IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A SUPERVISED MACHINE LEARNING BASED CLASSIFICATION TECHNIQUE FOR LOAN APPROVAL PREDICTION IN BANKING SECTOR

[1]Gaurav Raj Baser, [2]Dr. Sadhna K. Mishra

[1]M Tech Scholar, [2]HOD

[1] Computer Science and Engineering

[1]L.N.C.T., Bhopal, India

Abstract: The banking system claims that everyone's primary source of income comes from loans. A bank's primary source of revenue is, therefore, loans. One of the greatest threats to a bank's viability and profitability in today's cutthroat market is the difficulty of accurately assessing the risk associated with a loan application. Many individuals visit different banks every day to seek for loans. Every applicant does not get approval. Most banks utilize their risk assessment and credit scoring systems to determine whether or not to approve a customer's loan application. Getting a loan approved is a tedious process for bank employees. A more efficient, accurate, and quick loan approval procedure is possible with the use of modern technologies like machine learning models. Predicting whether or not a borrower will get a loan is an important yet long-standing problem for the financial services sector. Historically, banks and other lenders relied on manual processes and subjective criteria to evaluate loan applications, which often led to inconsistent decisions and increased risk of loan defaults. With the rise of ML algorithm, there is now an opportunity to create more accurate and reliable predictive models that can help financial institutions make better lending decisions. This research uses ML techniques to identify trends in a shared dataset of loan-eligible individuals. The dataset is prepared by performing exploratory data analysis and data balancing. The explored algorithms include Cat Boost, Gradient boosting and XGBoost. Preprocessing data, selecting features, balancing data, training and testing, classification, and comparing performance using accuracy and precision as classification metrics are the main goals of the project. Recall and f1-score. Following Gradient Boost Classifier (86.39%) and XGBoost (84.61%), the results reveal that Cat Boost Classifier had the greatest accuracy at 88.16%. The findings show that ML algorithms have the ability to make loan approvals better and decrease the likelihood of defaults.

Index Terms **- Component, formatting, style, styling, insert.**

CHAPTER- 1

INTRODUCTION

An Overview of ML Models for Bank Loan Eligibility Predictions Is Provided in Chapter 1. It covers a topic like bank credit risk and risk management, 7C's of credit analysis, types of credit risk the nature and security of loans, and factors affecting bank loans. The research issue statement, research questions, research motivation, and research aims and goals are all laid forth in this chapter. Finally, provide the abstracts of the theses.

Introduction

Due to the daily expansion of data caused by the financial sector's digitization, individuals prefer to submit loan applications via the Internet. As an increasingly prevalent instrument for data analysis, artificial intelligence (AI) is gaining in prominence. Applying their industry expertise, personnel from various businesses are utilising AI calculations to resolve issues. The approval of loans is becoming increasingly problematic for banks. The likelihood of an error occurring is considerable, as bank employees are confronted with a substantial volume of applications on a daily basis. Distribution of loans is a basic function of almost all banks. Divide the earnings from the loans across the bank's accounts. Therefore, a bank might suffer a huge loss due to a single

error (Gupta *et al.*, 2020). The banking industry's primary goal is to guarantee the safety of its customers' money. Nowadays, a lot of banks and other financial organisations only provide loans to those who pass a rigorous verification and validation procedure; nevertheless, this does not ensure that the recipient is the most worthy candidate among all of them. Our approach uses ML to automate the feature validation process and predict if a certain candidate is safe or not. Loan Prediction is a game-changer for everyone involved, from bank workers to borrowers (Kumar, Arun, Garg Ishan, 2016).

Machine learning can help automate loan eligibility prediction by analyzing vast amounts of data and identifying patterns and trends to eliminate human error. The ability to draw conclusions about the future based on information already collected is a major strength of ML (Srivastava, 2018). Banks may create models that reliably forecast loan eligibility depending on a number of criteria, including applicant qualification, gender, work status, and others, by training ML algorithms on past loan data and results. Additionally, the loan application process may be improved with the use of ML models by recognising high-risk applications and requiring human review based on crucial risk variables. By making loan decisions more quickly and accurately, eliminating the need for manual review, and making personalized loan recommendations, machine learning can also enhance the overall client experience. Overall, ML has the potential to improve the loan application process for both lenders and borrowers by making it faster and more accurate (Singh *et al.*, 2021).

Banking industry

The banking sector is a crucial component of any economy. Banks serve as intermediaries between savers and borrowers and are responsible for providing financial services to individuals, businesses, and governments. Lending money is a crucial function that banks provide. Loans are essential for the growth of the economy. Companies and individuals may use them to finance investments, house and automobile purchases, and other large-ticket items that would be impossible to fund otherwise. Loans also help businesses to expand their operations, hire more employees, and ultimately contribute to a growth of an economy. In addition to providing access to credit, banks also play a critical role in managing risk. Banks carefully evaluate loan applications to determine the creditworthiness of the borrower and assess the risks associated with the loan (Samreen, Zaidi and Sarwar, 2013). This lowers the default risk by directing loans to more probable repayers. Banks create lending policies, and according to such regulations, loans are approved based on the status of the applicant. Loan applications are reviewed in accordance with the standing of the applicant under the lending policies established by banks. Loans are frequently only approved by banks after a thorough evaluation of the applicant’s condition, either through methodically examining submitted documents or through direct asset verification. The person chosen from among all the applicants may not be the greatest choice, nevertheless (Sheikh, Goel and Kumar, 2020). There are many components that describe in loan eligibility.

- **Principal:** This amount of money was the initial loan amount by a bank.
- **Loan Term:** An amount of time of a loan repayment period for the borrower.
- **Interest Rate:** An APR, which is often utilized to express how quickly an amount owed is increasing.
- **Loan Payments:** The sum that has to be paid on a weekly or monthly basis after pay off a debt. In accordance with a loan's principal, length, and interest rate. In order to get a loan, a prospective borrower has to prove they can pay it back and are financially responsible.
- **Income:** In order to prevent borrowers from experiencing financial hardship, banks may establish a minimum income requirement for bigger loans. Particularly for mortgages, a steady income for a number of years could be required.
- **Credit Score:** Credit scores are numerical indicators of a borrower's reliability that are calculated from their payment and borrowing patterns (Boddepalli, 2022). The credit score of an individual might take a major hit in the event of bankruptcy or missed payments.
- **Dept-To-Income Ratio:** Borrowers' income and credit history are two factors that lenders consider when deciding how many loans they have available at any one time. When a borrower has a lot of debt, it could be hard for them to pay it back.

Types of risks in banking

Credit, market, and operational risks were classified by the RBI as a three main categories of banking-related hazards (Krishn A., 2010). Financial risks and non-financial risks are two broad categories into which the hazards could fall. The credit and market risks that make up financial risk are distinct from the operational risks that are considered non-financial. Below is fig.1.1, which shows the three primary categories of risk classification. Credit risk includes risks such as those involving counterparties or borrowers, as well as those pertaining to the industry, the portfolio, or concentration. Risks associated with interest rates, liquidity, hedging, and foreign exchange (Forex) or currency are all examples of market risk. Nonfinancial risks include, but are not limited to, funding, strategy, legal, and political risks.

1) Credit risk

(Kargi, 2014) views the generation of credit as the principal means by which banks generate revenue. It indicates that banks take on and accept risks in order to generate profit by granting credit. Because of this, credit risk management is still essential to the expansion and survival of banks, since credit operations have the potential to generate financial instability.

2) Counterparty/ Borrower Risk

Each party to a contract has the risk that the counterparty will not uphold their end of the bargain. This is known as counterparty or borrower risk. When designing and executing contract contracts, counterparty risk must be taken into account since it poses a risk to all parties.

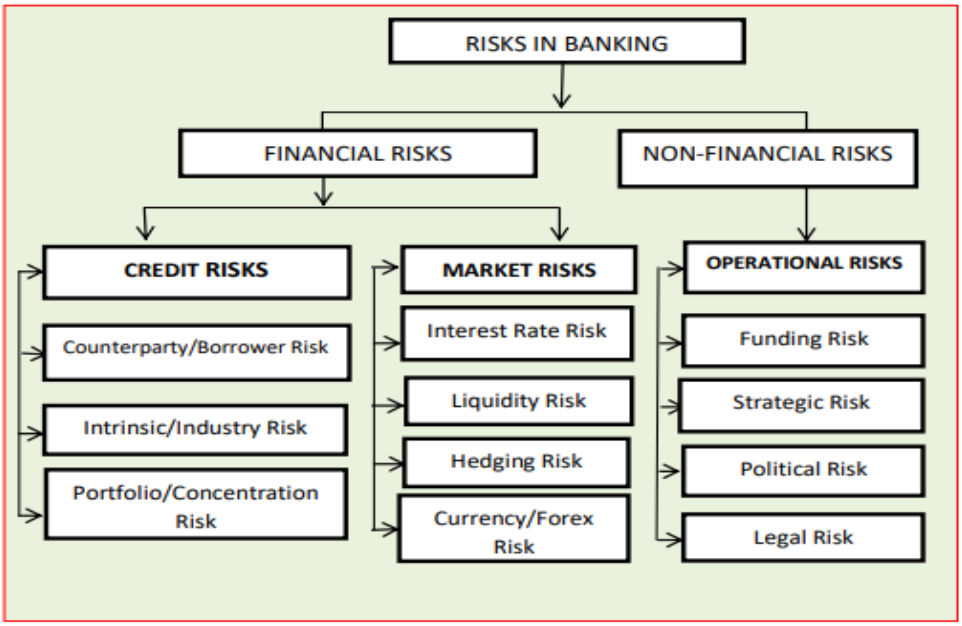


Figure 1.1: Types of risk in banking

3) Intrinsic/ Industry Risk

The two primary sources of credit risk are default risk and portfolio risk. Investment portfolios have both inherent and concentrated risks. An organization's loan portfolio is vulnerable to credit risk due to both internal and external causes. Some examples of external influences include government regulations, trade restrictions, interest rates, currency exchange rates, stock price fluctuations, and economic penalties. Factors inherent to the problem include appraisal flaws, an over-reliance on collateral and low-risk pricing, a lack of a mechanism to review loans, and a failure to monitor funds after they have been disbursed. In addition, credit committees and officials do not have access to well-defined lending limitations.

4) Portfolio/Concentration Risk

The concentration risk is a probability that a certain percentage of borrowers may default on a bank's loans. A concentration ratio is used to assess the potential dangers of concentration. Each bank loan's share of the total accounts is shown there. The default rate will be too high for the bank to handle if the economy slows down in a certain sector where the majority of its assets are located. Therefore, it's smart to lend to many parts of the economy to spread out the risk.

5) Market risk

The potential for financial loss as a result of fluctuations in market prices is known as market risk. Due to changes in the market price, an investment might be subject to market risk. Here, the possibility of a decline in the investment's value exists. The phrase could apply to a particular commodity or currency and is also called systemic risk.

6) Liquidity Risk

If a financial organisation can pay its debts as they become due, it is considered liquid. The capacity of a bank to handle withdrawals of deposits and other obligations, as well as to finance the expansion of its loan portfolio and off-balance sheet claims, is known as liquidity. A financial risk brought on by erratic liquidity is liquidity risk. It results from using short-term obligations to support long-term assets, which exposes the liabilities to rollover or refinancing risk.

7) Hedging risk

A simple way to think of hedging is as insurance. Hedging is like purchasing insurance against anything bad happening. Although hedging cannot eliminate the possibility of unfavourable events, it can reduce their severity in the event that they do occur. Companies, portfolio managers, and individual investors all utilise hedging strategies to lessen the impact of potential losses. When it comes to financial market hedging, it's not as simple as making an annual payment to an insurance provider.

8) Interest-rate Risks

Negative changes in interest rates pose a danger to a bank's bottom line and other financial transactions. One potential consequence is interest rate risk, which might affect net interest income. The term refers to the gap between the interest that is generated on loans and the interest that is paid on deposits. Financial losses associated with asset and obligation management due to changes in market interest rates constitute this form of risk.

9) Operational risk

In an operational risk assessment, "direct or indirect loss stemming from the inadequate or failing internal process, people and systems or external events" is a tangible possibility. operational risk as failures in the reporting system, information systems, internal monitoring policies, and processes that are meant to address issues promptly or ensure adherence to the internal risk policy.

10) Legal and Regulatory Risk

There are a great number of contractual agreements that financial institutions engage into every day as a result of their normal business operations. It follows that banks often find themselves entangled in complex legal battles with their counterparties. Legal risk refers to the possibility that a company's activities may be severely interrupted due to litigation brought by dissatisfied customers, employees, or regulators over issues such as workplace restrictions, poor recordkeeping, or environmental damage.

11) Strategic Risk

This danger might develop if an ineffective company strategy is pursued. Examples of potential sources of strategic risk include management's inability to adapt to a changing business environment, insufficient allocation of resources, and sloppy decision-making.

12) Funding Risk

This is the risk of not having enough money to finish a project or having to pay for it at a higher rate. Inadequate finances to run the firm are the consequence of a project's cash flow being strained due to a high cost of capital (Boateng, 2020).

Bank credit risk and risk management

Banks have ranked credit risk as their top management concern. As their core competency, banks take on financial risks in return for advantages, hence the quality of their credit risk management is crucial. The following is an explanation of credit risk: it occurs when entrepreneurs face deterioration or other factors that prevent them from fulfilling their contractual obligations, such

as entanglements between firms. As a result, there is a risk of agreement violation and financial loss. In general, there are two subsets of credit risk based on various items and behaviours:

1. Issuer Risk, which is a kind of lending risk. The failure of bond issuers or borrowers to repay their obligations or a decline in their creditworthiness poses this risk since it is a breach of agreement. Debt credit conditions of bond issuers and borrowers, as well as the degree to which financial products are risk sensitive, are common factors that contribute to lending risk or issuer risk.
2. After first-party risk, there is second-party risk, which may be further classified into pre-settlement risk and settlement risk. A danger that the bank may lose the equality principle because the counterparty does not satisfy their contract obligations by the due date is known as settlement risk. There is a possibility that the bank may incur a risk of contract breach prior to the final settlement day, which is known as pre-settlement risk.

Credit risk management departments and organisations at banks may take several shapes and sizes. After achieve an objective of credit risk management and supervision, a bank must make sure that connected authorities and official positions operate freely and responsibly, without relying on surface independence (Aebi, Sabato and Schmid, 2012) (Jiang and Lo, 2014) (Ferreira and Oliveira, 2014) Swami 2014):

- Avoiding conflicts of interest requires keeping business operations apart from credit giving and verification processes.
- To ensure that the credit result report is fair and objective, credit verification processes have to be separate from credit providing operations.
- To avoid issues like fraud and malpractice, it's crucial to maintain accounting activities apart from business processes & credit granting/verification functions.
- There should be complete separation between the credit-providing activities and the unit responsible for the credit risk measurement system. This will ensure that this unit is free from any other potential disruptions.
- If you want to reduce the likelihood of mistakes in the credit risk assessment system, it's a good idea to have a separate office worker responsible for checking the system from the one who designed or selected it.
- Authorities should follow the rules that limit who may own shares in the bank.
- Verify again the employees who have a vested stake in the bank's creditworthiness, including the general manager and senior officer.
- Credit giving must be in conformity with the plans and policies of the bank's credit risk management, which must be reviewed regularly (at least once a year) to ensure that the top-level managers are effectively implementing the requirements. This is to ensure that senior management is held fully accountable for developing and maintaining an adequate system for managing credit risk (Themba and Narayana, 2014).
- Verify the appropriateness and adequacy of the bank's capital by conducting routine inspections of management records and by considering appropriate credit risk methods.

7C's of credit analysis

In order to approve a proposal, banks often look at the 7C's of credit, which pertain to the borrower (By, Karki and Dev Campus, 2008):

1. **Character:** An important factor in a lender's selection is the borrower's character, which includes attributes like honesty, dependability, and integrity. The bank takes the borrower's genuine intent, honesty in responding to their questions, sense of duty, and commitment to repaying the loan seriously.
2. **Collateral:** Banks always demand collateral or securities from borrowers, regardless of their financial situation, to protect their hazardous assets in the event of a default. Collateral may be physical assets like land or buildings, or it can be a kind of operating capital like accounts receivable or inventory.

3. **Condition:** There are a lot of societal and economic factors, such security and political climates, that impact businesses, and these factors are beyond the control of the borrower. When a lending authority has a positive outlook on the borrower's business model, they are more likely to provide a loan.
4. **Cash Flow:** Credit bureaus often look at a company's cash flow to see whether they can repay the loan plus interest. They provide credit to customers whose figures reflect a good reaction.
5. **Capacity:** The bank considers two factors. The applicant's legal competence to borrow money is the first thing the bank checks. The second consideration is whether or not the borrower can sustainably earn enough money to pay back the loan. A customer is considered to have high capacity and the bank will offer a loan based on factors such as excellent management and solid market value. As a result, appropriate ratios are examined using forecasted and past financial data, including liquidity, leverage, profitability, and efficiency.
6. **Capital:** The borrower's net value is called capital. Borrowers with insufficient capital will have a high leverage ratio. In order for a bank to provide credit, the borrower must either meet their capital requirements or have a leverage ratio that meets their standards.
7. **Credit Information:** A borrower should verify a sort of loan they need and a bank should provide them all the necessary credit details in advance.

Nature and security of loans

An ability of borrowers to repay a loan amount is the primary consideration for banks when lending money. The reliability, stability, and capacity of the borrower are therefore the most significant factors for banks. However, the borrower's physical asset serves as security for the loan since the bank is unable to bear any risk in this area. The bank may seize the borrower's assets to recoup the loan amount in the event that the borrower defaults on the loan. It may get the money by selling the assets that were pledged as security. So, we may classify loans as either (a) secured or (b) unsecured based on the level of security they provide. A loan that does not need collateral in the form of physical assets is known as an unsecured loan. These loans are given to businesses or organizations in exchange for the owner's, manager's, or director's personal assets. The opposite is true with secured loans, which are backed by physical assets like real estate or stocks in trade. The bank creates a charge against the borrower's assets in favour of itself when it lends money against certain assets. In the event that a borrower is unable to repay a loan, the bank may collect the outstanding balance from the client using the money that is made from the sale of the assets. While many different kinds of collateral may be put up to secure loans, not all of them are acceptable to banks. The following are examples of the kinds of assets that the bank often accepts:

a. Tangible assets such as plant and machinery, motor-van, etc.
b. Documents of title to goods, like Railway Receipt (R/R), Bills of exchange, etc.
c. Financial securities (Shares and Debentures)
d. Life-Insurance Policy.
e. Real Estates (Land, Building, etc).
f. Fixed Deposit Receipt (FDR).
g. Gold Ornaments, jewellery etc.

Factors Affecting Bank Loan

Funds for lending may be found in three places: reserves, deposits, and capital. A variety of variables may impact each of these sources, which in turn would affect lending (Zewdu Seyoum, 2010). Considering that lending is the main function of the banking business, it is imperative that bank management promptly attend to, analyse, and address any internal and external variables that impact or restrict lending. Banks' ability to stay in business would be severely threatened if they stopped lending money, particularly interest revenue. If this is the case, the variables that affect banks' loan portfolios also affect their NPLs, which are a direct result of low asset quality (Rawlin, Sharan and Lakshmipathy, 2012). Consequently, the following are some of the banking loan-influencing elements that may have an effect on nonperforming loans:

A. **Capital position:** Bank capital is a standard measure for the safety of depositors' money. A bank's ability to take on risk is proportional to its capital to deposit ratio. Loans with longer maturities and higher credit risk may be made by relatively big capital structures.

B. **Profitability:** Different banks may place a greater emphasis on earning potential. Banks that are in a stronger financial bind may adopt more stringent lending standards. A policy that is too aggressive could refer to consumer loans, which often have higher interest rates than short-term loans.

C. **Stability of deposits:** A deposit type and the fluctuation rate need to be thought about. After sufficient funds have been set aside for reserves, the bank may then begin lending. Although these reserves are set up to handle expected changes in deposit levels and loan demand, banks nonetheless factor in deposit stability when making lending decisions due to unexpected demand.

D. **Economic conditions:** A flexible lending policy works better in an economy that doesn't experience cyclical or seasonal fluctuations. The deposit of a food-deficit economy is more prone to wild swings than that of a stable economy. The country's economy has to be thought about. If a factor has a significant negative impact on the country as a whole, it may have a trickle-down effect on local situations.

E. **Influence of monetary and fiscal policies:** When government spending and interest rates are low and flexible, commercial banks have more money to give out. With these rules in place, financial institutions will be able to lend money more freely.

F. **Ability and experience of bank personnel:** As a bank formulates its loan policy, the knowledge and experience of its lending staff is crucial. The shortage of qualified workers was likely a factor in the banking industry's reluctance to go into consumer lending.

G. **Credit needs of the area served:** financial institutions have expertise in a variety of loan kinds, including mortgage real estate. One of a primary purpose of chartered banks is to offers the community's credit requirements. Lenders have an ethical obligation to provide credit to those who can demonstrate that their loan demands are reasonable and economically viable. Banks' lending and investment decisions are influenced by other variables as well.

H. **The interest rate:** symbolizes the potential rates of return from several alternative investment and financing endeavors. Finding the sweet spot between profit and risk is a perennial challenge for bank managers.

I. **The liquidity of fund:** The term refers to the total value of liquid assets that are invested or loaned out. A bank's ability to keep its cash on hand depends on its vigilant defense against unsustainable losses in lending and investment. The value of the bank's assets may fall below its obligations if it made an excessive number of poor loans.

J. **Tax:** There are a few ways in which the corporate income tax rate impacts bank loans. Firstly, when banks have a heavy tax load, they may choose to either charge higher interest rates on loans or fees, or they can offer lower interest rates on deposits. The second thing to think about is that a rate of corporate income tax affects input substitution and output. According to the output substitution impact, the integrated sectors' production drops as the CIT rate rises. As a result, fewer people are looking to take out loans, and the input substitution effect shows how equity is being replaced with other inputs, including debt. Banks are able to lower their tax burden by strategic portfolio allocation, which is a key component of banking industry taxation. Customers may see increased fees and interest spreads as a result of the bank passing the cost of the tax on to them. Interest rates on loans are increased while rates on savings are decreased in direct proportion to the quantity of NPLs, which shifts the tax burden onto consumers (Khan *et al.*, 2011). Additionally, corporations shift their tax burden on other individuals or entities due to the existence of double taxation.

Organization of The Thesis

The primary premise is outlined below:

Chapter 1 gives an overview of ML models utilized to forecast who would be eligible for a bank loan. Research questions, a description of the issue, and an explanation of why this study is necessary are all laid forth in this chapter. Finally, provide the abstracts of the theses.

Chapter 2 includes background information, a summary of the literature, and previous instances of similar work to set the stage for the investigation. Machine learning and deep learning-based eligibility prediction for bank loans is also covered in this chapter. Finding the study Gap and the obstacles to doing this study may be aided by this survey.

Chapter 3 refers to the methodology used in the study that produced this system to forecast loan eligibility from financial institutions. Include a process diagram and a description of our proposed paradigm. A data set describing the process of feature selection and data pre-processing was available. Our model and all of its details were briefly described.

Chapter 4 analyses the results of putting this work into action. The experimental design, data sets, and application of the EDA paradigm are all covered in this chapter.

Chapter 5 reviews a literature on a topic of predicting bank loan eligibility, discusses the ways that have been suggested, and assesses the efficacy of these methods; finally, uses a model that performs a best to make these predictions. Checking the predictions against a previous model's predictions is the next stage.

Chapter 6 The final chapter concluded with Limitations and Further Research.

Problem Statement

An essential function of banking is the estimation of loan risk. Because of this, studying the potential gains from using machine learning is intriguing. The potential advantages of using machine learning to default prediction have been covered in several scholarly articles. When compared to more traditional approaches, machine learning has the potential to provide more accurate default predictions. The number of methods compared is, however, somewhat small in most of the studies. Furthermore, it is difficult to compare the findings since almost all of the publications employ distinct data sets. This opens up a fascinating possibility for comparing the results of several algorithms using the same datasets.

Research Questions

This study's central question is stated as:

- **RQ1:** When it comes to loan prediction, how well does machine learning perform?
- **RQ2:** Which machine learning methods are appropriate for classifying objects into binary categories?
- **RQ3:** What criteria should be utilized to objectively and properly evaluate the performance of various ML algorithms?

Aim and Objectives

An overarching goal of this research is to create a model for financial loan eligibility prediction that makes use of ML. For a purpose of default loan forecast, we will run ML models and calculate several performance metrics. Some key objectives as:

- To collect the Loan Eligible Dataset by a Kaggle for studying and prediction of loans eligibility.
- To implemented machine learning model for the loan prediction using some classification measures.
- To improve a classification performance of a loan prediction using ML model.
- To evaluated with a standard metrics and compared with each other on base and propose models.

Significance of the Study

In contrast to the systemic change, there has been a dearth of research focusing on the banking sector. Performance evaluation has been the primary focus of most studies involving commercial banks. The research fell short of what was needed to bring about a paradigm shift in the loan and advance industry. Since loans are essential to the functioning of the financial system, thorough and precise research is required. The whole financial system is based on lending. That is why it is so important to do research like this on commercial banks' loan and advance policies. Anyone curious in the state of bank loans and advances would do well to examine this report. These findings may also be useful for the chosen banks' credit departments.

Research gap

Numerous issues remain uncured. To start, many datasets are not well-suited to a straight line. As an example, consider a quadratic connection where the value of y fluctuates substantially with respect to x but too little with respect to x. A lender's review of a borrower's credit history is a common part of the assessment process. Many factors go into deciding whether or not to approve a loan. These include the borrower's occupation, assets, credit history, and the amount of the loan. If prior borrowers who met your requirements have also paid their loans on time, your application will have a better chance of approval. An example of a data science challenge that might benefit from ML methods is the prediction of a new applicant's loan status using the same set of criteria as past applications, which heavily relies on previous information and comparisons.

CHAPTER- 2

LITERATURE REVIEW

Chapter 2 provides background information, a literature review, and examples of previous work that is relevant to the inquiry. Additionally, this chapter covers the use of DL and ML for the prediction of loan eligibility. Moreover, details pertaining to bank loans are detailed. Finding the study Gap and the obstacles to doing this study may be aided by this survey.

Background study

Many industries are considering using machine learning and data mining methods because of their recent advancements. An interest in improving current techniques of risk estimate has arisen due to the rising demand for effective risk management by financial organisations, which includes the banking industry. One possible outcome of using machine learning methods is that banks' financial risks may be better quantified. In the field of credit risk, the Basel accords have evolved throughout time to establish norms for supervisory practices and methods for managing risk; they serve as guidelines for how banks should handle and measure their own risks. Basel II (IRB) incorporates two approaches to establish the minimum capital requirement: the standardised technique and the internal ratings-based method (BIS, 2014). Banks evaluate the potential harm they may face in the future by considering a variety of risk factors. In an event of a client default, one of these metrics would be the expected loss (EL) that the bank would incur. The likelihood that a certain customer may default is one factor that goes into EL estimate. When customers are in default, it indicates they haven't paid their bills as agreed upon and may have trouble paying back their debts. The acquisition of a model capable of predicting which customers would default is being considered. A popular method for determining the likelihood of customer default is Logistic Regression (Tong, Mues and Thomas, 2012). This research will look at a variety of machine learning approaches to see if any of them can compete with the more conventional ways. An individual's forecast on the future is called a prediction.

Related work

As in other fields where ML algorithms have proven effective, such as data mining and decision support systems, they have also shown promise in the prediction of bank loans. Using both conventional and deep learning techniques, please detail the machine learning-based loan eligibility prediction here. An use of ML for a purpose of predicting loan eligibility has been the subject of much research. Nevertheless, as will be shown later on, very few have really performed a comprehensive literature assessment on the subject.

Predicting Bank Loan Eligibility using machine learning

This study (Prasanth *et al.*, 2023) proposed an active system that uses data from past events to rank customer wait times. New loan application systems appear every day. The eligibility and criteria of the client are the primary factors considered by the bank when deciding whether to grant a loan request. In this case, the final clearance is determined by using machine learning algorithms to predict previous customer data. An effective method for precisely forecasting and documenting the customer's loan payback is the RF model. While reducing a temporal complexity, a suggested algorithm enhances accuracy.

This study (Rahman, Purno and Mim, 2023) utilise a variety of machine learning algorithms to sift through massive volumes of data in search of trends that might be used to forecast the likelihood of loan approval and a borrower's ability to repay a loan. A secondary objective of this research is to show how the financial industry may improve its predictive modelling tool with the use of machine learning. One thousand records from a reputable bank in Dhaka, Bangladesh, including sixteen characteristics, were

used to compile the data used in this study. Finally, seven data mining approaches are used to ensure correctness. The following are some suggestions for dramatically streamlining the loan approval process to save time and money: An assortment of ML methods include LR, DT, gradient boosting, RF, naive bayes, and SVM. According to the results, the Random Forrest approach is the most accurate.

In this work (Swapnesh, Nayak and Swarnkar, 2023) deployed various ML algorithms to identify the loan approval status and compare the performance of implemented models. The implemented models will attempt to predict our target column on the test dataset using information from the loan eligibility prediction dataset obtained from Kaggle, which includes features like loan amount, number of dependents, and education. The parameters like accuracy, confusion matrix, ROC curve, and precision are measured for specific models whose performance is significant.

In (Uddin et al., 2023) provide a novel loan prediction system that uses machine learning (ML) to automatically find people who are eligible to get a loan. Covered in this comprehensive study are data preprocessing and efficient data balancing using SMOTE, as well as the application of various ML models. These models include LR, DT, RF, Extra Trees, SVM, KNN, Gaussian NB, AdaBoost, GB, and advanced DL models such as DNN, RNN, and LSTM models. Accuracy, recall, and F1_score are utilized to rigorously evaluate a model's performance. The Extra Trees operates better than its competitors, according to our trial results. In addition, we outperform the Extra Trees by 0.62% when using an ensemble voting model that incorporates a 3 best ML models to forecast who would default on bank loans. A desktop-based programme that is easy to use has been designed to help users engage. Our findings demonstrate that the voting-based ensemble model achieves an impressive accuracy of 87.26%, surpassing both standalone ML models (including Extra Trees) and existing modern approaches. Both banks and borrowers stand to gain from this cutting-edge system's ability to expedite and simplify the loan approval process.

The study (Dhruv et al., 2023) require the creation of two independent prediction models: one to forecast future income and another to forecast future repayment of bank loans. Based on users' unique financial situations, the income prediction algorithm will project their potential future income. Financial institutions may use the borrower's financial history and present financial situation to inform their loan repayment prediction model. Using this predictive technique, they can get a better understanding of people's financial stability and the creditworthiness of borrowers. It will help individuals plan for their financial futures in many ways, including saving for retirement and investing in the stock market. Financial institutions will be able to make better lending choices, default rates will be lower, and the sector as a whole will be healthier as a result. Individuals and banks alike stand to benefit from the insights and tools made possible by a well-developed predictive framework for income and bank loan repayment. By using precise income and loan payback projections, one may make educated financial choices, enhancing the overall financial stability and welfare of everyone concerned. Based on customer questions, its user-friendly financial bot may give basic explanations of financial concepts.

In (Zhu et al., 2023) An explanatory model is necessary to improve the user's trust in the model by making the rules of the prediction model more comprehensible. To forecast when a loan would go into default, we use models like XGBoost, LightGBM, DT, and LR. According to a prediction findings, XGBoost and LightGBM are the most effective models, surpassing decision tree and logistic regression. At 0.7213, LightGBM has an area under the curve. LightGBM and XGBoost have accuracy levels higher than 0.8. XGBoost and LightGBM both have precisions higher than 0.55. We also conducted an explainable study of the prediction results simultaneously, using the local interpretable model-agnostic explanations technique. The outcomes that are expected to happen depend on variables like the amount of a loan, a length of a loan, a grade of a loan, and the borrower's credit score.

This study (Rodrigo, Sandanayake and Silva, 2023) proposed an alternative to the conventional approaches used by banks to determine creditworthiness, one that compares five ML algorithms to predict the likelihood of personal loan defaults using the debt-to-income ratio. Customers' debt-to-income ratios were used to classify risk groups by ensemble clustering. The influence of debt payments on loan defaults was also explored. The experimental findings showed that when predicting loan defaults, ensemble clustering outperformed traditional classification models.

In this study (Gao, Ju and Yang, 2023) They demonstrate that three ML algorithms—ANNs, GB Trees, and RF—are effective at predicting loan defaults with accuracies of 70%, 74%, and 81%, respectively, using a combination of U.S. severe weather data and data extracted from Lending Club. Further proof that extreme weather and other explanatory factors are economically relevant may be found in the Shapley Additive Explanations (SHAP) model.

In this study (Mamun, 2022) The purpose of using ML algorithms is to forecast who will be eligible for a loan based on patterns extracted from a shared dataset of loan approvals. Age, income type, loan annuity, latest credit bureau report, kind of organisation, and duration of employment are some of the customer data points that will be used for the research. The most important features, or those that significantly affect the prediction result, were identified using ML approaches including Decision tree, XGBoost, RF, Adaboost, LGBM, and KNN. Using industry-standard measures, we compare and evaluate the aforementioned algorithms. Logistic Regression outperformed the others with a 92% success rate. With an F1-Score of 96%, it outperformed all other machine learning approaches and was therefore named the best model.

This paper (Li et al., 2021) retrieves the dataset containing loan defaults from lending club. The dataset's class imbalance is addressed using the ADASYN approach. This research makes use of the Blending method to improve prediction accuracy by combining three models: LR, RF, and CatBoost. The experimental results demonstrate that the proposed fusion model performs better than the three models mentioned earlier (LR, RF, and CatBoost) when it comes to forecasting the probability of customer loan default employing the dataset for training and reducing the external risk that online lending platforms encounter when handling customer loan default.

In (AKÇA and SEVLİ, 2022) anticipate if the bank loan proposals will be accepted using the SVM algorithm. The results were determined by using a SVM with 4 SVM kernels, a grid search method to enhance prediction, and cross validation to ensure the results were very reliable. According to a data, a sigmoid kernel had the lowest success rate at 83.3% accuracy, while a poly kernel had the highest results at 97.2% confidence. Due to an imbalanced dataset, certain accuracy and recall numbers are lower than typical, such as 0.108 and 0.008, respectively. For every 1 true value, there are 9 negative values, representing 9.6% of the true value. This research suggests that SVC should be used in the banking sector to better anticipate whether customers would accept loan offers.

In (Tejaswini, 2022) proposed a vigorous predictive approach to decide whether an applicant should be accepted or rejected. An efficient, simple, and fast way to choose eligible applications was the goal of this research. Many private companies contributed to a dataset. Separate datasets were utilized for training and testing purposes; a former was utilised to instruct the model, while the latter was employed for model validation. In order to forecast whether a consumer would be approved for a loan, this research used three different kind of machine learning algorithms. The evaluation outcomes suggest that the Decision Tree algorithm has outperformed all models with the highest accuracy of 82%.

In this study (Kumar et al., 2022) A number of ML algorithms are used to forecast whether or not the consumer is eligible for a loan. To incorporate machine learning methods, it is necessary to gather customer data from many banks and access client profiles in order to analyse the data using the parameters. Machine learning is a step ahead of the curve compared to older methods of loan approval that relied on data analysis and client profiles to make lending decisions. Project objectives include data cleansing, appropriate characteristic selection, and a comparison of AdaBoost, RF, SVM, KNN, and DT as ML approaches for predicting a customer's loan eligibility. During this procedure, two datasets are utilised: one for training the model and the other for testing its efficacy. According to the findings, the adaboost-based ensemble model DT performed better than the other models that were implemented.

This paper (Orji et al., 2022) offers a suite of six(6) ML technique—LR, RF, GB, DT, SVM, and KNN—to assist in determining loan eligibility. The models were trained using the "Loan Eligible Dataset," a historical dataset that is available on Kaggle and licenced under Database Contents Licence (DbCL) v1.0. Python modules hosted on Kaggle's Jupyter Notebook cloud platform were used to process and analyse the dataset. Based on our study, the RF model achieved a best performance accuracy at 95.55%,

while the LR approach achieved a lowest performance accuracy at 80%. Both the accuracy and precision-recall metrics used to evaluate loan prediction models in the literature were surpassed by our Models.

This study (Meenaakumari et al., 2022) create an ML model that can analyse the user-provided data and determine, based on that data, if the user is qualified for a health loan. This procedure begins with retrieving a dataset from Kaggle that contains all of the required loan application parameters. Two approaches, encoding and the null value removal method, are used to preprocess the dataset once it has been gathered. At the same time, three distinct methods were used to generate three separate ML models. These three models are known as RF, NB, and LR, or Linear Regression. After that, the models will be trained using the preprocessed data. Next, they will evaluate the models' performance by comparing a few parameters. Both the accuracy and the error rate of the RF method are determined to be the best by the study. The RF algorithm not only forecasts loan eligibility with lower error values, but it also has an accuracy of 91%. Among the available algorithms for loan forecasting, a LR model is a least effective due to its large error value and poor accuracy values.

In (Kumar, Sharma and Mahdavi, 2021) to determine which ML-based models are most suitable for assessing credit risk and to compare them. The goal of the authors was to demonstrate the several ML algorithms used by academics to evaluate rural borrowers' creditworthiness, particularly for those with a poor credit history. The ML algorithms employed in this study were popular and effective, as shown by their findings. Financial institutions throughout the world are grappling with the negative effects of historically low loan payback rates and are actively seeking for better methods to manage the loan approval process.

In (Park et al., 2021) used accuracy, recall, and precision to evaluate six distinct machine-learning algorithms. Determining the likelihood of a loan's approval is the primary goal of this project. Random Forest displayed the most accuracy in their study, at 95.55%, whereas Logistic Regression displayed the lowest accuracy.

In this paper (Ambika and Biradar, 2021) for the sake of the bank's time and resources, it would be wise to minimise the danger of selecting the prudent person. This is achieved by first mining Big Data for details about previous loan recipients and then training the computer with a ML model that yields the most accurate findings, all based on these records and experiences. Determining whether or not it is safe to provide the loan to a certain person is the main objective of this research. There are four sections that make up the main body of this piece. Section I: Data Collection; Section II: Machine Learning Model Evaluation; Section III: System Training Based on the Most Promising Model; and Section IV: Testing

According to (Xu, Lu and Xie, 2021) utilized RF, XGBoost, GBM, and Neural Network ML models to assess the Chinese P2P market's ability to forecast loan defaults. All four of their models were more accurate than 90%, but RF was the best. This study's methodologies and algorithms are quite similar to ours; however, their focus was on predicting P2P loan default, while ours was on predicting whether or not a consumer would be eligible for a loan.

In this research, (Meshref, 2020) wide array of ensemble ML methods, including Boosting and Bagging, have been implemented. Compared to other modern loan forecsting models in a literature, our study indicated that the loan approval prediction model outperformed them by almost 25% with an accuracy of 83.97%. There was confidence in the models' predictions because of the research's model interpretation efforts, which helped shed light on several crucial scenarios that bank decision makers may face. For decision makers to grasp the delicate balancing act of keeping their financial lending system secure and reliable while still offering fair credit opportunities to clients, we think the obtained model accuracy, along with the interpretation information, is imperative.

The research by (Aphale and Shinde, 2020) researched real bank credit data to assist banks in developing an automated risk assessment system and to forecast consumers' creditworthiness. They used a variety of ML methods, like NN, NB, KNN, decision trees, and ensemble learning. They also had lower-than-usual model accuracy, with estimates ranging from 80% to 76%.

In (Aslam et al., 2019) examined the three most popular and trustworthy neural networks and ML algorithms for loan default prediction. Insightful discussion of the benefits and drawbacks of employing certain models was provided in the research. The authors pointed out that most research have concentrated on the accuracy of loan default predictions, while a small number have looked at the consequences of false negatives, which may be very harmful to lending organisations. Therefore, the authors concluded that lowering the rate of false negatives in loan lending forecasts should be a primary focus of future studies in this field.

In (Karthiban, Ambika and Kannammal, 2019) Algorithms for ML influence and govern almost all contemporary applications. Many researchers are developing new approaches to improve ML algorithms' precision and performance, but this remains an ongoing issue. A bank provided them with the data. This study compared the precision, recall, and f-measure of several classification algorithms for predicting loan approval decisions made by financial institutions. The most successful classifier, according to the classification matrices (accuracy, precision, recall, and F-1 score), was Gradient Boosting, which achieved an accuracy of 98.06% and an F-1 score of 99.20%.

In (Abakarim, Lahby and Attioui, 2018) research proposes employing a binary classification technique to determine loan acceptance in real-time. They are able to categorise loan applicants as good or poor risks using their deep neural network-based technique. Experimental results show that compared to conventional binary classifiers, the Real-Time model built on deep neural networks achieves higher levels of accuracy, recall, and precision.

This paper (Arutjothi and Senthamarai, 2017) construct a credit score model with credit data. A number of ML techniques are used in the development of the financial credit score model. Our proposal for analysing credit data is based on a machine learning classifier. They employ a K-Nearest Neighbour (K-NN) classifier in conjunction with Min-Max normalisation. The R software package's utility is used to accomplish the goal. By far, the most accurate data is provided by the proposed framework. For commercial banks, it's a machine learning classifier for loan status prediction.

The study (Asare-Frempong and Jayabalan, 2017) determine which classifiers work best for forecasting how customers will react to a bank's DM campaign. This research compared 4 classifiers—MLP, NN, DT, LR, and RF—using data collected by a random bank. Using ROC and classification accuracy metrics, the results demonstrated that the RFC outperformed an others with an 87% accuracy rate. A secondary objective was to determine which customer attributes were associated with previous term deposit subscribers who were most likely to resubscribe. This was accomplished by means of cluster analysis.

In (Jin and Zhu, 2015) During the modelling phase, utilise random forest to choose features based on the loan and applicant's qualities. Unlike competing risk prediction models, this one divides predictions into at least three distinct groups, rather than the usual two: default and non default. Afterwards, we evaluate the prediction results of five DM models: two DTs, two NNs, and one SVM using two metrics: area of the lift cumulative curve and average percent hit rate. Factors that contribute significantly to loan defaults, according to the empirical findings, include the loan's length, yearly income, amount, debt-to-income ratio, credit rating, and revolving line utilisation. Plus, the prediction abilities of SVM, CART, and MPL are almost identical.

Deep learning for bank loan eligibility

In (Vivek and Mahaveerakannan, 2023) check out the differences and similarities between LR and the RF method to determine which one enhances the accuracy of loan appraisal. Data was compiled using the following criteria: the mean, standard deviation, a 95% confidence interval, a 0.05 percent threshold, and the outcomes of the current study. The data was acquired from several sources. Two distinct methods are employed for data classification on a n=10 sample, with the G-power tool amplified by 80%. A remarkable 81.30% accuracy rate was achieved using the LR approach, in contrast to the RF method's 75.6 %. Distinct differences between the two sets of data are considered statistically significant when they are less than 0.05 or 0.001. This finding provides additional evidence that LR outperforms RF when it comes to estimating a customer's loan amount.

In (Kiran et al., 2023) built a system to automatically forecast loan repayment terms using ML techniques to address a problem. ML will be implemented using historical data. In order for the computer to analyse and understand the procedure. They will be notified of the findings after the algorithm has searched for a suitable candidate. This research utilises two algorithms—the RF algorithm and the LR algorithm—to forecast whether or not consumers will be approved for loans. Two methods will be tested on the same dataset to see which one achieves the highest level of accuracy while deploying the model.

This study (Annisa and Rusdah, 2022) construct a credit quality prediction model for first-stage credit applications based on 12 customer factors. Loans that are expected to perform well are referred to as performance loans (PL). The BPR Nusumma Group utilised data from 3338 clients between 2017 and 2020 for the non-performing loan projection. When evaluating credit applications from potential customers, this study's findings might serve as a useful management tool. Suppressing the increase of customers in the non-performing loan category is an advantage. DL algorithms, DT, ANNs, naive Bayes, and random forests were all compared during the modelling phase. Due to its superior performance compared to single classifier methods and ability to handle unbalanced datasets, the Random forest was chosen for the prediction model.

In (Sindhuraj and Patrick, 2023) Loan eligibility prediction using a deep neurofuzzy network is achieved by the use of the suggested optimisation technique for socially adaptive border collies. Due to its integration of Social Ski-driver (SSD), Border Collie Optimisation (BCO), and the adaptive approach, Adaptive SBCO outperforms all other methods. The proposed technique outperforms the existing one, with maximum sensitivity of 95.4%, specificity of 97.3%, and accuracy of 95%.

In this study (Zhang, Wang and Liu, 2023) An innovative model for profit-driven forecasting is put forward, which optimises the hyperparameters of the predictor-categorical boosting utilising a profit indicator as the goal of the Bayesian optimisation. An improved comprehension of the input variables' relationships to the projected values may be achieved by computing the SHAP value. Experimental findings and statistical testing performed on two datasets by Renrendai & Lending Club indicate that the suggested method performs better than the competitors with respect to review metrics relevant to profits. The two datasets show mean profit values of 5168.8762 and 352.9787, and mean extra profit rates of 3.0872% and 2.1858%, respectively.

Table 2.1: comparative analysis of related work on bank loan prediction

Author	Methods	Dataset	Key Findings
(Morad i and Mokhatab Rafiei, 2019)	Fuzzy Logic	Bank customers' actions in response to unique economic and political conditions	There should be certain qualitative variables included to the risk analysis, such as responsibility, dedication, honesty, standing, and morality.
(Goh et al., 2020)	Hybrid Model (HS-SVM and HS-RF)	German and Australian datasets that may be accessed openly via the UCI repository (https://archive.ics.uci.edu/).	Equivalent outcomes for credit scoring may be achieved using a Modified Harmony Search (MHS) model.
(Muha ngi, 2017)	Linear Regression	Fort Portal municipality, Western Uganda is home to six microfinance institutions and forty-eight loan officers and six credit managers.	Examining the difficulties encountered by credit officers throughout the loan evaluation process

(Bennouna and Tkiouat, 2018)	Fuzzy Logic	Client behaviour of microfinance institutions over time (descriptive, behavioural, and loan-specific variables)	A fuzzy logic technique to evaluating customer behaviour with the goal of lowering default rates on loans
(Islam, Arifuzzaman and Islam, 2019)	SMOTE, various classifiers (MLP, DT, RF, LR)	Bank Marketing data	SMOTE approach achieved 90.81% accuracy for Rotation Forest DT model. Various classifiers reached up to 95.3% accuracy.
(Papouskova and Hajek, 2019)	Ensemble machine learning techniques	P2P lending datasets	Focus on P2P lending platforms. Ensemble techniques used.
(Ruangthong and Jaiyen, 2015)	ANN, SVM, DT, logistic regression	Bank Marketing	Best results: ANN model achieved 79% AUC and 67.2% ALIFT. Valuable insights for successful long-term deposit marketing campaigns.
(Anicetto, Barboza and Kimura, 2020)	The AdaBoost & RF models are compared to an LR model-based benchmark.	A database containing the loan amounts and payback dates of 124,624 customers from a major Brazilian bank	The ML models that outperform the competition when it comes to classifying borrower adequacy are Random Forest and AdaBoost.
(Boughaci and Alkhawaldeh, 2018)	When selecting features, use LS, SLS, and VNS in conjunction with an SVM classifier.	German and Australian credit datasets	To evaluate the feature selection-based method's efficacy in conjunction with other machine learning approaches to credit scoring, further study is needed.

(Ozgur, Karagol and Ozbugday, 2021)	Evaluation of six ML methods in comparison to the gold standard, Linear Regression, including Tree Regression, Bagging, Boosting, Random Forest, Extra-Trees, and Xgboost	For the period 2002Q4-2019Q2, the data set includes nine variables particular to the bank, seven indicators of the economy as a whole, and three external factors that can affect the bank's lending practices. The data set covers 19 deposit banks in Turkey.	An outcomes show that a RF model achieves a best accuracy in its predictions.
(Assef and Steiner, 2020)	ANN-MLP, LR and SVM	543 different businesses (2600 customers who have not defaulted, 1551 who have, and 1281 who have just temporarily defaulted)	For more accurate outcomes, credit risk evaluations could use hybrid approaches.

The study (Owusu *et al.*, 2022) explores the issue of loan default in P2P lending on the internet by presenting a methodology to determine if a loan instance is in default or completely paid. An ADASYN method is utilized to oversample a minority class in order to rectify the data imbalance in a loan default dataset that was taken by Kaggle. The training and validation demands are met by use of a DNN. A 94.1% forecast accuracy was achieved. After doing several trials with varying batch sizes and time intervals, this performance consistently yields the best outcomes. Outcomes show that a method is really rather promising.

In (Awodele *et al.*, 2022) it was said that a 9% improvement in accuracy may be achieved by cascading Deep Learning networks with Support Vector Machines. In this research, an authors employed DNN to convert input data from a lower dimension to output features in a higher dimension, which wasfurther used to upskill a support vector machine-based classification model.

This work (Natasha, Prastyo and Suhartono, 2019) reduce the possibility of default by classifying customer risk. Parametric approaches like Discriminant Analysis and Binary LR have been used by the banking industry for credit evaluation in recent decades. Neural networks and SVM are two examples of non-parametric ML methods that have evolved in the last 20 years. In the last few decades, the banking sector has relied on parametric methods for credit appraisal, such as Binary LR and Discriminant Analysis. Recent 20 years have seen the development of many non-parametric ML approaches, such as neural networks and SVMs. The purpose of this research is to evaluate and contrast several approaches to consumer loan classification using parametric statistics and non-parametric machine learning. Using a DNN with 10 neurons in h1 and 3 neurons in h2 and an AUC of 0.638 on the test dataset is the best way to categorise consumer loans.

This study (Arutjothi and Senthamarai, 2017) The purpose of design is to forecast which clients will pay back loans, since lenders must foresee the risk that borrowers won't be able to return the money. Among the three models studied, rating-based logistic regression outperformed the other two: random forests and decision trees. Anyone looking for poor credit is turned down, perhaps because they can just choose not to pay. Most of the time, a decrease that can pay back the loan is available to high-value applicants. It would imply that the corporation has no tolerance for certain marital statuses or sexual orientations.

Machine learning

Data access and autonomous learning are key features of contemporary AI applications like ML, which aim to automate formerly human-intensive processes. Learning is a key characteristic of AI. The term "ML" describes a method by which computers may acquire new skills and knowledge in real time. Training algorithms to improve accuracy is the main emphasis of ML, a subfield of AI. Big data, analysis, data science, and AI are frequently used interchangeably. After effectively handle such massive volumes of data, worldwide enterprises are turning to ML as a solution. In ML, the objective is to train computers to analyse data better on their own. Even after poring over the numbers, there are instances when we still don't get it. Then, we apply ML to the situation (Richert and Coelho, 2015).

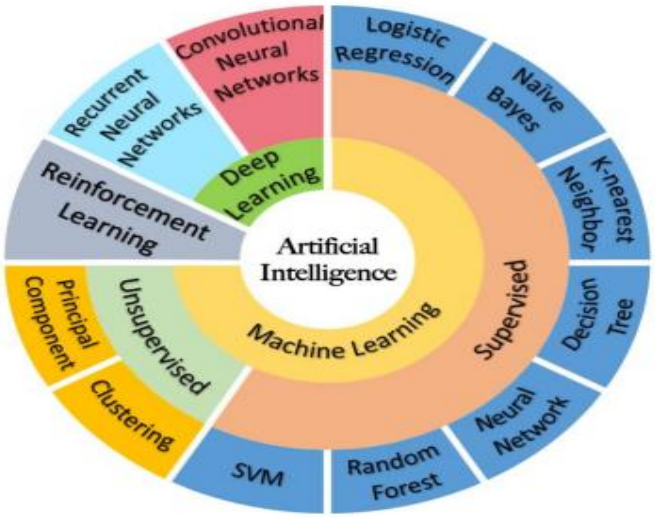


Figure 1.2: Machine learning

There are several subcategories of machine learning challenges. Classifying ML tasks according to their learning methods is the most common practice. Here they will go over the most common ways of learning. There is a mathematical representation, an example, and an explanation of each category.

Supervised learning

One approach is supervised learning, which involves teaching a model how to transform inputs into outputs given a set of known values for both the inputs and the outputs (Ethem, 2015). Inferring a function from labelled training data for use in classifying future unlabeled data is one way to characterise this problem. This usually implies that a training set is provided in a supervised learning situation. Each of the many examples in this collection has a class and a set of characteristics that have values.

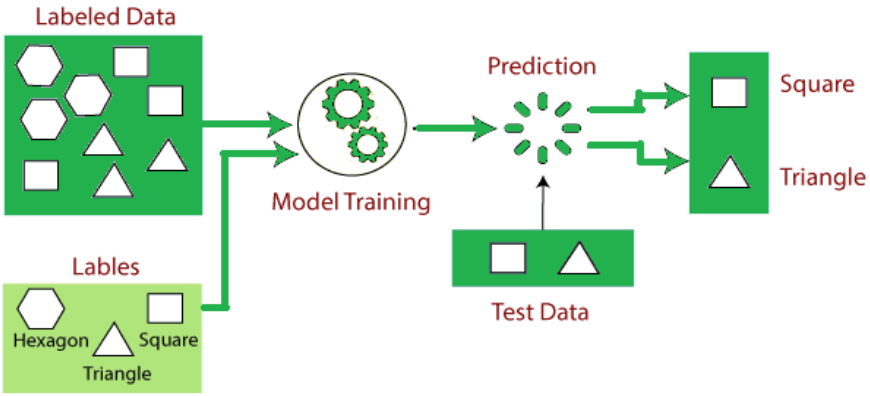


Figure 1.3: Supervised learning

Unsupervised learning

This technique involves training the algorithm with just an input set; it does not provide any output, intended outcomes, or feedback. By itself, the algorithm must discover data structure. Understanding how to distinguish between organised and unstructured noise is one way to look at unsupervised learning (Ghahramani, 2004). The use of unsupervised learning to the

problem of behavior-based network security detection shows great promise. No human being could ever hope to process all of the data produced. A data anomaly might be detected by an unsupervised learning algorithm even if the algorithm has never seen a breach before. Alerting IT security might be possible in the event that such an abnormality is identified.

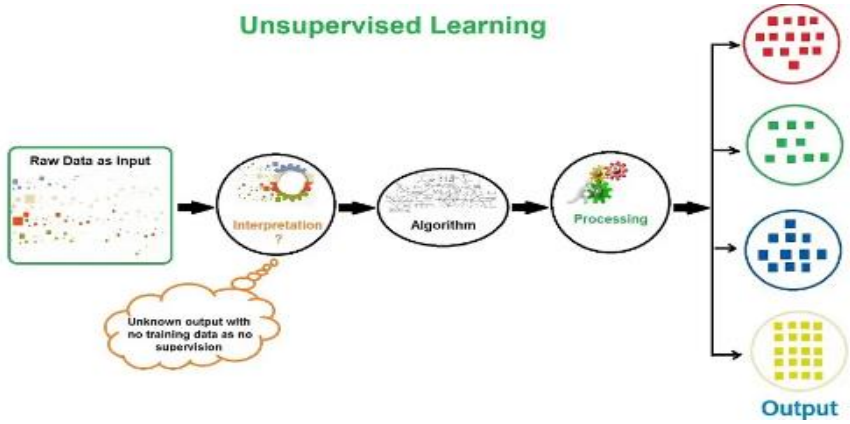


Figure 1.4: Unsupervised learning

Semi-supervised learning

The classification issue, in which only some observations have matching class labels, is taken into account in semi-supervised learning (Kingma *et al.*, 2014). This section will delve into unsupervised learning, which is a kind of learning algorithm that falls somewhere in the middle of supervised and unsupervised learning. Unlabeled data may seem to have little value at first, and one would assume that the task at hand is supervised learning and treat the data without labels as such.

Reinforcement learning

At beginning, a reinforcement-learning system does not have any labels. A system like this learns from its interactions with the environment by making changes to the environment. A positive or negative impact on the environment is the result. The algorithm's objective is to generate actions that maximise rewards or minimise punishments. Cars that can navigate themselves have recently been in the spotlight. Reinforcement learning may teach a vehicle to drive on such a path safely, without breaking any rules or colliding with other vehicles.

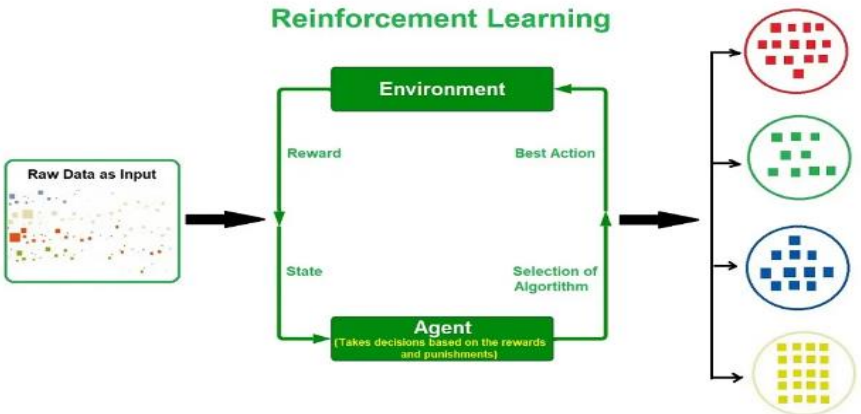


Figure 1.5: Reinforcement learning

CHAPTER- 3

SIMULATION TOOL

This chapter provide our simulation platform in this chapter. Python programming is the platform utilized for simulation. As an editor, Jupyter Notebook works. These are explained in detail below:

Hardware Tool

Research was carried out on many current hardware tools, such as:

- **Computer:** For this implementation a high-performance computer system is required to train the model smoothly and efficiently.
- **CPU:** To manage the computational load, use an advanced multi-core CPU (e.g., Intel Core i5 or above).
- **GPU:** Graphics processing units (GPUs) with CUDA technology integrated, such the GeForce GTX or RTX series from NVIDIA, that improve the speed of DL model training and inference.
- **RAM:** Sufficient RAM, such as 8 GB or more, is required to facilitate the training process of deep learning models and large datasets.
- **Storage:** Space for the dataset, model checkpoints, and intermediate results in performance must be sufficient.
- **Operating System:** The required software and libraries may be found in any common operating system, including Windows, macOS, or Linux.

Simulation Tool (Python)

The programming language known as Python is interpreted, general-purpose, high-level, and dynamic. Application development using the Object-Oriented programming (OOP) methodology is supported. With its abundance of high-level data structures and ease of learning, it is a great choice. Since it is both powerful and simple to learn, Python has become a popular option for app developers. Python is a great language for scripting and application development because of its syntax, dynamic typing, and interpreted nature. You may use Python with whatever programming style you choose, whether it is object-oriented, imperative, functional, or procedural. Python was never meant to be used for web programming or any other specific purpose. Its versatility in working with online, enterprise, 3D CAD, and other platforms is what gives it its name: multipurpose programming language. Because it is dynamically typed, we may assign a value of 10 to an integer variable without using data types to specify it. Due to its short edit-test-debug cycle and lack of a compilation phase, Python development and debugging are both accelerated. Because of its versatility, Python is used in nearly every area of software development. Python has already established a foothold in every new industry. This programming language is quite popular and has the ability to develop a huge range of applications (Butwall, Ranka and Shah, 2019)(Korkmaz, Sahingoz and DIri, 2020).

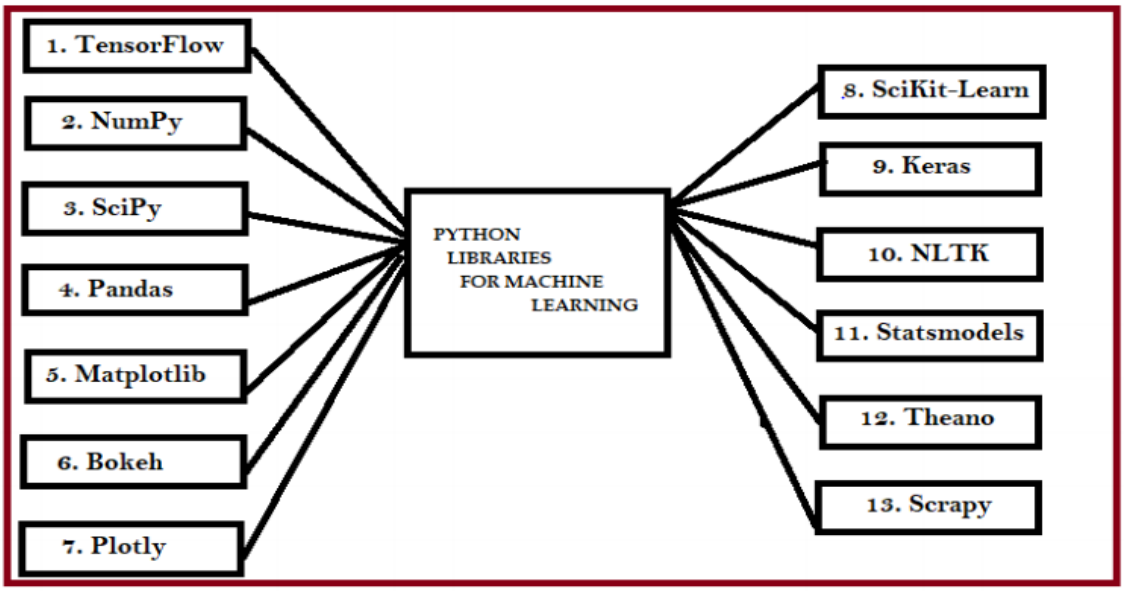


Figure 3.1: Python Libraries for Machine Learning

Python for Data Analysis

Python is an excellent choice for data analytics due to its many useful properties, such as its huge library, ease of learning, robustness, readability, scalability, integration with other languages, and active community and support system. Figure 5.1 shows a selection of top Python ML libraries; this section delves into a few of them. Python modules for statistical computation (Kumar, 2018)(P.Lee, 2017):

- **NumPy-** One such Python module is NumPy. Python, as in numerically. It is a library that includes array processing algorithms and multidimensional array objects. The mathematical software that forms the basis of the scientific computing stack is NumPy, which stands for Mathematical Python. A plethora of helpful features are provided for Python n-array and matrix operations. Using the library's vectoring functionality, mathematical operations on the NumPy array type may be executed more quickly and efficiently.
- **Pandas -** Pandas is an easy-to-use Python library for working with "labelled" and "relational" data. Pandas is an open-source Python package that offers a powerful data analysis and manipulation tool using free and open-source software. One econometric approach to multidimensional data is known as "panel data," and the term "pandas" describes it. Pandas was created in 2008 by developer Wes McKinney in response to a demand for a versatile, high-performance tool for data processing. Python was mostly utilised for data munging and preprocessing before Pandas. In terms of data analysis, it was mostly ineffective. Pandas figured it out. Pandas allows us to do the five standard processes of data processing and analysis: load, prepare, manipulate, model, and analyse. These procedures apply regardless of the data's origin.
- **Matplotlib-** It is easy to perform both simple and complex visualisations using Matplotlib, a Python library and another essential module of the SciPy stack. This fantastic programme, together with NumPy, SciPy, and Pandas, is helping to propel Python into the conversation as a competing scientific tool with MatLab and Mathematica. Using this library will need more code and work than utilising more high-level tools, but the time and effort is worth it in the end because of the complex visualizations you can do.
- **SciKit-Learn-** Image processing and ML are two examples of the many functions that may be enhanced with the help of Scikits, which are supplementary packages for the SciPy Stack. As for the second, scikit-learn is among the most well-known of these programmes. The math operations of SciPy are heavily utilised by the package, which is constructed on top of it. It is easy to integrate ML into production systems using scikit-learn because it provides a uniform and succinct interface to the most popular ML techniques. The de facto standard for Python ML, the library combines excellent performance, extensive documentation, ease of use, and quality code.
- **Plotly-** The web-based toolkit exposes APIs to several programming languages, including Python, and allows users to construct visualisations. The plot.ly website offers a variety of powerful, pre-made images. You have to configure your API key before you can use Plotly. There is a workaround, though, because the graphics will be processed on the server and then uploaded to the internet.

Jupyter Notebook simulation platform

In 2014, I Python gave birth to Jupyter Notebook. You may create and edit notebook documents, or "notebooks" for short, using this online programme that is built on the server-client architecture. As an integrated development environment (IDE), presentation tool, and data science environment, Jupyter Notebook is accessible to users of various programming languages. For individuals who are new to data science, it is an ideal beginning point. Markdowns are compatible with the Jupyter Notebook, so you may include HTML elements into media files like photos and movies. Matplotlib and Seaborn are data visualisation packages that we may employ to display graphs with our code in the same page. In addition to all of this, you have the option to export your finished work as a.py file, as well as to PDF and HTML files. Along with that, you may use your notebooks to do presentations and blogs (Butwall, Ranka and Shah, 2019).

The Jupyter Notebook is a widely used application for creating and observing interactive data science projects. Code or its output, together with images, writing, mathematics, and other rich material, can all be found in a single notebook page. Notebooks are becoming more and more of a contender for the core of contemporary information science, technology, and innovation thanks to

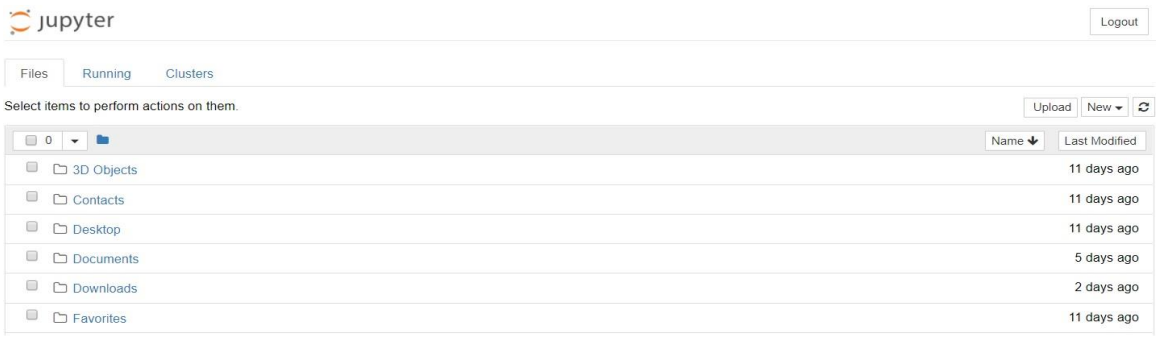
instinctive workflow, which promotes iterative yet quick setup. Jupyter is an open-source project, and the worst part is that it is free.

The Jupyter project was born out of the 2010 reference release of I Python Notebook, which it now follows. While the aforementioned work makes extensive use of Python, Jupyter Notebooks are compatible with a number of other computer languages as well.

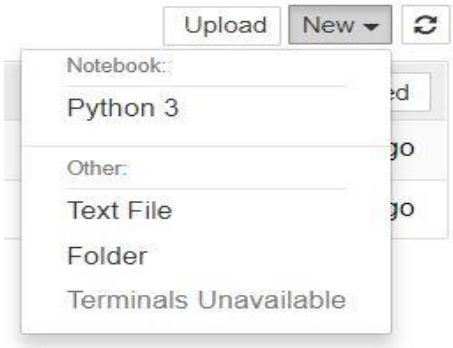
Pip is another option, but it should only be used by people who are very familiar with Python or who want to handle their packages manually.

Running Jupyter

To launch Jupyter on Windows, you may use the Anaconda shortcut in the start menu. This will open your usual browser in a new tab that mimics a screenshot.



A dashboard for your Jupyter Notebooks. This is tailor-made for you. Here is where you may start creating your journals, changing existing ones, or searching. The "New" option is on the top right of the page; choose "Python 3" from there.



See Untitled. The notebook will start running as soon as you see the green word when you return to the dashboard, Ipynb&.

The Notebook Interface

Jupyter is actually an expert word processor; I hope the interface does not seem too strange given that you have an open notebook in front of you. Why are not you looking at it? The list of instructions indicated by the small button with either symbol will appear after a few minutes (or when you press Ctrl + Shift + P), so be sure to check the menus.

Jupyter is similar to a word processor in that it helps you comprehend and work with cells and kernels, two terms that are likely new to you. Thankfully, these concepts are not hard to grasp.

- " There is a kernel called a "programme processor" that runs the code in a notebook.
- The transcript or code executed by the notebook kernel is shown in a cell.

Cells

Making ensuring cells are under control is a prerequisite before returning to kernels. A notebook's cells make up its main body. In the notebook screenshot up there, you can see an empty cell with a green contour and a box within. We address two main categories of cells:

- A code cell consists of code that runs in the kernel or produces the following output.
- The output of a Markdown cell is instantaneously shown together with the Markdown formatting.

The very first cell in any up-to-date notebook will always be a code cell. Using the age-old greeting as an example, let us give it a go.

world. Type print('Hello World!') Tap on the run button into cell 

in a toolbar above or press Ctrl + Enter. This must be an outcome:

The output of a running cell is seen below, and a label on the left has been updated from In[] to In[1]. This item displays the code cell output with the text. Ever since markdown cells do not have a label on the right side, it is possible to tell them apart from code cells. The "in" part of the label stands for "data," and the first cell is considered to be operating when the label no. is on a kernel. Now that the cell is the second most important part of the kernel, execute it and change its label to in [2]. We will go further into kernels to uncover their benefits. To create your first new code cell, follow these steps.

Although it operates for three seconds, this cell does not provide any output at all. Just in case you forgot, Jupyter indicates that a cell is now running once its label is changed [*].

For instance, the output of a cell type is determined by the value of the final line in the cell, which may be an attribute value, a function call, or something else completely, as well as all the text information explicitly provided during cell execution (How to Use Jupyter Notebook in 2020: A Beginner's Tutorial, 2020):

Datatype Conversion

It is common to need to convert between built-in formats. To distinguish between distinct categories, you may only utilise their names. There are multiple built-in functions for transferring between data types. These functions restore a new object that represents value converted. The few are listed below:

Table 3.1: Functions & Description

S.no.	Functions & Descriptions
1.	**int (x [, base])**; Conversion of x to integer. basis stipulates that x is number.
2.	**float(x)**; Transforms x into set of floating points.
3.	**Complex (real [, image])**; Produces an interesting no.
4.	**Str (x)**; Conversion of object x to a string file.
5.	**Repr (x)**; Turns item x to a string expression.
6.	**Eval (str)**; Returns an object to evaluate a string.
7.	**Tuple (s)**; Turns s into tuple.
8.	**List (s)**; Turns s into list.
9.	**Set (s)**; Turns s into set
10.	**Dict (d)**; Make dictionary. d shall be tuple (key, value) series (Python 3 - Variable Types, no date).

CHAPTER- 4

RESEARCH METHODOLOGY

Chapter 4 presents the research approach that we used for bank loan eligibility prediction. It covers the information about the dataset and how can it be preprocessed. It also presents the proposed algorithm and proposed flowchart.

Problem Statement

With the goal of improving accuracy and reducing fraud, we will develop and deploy a system that uses ML to forecast the user's likelihood of receiving a loan from a bank. All around the world, financial institutions and home finance companies facilitate a wide range of loan products, such as personal loans, mortgages, business loans, and more. These companies are located in urban, semi-urban, and rural areas. When a customer asks for a loan, these companies verify their eligibility. This research provides a method to automate the process by use of an ML approach. After that, they need to go online and apply for a loan. The applicant's marital status, credentials, dependents, yearly income, loan amount, and credit history are among the many topics covered by this paperwork. Using a machine learning system to make this process automatic.

Proposed Methodology

This section describes the method and algorithms that were employed throughout the implementation. Using a unique structure based on ML techniques, this research analyses the prediction of bank loan eligibility. The jupyter notebook simulation tool and the Python programming language were utilised for the study strategy. Make use of sklearn, Pandas, NumPy, seaborn, matplotlib, and warning, among other Python modules, in this investigation. The research approach is completed in various steps and phases. The first step is data collection, a publicly available Loan Eligible Dataset was used. This dataset is obtained from the Kaggle open source. The dataset needs to be analyzed comprehensively for filling NaN value and outlier detection. As part of the preparation step, you may assist the data interpretation process by visualising the data. the collected data was highly imbalanced,

handling imbalanced classification problems using oversampling techniques like (SMOTE). For the intent of data normalization of transform features apply a min-max scaler. After data collection and preprocessing, data need to be split for the training and testing subsets. The purpose of testing data is to assess the performance of trained models. Then, apply classification techniques like XGBoost, Gradient Boosting, and Cat Boost classifiers and for model building. The last thing to do is to test how well the classification algorithms worked using several performance matrices including F-measure (F1-measure), recall, accuracy, and precision. Input values are calculated using the scikitlearn package in Python; these values then form the entities of the confusion matrix. When we have finished all the necessary procedures, our categorization systems will work much better. The proposed flowchart shows the all steps that are followed:

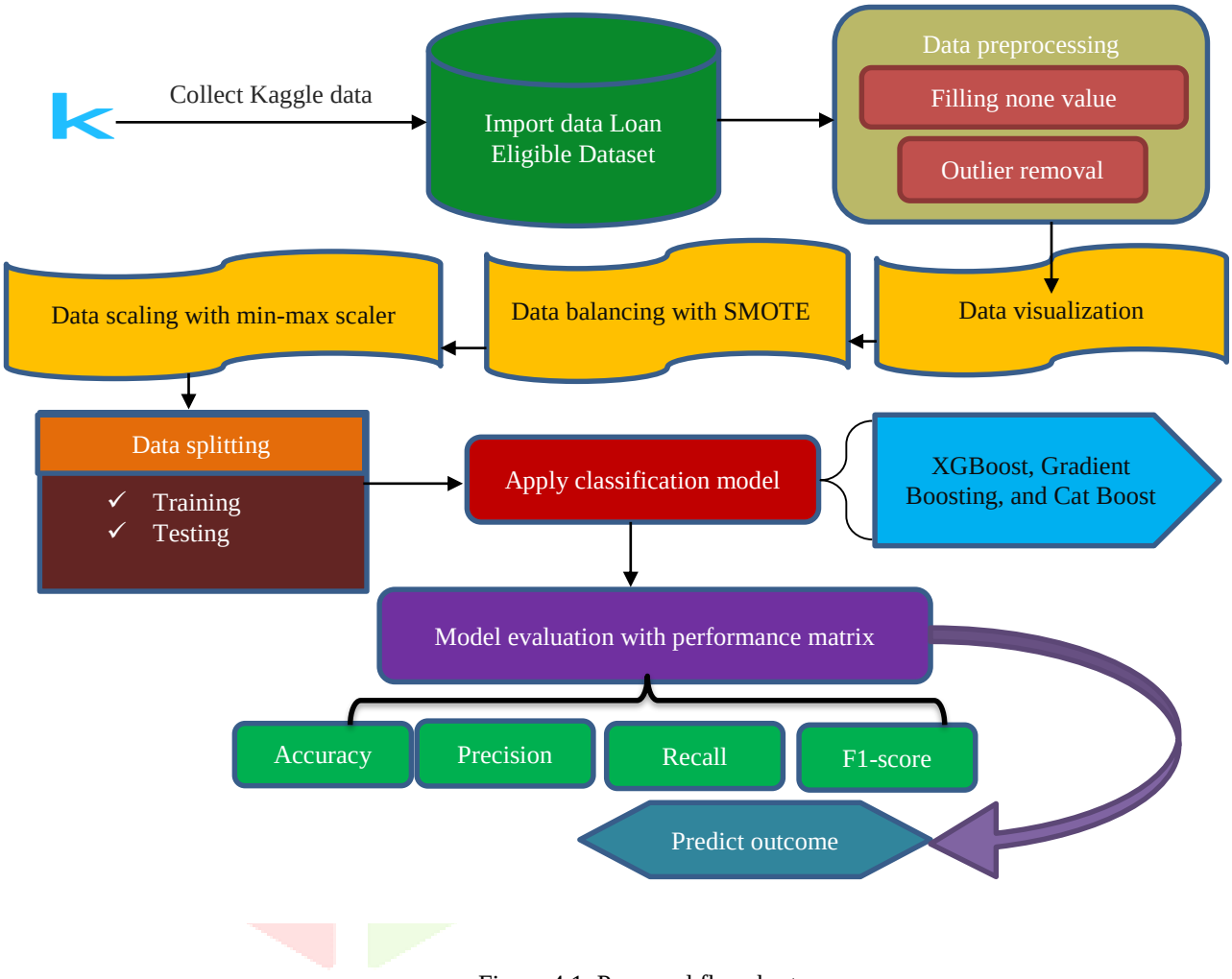


Figure 4.1: Proposed flowchart

Data collection

Collecting raw data from various sources is known as data collecting. In this research, collect a Loan Eligible Dataset[1] from the Kaggle web page. List of columns in this dataset: CoapplicantIncome, Gender, Loan_ID, Dependents, Self-employed, Education, Applicant Income, Loan Amount Term, Loan Amount, Married, Property Area, Credit History, and Loan Status.

Data Preprocessing

One way to prepare data for use by an ML model is through data preprocessing, which involves cleaning and organizing the raw data. There can be inconsistencies due to missing values in the obtained data. To improve the method's performance, data must be preprocessed. To prepare the dataset for analysis, performed several preprocessing steps that are as follows:

- **Filling Nan values:** The Fillna function in Pandas may be used to fill in the NA/Nan values according to the given manner. We use the Nan object to indicate missing data in the Pandas Data Frame.
- **Outlier removal:** For high-dimensional datasets to produce reliable results, data quality is paramount. One way to ensure that datasets are of high quality is with the help of outliers. The traditional method of identifying outliers

[1] https://www.kaggle.com/datasets/vikasukani/loan-eligible-dataset

disregards the dataset's data production process and does not take distribution tails into account (Paulheim and Meusel, 2015).

Data Visualization

It is recommended to do a high-level analysis of each attribute pattern and analyse the results using data visualisation. We can see from the data visualisation that we need to check a few attributes to make sure the model categorization is accurate. Data visualisation refers to the practice of visually representing information and data. Data visualisation tools make it easy to see patterns, trends, and outliers in large amounts of data via the use of visual components such as maps, graphs, and charts.

Data balancing with SMOTE

A credit card transaction dataset typically includes a data unbalanced problem, which might mislead the detection procedure to the misclassifying problem. Applying Random Under-Sampling SMOTE methods can help you get past the class imbalance issue. (Yen and Lee, 2009). When optimising for minority cases, SMOTE takes their K Nearest neighbours into account. To increase the quantity of occurrences in training data uniformly, one may use this statistical strategy.

A more balanced dataset will improve the minority class's forecasting accuracy. To address the issue of class imbalance, SMOTE is employed. Using operations in feature space, SMOTE generates synthetic samples that reflect a class that is underrepresented. All of the minority class samples, along with their KNNs, are aligned with the manufactured samples. If you want to make a fake instance of a minority class, you have to take the feature vector and compute the difference between it and its closest neighbour. Then, multiply that result by a random integer between zero and one. A simulated member of the minority group is generated by multiplying the result with the feature vector under consideration (Zulfiker *et al.*, 2021). Take into account that Zi is one of the KNNs of Zi and that fn is the feature vector of the minority class sample that is being considered. Equation (4.1) may be used to represent the synthetic sample that was created, $Znew$.

$$f_{new} = f_i + (f_i - f_{near}) \times R \dots \dots . (4.1)$$

Here, R is a random number between 0 and 1.

Data scaling (MinMaxScaler)

As a preprocessing step, data normalisation comprises scaling or otherwise modifying the data to guarantee that each attribute contributes at least as much as the others. Data redundancy is eliminated by the process of normalisation, which involves grouping data into numerous linked tables. If the data utilised to construct a thorough statistical model for the classification issue is inaccurate, then ML techniques will not work (Singh and Singh, 2020). Since the MinMaxScaler estimator is among the most used methods for this part, it was utilised to resolve the issue. To fit the training set's original values within the specified range (i.e., such that all values fall between the zero and one range), this approach scales and transforms each feature separately. Equation (4.2) shows the calculation for this operation:

$$X_{Scaler} = \frac{X_{Std}}{(max - \min) + min} \dots . (4.2)$$

Data splitting

Data splitting is an essential ML technique that divides a dataset into many subsets for different tasks including training, evaluating models, and tweaking hyperparameters. Create a "training" and "testing" set of data to use in the analysis.

Classification models

For the classification techniques, use ML classifier. ML is a method that enables frameworks to acquire new knowledge autonomously, without the intervention of a human specialist. Artificial intelligence (AI) has mostly focused on developing programmable computers with the ability to learn from their own experiences. On the other hand, ML is concerned with techniques that let machines learn from their information. Classifications of ML methods are based on the outcomes that are expected from them (Endut et al., 2022). In this study, use XGBoost, Gradient Boosting, and Cat Boost that describe in below:

XGBoost Classifier

As an algorithm, XGBoost was suggested by (Chen and Guestrin, 2016) based on the GBDT framework. To avoid overfitting, XGBoost alters the objective function by adding a regularisation term, unlike the GBDT. We construct the objective function as:

$$O = \sum_{i=1}^{n} L(y_i, F(x_i)) + \sum_{k=1}^{t} R(f_k) + C \ldots . (4.3)$$

The regularisation term at the iteration time k is denoted by $R(fk)$, and the constant term **O** can be optionally removed. The word for regularisation $R(fk)$ is written as:

$$R(f_k) = \alpha H + \frac{1}{2}\sum_{j=1}^{H} w_j^2 \ldots . (4.4)$$

α stands for a complexity of a leaves, **b** for a number of leaves, η for a penalty parameter, and wa for an output outcome of every leaf node. More specifically, the projected categories according to the classification criteria are represented by the leaves, and the node that cannot be severed from the tree is called the leaf node. Furthermore, although GBDT employs a first-order derivative for its goal functions, XGBoost employs a second-order Taylor series. We may get the objective function by assuming that a loss function is the MSE, which leads us to:

$$O = \sum_{i=1}^{n} [p_i w_{q(x_t)} + \frac{1}{2}(q_i w_{q(xt)}^2)] + \alpha H + \frac{1}{2}\eta \sum_{j=1}^{H} w_j^2 \ldots . (4.5)$$

A loss function's first derivative is represented by ci and its second derivative by hi. The notation "$c(xi)$" is used to represent a function that puts data points into the correct leaves. After adding together all the loss values, we get the ultimate loss value. The sum of the loss values at each sample node in the DT (the "leaf nodes") gives the overall loss value. Hence, another way to express the objective function is:

$$O = \sum_{j=1}^{T} [P_j w_j + \frac{1}{2}(Q_j + \eta) w_j^2] + \alpha H \ldots . (4.6)$$

The given expressions denote all samples in the leaf node a, where Pa=$\sum i \in Jbki$ and Qb=$\sum i \in \mathbf{O}bci$. Finally, finding the minimum of a quadratic function becomes the optimisation issue of the objective function. Moreover, XGBoost is now more resistant to overfitting because to the inclusion of the regularisation term (Liang *et al.*, 2020).

Cat Boost Classifier

The term Cat Boost is derived from the words "Categorical" and "Boosting" in combination. That open-source ML method that Yandex developed is used by popular languages such as Python and R (Hancock and Khoshgoftaar, 2020). The Cat Boost framework is an example of a GBDT, which primarily learns from symmetric decision trees. In addition to being generally efficient and accurate in processing class-type data, it has fewer parameters and supports class variables. Not only that, it lessens the likelihood of overfitting by fixing gradient bias and prediction shift (Hancock and Khoshgoftaar, 2020). As a criteria for node splitting in a DT, a label means (also called greedy target variable statistics) are represented as:

$$\widetilde{x_k^i} = \frac{\sum_{j=1}^{p-1} [x_{j,k} = x_{i,k}] \cdot Y_i}{\sum_{j=1}^{n} [x_{j,k} = x_{i,k}]} \quad \ldots . (4.7)$$

Feature data frequently contains more information than label data, which is the most obvious problem with this application. Features are made to be expressed as the average of the labels when the training and test datasets have distinct distributions and structures, and conditional bias develops. In order to enhance the performance of Greedy TS (Target-based Statistics), it is

customary to incorporate a priori distribution parameters into the formula. These parameters serve to alleviate the influence of noise and low-frequency category data that may affect the distribution of the data.

$$\widehat{x}_k^i = \frac{\sum_{j=1}^{p-1}\left[x_{\delta_{j,k}} = x_{\delta_{p,k}}\right]\cdot Y_{\delta_j} + a\cdot p}{\sum_{j=1}^{p-1}\left[x_{\delta_{j,k}} = x_{\delta_{p,k}}\right] + a} \quad \ldots. (4.8)$$

p stands for the additional previous term, and a is usually the weight coefficient that is bigger than 0. The antecedent term in classification problems denotes the probability of obtaining affirmative cases prior to the commencement of the experiment (Chang *et al.*, 2023). Furthermore, to enhance the expressiveness of the model, the algorithm will integrate category features into newly generated features automatically. Due to a classification nature of a blueberry ecological suitability dataset and the aforementioned benefits of the CatBoost algorithm, its application may learn more information to the fullest degree possible, allowing for even better model expression(Bentéjac, Csörgő and Martínez-Muñoz, 2021).

Gradient Boost Classifier (GBC)

One popular machine learning (ML) method is the Gradient Boosting Classifier (GBC), which is a member of the ensemble learning family. It is an effective prediction model that integrates boosting and gradient descent. Typically, DT is used by GBC to optimise an ensemble of basic prediction models via iterative processes. Alterations to the learner models and the loss function are both within the realm of possibility. Parameter predictions may present difficulties in implementation when a custom loss function (y,f) or base-learner h(x,f) is utilised. A novel function h(x,_t) was selected as the proposed solution, which has the following characteristics: -g-t (x_i) -N (i=1) -negative gradient

$$g_t(x) = E_y\left[\frac{\Psi(y,f(x))}{\partial f(x)}|x\right]_{f(x)=f^{t-1}(x)} \quad \ldots. (4.9)$$

It is desirable to choose a new function increment that is more closely connected to -g_t (x) as it provides a local solution to a boost increment in a function space. A less difficult least-squares minimization problem may therefore be used rather than the more difficult optimisation problem:

$$g_t(x) = E_y\left[\frac{\Psi(y,f(x))}{\partial f(x)}|x\right]_{f(x)=f^{t-1}(x)} \quad \ldots. (4.10)$$

The whole version of the original GB method, as published by Friedman (2001), is necessary, to put it simply. An individual's choice in constructing (y,f) and h(x,) will substantially affect the derived method's ultimate shape and all of the related equations (Natekin and Knoll, 2013).

Model evaluation

The procedures for assessing the efficacy of the models used to forecast loan risk should be detailed here. Confusion matrices and computing metrics like as recall, accuracy, precision, and F1-score are often used for performance evaluation in domains such as statistics, ML, data mining, and AI. An often-used method for assessing issues involving two classes is a confusion matrix (L.Gupta, K. Malviya and Singh, 2012). The research relied on the confusion matrix as an evaluation tool since it provides proof of the suggested prediction and classification model's actual and expected classification performance.

Proposed algorithm

This section presents the suggested method that was used for predicting eligibility for bank loans in this study.

Proposed Algorithm: bank loan eligibility prediction

Step 1: Install python simulation tool and jupyter notebook

- Pandas, matplotlib, NumPy, sklearn, and seaborn are some of the Python libraries that need be imported before implementation.

Step 2: Data Collection

- To collect a Loan Eligible Dataset from a Kaggle website.

Step 3: Data Preprocessing

- Preprocess the data for filling NaN value and outlier removal.

Step 4: Data balancing

- Use oversampling (SMOTE) techniques.

Step 5: Data normalization

- Apply min-max feature scaling technique.

Step 6: Data Splitting

- Training
- Testing

Step 8: classification models

- XGBoost, Gradient Boosting, and CatBoost

Step 9: Model Training

- To train the model on the preprocessed data

Step 10: Model Evaluation

- Accuracy
- Precision,
- recall
- F1-score

Step 8: predict outcome

Finish!!!!

The final outcome in the form of various figures, tables and graphs.

CHAPTER- 5

RESULTS ANALYSIS AND DISCUSSION

Chapter 5 presents the findings, analyses, and discussion of detecting phishing attacks via machine learning methods. The outcomes of a work are most effectively described by displaying them. Outcome precision is determined by the algorithms' accuracy, recall, precision, and F1_Score.

Dataset Description

The "Loan Eligible Dataset" comprises 613 rows and 12 columns, which were obtained from the Kaggle platform in support of a research investigation concerning the prediction of loan eligibility. Gender, Loan_ID, Dependents, Married Status, Self-Employment Status, Education Level, Applicant and Coapplicant Income, Credit History, Loan Amount Term, Loan Amount, and Property Area are among a critical element contained in the dataset. The dataset was obtained from Dream Housing Finance, an organization that provides residence loans across urban, semi-urban, and rural regions. The aim is to implement an automated system that determines loan eligibility in real-time, utilising consumer information submitted via online applications. Dream Housing Finance utilises various criteria, including credit history, education, gender, loan amount, education, marital status, and number of dependents, in order to ascertain which customer segments, qualify for particular loan amounts. By facilitating targeted consumer outreach, this automation will improve the overall efficacy of the loan approval process.

Exploratory Data Analysis (EDA)

EDA constitutes an initial and critical phase of any research investigation. The primary aim of performing exploratory analyses is to attain the greatest possible understanding of the data through the utilization of diverse methodologies (Lord, Qin and Geedipally, 2021). The primary objectives of exploratory data analyses are to identify outliers, discover variable patterns, test hypotheses, and evaluate data assumptions through the use of summary statistics and visual representations. Exploratory analyses leverage data visualization to unveil the hidden characteristics of the data and facilitate the identification of one or more models that will be employed for data analysis.

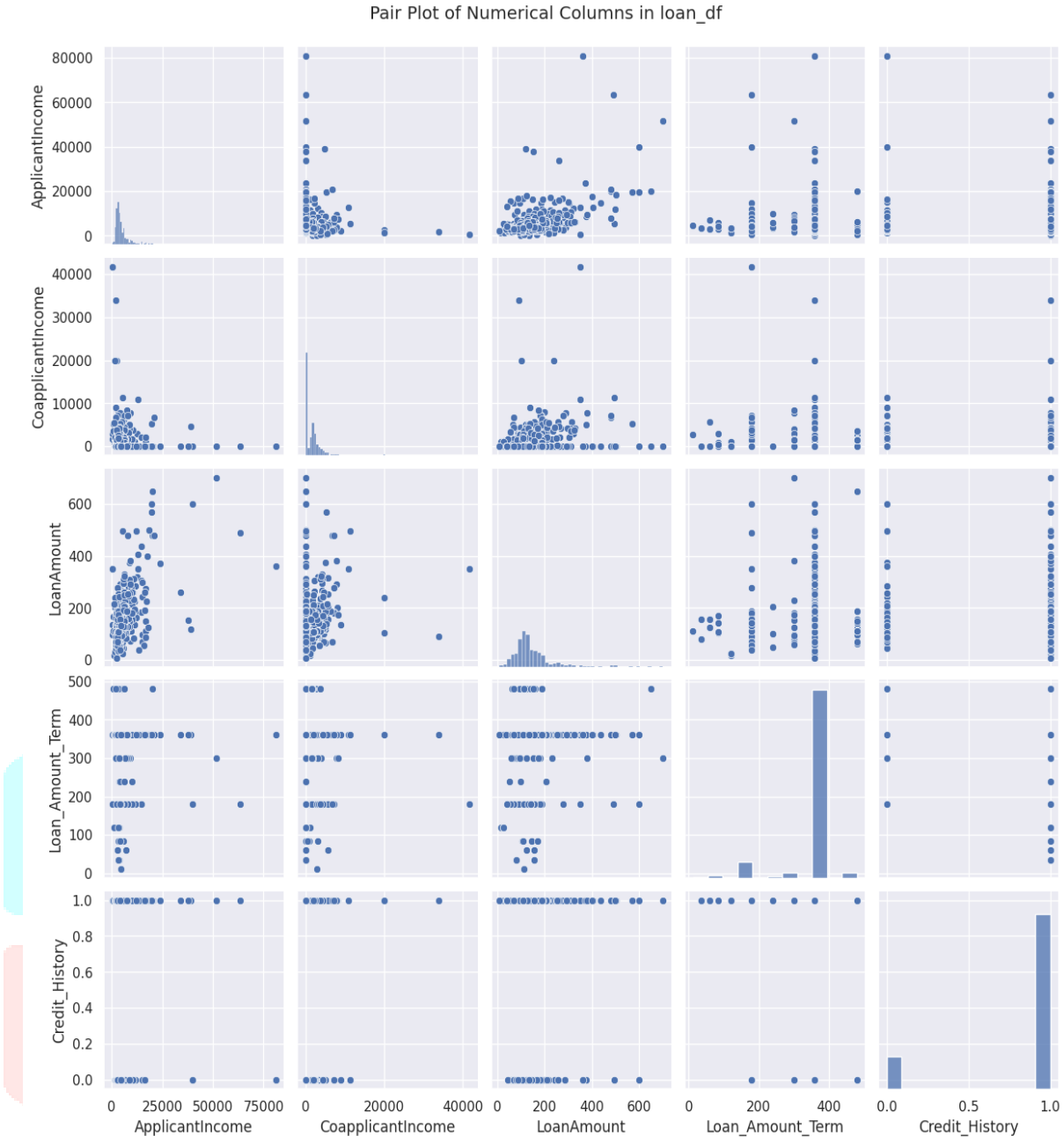


Figure 5.1: Pair Plot of the Loan Eligible Dataset's attributes

Figure 5.1 shows the association among the dataset's properties in a pair plot. Pair plots are used to visually represent the correlations between pairs of variables within a dataset. The scatterplots display trends, relationships, and outliers by representing each combination of characteristics. Pair plots are very useful for analysing datasets with several qualities, since they provide a succinct picture of the connections between the variables. It is a hybrid of a bar plot and scatter plot.

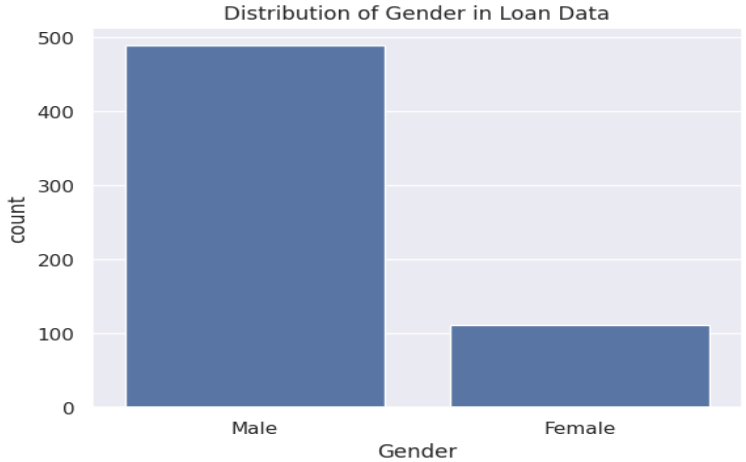


Figure 5.2: Count Plot for the Gender Value Distribution

Figure 5.2 illustrates a count plot representing the distribution of gender in the dataset. The x-axis of the plot indicates the gender, with two distinct classes: Male and Female. Meanwhile, a y-axis depicts count values ranging between 0 and 500. In a given figure, the Male class has a count over 400, while the Female class has a count exceeding 100, respectively.

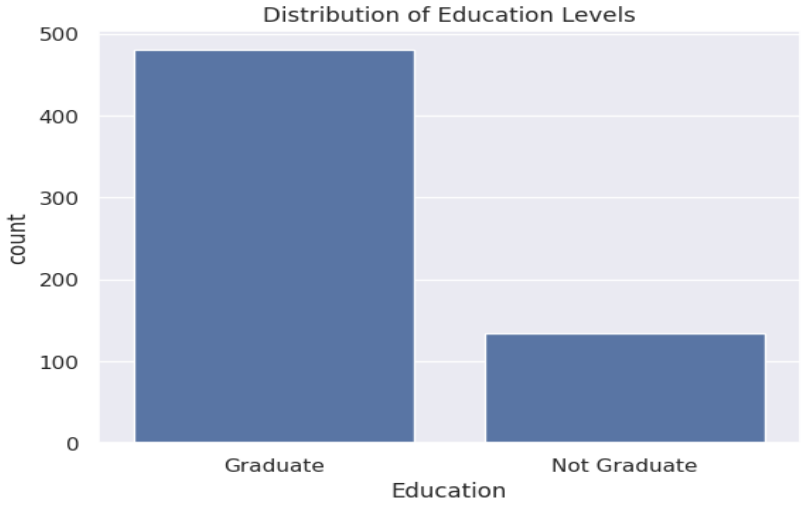


Figure 5.3: Count Plot for the Education levels

Figure 5.3 shows a count plot illustrating a distribution of an education levels in a dataset. A x-axis of a figure depicts a variable " Education," which is divided into two categories: "graduate" and "not graduate." This graph shows that the not graduate class has a count value over 200, indicating that there are more than 200 not graduate applicants. On the other hand, the graduate class has count values above 350, indicating that there are more than 350 graduate applications.

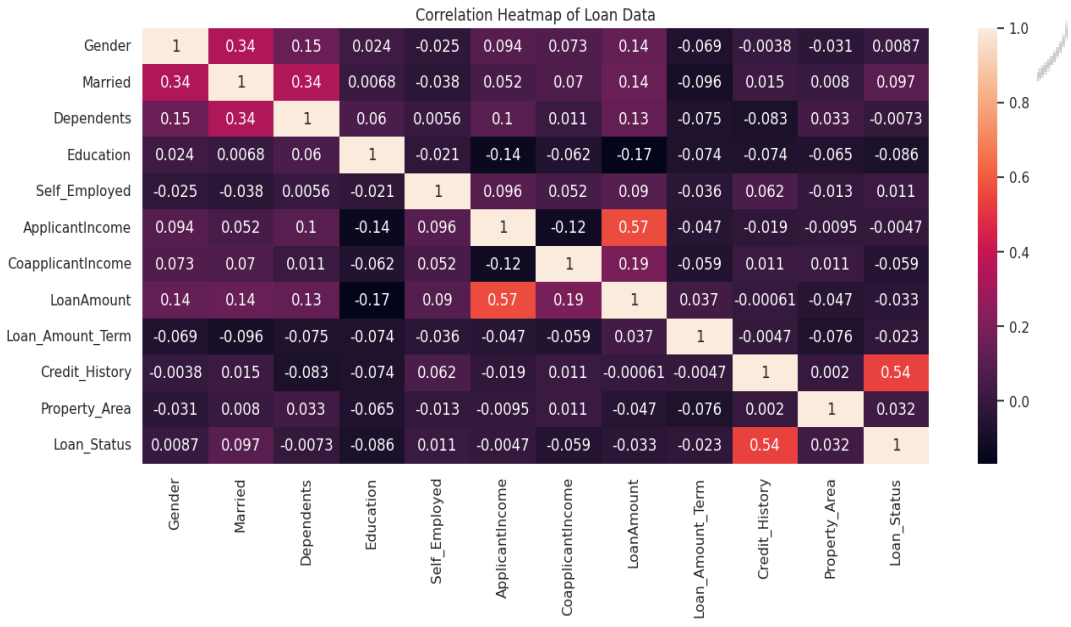


Figure 5.4: Correlation Heatmap Matrix for the Loan Eligible Dataset

Figure 5.4 illustrates the correlation heatmap matrix of the loan eligible dataset, illustrating the correlation between features in the dataset. The x-axis of the matrix reflects the features of the dataset, including Loan Amount, Loan Amount Team, Applicant Income, Coapplicant Income, and Credit History etc. Similarly, a y-axis likewise represents these same attributes.

Performance Parameters

The research study uses a wide range of performance measures to check how well the machine learning model predicts who will be eligible for loans. The confusion matrix is part of the test and gives information about true positives, true negatives, false positives, and false negatives. A model's overall correctness is shown by the accuracy score, its ability to avoid false positives by the precision score, its ability to capture genuine positives by the memory score, and its overall efficacy is blended by the F1-

score, which combines the recall and precision values. The ROC curve shows the trade-off between the TPR and the FPR, and it also provides a thorough description of the model's predicting skills. Detailed explanations of the performance metrics follow.

Confusion Matrix

One of the most often used techniques for validating an accuracy of a accomplished classification is a confusion matrix. Tables containing various combinations of projected and actual values for decision classes make up confusion matrices for binary decision classes. Predicted classes are in the columns, while current classes are in the rows. The properly identified samples for both classes are located on the diagonal of the confusion matrix, while the mistakes are represented by the cells that are off the diagonal (Kozak *et al.*, 2022). The kind and representation of the classifier's mistakes are shown in the confusion matrix. It is a thorough analysis of the responses taking into account the proportion of accurate and inaccurate classifications for each class. Figure 5.5 below shows a confusion matrix.

Total	Class 1 (Predicted)	Class 2 (Predicted)
Class 1 (Actual)	TP	FN
Class 2 (Actual)	FP	TN

Figure 5.5: Confusion Matrix

- **Accuracy:** An accuracy metric measures how often a classifier predicts outcomes correctly. The accuracy is determined using a formula that divides a total amount of predictions by the number of correct ones.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \dots\dots\dots\dots\dots (5.1)$$

- **Precision:** The accuracy metric measures the frequency at which predictions of outcomes are made by a classifier that are accurate. Dividing the sum of all forecasts by the number of accurate predictions yields the accuracy, according to a previously established formula.

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}} \dots\dots\dots\dots\dots\dots\dots\dots (5.2)$$

- **Recall:** Recall is a useful metric for gauging the classifier's accuracy in correctly identifying positive samples inside a given class. You may find the recall calculation algorithm down below.

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}} \dots\dots\dots\dots\dots\dots\dots\dots (5.3)$$

- **F1-Score:** "The $F1$ measure" is the name given to the harmonic mean plus accuracy. The formula is given below for calculate f1-score. The formula for calculating the F1-score is provided below.

$$F1 - Score = \frac{2(Precision * Recall)}{\text{Precision} + \text{Recall}} \dots\dots\dots\dots (5.4)$$

Experimental Analysis

The research study utilises an advanced method for predicting loan eligibility, including powerful ML methods such as Gradient boosting, CatBoost and XGBoost classifier. Multiple performance measures, like F1-score, recall, accuracy, and precision, are used to thoroughly assess an effectiveness of these classifiers. This thorough assessment guarantees a strong comprehension of the models' ability in properly forecasting loan eligibility.

Results of the Proposed XGBoost Classifier

This section presents a results of a proposed XGBoost classifier to showcase the model's effectiveness in predicting loan eligibility.

	precision	recall	f1-score	support
0	0.82	0.88	0.85	83
1	0.88	0.81	0.84	86
accuracy			0.85	169
macro avg	0.85	0.85	0.85	169
weighted avg	0.85	0.85	0.85	169

Figure 5.6: Classification Report of the XGBoost Classifier

Figure 5.6 displays the classification report generated by the XGBoost classifier. This report was used to showcase the proposed model's performance during testing. Binary digits, such 0 and 1, are used to categorise the classifier. Class 0 performed well in the classification report, with a f1-score of 85%, recall of 88%, and precision of 82%. In contrast, class 1 achieves 88% precision, 81% recall, and 84% f1-score.

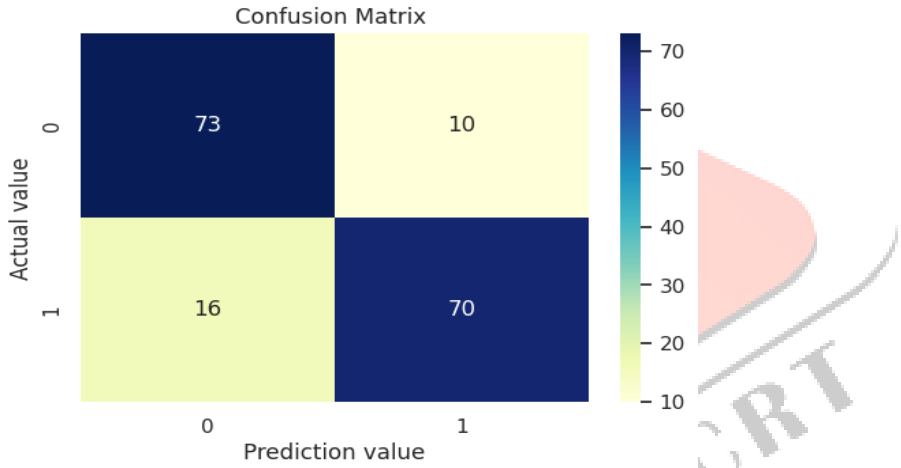


Figure 5.7: Confusion Matrix of the XGBoost Classifier

Figure 5.7 shows the confusion matrix of the XGBoost classifier. This shows how well a suggested model works. The predicted value, which is split into two distinct categories, such as 0 and 1, is displays on a x-axis of a confusion matrix. An actual value, which is also split into two distinct categories, is shown on the y-axis. The figure shows that class 0 correctly predicted that 143 of the applicants would not be eligible for a loan, while class 1 correctly predicted that 26 of the applicants would be eligible for a loan.

Table 5.1: XGBoost model performance with parameters on loan dataset

Parameters	**Test model**	**Train model**
Accuracy	84.61	100
Precision	87.5	100
Recall	81.39	100
F1-score	84.33	100

Table 5.1 shows the outcomes of applying the XGBoost model to a loan dataset. It includes the F1-score, recall, accuracy, and precision of both the train and test models. The training model is so good at learning from the training data that it achieves a flawless prediction on the training dataset and a 100% success rate across all measures. On the other hand, the test model operates

admirably, achieving an F1-score of 84.33%, recall of 81.39%, precision of 87.5%, and accuracy of 84.61%. Although these numbers are slightly lower than the training performance, they nonetheless indicate that the model is effective in generalising to new data. To further understand the model's predictive capabilities, Figure 5.8 shows a bar graph comparing its performance on the training and test datasets. Overall, the XGBoost model shows promise for accurate loan scenario predictions thanks to its balanced F1-score, recall, and precision values and good accuracy.

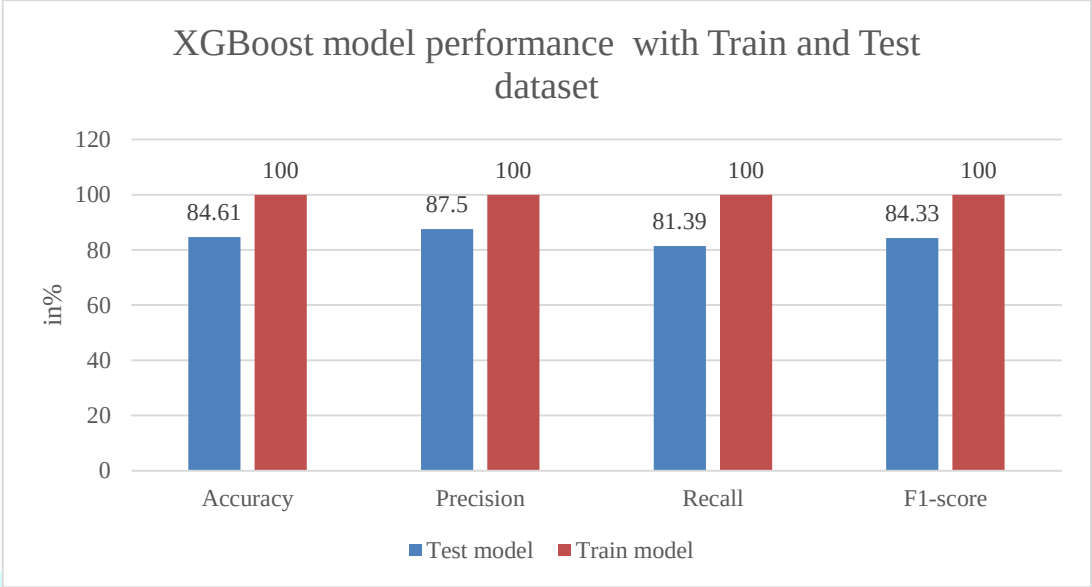


Figure 5.8: Bar graph XGBoost model performance with train and test dataset

Results of the Proposed Gradient Boosting Classifier

This section presents a results of a suggested Gradient Boosting classifier to showcase the model's effectiveness in predicting loan eligibility.

	precision	recall	f1-score	support
0	0.85	0.88	0.86	83
1	0.88	0.85	0.86	86
accuracy			0.86	169
macro avg	0.86	0.86	0.86	169
weighted avg	0.86	0.86	0.86	169

Figure 5.9: Classification Report of the Gradient Boosting Classifier

Figure 5.9 displays the Gradient Boosting classifier's classification report, which demonstrates the proposed model's performance during testing. Two categories, like 0 and 1, make up the classifier. Class 0's certification report has an impressive f1-score of 86%, recall of 88%, and precision of 85%. Instead, class 1 achieves 88% precision, 85% recall, and 86% f1-score.

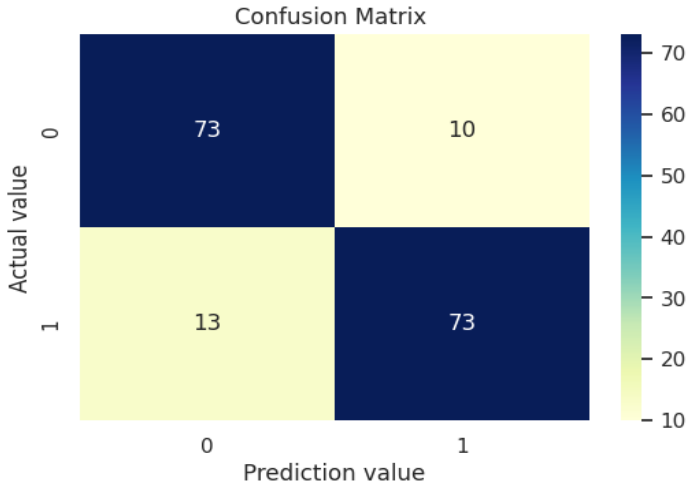


Figure 5.10: Confusion Matrix of the Gradient Boosting Classifier

Figure 5.10 shows the confusion matrix of the Gradient Boosting classifier. This shows how well the suggested model works. The figure shows that class 0 correctly predicted that 146 of the applicants would not be eligible for a loan, while class 1 correctly predicted that 23 of the applicants would be eligible for a loan.

Table 5.2: Gradient Boosting model performance with parameters on loan dataset

Parameters	Test model	Train model
Accuracy	86.39	100
Precision	87.95	100
Recall	84.88	100
F1-score	86.39	100

The F1-score, recall, accuracy, and precision for the test and train models of a Gradient Boosting model applied to a loan dataset are displayed in Table 5.2. On a training dataset, a training model achieves perfect scores of 100% across all criteria, demonstrating faultless prediction. With an F1-score of 86.39%, recall of 84.88%, precision of 87.95%, and accuracy of 86.39%, the test model, on the other hand, shows strong performance. Taken as a whole, these numbers show how effectively the model can generalise to new data while keeping its prediction power strong. By comparing a model's performance on a training and test datasets, Figure 5.11 provides a visual representation of a model's performance. Outcomes show that a Gradient Boosting model is quite accurate and has balanced precision, recall, and F1-score values when it comes to forecasting loan outcomes.

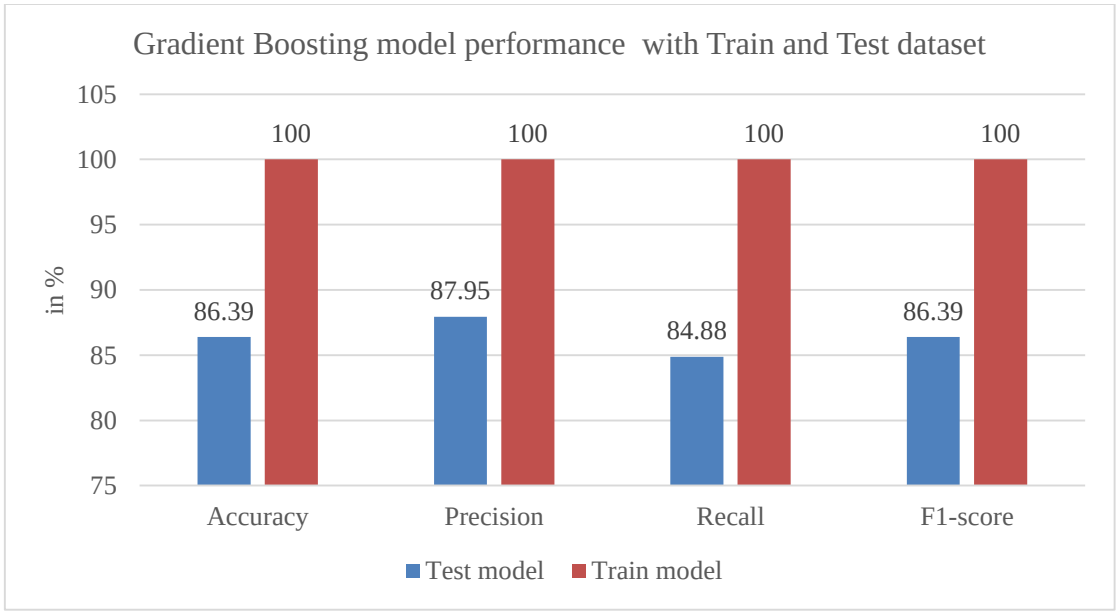


Figure 5.11: Bar graph Gradient Boosting model performance with train and test dataset

Results of the Proposed Cat Boost Classifier

As proof that the model can accurately predict who would be eligible for a loan, this section displays the outcomes of the suggested Cat Boost classifier.

	precision	recall	f1-score	support
0	0.85	0.88	0.86	83
1	0.88	0.85	0.86	86
accuracy			0.86	169
macro avg	0.86	0.86	0.86	169
weighted avg	0.86	0.86	0.86	169

Figure 5.12: Classification Report of the Cat Boost Classifier

To show how well the suggested model performed during testing, Figure 5.12 shows the Cat Boost classifier's classification report. The classification report shows that class 0 has a f1-score of 86%, recall of 88%, and precision of 85%. In comparison, Class 1's f1-score, recall, and precision are 86%, 85%, and 88%, respectively.

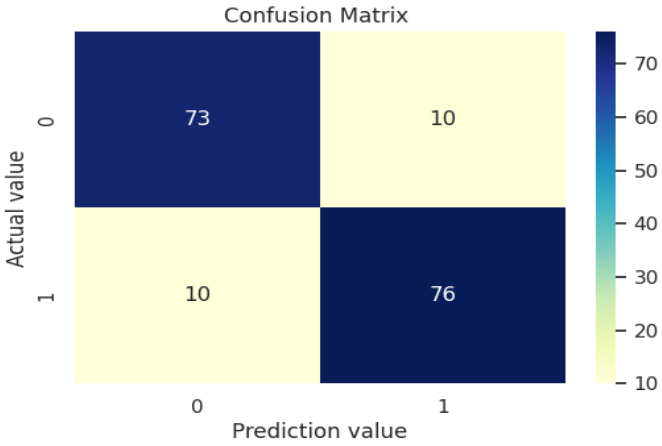


Figure 5.13: Confusion Matrix of the Cat Boost Classifier

Figure 5.13 demonstrate a Cat Boost classifier's confusion matrix. This shows how well the suggested model works. The figure shows that class 0 correctly predicted that 149 of the applicants would not be eligible for a loan, while class 1 correctly predicted that 20 of the applicants would be eligible for a loan.

Table 5.3: Cat Boost model performance with parameters on loan dataset

Parameters	Test model	Train model
Accuracy	88.16	100
Precision	88.37	100
Recall	88.37	100
F1-score	88.37	100

For a comprehensive analysis of the CatBoost model's performance on a loan dataset, see Table 5.3 for the values of accuracy, precision, recall, and F1-score for both the train and test models. All four metrics— F1-score, recall, accuracy, and precision — reach 100%, demonstrating that the training model makes faultless predictions on the training dataset. In contrast, the test model continues to perform admirably, with a recall of 88.37%, an accuracy of 88.16%, and a precision of 81.37%. According to these findings, the model shows great accuracy and balanced predicting skills on both the training and test datasets. Figure 5.14 provides a visual representation of these findings by showing, via a bar graph, how the model performed on the training and test datasets in comparison to one another. This extensive examination confirms that the model is capable of reliably forecasting loan results.

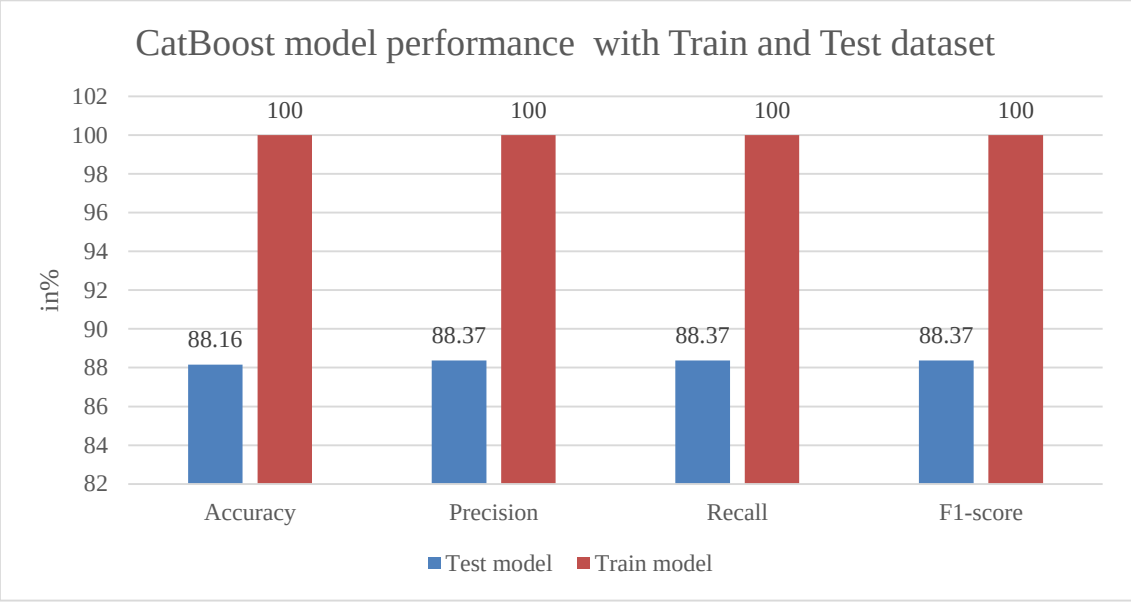


Figure 5.14: Bar graph Cat Boost model performance with train and test dataset

Comparison between base and proposed model

From Gradient Boosting and Cat Boost to Random Forest and GNB and KNN, Table 5.4 compares and contrasts all of the basic models used on a loan dataset with the suggested XGBoost model. The XGBoost model is on par with another base models according to accuracy, coming in at 84.61%. This is only below the Cat Boost model, which achieves 88.16%. Falling in the middle between Cat Boost and Gradient Boosting, the XGBoost model's accuracy of 87.5% is competitive. The model outperforms other basic models with an impressive recall of 81.39%, demonstrating its capacity to catch a substantial fraction of positive cases. With an F1-score of 84.33%, recall and accuracy are both well addressed.

Table 5.4: XGBoost model performance with parameters on loan dataset

Parameters	**Proposed models**			**Base models**		
	XGBoost	**B**	**Cat Boost**	**RF**	**GNB**	**KNN**
Accuracy	84.61	6.39	88.16	73	80	73
Precision	87.5	7.95	88.37	78	78	76
Recall	81.39	4.88	88.37	85	75	89
F1-score	84.33	6.39	88.37	81	87	82

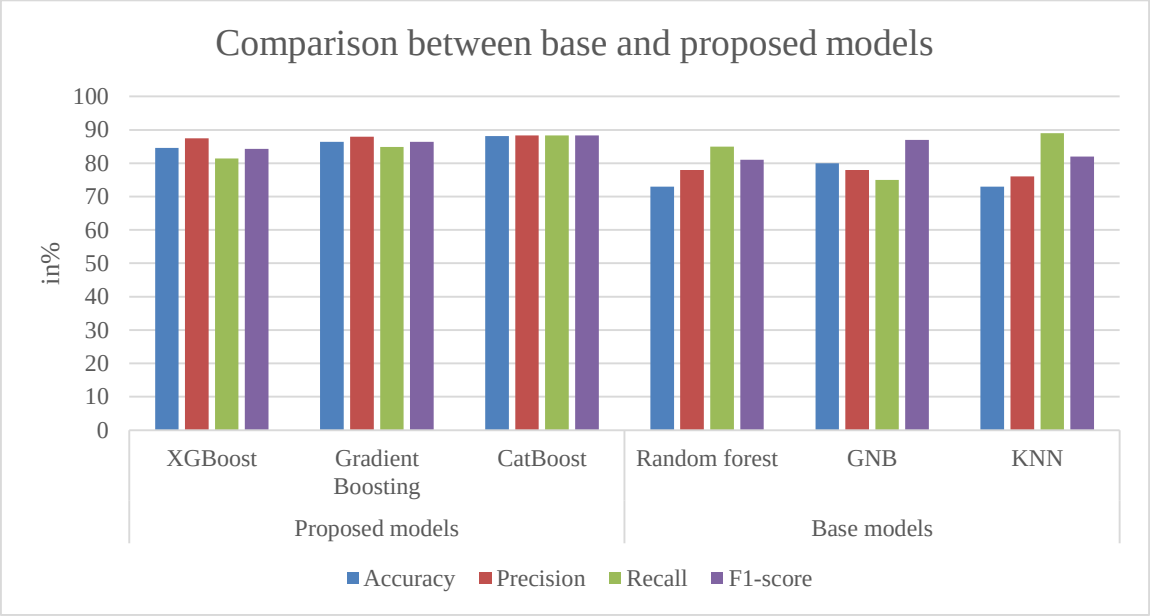


Figure 5.15: Bar graph of comparison between base and proposed models

When these measurements are compared to the basic models, XGBoost demonstrates performance that is exceptionally competitive across the board. Particularly noteworthy is the fact that it excels RF, GNB, and KNN according to accuracy, precision, and recall, which demonstrates its capabilities in forecasting the outcomes of loans. The comparison is depicted graphically in Figure 5.15 in the form of a bar graph, which offers a clear example of how the proposed XGBoost model compares to the base models. The findings indicate that the XGBoost model is a good contender for forecasting the outcomes of loans. It provides a balanced trade-off between F1-score, recall, accuracy, and precision, which is a significant improvement over other model.

CHAPTER- 6

CONCLUSION AND FUTURE WORK

Chapter 6 concludes the study, discusses its limits, and suggests directions for further research on loan prediction.

Conclusion

The prediction of loan approval is a crucial task for financial institutions, and has been a longstanding challenge in the industry. Historically, banks and other lenders relied on manual processes and subjective criteria to evaluate loan applications, which often led to inconsistent decisions and increased risk of loan defaults. With the rise of ML techniques, there is now an opportunity to develop more accurate and reliable predictive models that can help financial institutions make better lending decisions. This research suggests comparing several machine learning algorithms that might foretell if a loan will be approved.

A purpose of this research was to create and evaluate ML models that predict the likelihood of loan approval. Before diving into the loan approval process, we did exploratory data analysis to get a feel for the dataset. We used appropriate values from the data distribution to impute missing values in order to handle them. We also scaled and transformed the data logarithmically to prepare it for modelling. We then trained and evaluated a number of classification models, such as the XGBoost Classifier, Gradient Boost Classifier, and Cat Boost Classifier. To determine how well these models worked, we utilised accuracy as our metric of choice. Following that, three different classification models, namely XGBoost, Gradient Boosting, and Cat Boost, were presented and scrutinised for their effectiveness. With an accuracy of 88.16 percent, precision of 88.37 percent, recall of 88.37 percent, and F1-score of 88.37 percent, the test results showed that Cat Boost outperformed the other models. Taking everything into consideration, Cat Boost has shown that it has higher predictive performance, which is why it is the model that is suggested for predicting loan eligibility in this dataset. After analysing the results, we found that the Cat Boost Classifier had the highest accuracy of 88.16% on the test set, surpassing all other models. We may so infer that the Cat Boost model accurately predicts loan approvals using the traits supplied. These findings demonstrate how machine learning algorithms have the ability to enhance loan approvals and decrease default rates. In sum, the findings of this research may help financial institutions make better decisions by

shedding light on the relative merits of several ML technique for a purpose of loan approval forecasting. If classification is an important job in another area, the suggested technique may be used there as well.

Limitations and Future work

Although our models have shown promising outcomes, more study and development are necessary. A few possible future directions for this project are as follows:

- **Feature Engineering:** They may explore additional feature engineering methodologies to enhance existing features with more useful data. It may be necessary to include domain-specific data, build interaction terms, or use polynomial features after enhance a models' prediction capabilities.
- **Model Optimization:** After find an optimal values for a model's hyperparameters, we may use techniques like grid search or randomised search. More precise predictions could be the outcome of this improvement to the models' capability.
- **Handling Class Imbalance:** When dealing with a loan approval dataset that shows class imbalance—that is, a large difference among a number of approved loans and a number of rejected loans—we can employ methods like oversampling or under sampling, or we can use evaluation metrics like precision, recall, or F1 score.
- **Ensemble Approaches:** Stacking, boosting, and bagging are ensemble techniques that we may look at to combine the predictions of several models and perhaps enhance performance.
- **External Data Sources:** Further data sources, such as credit ratings or economic indicators, might be used to provide more comprehensive information for loan approval forecasts.
- **Deployment and Monitoring:** It is possible to automatically anticipate loan approvals in a production setting when a model has been selected. By regularly retraining the model and frequently evaluating its performance, its accuracy and correctness may be preserved.

REFERENCES

Abakarim, Y., Lahby, M. and Attioui, A. (2018) 'Towards An Efficient Real-time Approach to Loan Credit Approval Using Deep Learning', in *9th International Symposium on Signal, Image, Video and Communications, ISIVC 2018 - Proceedings*. Available at: https://doi.org/10.1109/ISIVC.2018.8709173.

Aebi, V., Sabato, G. and Schmid, M. (2012) 'Risk management, corporate governance, and bank performance in the financial crisis', *Journal of Banking and Finance* [Preprint]. Available at: https://doi.org/10.1016/j.jbankfin.2011.10.020.

AKÇA, M.F. and SEVLİ, O. (2022) 'Predicting acceptance of the bank loan offers by using support vector machines', *International Advanced Researches and Engineering Journal* [Preprint]. Available at: https://doi.org/10.35860/iarej.1058724.

Ambika and Biradar, S. (2021) 'Survey on Prediction of Loan Approval Using Machine Learning Techniques', *International Journal of Advanced Research in Science, Communication and Technology* [Preprint]. Available at: https://doi.org/10.48175/ijarsct-1165.

Aniceto, M.C., Barboza, F. and Kimura, H. (2020) 'Machine learning predictivity applied to consumer creditworthiness', *Future Business Journal* [Preprint]. Available at: https://doi.org/10.1186/s43093-020-00041-w.

Annisa, M. and Rusdah (2022) 'Prediction of Non-Performing Loans for Credit Application Analysis of Rural Bank Using Random Forest', in *International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*. Available at: https://doi.org/10.23919/EECSI56542.2022.9946628.

Aphale, A.S. and Shinde, D.S.R. (2020) 'Predict Loan Approval in Banking System Machine Learning Approach for Cooperative Banks Loan Approval', *International Journal of Engineering Research & Technology* [Preprint].

Arutjothi, G. and Senthamarai, C. (2017) 'Prediction of loan status in commercial bank using machine learning classifier', in *2017 International Conference on Intelligent Sustainable Systems (ICISS)*, pp. 416–419. Available at: https://doi.org/10.1109/ISS1.2017.8389442.

Asare-Frempong, J. and Jayabalan, M. (2017) 'Predicting customer response to bank direct telemarketing campaign', in *2017 International Conference on Engineering Technology and Technopreneurship, ICE2T 2017*. Available at: https://doi.org/10.1109/ICE2T.2017.8215961.

Aslam, U. *et al.* (2019) 'An empirical study on loan default prediction models', *Journal of Computational and Theoretical Nanoscience* [Preprint]. Available at: https://doi.org/10.1166/jctn.2019.8312.

Assef, F.M. and Steiner, M.T.A. (2020) 'Machine Learning Techniques in Bank Credit Analysis', *International Journal of Economics ...* [Preprint].

Awodele, O. *et al.* (2022) 'Cascade of Deep Neural Network and Support Vector Machine for Credit Risk Prediction', in *Proceedings of the 5th International Conference on Information Technology for Education and Development: Changing the Narratives Through Building a Secure Society with Disruptive Technologies, ITED 2022*. Available at: https://doi.org/10.1109/ITED56637.2022.10051312.

Bennouna, G. and Tkiouat, M. (2018) 'Fuzzy logic approach applied to credit scoring for micro finance in Morocco', in *Procedia Computer Science*. Available at: https://doi.org/10.1016/j.procs.2018.01.123.

Bentéjac, C., Csörgő, A. and Martínez-Muñoz, G. (2021) 'A comparative analysis of gradient boosting algorithms', *Artificial Intelligence Review* [Preprint]. Available at: https://doi.org/10.1007/s10462-020-09896-5.

BIS (2014) 'Basel Committee on Banking Supervision: A brief history of the Basel Committee', *Bank for International Settlement Paper* [Preprint].

Boateng, K. (2020) 'Credit Risk Management and Profitability in Select Savings and Loans Companies in Ghana', *International Journal of Advanced Research* [Preprint].

Boddepalli, R. (2022) 'Loan Eligibility Criteria using Machine Learning', *International Journal of Research Publication and Reviews*, 3.

Boughaci, D. and Alkhawaldeh, A.A. (2018) 'Three local search-based methods for feature selection in credit scoring', *Vietnam Journal of Computer Science* [Preprint]. Available at: https://doi.org/10.1007/s40595-018-0107-y.

Butwall, M., Ranka, P. and Shah, S. (2019) 'Python in Field of Data Science: A Review', *International Journal of Computer Applications*, 178(49), pp. 20–24. Available at: https://doi.org/10.5120/ijca2019919404.

By, S., Karki, A. and Dev Campus, S. (2008) 'A STUDY ON CREDIT MANAGEMENT AND ANALYSIS OF COMMERCIAL BANK (A Case Study of Everest Bank Limited)', (November), pp. 1–107.

Chang, W. *et al.* (2023) 'An Improved CatBoost-Based Classification Model for Ecological Suitability of Blueberries', *Sensors* [Preprint]. Available at: https://doi.org/10.3390/s23041811.

Chen, T. and Guestrin, C. (2016) 'XGBoost: A scalable tree boosting system', in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Available at: https://doi.org/10.1145/2939672.2939785.

Dhruv, C. *et al.* (2023) 'Framework for Bank Loan Re-Payment Prediction and Income Prediction', in *ICSCCC 2023 - 3rd International Conference on Secure Cyber Computing and Communications*. Available at:

https://doi.org/10.1109/ICSCCC58608.2023.10176363.

Endut, N. *et al.* (2022) 'A Systematic Literature Review on Multi-Label Classification based on Machine Learning Algorithms', *TEM Journal* [Preprint]. Available at: https://doi.org/10.18421/TEM112-20.

Ethem, A. (2015) *Introduction to Machine Learning Second Edition Adaptive Computation and Machine Learning*, *Massachusetts Institute of Technology*.

Ferreira, N. and Oliveira, M.M. (2014) 'An analysis of equity markets cointegration in the european sovereign debt crisis', *Open Journal of Finance*, 1(1), pp. 40–48.

Gao, W., Ju, M. and Yang, T. (2023) 'Severe weather and peer-to-peer farmers' loan default predictions: Evidence from machine learning analysis', *Finance Research Letters* [Preprint]. Available at: https://doi.org/10.1016/j.frl.2023.104287.

Ghahramani, Z. (2004) 'Unsupervised Learning BT - Advanced Lectures on Machine Learning', *Advanced Lectures on Machine Learning* [Preprint].

Goh, R.Y. *et al.* (2020) 'Hybrid harmony search-artificial intelligence models in credit scoring', *Entropy* [Preprint]. Available at: https://doi.org/10.3390/e22090989.

Gupta, A. *et al.* (2020) 'Bank loan prediction system using machine learning', in *Proceedings of the 2020 9th International Conference on System Modeling and Advancement in Research Trends, SMART 2020*. Available at: https://doi.org/10.1109/SMART50582.2020.9336801.

Hancock, J.T. and Khoshgoftaar, T.M. (2020) 'CatBoost for big data: an interdisciplinary review', *Journal of Big Data* [Preprint]. Available at: https://doi.org/10.1186/s40537-020-00369-8.

How to Use Jupyter Notebook in 2020: A Beginner's Tutorial (2020).

Islam, Md Shifatul, Arifuzzaman, M. and Islam, Md Saiful (2019) 'SMOTE Approach for Predicting the Success of Bank Telemarketing', in *TIMES-iCON 2019 - 2019 4th Technology Innovation Management and Engineering Science International Conference*. Available at: https://doi.org/10.1109/TIMES-iCON47539.2019.9024630.

Jiang, G.J. and Lo, I. (2014) 'Private information flow and price discovery in the U.S. treasury market', *Journal of Banking and Finance* [Preprint]. Available at: https://doi.org/10.1016/j.jbankfin.2014.06.026.

Jin, Y. and Zhu, Y. (2015) 'A data-driven approach to predict default risk of loan for online peer-to-peer (P2P) lending', in *Proceedings - 2015 5th International Conference on Communication Systems and Network Technologies, CSNT 2015*. Available at: https://doi.org/10.1109/CSNT.2015.25.

Kargi, H.S. (2014) 'Credit risk and the performance of Nigerian banks', *Acme Journal of Accounting Economics and Finance* [Preprint].

Karthiban, R., Ambika, M. and Kannammal, K.E. (2019) 'A Review on Machine Learning Classification Technique for Bank Loan Approval', in *2019 International Conference on Computer Communication and Informatics, ICCCI 2019*. Available at: https://doi.org/10.1109/ICCCI.2019.8822014.

Khan, F. *et al.* (2011) 'Determinants of bank profitability in Pakistan: A case study of Pakistani banking sector', *World Applied Sciences Journal* [Preprint].

Kingma, D.P. *et al.* (2014) 'Semi-supervised learning with deep generative models', in *Advances in Neural Information Processing Systems*.

Kiran, M.V.S. *et al.* (2023) 'Loan Eligibility Prediction Using Machine Learning', *International Journal for Research in Applied Science and Engineering Technology*, 11(8), pp. 55–60. Available at: https://doi.org/10.22214/ijraset.2023.55132.

Korkmaz, M., Sahingoz, O.K. and DIri, B. (2020) 'Detection of Phishing Websites by Using Machine Learning-Based URL Analysis', in *2020 11th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2020*. Available at: https://doi.org/10.1109/ICCCNT49239.2020.9225561.

Kozak, J. *et al.* (2022) 'Preference-Driven Classification Measure', *Entropy* [Preprint]. Available at: https://doi.org/10.3390/e24040531.

Krishn A., G. and S.A. (2010) 'Risk management in Indian banks: Emerging issues and challenges', 1(1), pp. 102-109.

Kumar, Arun, Garg Ishan, and K.S. (2016) '"Loan approval prediction based on machine learning approach.'

Kumar, A., Sharma, S. and Mahdavi, M. (2021) 'Machine learning (Ml) technologies for digital credit scoring in rural finance: a literature review', *Risks* [Preprint]. Available at: https://doi.org/10.3390/risks9110192.

Kumar, C.N. *et al.* (2022) 'Customer Loan Eligibility Prediction using Machine Learning Algorithms in Banking Sector', in *7th International Conference on Communication and Electronics Systems, ICCES 2022 - Proceedings*. Available at: https://doi.org/10.1109/ICCES54183.2022.9835725.

Kumar, V.H. (2018) 'Python Libraries , Development Frameworks and Algorithms for Machine Learning Applications', *International Journal of Engineering Research & Technology (IJERT)* [Preprint].

L.Gupta, D., K. Malviya, A. and Singh, S. (2012) 'Performance Analysis of Classification Tree Learning Algorithms', *International Journal of Computer Applications* [Preprint]. Available at: https://doi.org/10.5120/8762-2680.

Li, X. *et al.* (2021) 'Prediction of loan default based on multi-model fusion', in *Procedia Computer Science*. Available at: https://doi.org/10.1016/j.procs.2022.01.094.

Liang, W. *et al.* (2020) 'Predicting hard rock pillar stability using GBDT, XGBoost, and LightGBM algorithms', *Mathematics* [Preprint]. Available at: https://doi.org/10.3390/MATH8050765.

Lord, D., Qin, X. and Geedipally, S.R. (2021) *Highway Safety Analytics and Modeling, Highway Safety Analytics and Modeling*. Available at: https://doi.org/10.1016/B978-0-12-816818-9.01001-5.

Mamun, M. Al (2022) 'Predicting Bank Loan Eligibility Using Machine Learning Models and Comparison Analysis', pp. 1423–1432.

Meenaakumari, M. *et al.* (2022) 'Loan Eligibility Prediction using Machine Learning based on Personal Information', in *2022 5th International Conference on Contemporary Computing and Informatics (IC3I)*, pp. 1383–1387. Available at: https://doi.org/10.1109/IC3I56241.2022.10073318.

Meshref, H. (2020) 'Predicting loan approval of bank direct marketing data using ensemble machine learning algorithms', *International Journal of Circuits, Systems and Signal Processing* [Preprint]. Available at: https://doi.org/10.46300/9106.2020.14.117.

Moradi, S. and Mokhatab Rafiei, F. (2019) 'A dynamic credit risk assessment model with data mining techniques: evidence from Iranian banks', *Financial Innovation* [Preprint]. Available at: https://doi.org/10.1186/s40854-019-0121-9.

Muhangi, R.A.B.W. (2017) 'The Effect of Loan Appraisal Process Management on Credit Performance in Microfinance Institutions (MFIs): A Case of MFIs in Uganda', *International Journal of Science and Research (IJSR)* [Preprint]. Available at: https://doi.org/10.21275/ART20172815.

Natasha, A., Prastyo, D.D. and Suhartono (2019) 'Credit scoring to classify consumer loan using machine learning', in *AIP Conference Proceedings*. Available at: https://doi.org/10.1063/1.5139802.

Natekin, A. and Knoll, A. (2013) 'Gradient boosting machines, a tutorial', *Frontiers in Neurorobotics* [Preprint]. Available at: https://doi.org/10.3389/fnbot.2013.00021.

Orji, U.E. *et al.* (2022) 'Machine Learning Models for Predicting Bank Loan Eligibility', in *Proceedings of the 2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development, NIGERCON 2022*. Available at: https://doi.org/10.1109/NIGERCON54645.2022.9803172.

Owusu, E. *et al.* (2022) 'Loan Default Predictive Analytics', in *Proceedings - 2022 IEEE World Conference on Applied Intelligence and Computing, AIC 2022*. Available at: https://doi.org/10.1109/AIC55036.2022.9848906.

Ozgur, O., Karagol, E.T. and Ozbugday, F.C. (2021) 'Machine learning approach to drivers of bank lending: evidence from an emerging economy', *Financial Innovation* [Preprint]. Available at: https://doi.org/10.1186/s40854-021-00237-1.

P.Lee, K. (2017) *Introduction to Python Data Analytics*.

Papouskova, M. and Hajek, P. (2019) 'Two-stage consumer credit risk modelling using heterogeneous ensemble learning', *Decision Support Systems* [Preprint]. Available at: https://doi.org/10.1016/j.dss.2019.01.002.

Park, M.S. *et al.* (2021) 'Explainability of Machine Learning Models for Bankruptcy Prediction', *IEEE Access* [Preprint]. Available at: https://doi.org/10.1109/ACCESS.2021.3110270.

Paulheim, H. and Meusel, R. (2015) 'A decomposition of the outlier detection problem into a set of supervised learning problems', *Machine Learning* [Preprint]. Available at: https://doi.org/10.1007/s10994-015-5507-y.

Prasanth, C. *et al.* (2023) 'Intelligent Loan Eligibility and Approval System based on Random Forest Algorithm using Machine Learning', in *International Conference on Innovative Data Communication Technologies and Application, ICIDCA 2023 - Proceedings*. Available at: https://doi.org/10.1109/ICIDCA56705.2023.10100225.

Python 3 - Variable Types (no date).

Rahman, A.T., Purno, M.R.H. and Mim, S.A. (2023) 'Prediction of the Approval of Bank Loans Using Various Machine Learning Algorithms', in *2023 IEEE World Conference on Applied Intelligence and Computing (AIC)*, pp. 272–277. Available at: https://doi.org/10.1109/AIC57670.2023.10263880.

Rawlin, R., Sharan, S. and Lakshmipathy, P. (2012) 'Modeling the NPA of a Midsized Indian Nationalized Bank as a Function of Advances', *European Journal of Business …* [Preprint].

Richert, W. and Coelho, L.P. (2015) *Building Machine Learning Systems with Python, Second Edition*, *Book*. Available at: https://doi.org/10.1007/s13398-014-0173-7.2.

Rodrigo, K.L.S., Sandanayake, T.C. and Silva, A.T.P. (2023) 'Personal Loan Default Prediction and Impact Analysis of Debt-to-Income Ratio', in *2023 8th International Conference on Information Technology Research (ICITR)*, pp. 1–6. Available at: https://doi.org/10.1109/ICITR61062.2023.10382822.

Ruangthong, P. and Jaiyen, S. (2015) 'Bank direct marketing analysis of asymmetric information based on machine learning', in *Proceedings of the 2015 12th International Joint Conference on Computer Science and Software Engineering, JCSSE 2015*. Available at: https://doi.org/10.1109/JCSSE.2015.7219777.

Samreen, A., Zaidi, F.B. and Sarwar, A. (2013) 'Design and development of credit scoring model for the commercial banks of Pakistan: forecasting creditworthiness of individual borrowers', *International Journal of Business and Social Science* [Preprint].

Sheikh, M.A., Goel, A.K. and Kumar, T. (2020) 'An Approach for Prediction of Loan Approval using Machine Learning Algorithm', in *Proceedings of the International Conference on Electronics and Sustainable Communication Systems, ICESC 2020*. Available at: https://doi.org/10.1109/ICESC48915.2020.9155614.

Sindhuraj, I.C.G.L. and Patrick, A.J. (2023) 'Loan eligibility prediction using adaptive hybrid optimization driven-deep neuro fuzzy network', *Expert Systems with Applications* [Preprint]. Available at: https://doi.org/10.1016/j.eswa.2023.119903.

Singh, D. and Singh, B. (2020) 'Investigating the impact of data normalization on classification performance', *Applied Soft Computing* [Preprint]. Available at: https://doi.org/10.1016/j.asoc.2019.105524.

Singh, V. *et al.* (2021) 'Prediction of Modernized Loan Approval System Based on Machine Learning Approach', in *2021 International Conference on Intelligent Technologies, CONIT 2021*. Available at: https://doi.org/10.1109/CONIT51480.2021.9498475.

Srivastava, S. (2018) 'Analysis and Comparison of Loan Sanction Prediction Model using Python', *n. International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR)*, [Preprint].

Swapnesh, D., Nayak, K. and Swarnkar, T. (2023) 'LOAN ELIGIBILITY PREDICTION USING MACHINE LEARNING : A Global Journal of Modeling and Intelligent Computing (GJMIC) LOAN ELIGIBILITY PREDICTION USING MACHINE LEARNING : A', (July).

Tejaswini, J. (2022) 'Accurate loan approval prediction based on machine learning approach', . *Journal of Engineering Science* [Preprint].

Themba, J. and Narayana, S.B. (2014) 'The Impact of liberalization of regulations in Banking Sector : Case Study of Botswana Banking Sector', *Open Journal of Finance*, 1(1), pp. 15–26.

Tong, E.N.C., Mues, C. and Thomas, L.C. (2012) 'Mixture cure models in credit scoring: If and when borrowers default', *European Journal of Operational Research* [Preprint]. Available at: https://doi.org/10.1016/j.ejor.2011.10.007.

Uddin, N. *et al.* (2023) 'An ensemble machine learning based bank loan approval predictions system with a smart application', *International Journal of Cognitive Computing in Engineering*, 4, pp. 327–339. Available at: https://doi.org/https://doi.org/10.1016/j.ijcce.2023.09.001.

Vivek, R. and Mahaveerakannan, R. (2023) 'Analyze the Lack of Accuracy in Loan Prediction using Logistic Regression Compared with Random Forest to Improve Accuracy', in *Proceedings of 8th IEEE International Conference on Science, Technology, Engineering and Mathematics, ICONSTEM 2023*. Available at: https://doi.org/10.1109/ICONSTEM56934.2023.10142597.

Xu, J., Lu, Z. and Xie, Y. (2021) 'Loan default prediction of Chinese P2P market: a machine learning methodology', *Scientific Reports* [Preprint]. Available at: https://doi.org/10.1038/s41598-021-98361-6.

Yen, S.J. and Lee, Y.S. (2009) 'Cluster-based under-sampling approaches for imbalanced data distributions', *Expert Systems with Applications* [Preprint]. Available at: https://doi.org/10.1016/j.eswa.2008.06.108.

Zewdu Seyoum (2010) 'Impact of reducing loan by Ethiopian banks on their own performance';

Zhang, L., Wang, J. and Liu, Z. (2023) 'What should lenders be more concerned about? Developing a profit-driven loan default prediction model', *Expert Systems with Applications* [Preprint]. Available at: https://doi.org/10.1016/j.eswa.2022.118938.

Zhu, X. *et al.* (2023) 'Explainable prediction of loan default based on machine learning models', *Data Science and Management* [Preprint]. Available at: https://doi.org/10.1016/j.dsm.2023.04.003.

Zulfiker, M.S. *et al.* (2021) 'An in-depth analysis of machine learning approaches to predict depression', *Current Research in Behavioral Sciences* [Preprint]. Available at: https://doi.org/10.1016/j.crbeha.2021.100044.