COMPARING MACHINE LEARNING ALGORITHM FOR PREDICTING LOAN APPLICATION FOR PERFORMANCE ENHANCEMENT

IN PARTIAL FULFILLMENT FOR AWARD

OF

MSC BUSINESS ANALYTICS

SUPERVISOR: CHARLES NWANKIRE

SUBMITTED BY: ELIZABETH OPARAOCHA

STUDENT ID: 20004831

AT DUBLIN BUSINESS SCHOOL

APRIL 2023-MAY 2024

DECLARATION

This is to certify that the project entitled " Comparing Machine Learning Algorithm for Predicting Loan Application For Performance Enhancement" is my original work has been read, approved and adjudged to meet the requirement for the award of a degree in Business Analytics (MSC) in Dublin Business. I am responsible for the content and every contribution has been acknowledge.

School, Dublin, Ireland

Signed: Elizabeth Oparaocha

Student No: 20004831

Date: 20th May, 2024

ACKNOWLEDGEMENT

I want to sincerely thank Mr. Charles Nwankire, who oversaw my thesis, for all of his help, encouragement, and support during this research project. I really like the advice and ideas from the academic experts. A special thank you to my friends, Akada Gamaliel, Florence Bosibori, Rita Nwankwo, Peter Ibeabuchi, Bitrus Oparaocha, Athena Ogazi and family (Mr and Mrs Oparaocha) for their understanding and continuous support throughout this difficult time. Finally, I would want to express my gratitude to all the participants and sources who provided the information and data needed for this study.

TABLE OF CONTENTS

List of Figures

ABSTRACT

This study evaluates various machine learning algorithms for predicting bank loan status using the CRISP-DM methodology. The research utilizes a dataset from Kaggle, focusing on algorithms such as Decision Trees, Support Vector Machines, AdaBoost, Random Forests, K-Nearest Neighbours, and Logistic Regression. Models were optimized using GridSearchCV, with recall as the primary metric to highlight the importance of accurately detecting loan repayment patterns.

The findings demonstrate that the Support Vector Machine model was the most effective, achieving a recall score of 0.99and an F1 score of 0.96. Ensemble methods, which combine multiple models, notably improved prediction accuracy while maintaining interpretability. This study identifies the most effective algorithms and provides insights into factors influencing loan decisions, offering practical recommendations for banks to reduce bad loan rates and promote sustainable lending practices through advanced machine learning techniques.

CHAPTER 1

1.0 Introduction

One of the major reasons a bank is set up are to lend money to the community and to grant funds. Loan distribution represents a pivotal aspect of banking operations Most of the bank income is derived from interest charged on loan extended to consumers. Loan is ever increasing from study loan, property loan, business loan etc. when a loan is granted the interest is high and the bank makes additional money, and it is considered a good loan but a bad loan the bank loss both the money principal and the interest. To avoid bad loan which is a situation where a loan was predicted and granted as a good loan. Consequently, there has been a pressing need to modernise and optimise loan approval processes to ensure accuracy, efficiency, and risk mitigation. In response to the challenges face with loan distribution, machine learning offers a processing solution to automate and enhance loan distribution operations.

This study will do an analysis on the best machine learning model that can predict a good and bad loan. The data collected was from Kaggle had noise and outlier. which were removed at preparation. The data set will also be visualised using heat-map to see the coefficient and correlation between the variables with the help of machine learning tools.

The banking system has lots of data mining tools which has assisted in making informed investment decisions for predicting the market trends, analysing company performance, and understanding global economic shifts. This is because the banking industry work with big data and data needs to be converted into something meaningful. The main aim of this report is to evaluate and compare some algorithms and investigate the best predicting loan applications for performance enhancement.

In this report the Random Forest classifier, Support vector Classifier, Logistic Regression and KNN will be compared, and the aim of comparison is for a better performance. This is a supervised learning (data is labelled) and classification comes under supervised learning and the data is categorical.

In previous year 2023 USA bank suffered almost $19billon of soured loan in the second quarter. Which was the highest level in more than three years as lenders? contend with rising defaulters (Bhattacharya, 2023). This has driven various analysts to carry out investigations on the grant of loan. The importance is to identify a better algorithm that will help in predicting loan. Although no algorithm can grant 100% that a client is secure or not. The technique will examine the various variables and predict how secure a client is for payment. The output variable is good loan and bad loan. If the answer is good then the client is capable of payment if the answer is not then the client is not capable of paying back the loan. The strategy and technique of machine learning to recognise bad and good loan are ensemble techniques which consist of a set of learners, logic regression which is using a function called logic. Furthermore, each classifier has accuracy, precision, recall and are respectively compared. If predicted that a person will pay a loan back it will be given in precision and recall. The true positive and false positive will also be plotted.

1.1 Statement of problem

With the current increase of bank loan application, there is a need for banks to be careful who they grant loans. A bad loan is a bad investment to the bank and can cause a huge loss to the entire banking sector. Over the years, bank records more losses on loan as defaulters increases. The banking sector need a good machine learning algorithm that can predict if the loan is a good loan or bad loan. A good loan is of great value to the bank and bad loan is a loss to the bank. Although a machine learning algorithm is not 100% accurate that the loan is good. It will only give 98% to 99% that the loan is a good or bad loan. Therefore, the main goal of this research is to determine which machine learning algorithms are the most accurate and efficient at predicting loan applications. This study compares, looks at, and assesses the efficacy of several machine learning algorithms utilised in loan prediction in order to determine the best for performance enhancement, the objective of this research is also to introduce text mining algorithms to loan prediction. While loans plays a crucial role in facilitating economic growth and individual prosperity, effective risk management are essential to ensure responsible lending practices and mitigate the adverse consequences of loan default. Leveraging data mining algorithms and machine learning techniques can significantly enhance the ability of financial institutions to assess risk accurately and make informed lending decisions.

1.2 Research Questions

1. Which machine learning algorithm is best suited for predicting loan

applications

The aim here is to compare the performance of all the machine learning algorithms used in this research.

2. How ensemble methods and individual methods algorithm work for

performance enhancement.

The aim is to evaluate the performance of ensemble learning techniques like

bagging and boosting to improve the performance of loan application.

1.3 The research aims and objective.

The purpose of this study is to investigate the best machine learning algorithm for predicting loan for performance enhancement.
The aim is to evaluate and compare various evaluations metrics like precision, recall, F1-score, and Accuracy of each machine learning algorithm. And also, the aim is to give an insight on the most effective algorithms used for loan prediction.

The aim is also to introduce text mining algorithms to loan prediction.

algorithms.

The objective includes.

1. To understand and examine the dataset collected which includes identifying the influential variables and important features.

2. To implement a range of machine learning algorithms which include Support

Vector Machine, Random Forest, KNN, Logistics Regression

- To evaluate the performance of the algorithms using their Recall, Precision,

F1-score and Accuracy score.

- To analyse and compare the results of each algorithm.
- To evaluate the impact of the feature importance and their performance on

loan prediction.

- To compare and give a conclusion and provide recommendations of the best.

algorithms for loan prediction

- To Analysis text mining algorithm to a loan prediction algorithm

1.3 RESEACH SCOPE

The main force on this research work is doing a comprehensive analysis on machine learning algorithms to predict loan applications in the financial sector. The study will use a dataset that encompasses many variables such as Purpose of Loan, mode of payment, property, income etc. the primary aim is to determine the machine learning algorithms that have the highest level of accuracy. While offering valuable insights of the variables that contribute to predicting loan applications. (Pandit, 2016)

1.4 LIMITATIONS

The limitation of these Research work is that it will depend on dataset on previous loan applications dataset. Which may not be adequate because of the ever-changing financial quest of loans. The Research is limited to a time frame and a particular dataset which may have implications on providing the best results as different banks has different variables for a loan application dataset. The external factors that could complement on a loan application which includes Government policy, foreign exchange policy, international trade policy was not considered in this research. Also, some machine learning restricted the interpretability of some results.

CHAPTER 2
LITERATURE REVIEW

2.0 INTRODUCTION

Loan prediction is a challenge in the banking sector, In the banking sector millions of funds are lost due to loan defaulters. As a result, a bank has developed serval algorithm to predict loan applications to identify a good and bad loan. The goal of the loan prediction algorithm is to predict which loan application is good and which is bad because of the high interest rate gotten from loan payment, it is very important to identify a bad loan that will have a huge loss to the bank. Loan interest rate is one of the ways banks generate fund therefore it is important that loan defaulters is sported out. Loan defaulters can result to significant losses for the bank and even lead to banking collapse that can affect the economy. (Dansana D, 2024)

2.1OBJECTIVES

The objective of this literature review is to comparing loan prediction algorithm in machine learning. In the banking sector, loan is one of the major ways that the banking sector generate fund, the banking sector has generated various machine learning algorithm to predict good loan and bad loan. Various researcher has evaluated the performance of various machining learning approach. The aim of this literature review is to investigate various machining learning method and models that has be used by previous researcher in predicting loan applications. Examine machining learning algorithms like Logistic regression, decision tree, KNN, random forest, gradient boosting, and neural method in other research work. The literature review will examine the straight and weakness of various algorithm and their best algorithm in predicting loan application. The review will also look at the evaluation metrics and the methodology of the loan prediction by showing their accuracy, precision, recall and F1 score.

2.2 REVIEW AND ABSTRACT FROM THE PAST RESEARCH

Over the years different algorithm have been using to predict loan approval status. The first loan application algorithm made use of decision tree and random forest. The random forest was adopted by Lin Zhu in his paper on Predicting loan default based on Random Forest algorithm. Zhu paper looked at the challenges and the opportunities of a lending platform due to the risk of loan defaulters, Zhu also used Random Forest to build loan application model. And used SMOTE method to address class imbalance, data cleaning and dimensional reduction are conducted. Zhu research also suggested that random forest outperformed other machine learning algorithm in predicting loan application., random forest accuracy was (98%) other algorithms like logistic regression (3%), decision tree (95%) and support vector machine (75%). The result concludes the random forest algorithm was one of the best options on credit risk prediction. The paper model of evaluation and assessment metrics was accuracy, F1-score and recall were used to compare the performance of the random forest algorithm with other machine learning approaches. The paper also looked at risk management algorithms like Artificial Neural Network (ANN) and Deep Neural Network(DNN) which were used to predict stock Prices. Zhu also review method like logistic regression, k-nearest Neighbour, stochastic gradient boosting, naïve Bayes but in conclusion Random Forest was the best. (Lin Zhu, 2019).

Pidikili Supriya paper on" loan prediction by using machine learning" . To predict the creditworthiness of loan applicant and to identify safe candidate for loan approval thereby minimizing the risk of the bank. The study analysed a dataset using a variety of machine learning techniques, such as decision trees, logistic regression, gradient booting, and classification. The study employed decision trees as a machine learning method to forecast loan applications after cleaning, preprocessing, and filling in missing values in the data input and data analysis and model building and evaluation and their accuracy was 81%, in their conclusion loan applicants with credit history not passing fails to get approved. Probability because they have the probability of not paying back the loan. In their investigation, it appears that the factors of gender and marital status are not taken into company. (Pidikiti Supriya, 2019).

On smart banking and credit risk assessment Syed Zamail evaluated four supervised learning algorithm, logistic Regression (LR), Extreme Gradient Boosting, Random Forest, and support vector machine for the performance of a credit risk assessment. The hyperparameter for each for each of the model used Grid search CV and model performance in accessing various evaluation metrics such as recall, precision, F1-scoreand accuracy. In their result SVM and XGB model outperformed other model in team of classification of accuracy and overall performance. In concluded that support vector machine is capable of outperforming other models like logistic regression, Random Forest, Gradient boosting, deep learning etc. was used for performance analysis. The concluded that the choice of feature and algorithm are two major aspects when deciding to give an individual a loan or not. (Syed Zamil Hasan Shoumo, 2018)

Hamid and Ahmed used data mining model to build a classifying a loan, they used j48, bayes net, naïve bayes. The paperwork j48 algorithm enhance version of c4.5 generated tress based on input data and splitting the measure like the information gain. Nayes net algorithm constructed Bayesian network using probability while Naïve bayes Bayesian principles. In their paper the best algorithm for the purpose with accuracy of 78% and error of 0.3 was j48 algorithm, in their conclusion they use predictive analysis to investigate customer behaviours and predict pay back credit (Aboobyda Jafar Hamid, 2016)

To predict loan, Aditi and Nidhi in their loan sanctioning prediction system used data mining techniques to predict whether a loan should be approved for an applicant based on various attributes, Naïve Bayesian algorithm for their model and KNN to improve their classification accuracy and quality of data, they also used KNN to deal with Missing Values, they used Binning Algorithm to remove anomalies which improve their efficiency. The paper looked at the certain attributes of where a person to be granted loan or not. (Aditi Kacheria, 2016).

Martine vojtek and Evzen Kocenda on their research on local banks in Czech Republic and Slovakia CART and Natural Network as a support tool in their variable selection and model quality evaluation and KNN was not used. In their conclusion CART and neural network where the best for predicting credit scoring. (Martin VOJTEK, 2006).

X.Francis and Sumathi in their paper on loan prediction used Exploratory Data Analysis (EDA) to analysis their dataset and summarise the main characteristics with visual method. The paper conducted EDA on a dataset to understand the nature of loan applicant better. Its present visualisation and interpretation of graph depicting relationship between factors like annual income, purpose of loan, credit characteristic. There aim use an underly structured of a relatively large set of variables using visualizing techniques. (X.Francis Jency, 2018).

For credit forecasting system the paper' Loan prediction using machine learning' used logistic regression and random forest for credit prediction along with cross-validation techniques to assess model performance. They concluded that logistic regression algorithm is the best in predicting loan status accuracy of around 82%. (Anant Shinde, 2022).

The study 'The Evaluation of Consumer loans using Support Vector Machine' addresses the pressing need for accuracy in the loan prediction in the financial industry. The study uses Support Vector Machine (SVM) and compare it with multilayer perception (MLP). It also compares SVM and MLP through cross-validation and T-test. The result demonstrates that SVM outperformed MLP in terms of generalization performance and visualization. In conclusion the study presents a comprehensive approach to consumer loan evaluation leveraging SVM technology to meet evolving requirement of the financial industry. Also, the study analysis of misclassification error provides valuable insight into optimizing SVM parameter. (Sheng-Tun LI, 2006)

To improve the performance of credit scoring. There next paper used the traditional statistical models like Linear Discriminate Analysis (LDA) and logistic Regression (LR) was used for credit scoring. The study used both Supervised and Unsupervised ML algorithms. The supervised ML was SVM, Decision Tree, K-nearest neighbour, Random Forest, Gradient Boosting Decision Tree (GBDT). For unsupervised learning includes K-means, hierarchical Clustering, DBSCAN and isolation Forest. It combined supervised and unsupervised technique to enhance the classification performance of credit scoring model. By integrating unsupervised and supervised ML technique shows promising result in improving the performance of credit scoring model. (Wang Bao, 2019).

This paper presents a study on model for car loan payment using supervised machine, the aim was to predict whether subscribers to car loans are likely to pay them back before the termination data. Logistic Regression model effectively predict prepayment risk for a car loan aiding banking institutions in decision making to minimize profit loss. (Zahi Sara, 2020).

The paper "Comparing classification Algorithm on predicting loan" compares the different machining learning algorithm for predicting the capability of a customer to purchase a loan. The study evaluates the effectiveness of several methods, including K-nearest neighbour, Random Forest, Decision Tree, Support Vector Machine, and Logistic Regression based on their accuracy, precision, recall and ROC curve. Random Forest accuracy was 98%, logistic regression was 95%, KNN accuracy was 63%, Decision tree was 98%. In conclusion the study concluded that random forest is the most suitable algorithm for predicting loan purchase followed by decision tree and SVC and KNN. (Agarwal Krishanu, 2022).

(Huang Zan, 2004) Discussed the application of machine learning techniques, specially SVM and Backpropagation neural network (BNN) for Credit rating predictions. The aim of the research was to provide a model with better explanatory information compared to traditional statistical method. It used SVM and BNN to predict credit rating and compare their performance and exploring method to enhance the interpretability of AI-based model in credit risk Analysis.

But neural networks have being used to predict risk assessment over the years. This next paper explores neural network (NNs) in the context of business risk based. NNs studies the brain and nervous system. It highlights the ability to handle nonlinearity and interdependencies among variable without requiring Strick distributional assumptions. Emphasizes the iterative nature of NNs model in learning from observation and minimize prediction errors (Calderon G Thomas, 2002).

This paper “Predicting Bank Loan Eligibility using machine learning model and comparison analysis” Developed model using techniques like Logistic Regression, Decision tree, Random Forest, KNN, LightGBM, XGBoost, Adaboost. Each algorithm's performance is assessed using common metrics including recall, accuracy, precision, and F1-score, Area Under Curve (AUC). LightGBM achiveve the highest accuracy 91.89% followed by ADAboost and Random Forest. AUC analysis also show LightGBM outperformed other classifiers. (Manum AL Miraz, 2022).

To predict loan approval in the banking sector. This paper used Logistic Regression, Decision Tree, Random Forest Machine leaning Algorithm. The result shows the prediction accuracy of the each of the machining learning algorithm. The key finding of the study and highlight the significant of machine learning algorithm particularly Decision tree. It under scores the potential of these technique to enhance decision making process and mitigate risk for financial institution. (Siva, 2022).

This paper aim was to teach student how to apply Logistic Regression to classify loan based on default probability. The logistic regression model is built using risk indicators to estimate the probability of default of each loan application. The graduate student explores more advanced classification method like compared to logistic regression. SVM also explore as an extension of Support Vector Classifier with student comparing their performance to logistic regression. (Min Li, 2018).

The paper discusses on approach for predicting loan approval using Machine Learning. Its focus on minimizing the impact of loan defaulters on individual investors on peer-to-peer lending platform. The paper aim was to design and implement a system using machine learning and data mining to predict the probability of loan approval for bank customers. The researchers used Decision tree algorithm for classification task related for loan approval predictions. Decision tree algorithm is employed for classification and regression task on predicting loan approval. The algorithm iteratively chooses the best attribute to split the dataset into subset until reaching a leaf node representing class label. (International Joural of Creative Research Thoughts., 2021).

The use of explainable Artificial Intelligence (XAI) techniques was used in this next paper. A case study of the Egypt banking sector. This paper propose model to predict loan approval using XAI. It introduces XAI as a field of AI that allows humans to understand and interpret AI decisions. The model is tested using fairness metrics including demographic parity, equality of opportunity and equalized odd. (Omar Mohib Ayad, 2023)

Table 1: Past Research on loan Prediction Using Machine Learning

Research paper	Author	Year	Algorithm	Accuracy
A study on predicting loan default based on Random Forest	Lin Zhu, Defeny Qui, Daji, Ergu	2019	Random Forest	98%
Loan Prediction by using machine learning models	Pidikiti Supriya, Myneedi Pavani.	2019	Decision tree	81%
Application of machine learning in credit risk assessment.	Syed Zamil Hasan Shoumo, Mir Ishak Maheer Drube	2018	XG boost	98%
Developing prediction model of loan risk in bank using data mining	Aboobyda Jafa Hamid, Tarij Mohammed Ahmed	2016	J48	78%
Loan prediction Using Machine Learning	Anant Shinde, Yash Patil, Ishan Kotian	2022	Logistic Regression	82%
Comparing Classification Algorithm on Predicting loan	Agarwal Krishanu, Jain Mohit, Kamawat Ashok Kumar	2022	Random forest	98%
Predicting bank loan Eligibility using machine learning	Manum AL Miraz, Farjana Afia	2022	Light GBM	91.88%

2.4 Research Gap:

Over the year many researchers have forced on the structured dataset of predicting loan approval. There have been limited study on unstructured dataset to predict a loan approval. An unstructured dataset can be in the form of a letter of application for a loan. These letters will require a text mining algorithm to predict, as lots of letters are sent to the bank for approval daily. But text mining has been used to predict.

positive and negative Reviews and spam emails. Hence there is a gap in using. machine learning algorithms to predict a loan application when it is unstructured. form like a letter.

2.5 Conclusion:

Loan application is an important decision in the banking sector, many algorithms have being used to predict loan approval, the most commonly used machine learning for predicting bank loan are Random Forest, Decision Tree, SVM, Logistic Regression, C24, Adaboost where discussed in some papers among others. The evaluate their performance using Metric score (Accuracy, F1-score, Precision, Recall). The study also highlights and discuss the important of accurately predicting loan approval to minimize the risk of default and financial losses for the bank. Comparison has been done of different algorithms performance across multiple studies and Random Forest has outperformed must machine learning algorithm. The study will also explore the unstructured data sources and provide valuable insight into predicting a solid foundation for further research in the field of loan prediction.

CHAPTER 3
DESIGN & METHODOLOGY

3.0 Introduction:

The methods employed to achieve the goals and objectives of research are covered in this section. The research is quantitative in character and employs machine learning algorithms. It focuses on numerical data, impartial measuring and statistical analysis. By spotting patterns and trends in data, machine learning draws conclusions that help with decision-making. Since the data for this study was gathered by a party other than the researcher, it falls under the category of secondary data collection. The strategy makes use of the CRISP-DM method.

3.2 CRISP-DM Methodology

In order to anticipate loan acceptance, this thesis contrasts CRISP-DM with alternative machine learning techniques. The well-documented CRISP-DM (Cross-Industry Standard Process for Data Mining) method is used to manage machine learning and data mining projects. This technique offers a methodical approach to tasks involving predictive modelling and complicated data processing. Professionals are guided through the entire data mining process using CRISP-DM, from comprehending the

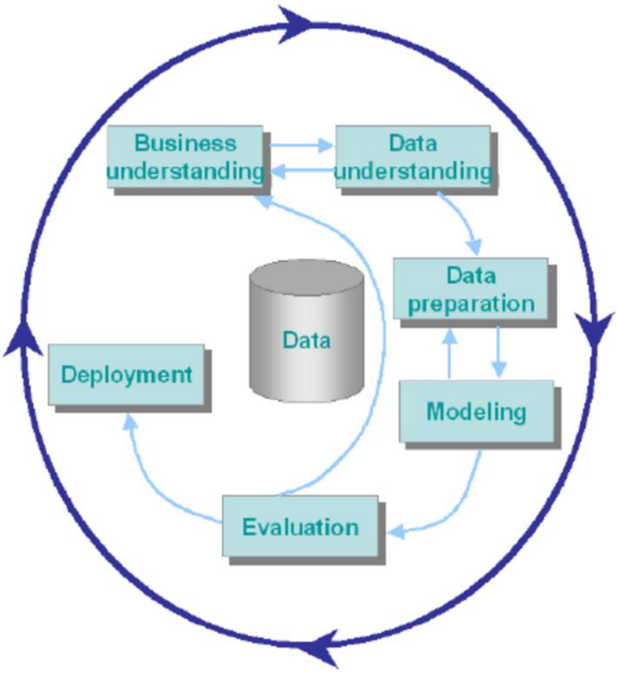


Figure 1: CRISPDM Architecture by Rafael Timoteo de

business challenge to putting the finished model into practice. This is a summary of the CRISP-DM process.

3.2.1 Business Understanding

This phase aims to analyse the important of predicting loan application accurately using various machine learning algorithm the various algorithm evaluation metric score will also be compares and evaluated This stage involves collaborating with Bankers and industry experts to identify key business objectives and requirements for accurate loan prediction forecasting.

Understanding how loan prediction in the banking sectors is vital for their financial and long-term of the economy. The study will examine the financial impact loan prediction accuracy, including costs associated with predicting a bad loan as a good loan . This study aims to explore the benefits of precisely forecasting good loan, such as predicting good loan, spotting bad loans, and improving machine learning algorithm. These outcomes are expected to increase the value of a good loan. The company Understanding phase establishes the parameters for a comparison study, including the goals of the research, the importance of precise loan forecasting in the banking industry, and the possible influence of study results on loan approval for company strategy and decision-making.

3.2.2Data Understanding:

The development of a loan prediction model in the banking sector is to necessitate the interest rate of a good loan and to predict and avoid bad loan. Therefore, it's critical to get thorough information whenever possible. The dataset used in this investigation was downloaded from the Kaggle site. The loan dataset that was used is categorised and has 2134 rows and 18 columns in size.

Some of the following are included in these columns:

1. **YEAR:** the year the loan was requested for.
2. **Final_d**: the date the loan was approved.
3. **Emp_length_int**: This is the employment length of the client. Which value are between 0 and 10, with 0 means less than one year and 10 means ten or more years.
4. **Home_ownership:** the living situation of the client requesting for a loan, if they own, rent or mortgage their house.
5. **Home ownership cat**. The number of people that previously own their house.
6. **Income cat**: the level of income of the client if it low, medium, or high income.
7. **Annual income:** the yearly income of the client. That is all the salary for a year.
8. **Income cat:** the level of income, income level 1 means they are receiving below 33000 and level 2 means they are receiving above 33000
9. **Loan amount:** the amount of which the client is requesting the bank to loan them.

10. **Team:** this is the duration of how long the client must pay back the loan. The duration of the loan is from 36 months to 60 months.
11. **Purpose:** why the client needs the loan. Purpose includes car loan, house loan, wedding, school loan etc.
12. **Interest payment:** what the interest level of the loan. While the interest be high, or low interest.
13. **Interest rate.** What is the interest rate of the loan. That is the percentage of interest rate.
14. **Total payment.** What is the total payment they will get back from the loan, including the interest and the initial loan.
15. **Total rec princp:** this is the total payment towards the original amount of the loan that is owed.
16. **Instalment:** how much is the partial payment of the loan granted out.
17. **Region:** where was the loan requested for. What geographical area was the loan requested for.
18. **Loan conditions:** what was the condition of loan. And if the loan was a good loan or a bad loan.

Among the 2134 loan requested for, 1753 was a good loan and 380 was a bad loan. The target variable was loan conditions which indicate if the loan was a good loan or bad loan. The remaining qualities where the predictor attributes are factors that influence the target attributes. These were the features used to predict a good or bad loan.

3.2.3 Data preparation:

Data preprocessing is a very important step in developing any machining learning model and it has been shown that efficiently use of data can make massive difference in result. (Syed Zamil Hasan Shoumo, 2018). The database that had missing values due to improper data entry or data entry problem which could be an issue in the data analysis and it could affect the accuracy of the model. KNN was used to overcome the issue of missing data. Data refilling was also done on the dataset, this was because the dataset had outliers and noise present in the data. The outliers was removed in order to increase the accuracy of the system.

3.2.4Converting Categorical Features to Numerical:

From 18 features, 6 are categorical variable. The CSV file's category feature after being processed by Pandas, it could be seen as object format. The categorical feature are converted with the use of label encoder.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2133 entries, 0 to 2132
Data columns (total 18 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   year                2133 non-null   int64  
 1   final_d             2133 non-null   int64  
 2   emp_length_int      2133 non-null   float64
 3   home_ownership      2133 non-null   object 
 4   home_ownership_cat  2133 non-null   int64  
 5   income_category     2133 non-null   object 
 6   annual_inc          2133 non-null   int64  
 7   income_cat          2133 non-null   int64  
 8   loan_amount         2133 non-null   int64  
 9   term                2133 non-null   object 
 10  purpose             2133 non-null   object 
 11  interest_payments   2133 non-null   object 
 12  interest_rate       2133 non-null   float64
 13  total_pymnt         2133 non-null   float64
 14  total_rec_prncp     2133 non-null   float64
 15  installment         2133 non-null   float64
 16  region              2133 non-null   object 
 17  loan_condition      2133 non-null   object 
dtypes: float64(5), int64(6), object(7)
memory usage: 300.1+ KB
```

Figure 2: Dataset Data type. by author

Dataset data type

3.2.5 HEATMAP

To show the graphical representation of the data In which the value within are in a matric and colour coded, the study will use heat map. Heat map will help visualize the dataset to recognise patterns, trends and correlation of the dataset when color intensity corresponds to the numerical value.

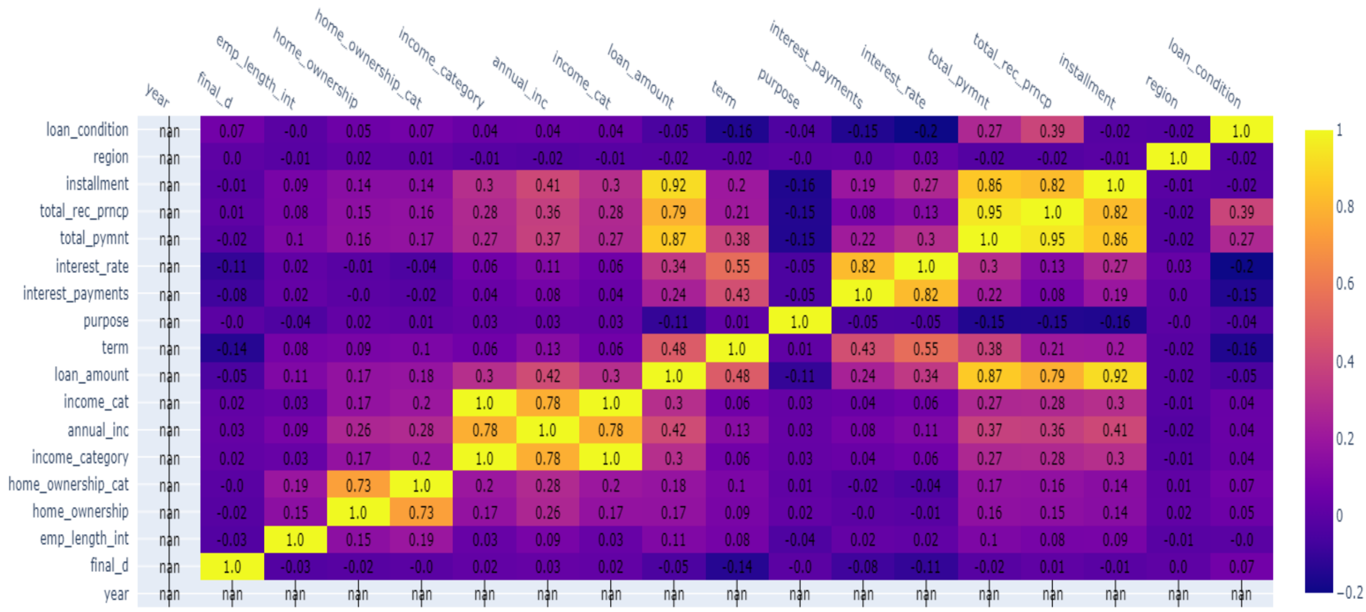


Figure 3: Correlation Heatmap

HEATMAP VISUALIZATION

3.2.6 SPLITING THE DATASET:

To validate the performance of the machine learning model the dataset was spitted into 70:30 ratio. The splitting was done to avoid overfitting of the model. The performance of the model is evaluated using potion of a dataset, which some will be used to train the model, and some will be used to test the model performance. The ratio was 70:30 train test split ratio and the Train test split function was imported from scikit learn library. This means that 70% of the dataset will be trained on the model and 30% will be used to test the model performance Which will give us a more balanced overall performance. This train test split is necessary to understand the generalisation of the model. Cross validation was also used to divide the entire dataset into folds and each folds will be used for train and test set.

3.2.7 SMOTE METHOD:

SMOTE (Synthetic Minority Oversampling Technique) was used on the target variable “Loan condition” this is because of the larger different in the number of good loan and bad loan. Which will cause an issue in the model learning. SMOTE was adopted to handle Sample imbalance problem. KNN was used to calculate the nearest Neighbour of the minority sample with Euclidean distance.

3.2.8 MODELLING:

The modelling phase is the utilisation of diversion of machining learning techniques to construct prediction models for a good loan in the banking sectors. After the data has been preprocess and feature engineering in the data preparation step and the data is split into train and test set. The result to be obtained is whether a loan was a good loan or a bad loan. Various machine leaning algorithms such as KNN, Random Forest, Decision tree etc are used to analysed and train the data.
The hyperparameter of each algorithm are optimised with the application of grid search and cross validation to improve efficiency. Next, the model's performance is evaluated using appropriate evaluation metrics, such as accuracy and recall, Precision and F1-Score. The model will be compared regarding there predictive accuracy to choose their best appropriate algorithms for predictability loan prediction. The interpretability is important because it will give and insight into the many elements that impact loan prediction. The object of the modelling phase is to finish the banking sector with a trustworthy predictive algorithm that can help predict the state of the loan. The chosen model's deployment and implementation are getting ready.

The following machine learning algorithm are used.

- Random Forest.
- Decision Tree
- Support Vector Machine
- Adaboost
- Logistic Regression
- K-nearest Neighbours

RANDOM FOREST:

Random forest is an extension of Decision tree technique. It is a method that amalgamate multiple Decision trees to create an accurate prediction that increase accuracy. So, it is an evaluation of Decision Trees. Random forest is based on a supervised model. It can be used for both Regression and Classification problem. The Hyperparameter are Max-depth, max feature, Min_sample_leaf, N_estimators, bootsrap and bagging. The advantage of random forest is it capability to manage intricate and high dimensional dataset The approach guarantees the generation of trees that capture the subtleties of the dataset by using a random subset of features. The best prediction is gotten by adding up output from all the Decision Trees. The advantage of random forest is that it is fast to train and predict, It only needs one or two turning parameter, it is used for high-dimensional problem, it is also used for visualisation, missing value and differential class weighting. The disadvantage of random forest is that random forest does not aid in interpretation as much as a single decision tree does. Moreover, random forests demonstrate scalability and computational efficiency, making them appropriate for managing large datasets with a significant number of observations. Random Forest will be used because it combines a high predictive accuracy with robustness and scalability. Random Forest is collection of decision trees that are not related. Selection metrics was use for splitting attributes in the decision tree is Gini index and the number of levels in each tree branch depends on the algorithm parameters. (Lin Zhu, 2019).

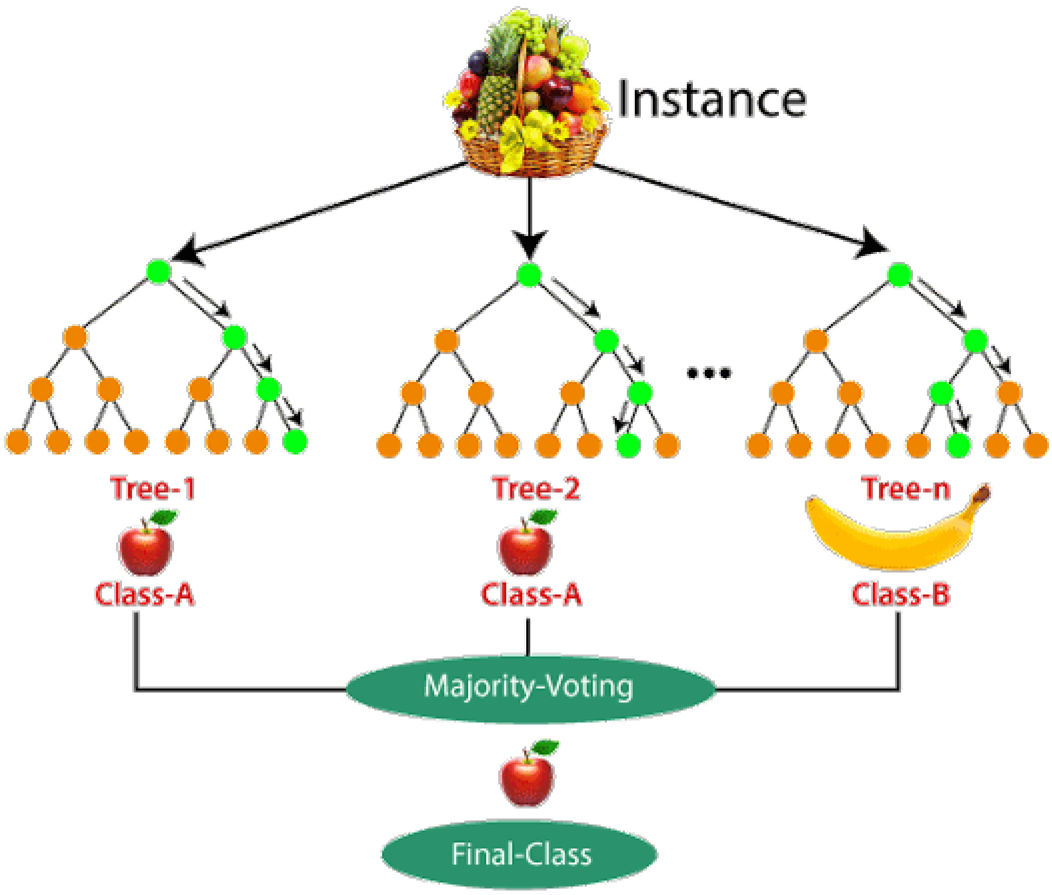


Figure 4:Random forest architecture (Sruthi, 2014)

DECISION TREE:
Decision tree is another Supervised learning algorithm that used to solve both Classification and regression problem. Decision Tree have external node and leaf node to solve the problem which are the attribute and class. On of the advantage of Decision tree is it simplicity, interpretability, and ability to handle categorical data and numerical data. But Decision tree is prone to overfitting, this happens when the tree grows deep, and the dataset has much noise. Pruning is used to limit the number of tree depth and improve the performance of the tree. Decision tree is an important machining learning algorithm that will be used in this study because it provides both predictive accuracy and interpretability which are essential in predicating loan application. Decision tree is use because it gives a good result for classification problem. Its prediction are interpretable and easy to implement and the result are unbiased and give accurate result

Support Vector Machine:
Support Vector Machine (SVM) is a supervised learning model that is used for classification problem. When training data is given to the model, the model use hyperplane to separate them into two categories. It also uses linear separation and non-linear separation are done in the SVM model. It uses Kernal trick on the non-linear classification. For this research the study uses grid search to point out the linear kernel to give a better classification accuracy, another reason for using SVM is because of the large dataset, SVM will show the complex relationships between features and target

variable. Advantage of SVM is robust non-linearity, Generalisation, and high-dimensional efficacy.

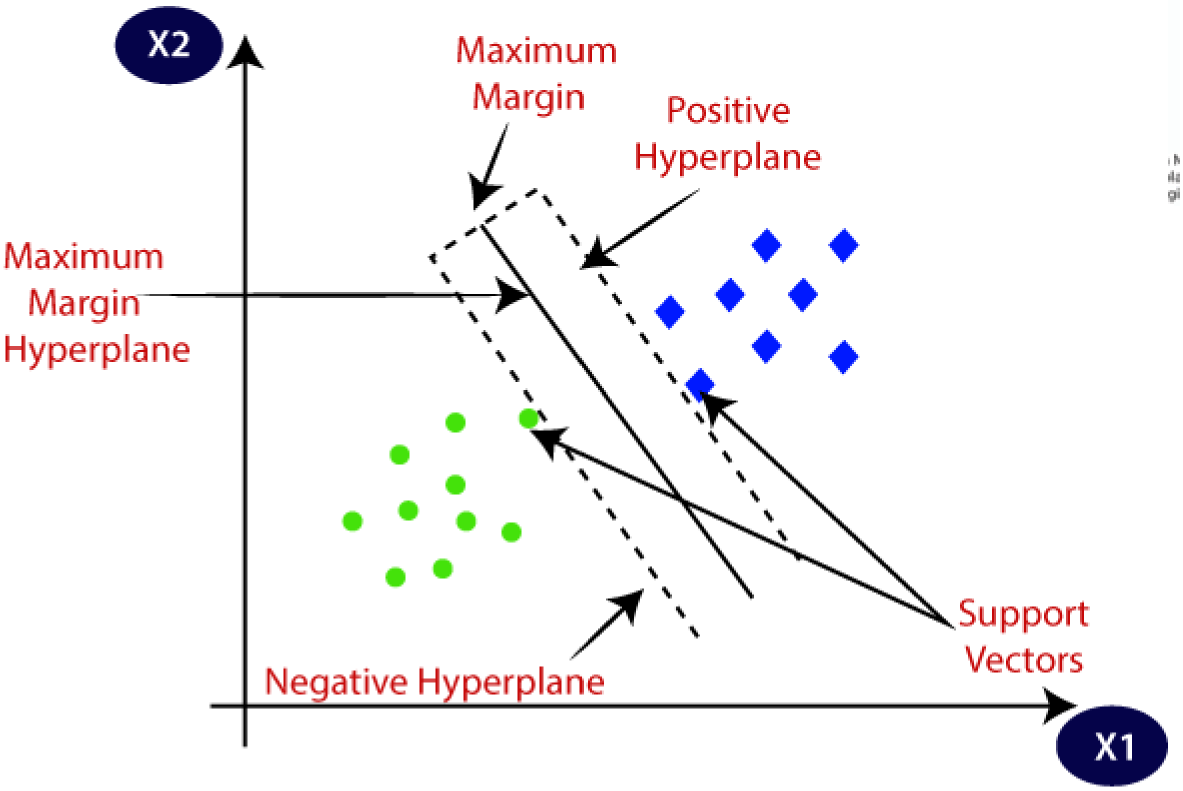


Figure 5: Support vector Machine (DataScience.com, 2020)

ADABOOST:

Adaptive boost classifier is one of the best classifiers for prediction. They provide accurate result by combining output of several weak learner. By weighting training data. Iteration makes use of decision stump as individual model. One advantage of Adaboost handle relationship especially complex data. Using Adaboost to capture non-linear patterns and for structure and unstructured dataset. Adaboost train multiple weak learners and adjust them by given them more weight. It then aggregates the weak leaner prediction. The benefit of Adaboost is that it handles unbalance data, versatility, reduce overfitting and feature importance.

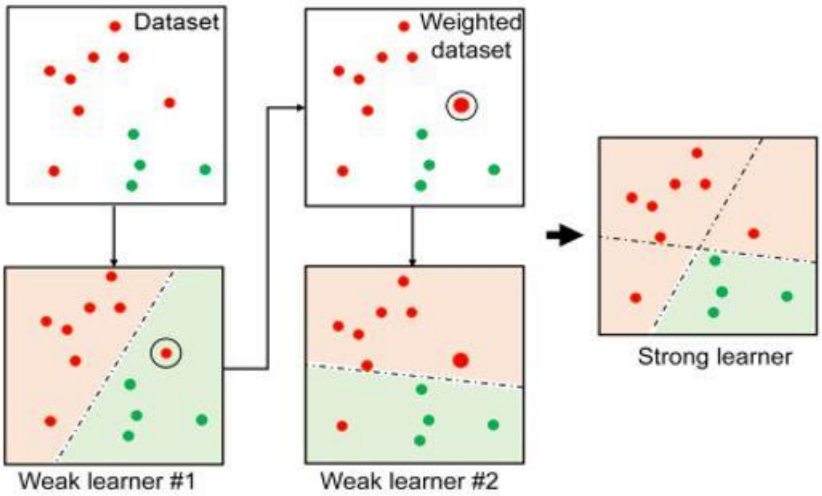


Figure 6: Adaboost Artchitecture (Sureshan, 2021)

Logistic Regression:
Logistic regression was used because it deals with a limited number of labels. The label in a logistic regression are not in continuous form. Logistics regression help to calculate the probability of a specific outcome and the probability are calculated using the logic function and the output label will be represented in a binary form (1 or 0)

$$\log\left(\frac{p(X)}{1-p(X)}\right) = a + bX$$

The function shows the estimated probability of a label. X represent the input of the independent variable p(X) represent the dependent variable and **b** represents the regression co-efficient. Also **P(X)/1-P(X)** is the odd ratio which is used to find the probability of the outcome. **C** will be used to tune the hyperparameter and L1 regularisation which uses Lasson Regression.

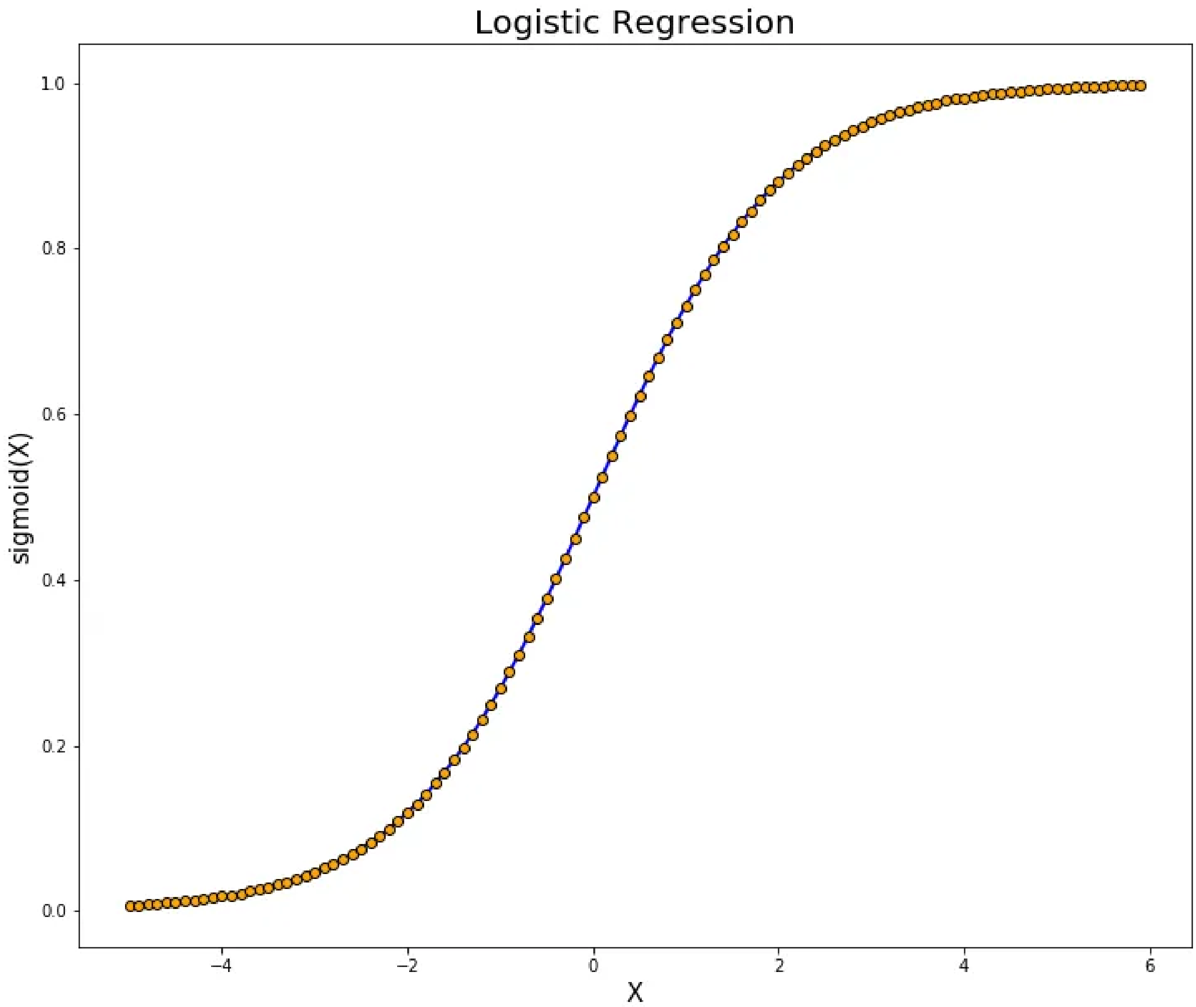


Figure 7: Logistic Regression Architecture (Thorn, 2020)

K-nearest Neighbour
KNN is a supervised machine learning approach that is used to classify categorical values. The Dataset are trained with selected features and the target variables is predicted. In KNN the target variable is categorical. KNN is basically used to predict true or false, yes, or no and many more scenarios. Because KNN is a non-parametric algorithm it cannot make assumptions like in the case of Logistic Regression, KNN assumes that all the data point are linearly separable therefore KNN is not used for predictions, like house prices, car prices and many more.

It classifies the data set based on the classification of its neighbour data.

Similarity/dissimilarity measures are used to classify the cases.

1. First, a value of KNN, simply K is chosen.
2. The distance between the known cases and the unknown cases is calculated (Euclidean Distance).

Establishing the KNN Framework:

n_neighbors=5, metric='minkowski', p=2, weights='uniform', algorithm='auto', leaf_size=30, metric_params=None, n_jobs=None) = knnClassifier = KNeighborsClassifier

With the following settings, a KNN classifier object is created with this line:

n_neighbors: The number of closest neighbours to take into account when forming a prediction is indicated by this parameter. It is set to 5 here.

metric: The distance metric that is utilised to calculate the separations between data points is specified by this parameter. In this instance, it is set to "minkowski," a generic class of distance metrics that include Manhattan distance (L1) and Euclidean distance (L2). The Euclidean distance metric, or 2, is represented by the parameter "p," which is set to 2.

weights: The weight function among the neighbours is specified by this option. In this case, it is set to "uniform," which assigns equal weight to each point.

Additional variables such as algorithm, leaf_size, metric_parameters, etc

Here, n_jobs are not explicitly set; instead, they have default values.

Adapting the Model:

The KNN model is trained using the training data in the line knnClassifier.fit(X_train, Y_train).

The training features, or independent variables, are represented by X_train.

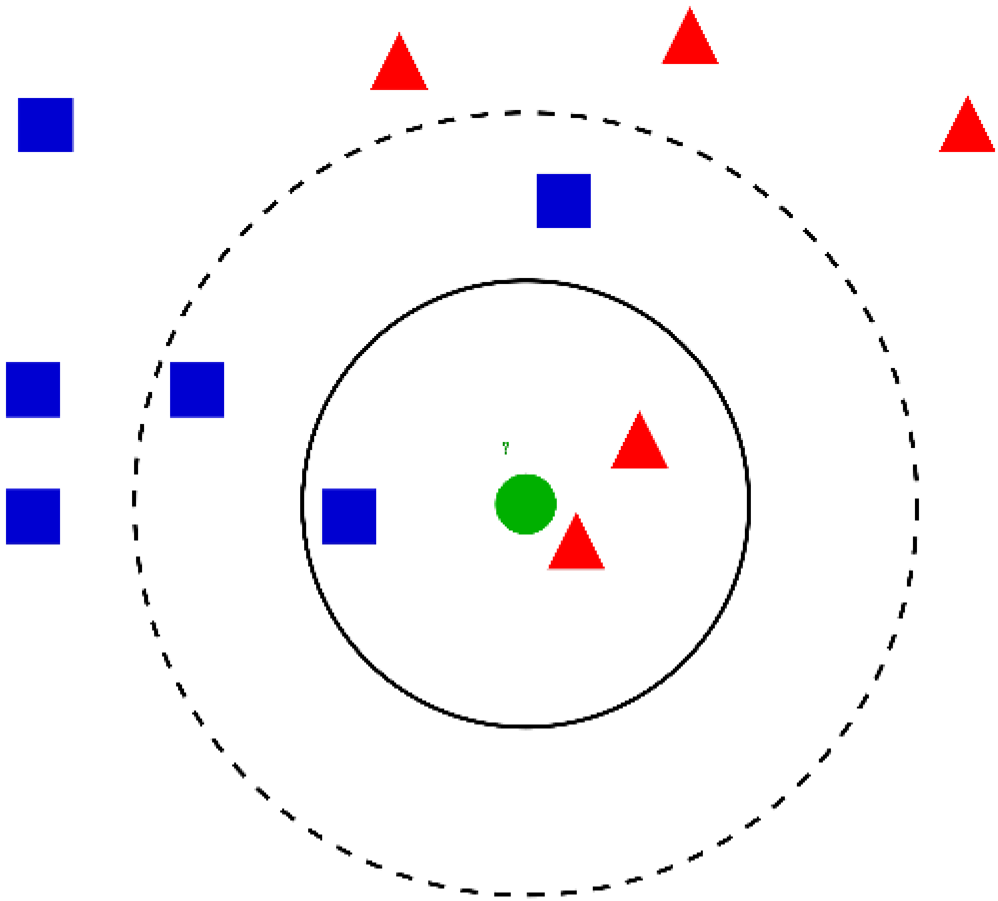
The training targets, or dependent variables, are represented by Y_train.

Forming Forecasts:

y_pred = knnPredictor.classifier(X_test): This line predicts the class labels for the test data points using the trained model.

The term "X_test" refers to the test features, or independent variables, for which the target class is to be predicted.

y_pred: This variable has the test data's anticipated class labels.



K-nearest Neighbour Architecture (Srivanstava, 2024)

3.2.9 MODEL TUNING AND EVALUATION

The hyperparameter was used in the machine learning algorithm to avoid overfitting and capacity for generalisation is known as model tuning. GridSearchCV was utilised in this work to tune the model, with recall serving as the primary metric. GridSearchCV ensures optimal model performance by systematically examining every conceivable combination of hyperparameters. Telecom businesses must measure recall, or the model's accuracy in identifying real consumers who have churned, to take appropriate action regarding retention.

In machine learning, Grid Search Cross-Validation (GridSearchCV) was used to determine the best hyperparameter to be used for each model. Hyperparameters that are established before to the model being trained, rather than being learned from the data.

GridSearchCV assists in automating the process of experimenting with every conceivable set of hyperparameters.

3.2.10 EVALUATION METRICS:

The major Evaluation Metrics used in this study are recall, precision, accuracy, and F1-score are the evaluation measures used.
Every one of them was examined for the different models. But as it will be preferable to reduce false negatives, the Recall Evaluation metric was selected as the primary evaluation method for this specific business case. Good loan was assigned a 0 (negative) during the conversion of the loan condition labels, and bad loan was assigned a 1 (positive). A false negative in this case would be a more serious mistake since I don't want a model to predict good loan when it should be bad loan (a circumstance where the algorithm predicted as a good loan but was a bad loan and the bank lost both money and interest). All the evaluation metric used are Accuracy: the proportion of samples that the classifier successfully classified to all samples provided in the test dataset.
F1-Score also called a balance F score the balance average of precision and recall.
Recall is the fraction of all the positive prediction out of the actual positive instance. The aim is to reduce the false negative.
Precision. The proportion of true positive predicted out of all positive prediction. The aim is to reduce false positive.
True Positive: we predicted as a good loan and it was a good loan, so the bank gained their money back along with interest.
False positive: predicted as a bad loan but was a good loan so the bank lost 40% interest.
False Negative: predicted as a good loan but was a bad loan and the bank lost both money and interest
True Negative: we predicted as a bad loan and it was bad loan, so the bank did not lose anything.

3.2.11 DEPLOYMENT:

The deployment is an essential part of the study. After developing and comparing different predictive model, the deployment is the implementation of the selected model into the business environment. This phase is necessary in the banking sector in order to identify good and bad loan by utilizing the model and predictions in the practical setting. In the deployment stage various considerations was made such as the existing business model, accuracy of the model, integrity of the model, data integrity, data flow and also the privacy and security. The stakeholder which includes the IT teams, business manager, data architect etc that have direct or indirect on the impact on the successful deployment of the model. The result of this thesis can be efficiently applied to the model to produce workable strategies that help banking sector increase loan prediction and the proactively increase interest rates.

3.3EXPLORATORY DATA ANALYSIS WITH POWER BI

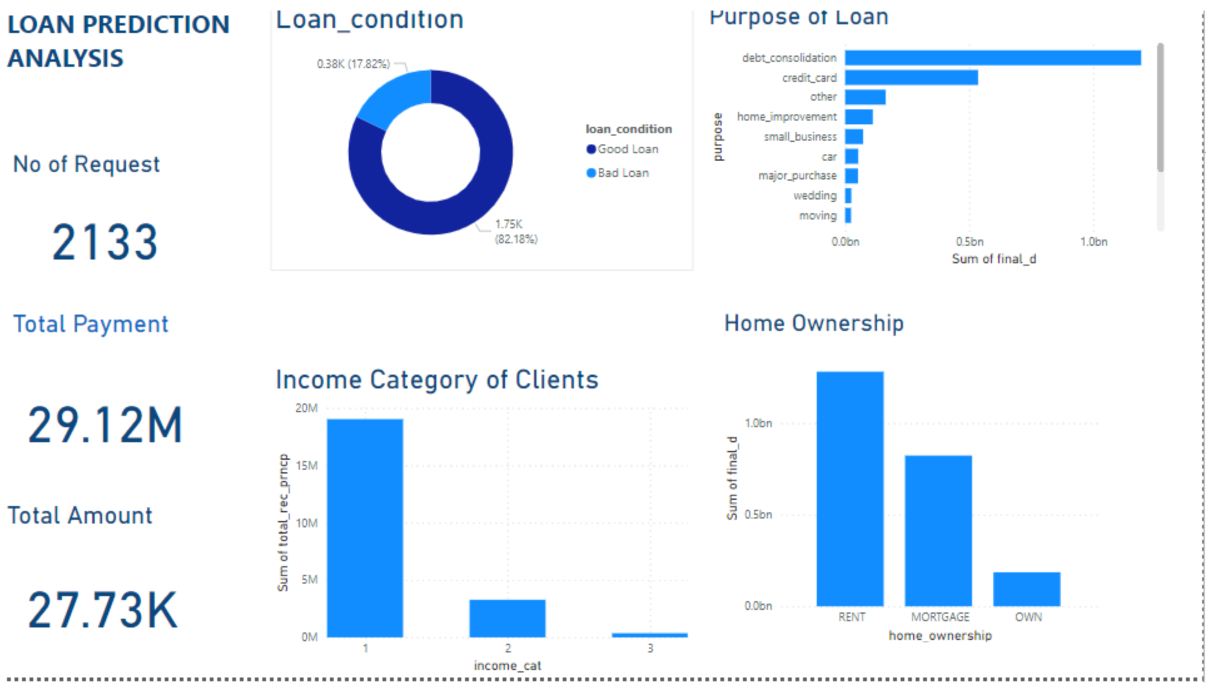
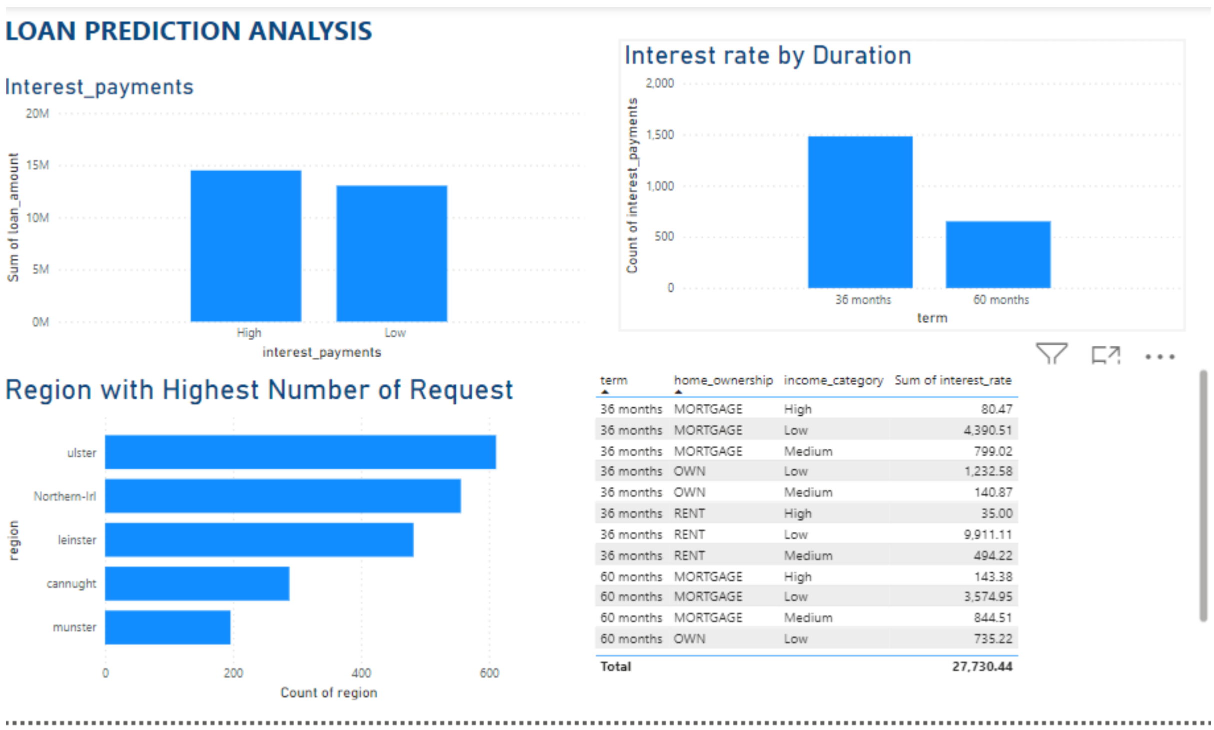


Figure 8: PowerBI visualization(EDA)



term	home_ownership	income_category	Sum of interest_rate
36 months	MORTGAGE	High	80.47
36 months	MORTGAGE	Low	4,390.51
36 months	MORTGAGE	Medium	799.02
36 months	OWN	Low	1,232.58
36 months	OWN	Medium	140.87
36 months	RENT	High	35.00
36 months	RENT	Low	9,911.11
36 months	RENT	Medium	494.22
60 months	MORTGAGE	High	143.38
60 months	MORTGAGE	Low	3,574.95
60 months	MORTGAGE	Medium	844.51
60 months	OWN	Low	735.22
Total			27,730.44

Figure 9: PowerBI Visualization(EDA)

The above image shows loan prediction analysis relating to the various variables of dataset which will give key insights and know areas that will need improvement. It shows that the total amount of loan that was loan out in the year in 2011. This is achieved using various visual chart like donut, key card, bar chart.
The total number of requests that was made was 2133 loan requested, and the total payment paid out from the bank was 29.12 million, and the total interest rate was 27.73k.
The pie chart shows that a number of loans that were requested for were good loans with 82.18% and the number of bad loans was just 17%. The statistic also hoes that the most requested reason or purpose of the loan was to buy House, other, debt consideration, home improvement, credit card etc followed respectively.
The chart also shows that income category of client with level 1 made more request for loan which the bank loan a total of 19 million to level one income category, the bank loan a total of 3,281,178.71 million to level two income category and the bank loan out a total of 359,587 thousand to level 3 income category. This shows that level three income category are led likely to asked for loan in the year 2011.
For home ownership. The chart shows that people that rent there home requested for more loan in 2011 than people that are mortgage or living in their own home. This was also proven by the purpose of loan chart which shows that house purpose was the highest reasons loan was requested for.
The second prediction analysis shows that when interest payment was high the loan amount obtained was €145,017,22 and when the interest rate was €130,380,72 was as the loan amount.
The interest rate by duration shows that the client prefers to make a short payback duration of 36months (payment plan) and that only a few clients payment plan was 60months. Region was the next, with the Ulster region making the most request for a loan with 611 requests. And Northern Irl made a request for a loan of 556, Leinster region requested for a loan was 482 requests. Cannught made a total of 288 requests, Munster had the lowest number of requests with 196 in the year 2011.
In the Metric chart, shows that interest rates gotten from the duration of the loan and the house ownership of the client and their income category. It shows that rented house owned with low-income rate and a duration of 36 months made a total interest of 9911.11 interest rate.
In conclusion, a good loan is predicted on the dashboard as a client with 36 months payment plan from which the interest rate is high- and the-income category is high.
Text Mining
These is the application of natural language processing and machine learning. These methods are used to undercover patterns, sentiments and trends that might not be immediately spotted by normal analysis. Text mining process large text data like letter to understand sentiments and relationships within documents. Text mining can be used to enhance context information. Text mining combine with statistical analysis can give insightful information.

CHAPTER 4

TOOLS, RESULT AND DISCUSSION

4.1 Random forest Algorithm result:

The Random forest algorithm showed a performance result on various evaluation metrics. It achieved and Accuracy of 0.97%. The precision was 1% and the recall was 0.80% respectively. The random forest can correctively identify true positive and false positive. F1-score which is the balance of precision and recall was 0.89%. The best estimator which is the number of trees was 10.

```
{'classification__n_estimators': 10}
0.6157894736842106
Accuracy:  0.9703125
[[544   0]
 [ 19  77]]
recall : 0.8020833333333334
percesion : 1.0
f1: 0.8901734104046243
```

[[544 0]],[[19 77]]. The confusion metric shows the model performance. It shows that the model correctly classified cases and the incorrectly cases. In the model, the predictive class were 544 and 0 were True positive, while the second row were predicted as Negative. 0 was the False Positive and 77 was the True Negative.
The positive outcome produces by Random Forest algorithm highlight. its potential to enhance loan prediction and identify good loan and bad loan in the banking sectors. This result has practical implication for the banking sector as they direct resources towards focused programme targeted as bad loan. However, ensemble model can be more difficult to understand than decision tree because it is a non-potential parametric model. The n-estimator parametric is 300 for the Random Forest model which indicates the number of decision tree used in the process. The model makes predictions by obtaining result from different trees, from the chosen data sample and then vote the best solution. The model as also shows the reliable estimate of feature important. The model works in the following ways.

- Randomly select the sample from the dataset.
- Create several Decision Tree for each sample and obtained the result prediction.
- Then voting is performed on all the predicted result.
- The best prediction is made by selecting result with most votes.

4.2Random Forest Algorithm Result with feature Importance list.

The top 10 features to identify as the most important feature from the feature importance to predict the loan application. The model Accuracy rate of 0.85% which indicates it ability to predict loan accuracy. The precision score suggests that while model successfully recognise false positive it cannot successfully predict false negative. This is not a good model performance.

Random Forest with the Feature Importance

total_rec_prncp	0.269079
final_d	0.133820
total_pymnt	0.124750
installment	0.111111
loan_amount	0.109952
interest_rate	0.045629
annual_inc	0.032090
term	0.031685
emp_length_int	0.025727
interest_payments	0.021985
home_ownership	0.017702
home_ownership_cat	0.015439

Figure 10: Relevant Features

The feature important list shows the relative significance of various attribute in predicting loan application. Total-rec-principal (0.269079) is the most variable, the second was Final_dat with a score of 0.133820,and the third most important feature was Total_payment 0.124750, the forth most important variable is Instalment with a score of 0.11111 and the fifth most important variable loan_amount is 0.109950. this suggest that the top variable are the strong indicators of loan prediction of a bad loan. Annual_income and terms had important scores of 0.032090 and 0.031685 respectively, this shows that a client yearly salary and duration of payment which is a term of 60 months or 36 months are Important in predicting loan application. Technically, emp_length_int and interest payment where other important variable used in predicting a loan application, there scoring was 0.025727 and 0.12985.
Features like Home_ownership and home_ownership_cat played less important role with scoring of 0.02527 and 0.21985 respectively. Another less important variable was purpose and region and income_category.

4.3 ADABOOST:

The result obtained from Adaboost algorithm shows good performance across the various metrics Accuracy is a broad indicator of the model's performance. The accuracy in this instance is 0.989, which is excellent. It indicates that 98.9% of the predictions made by the model were accurate.
Precision: This indicator counts the number of positive predictions made by the model that came true. With a precision of 1%, the model correctly predicted 1% of the positive predictions made by the model.
Recall: This metric counts the number of real positive cases that the model accurately predicted. A recall of 0.937 indicates that 93.7 percent of the real positive cases were properly identified by the model.
The F1-Score represents the harmonic mean of recall and precision. It's a method of striking a balance between the two measures to obtain a single indicator of the model's effectiveness. With an F1 score of 0.93, the model demonstrated good performance in

terms of recall and precision.
All things considered, the high accuracy, precision, recall, and F1-score values indicate that this model is operating effectively.

[[543 3],[[9 87]]. The model's performance is displayed by the confusion metric. It demonstrates how the model accurately and erroneously identified cases. The prediction class in the model was 543, with 3 being a true positive, and the second row being projected to be negative. The True Negative was 87, and the False Positive was 9.

4.4Feature importance for Adaboost:

The feature importance for Adaboost was slightly different from Random forest feature importance. In Adaboost total_rec_prncp was the most important variable 0.41 scoring and instalment variable was the next with a score of 0.30.

Loan_amount and final_d was also an important variable in Adaboost algorithm with a score of 0.10 and 0.06 respectively. Term and Interest_rate was also an important variable with a score of 0.02 and 0.02 each. But annual_income had less important role to play and it scoring was 0.02. other with less important role are interest_payment, total_payment, home_ownership_cat had low scores.

```
{'classification__n_estimators': 100}
0.9263157894736842
total_rec_prncp       0.41
installment           0.30
loan_amount           0.10
final_d               0.06
term                  0.02
interest_rate         0.02
annual_inc            0.02
region_2              0.01
interest_payments     0.01
total_pymnt           0.01
home_ownership_cat    0.01
region_0              0.01
emp_length_int        0.01
purpose_3             0.01
region_1              0.00
purpose_8             0.00
```

Figure 11 Adaboost feature performance

4.5 Support Vector Machine:

This model performance was fair. This model's drawback is that the model cannot be interpreted because it is non-parametric.
[[318, 226], [37, 59]]: This confusion metric. A classification model's performance is assessed using confusion matrices. They display the number of correctly classified cases and incorrectly classified cases.

The model's predictions for the positive class in this instance are probably represented by the first row. The model predicted 226 of the 318 real positive cases (True Positives) accurately, while misclassifying 37 of them as negative (False Negatives).
The model's predictions for the negative class are shown in the second row. The model

identified 226 false positives (False Positives) and correctly predicted 59 true negatives (True Negatives) out of 59 actual negative situations.
Accuracy: A value of 0.589, which is moderate accuracy.
Recall: 0.614584 In machine learning, recall is a metric that indicates how many real positive cases the model accurately predicted. Here, a recall of 0.615 indicates that 61.5% of the positive examples were picked up by the model.

Precision:0.20701754385964913 is the precision. Another parameter in machine learning that assesses how many of the model's positive predictions came true is called precision. With a precision of 0.207, the model only made accurate predictions in 20.7% of the positive predictions it made.

The F1-score: or harmonic mean of precision and recall, is expressed as f1: 0.962566844197862. When the classes are unbalanced, it provides a means of striking a balance between the two metrics and obtaining a single indicator of the model's performance. A high F1 score like 0.96 here suggests that the model is performing well on both precision and recall.

4.6 LOGISTIC REGRESSION:

Logistic Regression showed not so good performance when compared with all metrics used to evaluate the performance of the model. The breakdown of the evaluation metric.
 Accuracy: 0.5890625 - This metric indicates the proportion of correct predictions made by the model on the testing set. In this case, the model achieved an accuracy of roughly 59%.
 Precision: 0.20701754 - Precision reflects the ratio of true positives (correctly predicted loan conditions) to the total number of positive predictions. A low precision here suggests that the model makes many false positive predictions (predicting loan conditions that aren't true).
 Recall: 0.61458333 - Recall represents the proportion of true positives (correctly predicted loan conditions) to the total number of actual positive cases (loans with that condition). A high recall here means the model captures most of the true loan conditions.
 F1: 0.30971128 - F1 score is a harmonic mean between precision and recall, aiming to provide a balanced view of these two metrics
[[318, 226], [37, 59]] - This is a confusion matrix. Rows indicate actual classes, whereas columns represent expected classes. So, as an example:

The model correctly predicted the class for 318 of cases in the first actual class.
There were 226 of the first actual class that were wrongly projected to belong to the second class.
The model properly predicted 37 occurrences from the second real class, but wrongly forecasted 59 as belonging to the first class.
For the second class, the model exhibits high precision (20%) and accuracy (58%) but low recall (61%) for the same class and F1-score (30%). This indicates that the model

performs well in avoiding producing erroneous positive predictions for the second class.

However, it also overlooks many occurrences that are part of the second class.

4.7 K-NEAREST NEIGBOR:

For performing majority vote among the label k-nearest neighbours was used.
* Class 0(Good loan)'s f1-score is 0.92, recall is 0.88, and precision is 0.96. With 96% of predicted Class 0 being actual Class 0 and the model capturing 88% of all actual Class 0 data points, this indicates that the model performed well in detecting Class 0.
* Class 1(Bad loan)'s f1-score is 0.63, recall is 0.79, and precision is 0.53. By correctly detecting only 53% of projected Class 1 cases and capturing 79% of all actual Class 1 data points, the model demonstrated poorer performance on Class 1.

```
Classification              precision    recall  f1-score    support

          0       0.96       0.88       0.92       544
          1       0.53       0.79       0.63        96

   accuracy                             0.86       640
  macro avg       0.74       0.83       0.77       640
weighted avg      0.89       0.86       0.87       640
```

Figure 12: KNN Classification Report

4.8 DECISION TREE:

Decision tree. Had the following result
In the real Class 0, there were 476 data points that were accurately identified as Class 0 (predicted as Class 0).

In the real Class 0, there were 68 data points that were mistakenly assigned to the Class 1 category.
The real Class 1 contained 2076 data points that were appropriately categorised.
In the real Class 1, there were 14 data points that were mistakenly assigned to Class 0.

Measures:

The model's overall accuracy is 0.8625, which indicates that 86.25% of the time it properly classified cases.

Recall: The average recall for the two classes is 0.7917. The ratio of true positives to all actual positives is known as recall. In this instance, it indicates that the model generally captured. This indicates that, on average, 79.17% of the real positive cases were caught by the model.

Precision: The average precision for the two classes is 0.5278. The ratio of actual positive results to all projected positive results is known as precision. In this case, it indicates that, on average, 52.78% of the total instances that the model predicted as positive actually were.

The harmonic mean of recall and precision is used to create the F1-score, which combines the two metrics into a single score. It is 0.6333.

These measures (accuracy, recall, precision, and F1-score) are typically used in conjunction with a confusion matrix to assess a classification model's performance on a dataset.

4.9 COMPARISON OF MODEL PERFORMANCE:

MODEL	ACCURACY	PRECISION	RECALL	F1-SCORE
Random Forest with important features	**0.970**	**1**	**0.8020**	**0.890**
Adaboost	**0.987**	**0.96**	**0.90625**	**0.93554**
Support Vector Machine	**0.58**	**0.20**	**0.614**	**0.9354**
Logistic Regression	**0.5890**	**0.20**	**0.614**	**0.309**
KNN	**0.89**	**0.86**	**0.87**	**0.640**
DECISION TREE	**0.8625**	**0.5278**	**0.7916**	**0.633**

Table 2: Comparison of model performance

The model shows the various performance in predicting loan applications using various Machine Learning models. The first was Random forest and Adaboost which had a high Accuracy and Recall score. Random forest gave good result because the model used feature important to reduce the dataset and use only the important features. These fewer features affected the Precision as the model was slightly over fitting with a score of 1%. Support Vector Machine, Logistic Regression, KNN, Decision Tree also show good performance.

For this study **RECALL EVALUATION METRIC** was used because it is important to reduce the False Negative. False Negative here is a loan predicted as a good loan but was a bad loan and the bank loss the Interest and Capital. Hence False Negative is a worse error than False Positive So in order to reduce the false Negative, I assigned Good loan as 0 and Bad loan as 1. This was done because I wanted the model to highlight the bad loans, So I used recall evaluation metrics. Comparing the model performance Adaboost, KNN, SVM, Decision Tree also had better Recall performance. The Recall Score indicate that the model can capture most bad loan minimizing False Negative. Recall is important because it identify potential bad loan that will enable the bank to be proactive.

4.10 HYPER-PARAMETER TUNING

HYPER_PARAMETER TURNING

MODEL	HYPERPARAMETER	OPTIMUM VALUE	BEST RECALL PERFORMANCE SCORE
Random Forest	n-estimators	300	0.85
Adaboost	n-estimators	100	0.95
Support Machine Vector	C, Kernel	C: 0.001, Kernel: Ploy	0.99
Logistic Regression	alpha, eta0, l1_ratio, max_iter	alpha:0.001, eta0:0.001, l1_ratio:1, max_iter:100	0.94
KNN	n-estimators	5	0.86
Decision Tree	Criterion, max_depth,	Criterion: Entropy, max_depth: 20	0.90

Table 3: Hyper-Parameter Tuning

The optimum performance of Random Forest algorithm was n_estimators was set to [250,300,350,400,450]. The max feature ‘auto’ with a performance score 0.85 after the use of feature importance, this was achievable by using the to 10 important features. It was notice that adding the top features made the model to be slightly overfitting. For Adaboost the n_estimator was set [150, 200,250,300,350,400,450,500]. The optimum value was gotten was as 450 with performance score of 0.94.
Classification__eta0': [0.001, 0.01, 0.1, 1, 10, 100], and Classification__max_iter': [100, 500, 1000] were the hyperparameter settings for logistic regression.

'classification__l1_ratio': [0, 0.3, 0.5, 0.7, 1], 'classification__alpha': [0.0001,0.001, 0.01, 0.1, 1, 10, 100]
{'classification__alpha': 0.0001, 'classification__eta0': 0.001, 'classification__l1_ratio':

1, 'classification__max_iter': 100} were the optimal settings. The fact that the l1_ratio value is 1, as can be seen, indicates that the lasso regression penalty was applied by the model. To determine the memory performance score, 0.8761.

4.11TEXT MINING:

The dataset was a structured dataset with numerical data. To convert the data into format suitable for text mining, I identify the text column that contains textual data. Column like Purpose, loan_condition, term was considered. After identifying the columns, I clean the text data removing punctuation, converting all text to lowercase and removing stop words. The next was feature engineering, this is the creation of new features that indicates for example whether loan application letter mention ”debt consolidation ” in the purpose text. For the text mining technique, Sentiment Analysis was used to determine where a text was a good loan and bad loan.

CHAPTER 5
CONCLUSION AND FUTURE WORK

This study uses of traditional machine learning algorithm classifiers to compare the performance on predicting if a loan is a good loan or bad loan. while Support Vector Machine model perform well over all other in evaluation metrics. Recall which was the chosen evaluation metrics because we wanted to reduce the false Negative which are loans predicted as good loans but where bad loans and the company lost but their interest and capital.
The Random Forest and Adaboost model also performed well among other classifiers and one of the best classifiers. Random forest performance is a result of combining various Decision Trees and then voting is carried out on the best decision tree. The ensemble approach improve generalisation and prediction performance by reducing bias and variance. Adaboost is also an example of ensemble learning method that combine multiple weak classifiers to create stronger classifier.
This thesis uses machine learning approaches to provide a comprehensive and insightful comparative examination of bank loan prediction. The research project began with gaining an understanding of the problems the banking industry faced in connection to loan condition and the ensuing consequences for income, market competitiveness, and good loan. The CRISP-DM technique made it easier to create a methodical framework for data analysis, predictive model building, and efficacy evaluation. Many machine learning approaches, including Decision Trees, Support Vector Classifiers, Ada-boost, Random Forests, KNN, and Logistic Regression, were examined in the study. GridSearchCV was used to fine-tune the models using a variety of hyper-parameters. Additionally, memory was chosen as the primary assessment metric due to its importance in detecting trends in loan patterns and how to detect good loan and bad loan. The comparative study's findings have shed light on the relative advantages and disadvantages of each method for precisely predicting loan prediction. By combining many methods, the ensemble model method shown the ability to improve forecast accuracy while maintaining the ability to examine the model's results. The study identified the most accurate algorithms and produced important new information on the factors influencing loan decisions.

The recommendation is that the total_rec_principle which is the total amount of money that has been paid towards the loan. The higher the total-rec-pricpl the probability that the loan is a good loan the less interest rate which is significant to bad loan.
The thesis makes contributions by providing banking sector with useful recommendations for reducing that are based on advanced machine learning methods. With the help of this study, banks can better implement targeted good loans, which bad loan rates and fosters long-lasting loan prediction. However, it is crucial to recognise that there are certain limits to this research. This thesis, study loan predicting using machine learning but limited to banking industry. Finally, it is recommended that bank’s top 10 feature important are the feature that the model consideded in making it prediction both Random forest and Adaboost listed total_rec_priciple as the most important feature in predicting a good loan and bad loan.

References

Aboobyda Jafar Hamid, T. M. A., 2016. *DEVELOPING PREDICTION MODEL OF LOAN RISK IN BANKS USING DATA MINING.* [Online]
Available at: https://typeset.io/papers/developing-prediction-model-of-loan-risk-in-banks-using-data-47iguxfd5f
[Accessed 3 4 2024].

Aditi Kacheria, N. S. S. S. A. G., 2016. *Loan Sanctioning Prediction System.* [Online]
Available at: https://www.ijsce.org/wp-content/uploads/papers/v6i4/D2904096416.pdf
[Accessed 20 4 2024].

Agarwal Krishanu, J. M. K. K. A., 2022. *Comparing Classification Algorithms on predicting loan.* [Online]
Available at:
https://www.researchgate.net/publication/354375396_Comparing_Classification_Algorithms_on_Predicting_Loans?enrichId=rgreq-7bc8658a57a5caa1eb685aa1d8bca05e-XXX&enrichSource=Y292ZXJQYWdlOzM1NDM3NTM5NjtBUzoxMTQzMTI4MTExODU2NTg2N0AxNjc1Nzk1MDEyMzIy&el=1_x_2&_
[Accessed 4 4 2024].

Anant Shinde, Y. P. I. K. A. S., 2022. *Loan prediction using Machine Learning.* [Online]
Available at: https://www.itm-conferences.org/articles/itmconf/pdf/2022/04/itmconf_icacc2022_03019.pdf
[Accessed 5 4 2023].

Anon., n.d. [Online]
Available at: https://www.kaggle.com/datasets/vedato/bank-customers-classification

Anon., n.d. [Online]
Available at: https://www.kaggle.com/datasets/hashemialii/bankloan

Bhattacharya, A., 2023. *US banks are bracing for losses on bad loans by filling their coffers.* [Online]
Available at: https://qz.com/us-banks-are-bracing-for-losses-on-bad-loans-by-filling-1850620813
[Accessed 3 3 2024].

Calderon G Thomas, C. J. J., 2002. *A roadmap for future neural networks research in auditing and risk assessment.* [Online]
Available at:
https://www.sciencedirect.com/science/article/abs/pii/S1467089502000684
[Accessed 14 4 2004].

Dansana D, P. S. M. B. P. V. R. A. W. A., 2024. *Analyzing the impact of loan features on bank loan prediction using Random Forest algorithm,* india: Engineering Reports. john wiley and sons.

DataScience.com, 2020. *Support Vector Machine Theory.* [Online]
Available at: http://www.datasciencelovers.com/machine-learning/support-vector-machine-theory/
[Accessed 4 4 2024].

Huang Zan, C. H. H. C.-J. W.-H. c. W. S., 2004. *Credit rating analysis with support vector machines and neural networks: a market comparative study.* [Online]
Available at:
https://www.sciencedirect.com/science/article/abs/pii/S0167923603000861
[Accessed 5 04 2024].

International Joural of Creative Research Thoughts., 2021. *An Approach For Prediction Of Loan Approval.* [Online]
Available at: https://ijcrt.org/papers/IJCRT2106313.pdf
[Accessed 25 4 2024].

Lin Zhu, D. Q. D. E. C. Y. K. L., 2019. *A study on predicting loan default based on the random forest.* [Online]
Available at: https://pdf.sciencedirectassets.com/280203/1-s2.0-S1877050919X00186/1-s2.0-S1877050919320277/main.pdf?X-Amz-Security-Token=IQoJb3JpZ2luX2VjELb%2F%2F%2F%2F%2F%2F%2F%2F%2F%2FwEaCXVzLWVhc3QtMSJIMEYCIQDIpW7k54wczLs1jPkW7v8ZkQJAOyv0lo3gFtY5s%2FTE5AIhAKJRaeJ

CE5
[Accessed 2 3 2023].

Manum AL Miraz, F. A. M. M., 2022. *Predicting Bank Loan Eligibility Using Machine Learning.* [Online]
Available at: https://ieomsociety.org/proceedings/2022orlando/328.pdf
[Accessed 4 3 2024].

Martin VOJTEK, E. K., 2006. *Credit Scoring Methods.* [Online]
Available at: https://journal.fsv.cuni.cz/storage/1050_s_152_167.pdf
[Accessed 4 4 2024].

Min Li, A. M. T. S., 2018. *"Should This Loan be Approved or Denied?": A Large Dataset with Class Assignment Guidelines.* [Online]
Available at: https://www.tandfonline.com/doi/full/10.1080/10691898.2018.1434342
[Accessed 18 4 2024].

NADEEMFEROZI, M., n.d. *Loan Data.* [Online]
Available at: https://www.kaggle.com/datasets/mrferozi/loan-data-for-dummy-bank

Omar Mohib Ayad, A.-E. H. A. D., 2023. *A proposed model for loan approval prediction using XAI.* [Online]
Available at:
https://njccs.journals.ekb.eg/article_344392_78ac9c5216f909224d9850463a0cf44c.pdf
[Accessed 15 4 2024].

Pandit, A., 2016. *DATA MINING ON LOAN APPROVED.* [Online]
Available at:
https://www.cs.rit.edu/usr/local/pub/GraduateProjects/2161/asp3948/Report.pdf
[Accessed 4 4 2024].

Pidikiti Supriya, M. P. N. S. V. K. K. V., 2019. *Loan Prediction by using Machine Learning Models.* [Online]
Available at: http://www.ijetjournal.org/volume5/issue2/IJET-V5I2P28.pdf
[Accessed 2 4 2024].

Sheng-Tun LI, W. S. M.-H. H., 2006. *The evaluation of consumer loans using support vector machines.* [Online]
Available at:
https://www.sciencedirect.com/science/article/abs/pii/S0957417405001739?via%3Dihub
[Accessed 21 4 2024].

Siva, A., 2022. *Dynamic Prediction of Loan Approval using Machine Learning.* [Online]
Available at: https://ijrpr.com/uploads/V3ISSUE11/IJRPR7750.pdf
[Accessed 15 4 2024].

Srivanstava, T., 2024. *A Complete Guide to K-Nearest Neighbors (Updated 2024).* [Online]
Available at: https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/
[Accessed 4 5 2024].

Sruthi, 2014. *understand Random Forest Algorithim with example.* [Online]
Available at: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/
[Accessed 12 5 2014].

Sureshan, S., 2021. *Introduction to AdaBoost for Absolute Beginners.* [Online]
Available at: https://www.analyticsvidhya.com/blog/2022/01/introduction-to-adaboost-for-absolute-beginners/
[Accessed 5 4 2024].

Syed Zamil Hasan Shoumo, M. I. M. D. N. H. G. S. H., 2018. *Application of Machine Learning in Credit Risk Assessment.* [Online]
Available at:
https://dspace.bracu.ac.bd/xmlui/bitstream/handle/10361/11410/14201029,15101009,15101007,15101064_CSE.pdf;jsessionid=17EE507C239870965819CAD16D39F50F?sequence=1
[Accessed 02 04 2024].

Thorn, J., 2020. *Logistic Regression Explained.* [Online]
Available at: https://towardsdatascience.com/logistic-regression-explained-9ee73cede081
[Accessed 12 4 2024].

Wang Bao, N. L. K. Y., 2019. *Integration of unsupervised and supervised machine learning algorithms for credit risk assessment.* [Online]
Available at:
https://www.sciencedirect.com/science/article/abs/pii/S0957417419301472?via%3Dihub
[Accessed 18 4 2024].

X.Francis Jency, V. J. S. S., 2018. *An Exploratory Data Analysis for Loan Prediction.* [Online]
Available at: https://www.ijrte.org/wp-content/uploads/papers/v7i4s/E2026017519.pdf
[Accessed 4 4 2024].

Zahi Sara, A. B., 2020. *Modeling car loan prepayment using supervised machine learning*. [Online]
Available at:
https://www.sciencedirect.com/science/article/pii/S1877050920304920?via%3Dihub
[Accessed 21 4 2024].