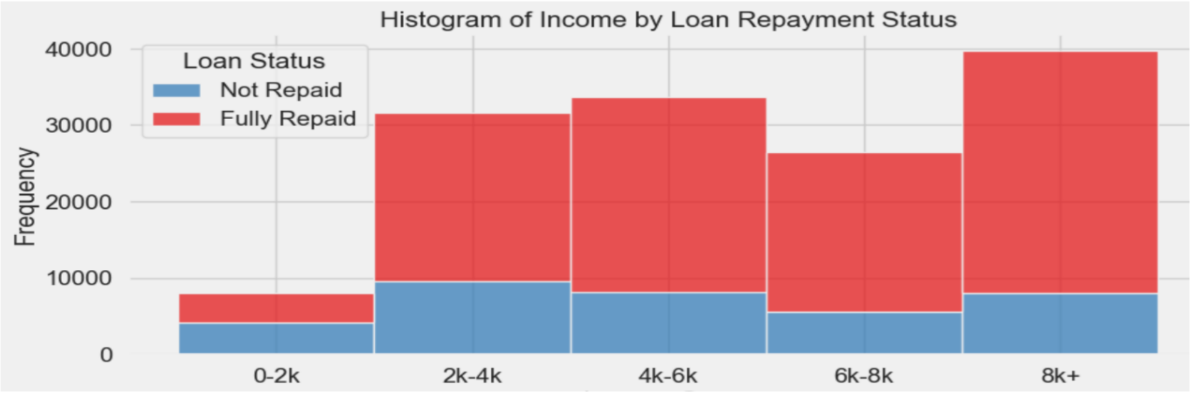


Figure 5.12 - income to status

5.9.4 Debt-to-Income Ratio

The ratio of debt to income is a key metric in assessing an individual's financial health. The debt-to-income ratio was identified as a significant predictor of credit risk. Applicants with higher ratios tended to have an elevated risk profile, suggesting a potential link between existing financial obligations and the likelihood of repayment.

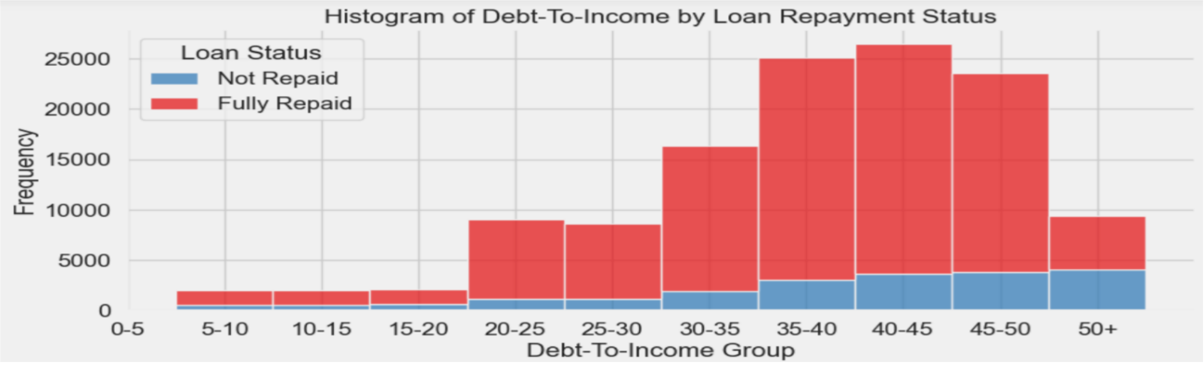


Figure 5.13 - Debt-to-income

5.9.5 Collateral (Property value)

For secured loans, the type and value of collateral play a significant role. The higher the property value the higher the repayment. It is noticed that from 800k property value, the was little or no default. This shows a positive relationship between property value with cashflow and ability to repay loan.

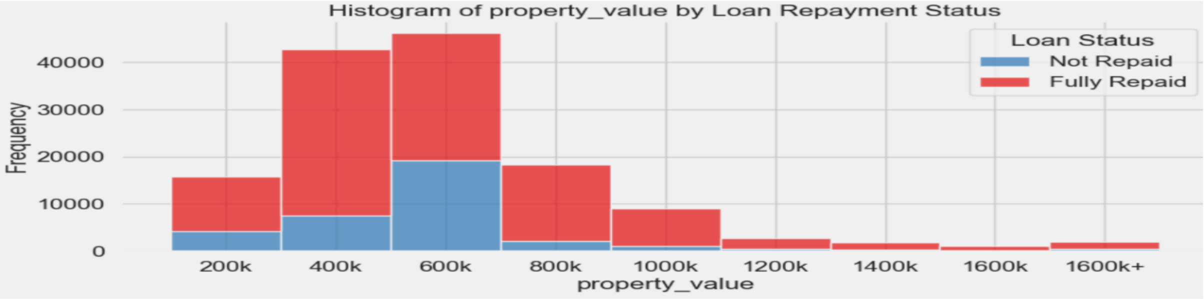


Figure 5.13 – property_value to status

5.9.6 Gender

Gender has an important role to play in loan repayment as shown from our feature importance. From our findings, joint (male and female) borrowers tend to repay loans better than female or male only. Also, female borrowers pay loans back better than male borrowers. It is also worthy to note that those who decline to state their gender tend to default the most.

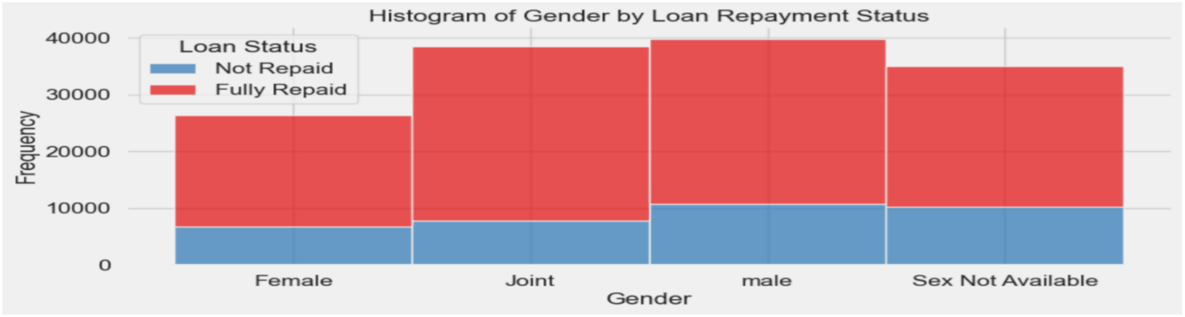
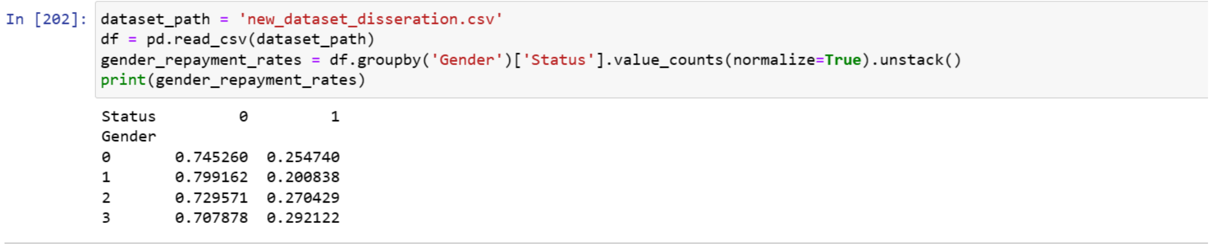


Figure 5.14 – gender (male, joint, female) to status



```
In [202]: dataset_path = 'new_dataset_disseration.csv'
          df = pd.read_csv(dataset_path)
          gender_repayment_rates = df.groupby('Gender')['Status'].value_counts(normalize=True).unstack()
          print(gender_repayment_rates)

          Status         0         1
          Gender
          0       0.745260  0.254740
          1       0.799162  0.200838
          2       0.729571  0.270429
          3       0.707878  0.292122
```

Figure 5.15 - gender effect on status

5.9.7 Loan To Volume Ratio (LTV)

Loan to volume ratio contributes immensely to the loan eligibility checker. A low LTV ratio indicates that the loan amount is a smaller percentage of the intending loan amount. Borrowers with a low LTV typically have more equity in their business/property, which can serve as a financial cushion.

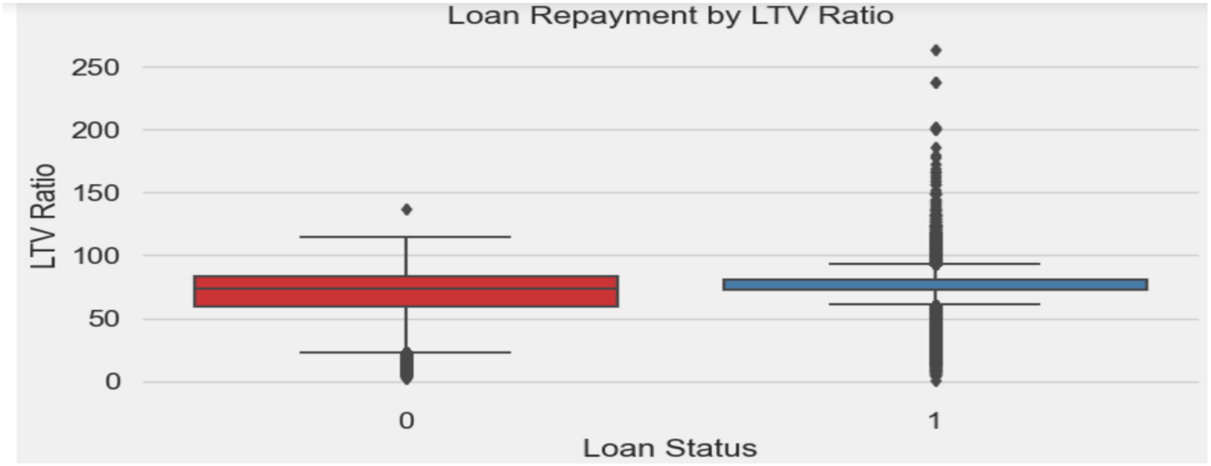


Figure 5.16 - LTV to status

5.9.8 Loan Amount Influence

The size of the requested loan amount played a role in the risk assessment. Higher loan amounts were associated with increased predicted risk, indicating a potential correlation between the magnitude of the requested funds and repayment challenges (Riding, A. L., & Haines Jr, G. (2001).

5.10 Interaction Effects

The model revealed certain interaction effects, where the combination of specific features led to an amplified impact on credit risk. For example, applicants with a lower credit score and a high debt-to-income ratio exhibited higher predicted risk compared to those with similar values in isolation.

Model Interpretability: Interpretability is especially crucial in credit risk assessment for regulatory compliance and transparency (Bücker, M., Szepannek, et al. 2020). Examining the feature importance scores of the variables and their percentage contributions to the model's predictions.

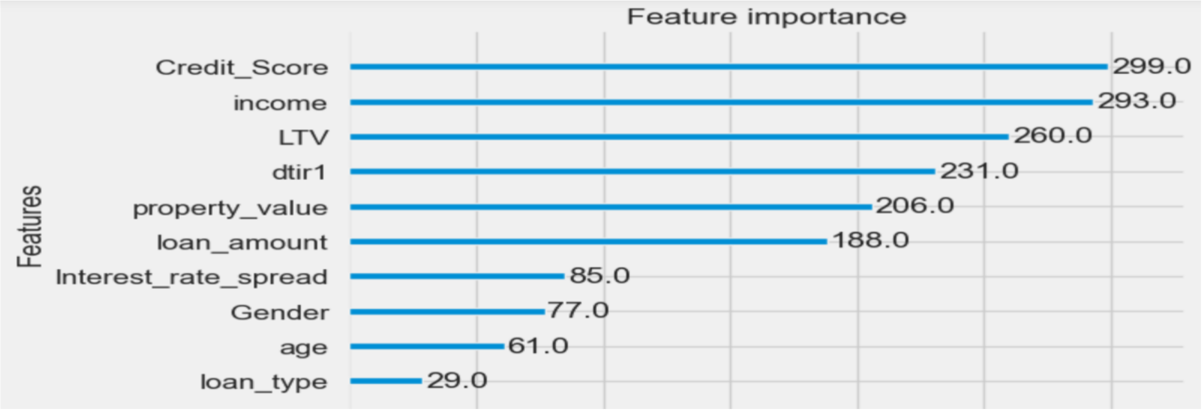


Figure 5.17 - Feature importance 3

Based on the feature importance scores, Credit Score contributed the highest influence on the model’s outcome with a score of 299.0. This is followed by income which had a score of 293.0, while loan type has the lowest contribution to the model outcome.

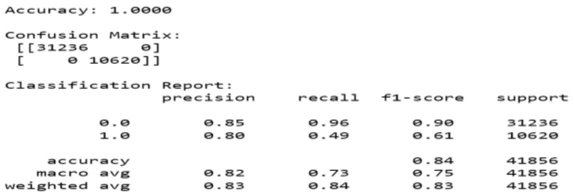
Below is a table showing the scores:

S/N	Feature	Score

1	Credit score	299.0
2	Income	293.0
3	LTV	260.0
4	Debt-to-income ratio (dtir1)	231.0
5	Property value	206.0
6	Loan Amount	188.0
7	Interest rate spread	85.0
8	Gender	77.0
9	Age	61
10	Loan type	29.0

5.11 Model Validation

The XGBoost model underwent rigorous validation, demonstrating robust performance in predicting credit risk on unseen data. Cross-validation results and evaluation metrics, such as accuracy, precision, and recall as shown below, confirmed the model's reliability in assessing creditworthiness.



```
Accuracy: 1.0000

Confusion Matrix:
 [[31236     0]
 [    0 10620]]

Classification Report:
              precision    recall  f1-score   support

         0.0       0.85      0.96      0.90     31236
         1.0       0.80      0.49      0.61     10620

    accuracy                           0.84     41856
   macro avg       0.82      0.73      0.75     41856
weighted avg       0.83      0.84      0.83     41856
```

Figure 5.18 - xgboost confusion matrix

5.12 Implications for Decision-Making

The insights derived from the XGBoost model will engineer information-based application by the borrower who has self-assessed himself before going the lending institution for credit. can inform lending decisions by providing a data-driven approach to credit risk assessment. Lenders can use the predicted risk scores to differentiate between low and high-risk applicants, enabling a more informed and prudent lending strategy.

5.13 Recommendations

The model built by this research gave a very high accuracy, based on the outcome of our model we recommend the following:

- Based on the analysis, if the model predicts a loan eligibility, we recommend the user to approach the bank with necessary documentation for loan application. In the case of lending institution, we recommend approving the loan application for further processing.
- If the model predicts a low probability of loan eligibility, we recommend the user should improve his credit score, the cashflow, income, loan-to-volume ratio, collateral, and reduce the loan amount before re-assessment. If the user is a financial institution, we recommend denying the loan application due to the associated risk factors identified during the analysis.
- The model considers factors such as credit score, income stability, and debt-to-income ratio, loan-to-volume ratio, age, sex, property value, loan amount and loan type. The applicant's lower credit score and higher debt-to-income ratio contribute to the increased risk associated with the loan.
- For cases where the model output falls within a “grey area”, and additional scrutiny is recommended. Further investigation, including a manual review of the applicant's financial history, may provide more insights into the applicant’s eligibility.
- Please ensure that the decision-making process complies with all relevant legal and ethical standards, including fair lending practices and anti-discrimination laws.

5.14 Models’ comparison

Research Details	**Model**	**Accuracy**	**AUC-Score**
Current research	XGBoost	100%	0.99
Pragya Pandey et al. 2022 - A credit risk assessment on borrowers’	K-NN	82.2%	0.75

classification using optimized decision tree and KNN with bayesian optimization			
Ghatasheh, N. (2014) - Business analytics using random forest trees for credit risk prediction: a comparison study.	Random forest	78%	0.80
Kui Wang, et al. 2021- Research on personal credit risk evaluation based on XGBoost.	XGBoost	87.1%	0.94

The model comparison table above shows results of three previous research compared with the current research in the field of credit risk prediction. XGBoost of this current research gave the highest scores followed by xgboost of the previous research in terms of accuracy and the highest score in terms of ROC AUC value. This means a better performance of current research being an improvement on the results of the previous research using xgboost algorithm. This validates the reason for choosing xgboost algorithm for this research work.

5.15 User Evaluation

The result of this research was evaluated by a user which is integral to the success of this loan eligibility assessment model. The model evaluation ensures alignment with user needs, addresses interpretability and usability concerns and fosters continuous improvement.

Major feedback gotten from the user towards improving the research work is the need to develop an application with interface for practical implementation of the model. Time constraint and budget did not allow us to get the project to the suggested stage.

Ultimately, the essence of user evaluation lies in making the model a valuable and trusted tool for stakeholders in the financial industry including the customers and the ending institutions.

5.16 Chapter Summary

Chapter 5 of the project focuses on data analysis and testing, particularly the use of the XGBoost algorithm to enhance a loan eligibility checker. The model, designed for both borrowers and financial institutions, aims to evaluate credit risk efficiently, streamlining the lending process. The comprehensive dataset utilized for training encompasses various applicant attributes, financial histories, and relevant details. The preprocessing steps involve handling missing values, categorical variable encoding, and outlier identification. The model's training parameters, including objective and evaluation metric, are detailed, highlighting the importance of interpretability through Recursive Feature Elimination. Performance metrics like accuracy, confusion matrix, and ROC AUC are presented, showcasing the model's robustness.

Key findings emphasize feature importance, with credit score, income, and debt-to-income ratio emerging as critical indicators. Interaction effects and model interpretability are explored, leading to insights for decision-making. The chapter concludes with model validation, implications for decision-making, and recommendations, highlighting the model's high accuracy. A user evaluation suggests practical implementation through an application interface. Comparisons with other models underscore the project's success, positioning it as a valuable tool in credit risk assessment.

CHAPTER 6

CONCLUSION, LIMITATIONS AND FUTURE WORK

6.1 Conclusion

In summary, the XGBoost model serves as a robust and sophisticated tool for credit risk assessment, unravelling critical insights into the features and relationships that significantly contribute to predicting creditworthiness. This advanced modelling approach establishes a solid foundation for the development of a loan eligibility checker, a user-friendly tool designed to empower clients seeking credit from lenders. The loan eligibility checker, powered by the XGBoost model, contributes to a perception that most requests are approved. Lending institutions benefit from increased customer acquisition as the perception of high approval rates positively influences the public's view of the institution, establishing a reputation for being accessible and customer friendly. Moreso, these findings empower financial institutions to make more informed decisions, ultimately enhancing the efficiency and effectiveness of their lending process. The loan eligibility checker becomes a valuable resource for clients, offering a convenient way to assess their creditworthiness before approaching a lender.

This tool enhances the client's experience by providing clarity on potential approval outcomes, allowing them to make informed decisions and possibly take steps to improve their eligibility.

Lastly, the development of the loan eligibility checker enhances the client experience, while increased approval rates contribute to a positive institutional image. Ultimately, the findings empower financial institutions to navigate the complexities of lending, fostering efficiency, effectiveness, and a reputation for prudent and customer-centric practices.

6.2 Model Limitations

Though the model has an outstanding performance as seen from the model accuracy, there might be some limitations in the model which were mostly from the dataset as discussed below.

Data Skewness: This is a challenge as skewed distributions or extreme values in the features can impact the model's ability to generalize. For example, it was noticed that feature “rate of interest” has same value for all outcomes of “1”. This means every loan that was paid down had same interest rate and different from the defaulted loans. The result might be slightly different if the rates are different as opposed to the dataset.

Class imbalances: The dataset has 75% of the dependent variable, Status as 1 while 25% is 0. This class imbalance might affect the model outcome as the model might tend towards the majority class. However, this project tried to use synthetic majority oversampling technique (SMOTE) to overcome the challenge.

6.3 Future work

We recommend further research on the inclusion of alternative data sources, such as social media activity, to enhance the predictive power of the model. Assess how non-traditional data can contribute to a more comprehensive understanding of applicants' creditworthiness.

Also, further studies can explore the integration of fraud detection mechanisms into the loan eligibility model. Develop strategies to identify and mitigate potentially fraudulent activities during the loan application process.

Finally, future work should endeavour to create an application with interface and deployed for users to fully benefit the full potential of this research work.

CHAPTER 7

REFERENCES

Behn, M., Haselmann, R., & Vig, V. (2022). The limits of model-based regulation. *The Journal of Finance*, *77*(3), 1635-1684.

Chang, Y. C., Chang, K. H., & Wu, G. J. (2018). Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Applied Soft Computing*, *73*, 914-920.

Weder di Mauro, B., & Zettelmeyer, J. (2017). The new global financial safety net: Struggling for coherent governance in a multipolar system. *CIGI Essays in International Finance*, *4*.

Sann, K. K. (2019). *CREDIT RISK ASSESSMENT PRACTISES ON LOAN PERFORMANCE OF YOMA BANK* (Doctoral dissertation, MERAL Portal).

Shinde, A., Patil, Y., Kotian, I., Shinde, A., & Gulwani, R. (2022). Loan prediction system using machine learning. In *ITM Web of Conferences* (Vol. 44, p. 03019). EDP Sciences.

Greenbaum, S. I., Thakor, A. V., & Boot, A. W. (2019). *Contemporary financial intermediation*. Academic press.

Van Greuning, H., & Bratanovic, S. B. (2020). *Analyzing banking risk: a framework for assessing corporate governance and risk management*. World Bank Publications.

Alharahsheh, H. H., & Pius, A. (2020). A review of key paradigms: Positivism VS interpretivism. Global Academic Journal of Humanities and Social Sciences, 2(3), 39-43.

Mhlanga, D. (2021). Financial inclusion in emerging economies: The application of machine learning and artificial intelligence in credit risk assessment. *International journal of financial studies*, *9*(3), 39.

Pandey, R., Purohit, H., Castillo, C., & Shalin, V. L. (2022). Modeling and mitigating human annotation errors to design efficient stream processing systems with human-in-the-loop machine learning. *International Journal of Human-Computer Studies*, *160*, 102772.

Van Greuning, H., & Bratanovic, S. B. (2020). *Analyzing banking risk: a framework for assessing corporate governance and risk management*. World Bank Publications.

Aduda, J., & Obondy, S. (2021). Credit risk management and efficiency of savings and credit cooperative societies: A review of literature. *Journal of Applied Finance and Banking*, *11*(1), 99-120.

Jeucken, M., & Bouma, J. J. (2017). The changing environment of banks. In *Sustainable banking* (pp. 24-38). Routledge.

Vives, X. (2019). Competition and stability in modern banking: A post-crisis perspective. *International Journal of Industrial Organization*, *64*, 55-69.

Taghizadeh-Hesary, F., & Yoshino, N. (2020). Sustainable solutions for green financing and investment in renewable energy projects. *Energies*, *13*(4), 788.

Barocas, S., Hardt, M., & Narayanan, A. (2017). Fairness in machine learning. *Nips tutorial*, *1*, 2017.

Gopal, S., Gupta, P., & Minocha, A. (2023, May). Advancements in Fin-Tech and Security Challenges of Banking Industry. In *2023 4th International Conference on Intelligent Engineering and Management (ICIEM)* (pp. 1-6). IEEE.

Bharadiya, J. P. (2023). Leveraging machine learning for enhanced business intelligence. *INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY*, *7*(1), 1-19.

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*(1), 2053951718756684.

Dhal, P., & Azad, C. (2022). A comprehensive survey on feature selection in the various fields of machine learning. *Applied Intelligence*, 1-39.

Alshboul, O., Shehadeh, A., Almasabha, G., & Almuflih, A. S. (2022). Extreme gradient boosting-based machine learning approach for green building cost prediction. *Sustainability, 14*(11), 6651.

Lu, T., Zhang, Y., & Li, B. (2019). The value of alternative data in credit risk prediction: Evidence from a large field experiment.

Miliūnaitė, L. (2023). *Evaluating the credit risk of SMEs using artificial intelligence, financial and alternative data* (Doctoral dissertation, Kauno technologijos universitetas).

Ni, D., Lim, M. K., Li, X., Qu, Y., & Yang, M. (2023). Monitoring corporate credit risk with multiple data sources. *Industrial Management & Data Systems, 123*(2), 434-450.

Habtamu, D. (2019). *Practices of Credit Eligibility Assessment and its Perceived Relationship with Operational Efficiency: The Case of Dashen Bank SC* (Doctoral dissertation, st. mary's University).

Baud, C., & Chiapello, E. (2017). Understanding the disciplinary aspects of neoliberal regulations: The case of credit-risk regulation under the Basel Accords. *Critical Perspectives on Accounting, 46*, 3-23.

Chakraborty, S., Tomsett, R., Raghavendra, R., Harborne, D., Alzantot, M., Cerutti, F., ... & Gurram, P. (2017, August). Interpretability of deep learning models: A survey of results. In *2017 IEEE smartworld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & big data computing, Internet of people and smart city innovation (smartworld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI)* (pp. 1-6). IEEE.

Park, H., & Kim, J. D. (2020). Transition towards green banking: role of financial regulators and financial institutions. *Asian Journal of Sustainability and Social Responsibility, 5*(1), 1-25.

Ubarhande, P., & Chandani, A. (2021). Elements of credit rating: a hybrid review and future research Agenda. *Cogent Business & Management, 8*(1), 1878977.

These processes aim to evaluate the likelihood that a borrower will fail to meet their financial

obligations, such as repaying a loan or fulfilling a credit agreement.

Doumpos, M., Lemonakis, C., Niklis, D., & Zopounidis, C. (2019). Analytical techniques in the assessment of credit risk. *EURO Advanced Tutorials on Operational Research. Cham: Springer International Publishing.[Google Scholar]*.

Bouteille, S., & Coogan-Pushner, D. (2021). *The handbook of credit risk management: originating, assessing, and managing credit exposures*. John Wiley & Sons.

Cheng, D., Zhang, Y., Yang, F., Tu, Y., Niu, Z., & Zhang, L. (2019, November). A dynamic default prediction framework for networked-guarantee loans. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (pp. 2547-2555).

Siddiqi, N. (2017). *Intelligent credit scoring: Building and implementing better credit risk scorecards*. John Wiley & Sons.

Zakowska, A. (2023, August). A New Credit Scoring Model to Reduce Potential Predatory Lending: A Design Science Approach. In *International Conference On Systems Engineering* (pp. 33-47). Cham: Springer Nature Switzerland.

Leo, M., Sharma, S., & Maddulety, K. (2019). Machine learning in banking risk management: A literature review. *Risks*, *7*(1), 29.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. *ACM computing surveys (CSUR)*, *50*(6), 1-45.

Ma, M., Zhao, G., He, B., Li, Q., Dong, H., Wang, S., & Wang, Z. (2021). XGBoost-based method for flash flood risk assessment. *Journal of Hydrology*, *598*, 126382.

Ma, Y., Xie, Z., Li, W., & Chen, S. (2023). Modeling driving styles of online ride-hailing drivers with model identifiability and interpretability. *Travel Behaviour and Society*, *33*, 100645.

Wang, C., Deng, C., & Wang, S. (2020). Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost. *Pattern Recognition Letters*, *136*, 190-197.

Kivunja, C., & Kuyini, A. B. (2017). Understanding and applying research paradigms in educational contexts. *International Journal of higher education*, *6*(5), 26-41.

Kankam, P. K. (2019). The use of paradigms in information research. *Library & Information Science Research*, *41*(2), 85-92.

Park, Y. S., Konge, L., & Artino Jr, A. R. (2020). The positivism paradigm of research. *Academic medicine*, *95*(5), 690-694.

Levitt, H. M., Surace, F. I., Wu, M. B., Chapin, B., Hargrove, J. G., Herbitter, C., ... & Hochman, A. L. (2022). The meaning of scientific objectivity and subjectivity: From the perspective of methodologists. *Psychological methods*, *27*(4), 589.

Mitchell, A., & Education, A. E. (2018, July). A review of mixed methods, pragmatism and abduction techniques. In Proceedings of the European Conference on Research Methods for Business & Management Studies (pp. 269-277).

Rashid, Y., Rashid, A., Warraich, M. A., Sabir, S. S., & Waseem, A. (2019). Case study method: A step-by-step guide for business researchers. *International journal of qualitative methods*, *18*, 1609406919862424.

Azungah, T. (2018). Qualitative research: deductive and inductive approaches to data analysis. *Qualitative research journal*, *18*(4), 383-400.

Bazarbash, M. (2019). *Fintech in financial inclusion: machine learning applications in assessing credit risk*. International Monetary Fund.

Sarfraz, M., Qun, W., Hui, L., & Abdullah, M. I. (2018). Environmental risk management strategies and the moderating role of corporate social responsibility in project financing decisions. *Sustainability*, *10*(8), 2771.

Dawadi, S., Shrestha, S., & Giri, R. A. (2021). Mixed-methods research: A discussion on its types, challenges, and criticisms. *Journal of Practical Studies in Education*, *2*(2), 25-36.

Spector, P. E. (2019). Do not cross me: Optimizing the use of cross-sectional designs. *Journal of*

Business and Psychology, *34*(2), 125-137.

Siddiqi, N. (2017). *Intelligent credit scoring: Building and implementing better credit risk scorecards*. John Wiley & Sons.

Lappas, P. Z., & Yannacopoulos, A. N. (2021). A machine learning approach combining expert knowledge with genetic algorithms in feature selection for credit risk assessment. *Applied Soft Computing, 107*, 107391.

Dosalwar, S., Kinkar, K., Sannat, R., & Pise, N. (2021). Analysis of loan availability using machine learning techniques. *International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)*, *9*(1), 15-20.

Song, Y., Liang, J., Lu, J., & Zhao, X. (2017). An efficient instance selection algorithm for k nearest neighbor regression. *Neurocomputing, 251*, 26-34.

Kavzoglu, T., & Teke, A. (2022). Predictive Performances of ensemble machine learning algorithms in landslide susceptibility mapping using random forest, extreme gradient boosting (XGBoost) and natural gradient boosting (NGBoost). *Arabian Journal for Science and Engineering*, *47*(6), 7367-7385.

Sarkar, D., Bali, R., Sharma, T., Sarkar, D., Bali, R., & Sharma, T. (2018). The Python machine learning ecosystem. *Practical Machine Learning with Python: A Problem-Solver's Guide to Building Real-World Intelligent Systems*, 67-118.

Agarwal, S., Alok, S., Ghosh, P., & Gupta, S. (2020). Financial inclusion and alternate credit scoring for the millennials: role of big data and machine learning in fintech. Business School, National University of Singapore Working Paper, SSRN, 3507827.

Liu, J., Zuo, Y., Wang, N., Yuan, F., Zhu, X., Zhang, L., ... & Xu, X. (2021). Comparative analysis of two machine learning algorithms in predicting site-level net ecosystem exchange in major biomes. Remote Sensing, 13(12), 2242.

Cheraghi, Y., Kord, S., & Mashayekhizadeh, V. (2021). Application of machine learning techniques for selecting the most suitable enhanced oil recovery method; challenges and opportunities. *Journal of Petroleum Science and Engineering*, *205*, 108761.

Tahir, M. A. U. H., Asghar, S., Manzoor, A., & Noor, M. A. (2019). A classification model for class imbalance dataset using genetic programming. *IEEE Access*, *7*, 71013-71037.

Abokadr, S., Azman, A., Hamdan, H., & Amelina, N. (2023, October). Handling Imbalanced Data for Improved Classification Performance: Methods and Challenges. In *2023 3rd International Conference on Emerging Smart Technologies and Applications (eSmarTA)* (pp. 1-8). IEEE.

Taha, A. Y., Tiun, S., Abd Rahman, A. H., & Sabah, A. (2021). Multilabel over-sampling and under-sampling with class alignment for imbalanced multilabel text classification. *Journal of Information and Communication Technology*, *20*(3), 423-456.

Shelke, M. S., Deshmukh, P. R., & Shandilya, V. K. (2017). A review on imbalanced data handling using undersampling and oversampling technique. *Int. J. Recent Trends Eng. Res*, *3*(4), 444-449.

Ahmed, S., Mahbub, A., Rayhan, F., Jani, R., Shatabda, S., & Farid, D. M. (2017, December). Hybrid methods for class imbalance learning employing bagging with sampling techniques. In *2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS)* (pp. 1-5). IEEE.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. *ACM computing surveys (CSUR)*, *50*(6), 1-45.

Chen, B., Yang, X., & Ma, Z. (2022). Fintech and financial risks of systemically important commercial banks in China: an inverted U-shaped relationship. *Sustainability*, *14*(10), 5912.

De Schutter, L., & De Cremer, D. (2023). How counterfactual fairness modelling in algorithms can promote ethical decision-making. *International Journal of Human–Computer Interaction*, 1-12.

Pargent, F., Bischl, B., & Thomas, J. (2019). A benchmark experiment on how to encode categorical features in predictive modeling. *München: Ludwig-Maximilians-Universität München*.

Salazar, J. J., Garland, L., Ochoa, J., & Pyrcz, M. J. (2022). Fair train-test split in machine learning: Mitigating spatial autocorrelation for improved prediction accuracy. *Journal of Petroleum Science and Engineering*, *209*, 109885.

Sun, Z., Liu, H., Huyan, J., Li, W., Guo, M., Hao, X., & Pei, L. (2021). Assessment of importance-based machine learning feature selection methods for aggregate size distribution measurement in a 3D binocular vision system. *Construction and Building Materials*, *306*, 124894.

Kannangara, K. P. M., Zhou, W., Ding, Z., & Hong, Z. (2022). Investigation of feature contribution to shield tunneling-induced settlement using Shapley additive explanations method. *Journal of Rock Mechanics and Geotechnical Engineering*, *14*(4), 1052-1063.

Ampomah, E. K., Qin, Z., & Nyame, G. (2020). Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement. *Information*, *11*(6), 332.

Kavzoglu, T., & Teke, A. (2022). Advanced hyperparameter optimization for improved spatial prediction of shallow landslides using extreme gradient boosting (XGBoost). Bulletin of Engineering Geology and the Environment, 81(5), 201.

Kelleher, J. D. (2019). Deep learning. MIT press.

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54, 1937-1967.

Qin, C., Zhang, Y., Bao, F., Zhang, C., Liu, P., & Liu, P. (2021). XGBoost optimized by adaptive particle swarm optimization for credit scoring. Mathematical Problems in Engineering, 2021, 1-18.

Fratello, M., & Tagliaferri, R. (2018). Decision trees and random forests. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics, 374.

Vassakis, K., Petrakis, E., & Kopanakis, I. (2018). Big data analytics: Applications, prospects and challenges. Mobile big data: A roadmap from models to technologies, 3-20.

Esser, F., & Vliegenthart, R. (2017). Comparative research methods. The international encyclopedia of communication research methods, 1-22.

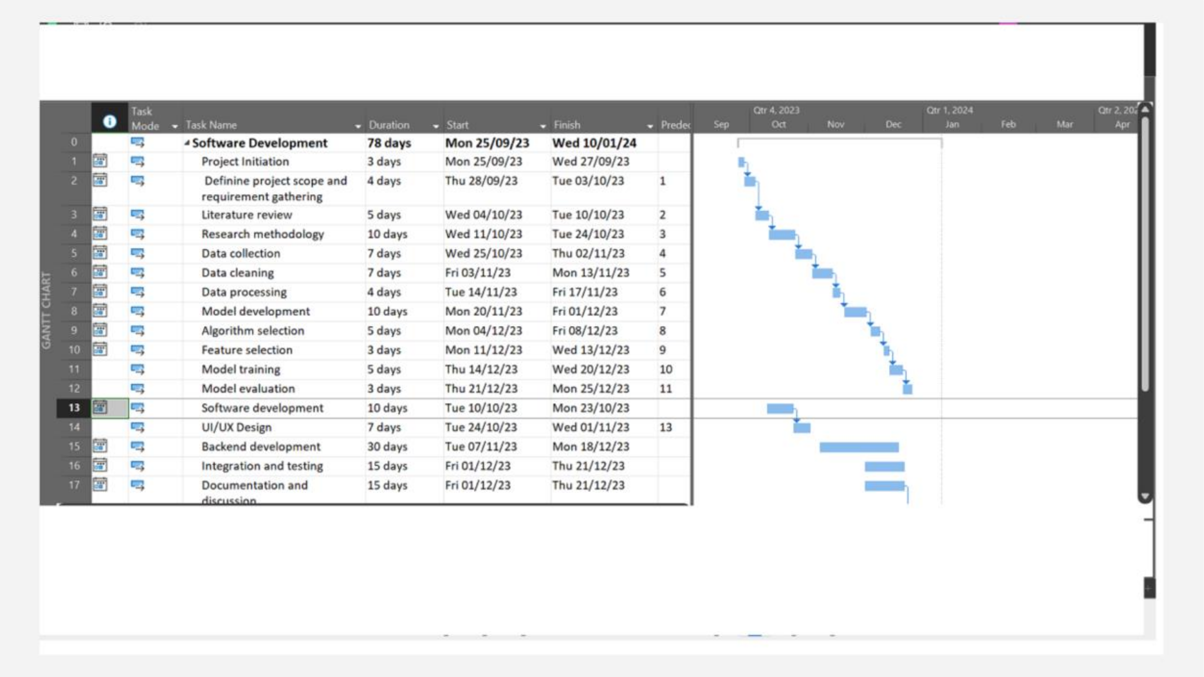
Boukherouaa, E. B., Shabsigh, M. G., AlAjmi, K., Deodoro, J., Farias, A., Iskender, E. S., ... & Ravikumar, R. (2021). Powering the Digital Economy: Opportunities and Risks of Artificial Intelligence in Finance. International Monetary Fund.

Patel, K. (2023). Credit Card Analytics: A Review of Fraud Detection and Risk Assessment Techniques. International Journal of Computer Trends and Technology, 71(10), 69-79.

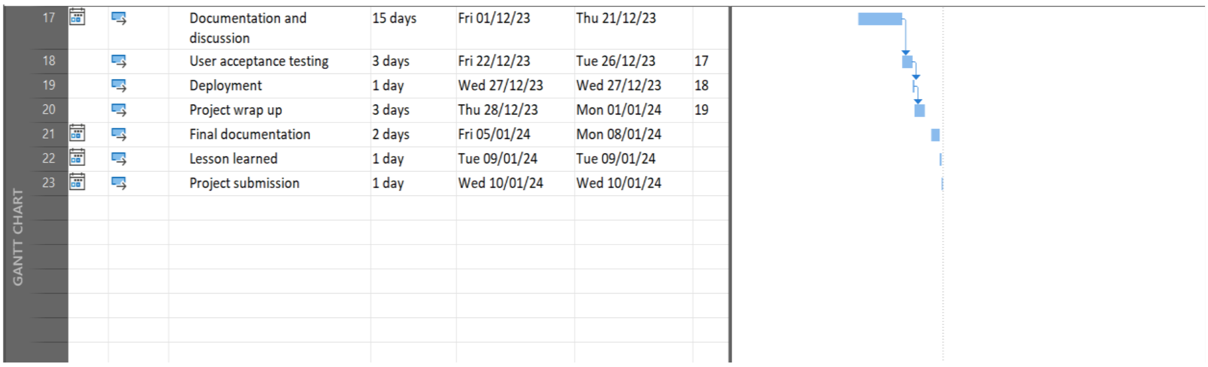
Xia, Y., He, L., Li, Y., Liu, N., & Ding, Y. (2020). Predicting loan default in peer-to-peer lending using narrative data. Journal of Forecasting, 39(2), 260-280.

Bhatore, S., Mohan, L., & Reddy, Y. R. (2020). Machine learning techniques for credit risk evaluation: a systematic literature review. Journal of Banking and Financial Technology, 4, S111-138.

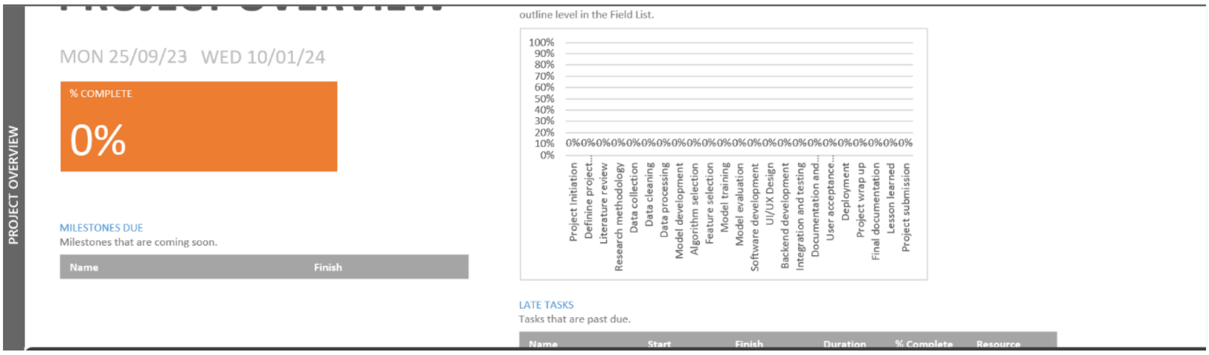
GANTT CHART



	Task Mode	Task Name	Duration	Start	Finish	Predec
0		**Software Development**	**78 days**	**Mon 25/09/23**	**Wed 10/01/24**	
1		Project Initiation	3 days	Mon 25/09/23	Wed 27/09/23	
2		Definine project scope and requirement gathering	4 days	Thu 28/09/23	Tue 03/10/23	1
3		Literature review	5 days	Wed 04/10/23	Tue 10/10/23	2
4		Research methodology	10 days	Wed 11/10/23	Tue 24/10/23	3
5		Data collection	7 days	Wed 25/10/23	Thu 02/11/23	4
6		Data cleaning	7 days	Fri 03/11/23	Mon 13/11/23	5
7		Data processing	4 days	Tue 14/11/23	Fri 17/11/23	6
8		Model development	10 days	Mon 20/11/23	Fri 01/12/23	7
9		Algorithm selection	5 days	Mon 04/12/23	Fri 08/12/23	8
10		Feature selection	3 days	Mon 11/12/23	Wed 13/12/23	9
11		Model training	5 days	Thu 14/12/23	Wed 20/12/23	10
12		Model evaluation	3 days	Thu 21/12/23	Mon 25/12/23	11
13		Software development	10 days	Tue 10/10/23	Mon 23/10/23	
14		UI/UX Design	7 days	Tue 24/10/23	Wed 01/11/23	13
15		Backend development	30 days	Tue 07/11/23	Mon 18/12/23	
16		Integration and testing	15 days	Fri 01/12/23	Thu 21/12/23	
17		Documentation and discussion	15 days	Fri 01/12/23	Thu 21/12/23	



17	Documentation and discussion	15 days	Fri 01/12/23	Thu 21/12/23	
18	User acceptance testing	3 days	Fri 22/12/23	Tue 26/12/23	17
19	Deployment	1 day	Wed 27/12/23	Wed 27/12/23	18
20	Project wrap up	3 days	Thu 28/12/23	Mon 01/01/24	19
21	Final documentation	2 days	Fri 05/01/24	Mon 08/01/24	
22	Lesson learned	1 day	Tue 09/01/24	Tue 09/01/24	
23	Project submission	1 day	Wed 10/01/24	Wed 10/01/24	



MON 25/09/23 WED 10/01/24

% COMPLETE

0%

MILESTONES DUE

Milestones that are coming soon.

Name	Finish

outline level in the Field List.

LATE TASKS

Tasks that are past due.

Name	Start	Finish	Duration	% Complete	Resource