An Ensemble Machine Learning Approach for Forecasting Credit risk of Loan Applications

C. L. PERERA, S. C. PREMARATNE
Information Technology Department, Faculty of Information Technology,
University of Moratuwa,
Katubedda, Moratuwa, 10400,
SRI LANKA

Abstract: - The business environment in Sri Lanka has become competitive with the development of the financial sector and the spread of the COVID-19 pandemic. The number of organizations and individuals applying for loans has increased. Lengthy authentication procedures are followed by financial institutes. However, there is no assurance whether the chosen applicant is the right applicant or not. Thus, this study proposed a methodology for assessing the credit risks associated with loans, to help make appropriate choices in the future. An Exploratory Data Analysis was performed to provide insights. This study focused on evaluating customer profiles based on the demographic and geographical data of the customers to forecast credit risks of loans using Machine Learning (ML) algorithms. Finally, the model performances were evaluated using evaluation metrics. The Stacking Ensemble outperformed the other techniques with the highest training and test accuracy of 0.99 and 0.78, respectively. The novelty of this study lies in performing a comprehensive data collection from a leading finance institution in Sri Lanka. The study highlights the importance of the choice of features, ML techniques, hyperparameters and evaluation criteria. Also, a novel ML technique, voting-based ensemble learning was proposed for enhancing performance.

Key-Words: - Authentication procedures, Credit risk, Ensemble Learning, Exploratory Data Analysis, Loan Applications, Machine Learning.

Received: May 15, 2023. Revised: August 26, 2023. Accepted: November 12, 2023. Available online: December 15, 2023.

1 Introduction

Due to intense competition at present, it is difficult for financial institutions to compete with each other to improve their overall business. Financial institutions have understood that customer retention and scam prevention must be tactical tools for strong rivalry, [1]. The accessibility of massive data, the formation of knowledge bases and the efficient use of data are helping financial institutions to open up effective delivery channels. Corporate choices can be improved through data mining and Machine Learning (ML), [2]. Customer segmentation, credit scoring and sanctions, forecasting loan amounts, improving stock portfolios, identifying deceitful transactions, and grading investments and promotions are some of the extents to which financial institutions can use data mining and ML techniques, [3].

Banks and financial institutions offer their customers various types of loans by lending money for specified periods at different interest rates, [4]. Loans can be broadly categorized into three types. The three types of loans are explained below.

1.1 Open-ended and Closed-ended Loans

Through open-ended loans, customers have the liberty to borrow money repeatedly, for example, using credit cards and credit lines subject to restrictions, which impose a limit on the maximum amount that can be borrowed at any instance, [4]. However, in the case of closed-ended loans, the customers have to settle the loans in full, to become eligible to borrow again; when a customer makes a repayment, the loan balance will decrease. Once the customer has settled the loan in full, if he/she wishes, he/she can apply for a fresh loan by submitting once again the full set of documents, required for checking his/her creditworthiness and obtaining the necessary

approvals, [4].

1.2 Secured and Unsecured Loans

In secured loans, collaterals, such as bonds, stocks, and personal assets, are accepted as guarantees. The cost of the assets offered as a guarantee is estimated before the loan is approved. If the debtor fails to recompense the loan, the creditor can seize the asset's ownership and recover the loan's balance amount. Two examples of secured loans are mortgages and auto loans. The borrowers of unsecured loans do not have to offer any assets as collateral. However, before approving the loan, the lender will assess the borrower's financial status to ascertain whether the borrower can repay the loan. Unsecured loans include education loans and personal loans, [4].

1.3 Conventional Loans

Conventional loans are not insured by any government organization. They have to conform to the rules set by Fannie Mae and Freddie Mac. However, non-conforming loans do not fulfill this requirement, [4].

Every day financial institutions obtain a vast number of credit requests from diverse customers. When approving a loan, the financial institutions initially authenticate their profile and documents, [4]. Figure 1 indicates the procedure of loan sanction, [4]. However, all loan applicants will not get the authorization of the financial institutions. Most financial institutions use their benchmarks of credit scoring models and risk evaluation practices when examining loan applications to decide whether to approve an application, [4].

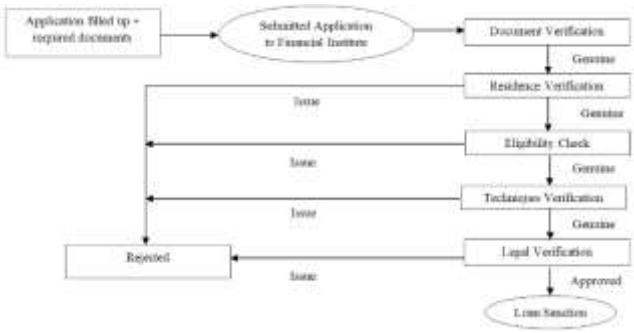


Fig. 1: Process for Loan Sanction

Various risks are associated with loan disbursements made by lenders. These risks include credit risks, which occur when the borrower does not repay the loan on time or when he/she does not pay it at all; liquidity risks, which occur when the lender faces a cash shortage after many customers have withdrawn large amounts of cash at short notice; and interest rate risks, which occur when the estimated interest rates are too low to earn Return on Investment (ROI), [5].

Lenders address these risks by assessing the creditworthiness and recompense ability of the borrowers, and the risks of loaning funds to them. Considering these assessments, lenders will estimate the amounts that can be lent to the borrowers, [5]. Risk management and measurement is every financial institution's core. Thus, the major challenge faced by financial institutions is the implementation of risk management systems to identify measure and control business exposure. There should be effective measures to identify and deal with these risks, based on advanced data mining and ML technologies.

The paper is structured as follows; Chapter I provides a general introduction to the study including the problem background. Chapter II indicates the literature review. Chapter III describes the involvement of ML and implementation technologies used to develop the models. Chapter IV presents a novel approach to predicting the credit risk of loan applications. Chapter V explains the implementation stage of the study. Chapter VI shows the research findings and evaluation. Chapter VII provides the conclusion and future works.

2 Literature Review

This section describes the exploration of the use of ML techniques in previous studies to analyze credit risk using ML techniques.

A study conducted in, [5], predicted loan approval or rejection of an applicant using Logistic regression, Decision Tree (DT) and Random Forest (RF) with input variables such as sex, marital status, education, dependents, earnings, loan amount, credit history and area of the property possessed. The best accuracy, 81.12%, was obtained with logistic regression. The Probability of Default (PD) on loan repayments was estimated in, [6], using K-Nearest Neighbor (KNN), RF, Artificial Neural Network (ANN) and Naïve Bayes (NB). RF demonstrated the best performance, with an accuracy of 0.998. The study, [7], used NB, DT, KNN, RF, Gradient Boosting, and other techniques to analyze loan repayment trends to predict

non-performing loans. The highest accuracy, 96.55%, was attained by RF.

A methodology to reduce the default risk was proposed in, [8], using DT, RF, Support Vector Machine (SVM), Linear Models, ANN, and Ada Boost ML techniques. The loan repayment ability was predicted by the study, [9], using Light Gradient Boosting (LGB), Multi-Layer Perceptron (MLP), RF, NB, and logistic regression. The best area in ROC curves was obtained by MLP. A study conducted in, [10], predicted the loan sanctioning process using logistic regression and algorithms such as DT, SVM, and NB. NB achieved the highest accuracy of 80.42%. The study, [11], performed credit categorization based on maturity period, credit spread and remaining credit, using KNN. In, [12], a prediction model for bank loan approvals was constructed using logistic regression, NB, and DT. NB achieved a higher accuracy of 80%. A methodology to predict the default risk of loan customers was presented in, [13]. The study used SVM, RF and Ensemble learning. Findings showed that the ensemble model gained the best results. Authors in the study, [14], developed SVM, DT, Bagging, Ada Boost and RF and compared the accuracy with Logistic Regression. Results revealed that RF and Ada Boost models achieved higher accuracy.

To investigate loan default, the study, [15], employed a DT as the base learner and contrasted it with ensemble learning strategies like RF, boosting, and bagging. The findings demonstrated that the ensemble model works better than individual models. The study, [16], used DT to predict loan sanctions. The best accuracy on the test set is achieved as 0.811. In the study, [17], classifiers based on ML and deep learning models were compared in predicting loan default probability. For this purpose, the most important features from various models were chosen. It was suggested that a financial institution develop an early warning system based on ML to help it increase its profitability. A new credit risk model was developed in the study, [18], using ordinal logistic regression (OLR) and increased the accuracy by using ANN, SVM and RF. The accuracy of the model improved from 68% using OLR to 82% when using ANN and above 90% when using SVM and RF. The PD on a loan was forecasted in the study, [19], using DT and RF. The RF algorithm yields the best predictive performance with an accuracy of 80%. The models produced by using a variety of training techniques, including one-step secant (OSS) backpropagation, Levenberg-Marquardt (LM) algorithm, scaled conjugate gradient (SCG) backpropagation, and an ensemble of SCG, LM, and OSS, were compared in the study, [20]. Findings revealed that training algorithms enhanced the loan default prediction model design, and ensemble models outperformed individual models. The study, [21], employed the k-Means algorithm to develop customer segmentation based on two features, the average amount of goods purchased by customers per month and the average number of customer visits per month. Four customer clusters were identified with 95% accuracy, namely, High-Buyers-Regular-Visitors (HBRV), High-Buyers-Irregular-Visitors (HBIV), Low-Buyers-Regular-Visitors (LBRV) and Low-Buyers-Irregular-Visitors (LBIV).

In the study, [22], binary classifiers were built based on ML and deep learning to predict loan default probability. The findings demonstrated that the tree-based models are more stable than the models based on multilayer artificial neural networks. The study, [23], estimated the PD on repayments of bank loans, using RF, NN, KNN and NB. The best predictive performance is obtained from the RF algorithm with an accuracy of 0.998. The financial status of an organization was forecasted in, [24], and it was discovered that the Tree Model for Genetic Algorithm is the best model with an accuracy of 81.75%. A methodology that combined the KNN, Binning, and NB algorithms was presented in, [25], to forecast the credible customers who have applied for loans. The C4.5 classification algorithm was employed in the study, [26], to estimate the risk percentage associated with lending. The study, [27], developed models using NB, J48 and Bagging algorithms to classify customers into ‘Safe’, ’More Safe’, ’Risk’ and ’More Risk’ categories. The bagging algorithm is best suitable for the credit risk with an accuracy of 85.84%. A loan credibility prediction system was proposed in the study, [28], to assist organizations in making the right decision to approve or reject the loan request of customers using the Decision Tree Induction Algorithm. The study, [29], used gradient boosting, DT, and logic regression to predict whether or not it would be safe to grant a loan to a specific individual. The best accuracy of 0.811 was obtained by gradient boosting. DT and ANN were used in the study, [30], to conduct a credit analysis. ANN attained the best accuracy of 97.07%.

3 Technologies Adopted in the Study

This section explains ML and implementation technologies to develop models, which were adopted for the study. Also, presents the usefulness of ML techniques that differentiate from the technologies applied in the existing literature.

3.1 ML Technologies Used to Develop Models

ML is a subcategory of AI which can be learnt from past data, builds the prediction models, and forecasts the output for it when it obtains new data. ML techniques used to develop models to predict credit risk are as follows;

1) Regression: It is a statistical technique used to build the relationship between dependent and independent variables. Equation (1) given below can be used to make the predictions using multiple regression.

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_iX_i \ldots b_kX_k + \varepsilon \quad (1)$$

where;
Y= Target variable
b_i = Polynomial coefficient of X_i
X_i = i^{th} independent variable
k = Number of independent variables
ε = Bias

2) Naïve Bayes: It is a supervised learning algorithm, based on the Bayes theorem. The equation for Bayes' theorem is given below;

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \quad (2)$$

where,
P(A|B): Probability of A given B
P(B|A): Probability of B given A
P(A): Probability of A happening
P(B): Probability of B happening

3) Decision Tree: It is a tree-structured classifier, where internal nodes denote the attributes of a dataset, branches denote the decision rules, and every leaf node denotes the result.
4) Random Forest**:** It is a supervised learning technique, which is based on ensemble learning. In this method, precision is improved, and overfitting is avoided. It forecasts considering majority votes of forecasts from each tree. In this method, precision is improved, and overfitting is avoided.
5) Artificial Neural Network (ANN): It is an adaptive system that varies the structure by the information transferring through the network in the learning stage. The feed-forward ANN shown in Figure 2 has three layers, composed of connected neurons. The three layers include the input layer which gets the external signal, hidden layers which process the internal operations, and the output layer which transfers the predictive outcome. The transfer function for a node is computed using Eq. (3).

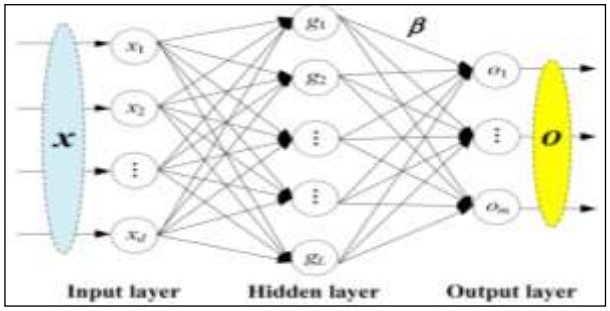


Fig. 2: Illustration of an ANN

$$Y = f\{\sum w\,x + b\} \quad (3)$$

where,
X = Input vector
Y = Output of the neuron
F = Transfer function of the neuron
W = Weight vector of the neuron
B = Bias of the neuron

6) Boosting Algorithms: The fundamental principle of functioning the boosting algorithm is to create several weak learners and integrate their predictions to form one strong rule (Figure 3).

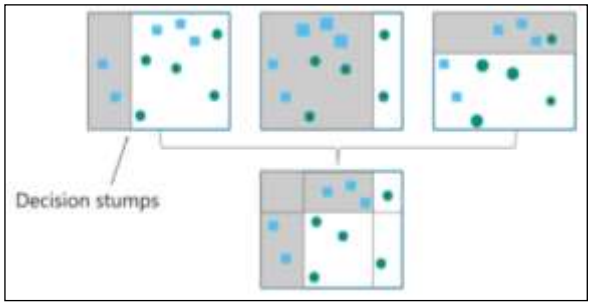


Fig. 3: Functioning of Boosting Algorithms

Categories of Boosting Algorithms

- Adaptive Boosting or Ada Boost

Ada Boost fits a series of weak learners on training data with different weights. It first forecasts and assigns the same weight to every outcome. If the forecast obtained from the first

learner is incorrect, then a higher weight is assigned. Learners are added until the number of models or accuracy reaches a limit.

- Gradient Boosting

Gradient boosting fits numerous models serially. Every model uses gradient descent to slowly minimalize the loss function of the whole system. The learning process constantly fits new models to offer more precise estimations of the target variable.

- Light Gradient Boosting

It is a gradient-boosting structure that exploits a tree-based learning algorithm. It is called "Light" because of its computational capability and efficient results. It needs less memory to run and is capable of handling large amounts of data.

7) Ensemble Learning Algorithms: It is a method which creates several ML models to explain a particular problem. The forms of ensemble learning are as follows;

I. Bagging Ensemble learning
The bagging ensemble is also known as Bootstrap Aggregation. Isolated models are trained with the bootstrapped samples and the forecasts of the sub-models are integrated to get the outcome.

II. Boosting Ensemble learning
Boosting ensemble trains the model and successive models are built considering the residual errors of the previous model. Then forecasts are ranked by accuracy and integrated to produce an outcome.

III. Voting Ensemble learning
A voting ensemble is built by binding the forecasts of the preceding models, which can be used to assign weights (Figure 4).

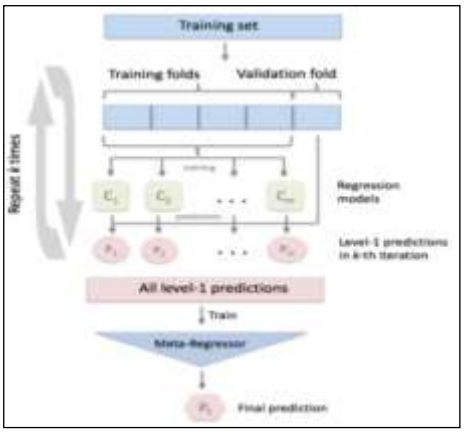


Fig. 4: Voting Ensemble learning

3.2 Implementation Technologies Used to Develop Models

Anaconda Software Distribution containing over 150 data science packages was used in the study. The packages were used to perform various ML tasks which are explained under the section on Implementation. Jupyter Notebook was the prime selection to write the Python code which enables running various experiments (Figure 5).

Fig. 5: Anaconda Navigator

Also, Power BI, a business analytics service by Microsoft was used to perform further analysis and provide interactive visualizations (Figure 6).

Fig. 6: Power BI

4 A Novel Approach for Forecasting the Credit Risk of Loans

This section explains the overview of the approach to predicting credit risk. In this scenario, the approach offered in this efficient and precise solution for the prediction of credit risk using ML techniques is highlighted.

4.1 Overview of the Novel Approach for Forecasting the Credit Risk of Loans

The study first explored the methods used by financial institutions in Sri Lanka to approve loans by analyzing the creditworthiness of loan applicants. It will identify the shortcomings of the methods and the obstacles faced by the institutions when implementing them.

Numerous cases are happening each year where debtors default the loan payments which causes financial institutions to bear huge losses. Therefore, Models were devised to evaluate the credit risks by evaluating customer profiles based on several aspects, such as demographic, geographic data of the customers and loan-specific data.

4.2 Conceptual Design

The CRISP-DM approach is used as the conceptual design to develop the models shown in Figure 7.

Fig. 7: CRISP-DM Approach

The stages are described as follows;

i. Business understanding: It is about getting to know the research background and how the study will accomplish the objectives.
ii. Data understanding: It needs the gathering of data intended for the study.
iii. Data preparation: It comprises preprocessing of the data.
iv. Modeling: It covers applying the modelling techniques.
v. Evaluation: The performances of the models are examined.
vi. Deployment: It is the deployment of the models.

4.3 Significance of the Study

Evaluating credit risk is important to a financial institution's achievement, as these aspects directly depend on profitability. Traditional techniques are incompetent and time-consuming. This study aims to explore the use of ML methods in predicting credit risk that are more robust and flexible.

In the study, various ML techniques such as Bagging Algorithms (DT and RF), Boosting Algorithms (Ada Boost, Gradient Boosting and Light Gradient Boosting) and ANNs were used to predict credit risk. Also, a novel ML method, voting-based ensemble learning was being used for enhancing performance.

The objective of the study is to exhibit the dominance of novel methods over conventional statistical models. It assesses and compares various ML techniques with ensemble learning techniques in predicting credit risk.

5 Implementation

This section explains the data collection, loading data and suitable libraries, data preprocessing, EDA and building prediction models.

5.1 Data Collection

This is the preliminary study which was carried out by performing a comprehensive data collection from a leading finance institute in Sri Lanka for the period 2010–2021, which consists of 169 branches located in 25 districts in Sri Lanka and 25 Facility Types. The collected data were related to the demographic and geographic data of the customers and loan-specific data.

5.2 Loading of Data and Libraries

The libraries include Pandas, a Python library, used to extract information from the dataset. For visualization, Matplotlib and Seaborne libraries were used to plot histograms and scatter plot graphs. Also, various other libraries were used to develop the models.

5.3 Feature Selection

The features were extracted based on evidence found in the literature and guidance provided by the business stakeholders. The features used in the study are as follows;

- Month: The month in which the loan is sanctioned.
- Year: The year in which the loan is sanctioned.
- Unit Price: The agreement price of one asset.

- Number of Equipments: The number of equipments in the loan/lease agreement.
- Period: The period of the agreement.
- Interest Rate: The interest rate of the agreement.
- Number of Rentals: The number of rentals of the agreement.
- Facility Amount: The total amount lent by the company to the borrower.
- Age: The age of the debtor.
- Gender: The gender of the borrower.
- Marital Status: The marital status of the borrower.
- Occupation: The occupation of the borrower.
- Facility Type: The type of the facility. E.g. LEASE, PERSONAL LOAN, VEHICLE LOAN etc.
- District: The district of the borrower.

The target variable is considered as the Customer Status (Active or Sink). Customers who were active on the maturity date were considered as 'Active' customers and those not active (either ceased or legal transfer) on the maturity date were considered as 'Sink'.

5.4 Data Preprocessing

The outliers were removed by the Interquartile range method. The box plot was obtained after removing outliers. Categorical features such as Branch, Facility Type, Gender, Marital Status etc. were converted into numeric values using Label Encoding. Feature scaling was done to convert the different scales of dimensions of variables into a single scale.

5.5 Exploratory Data Analysis (EDA)

Univariate, Multivariate and Correlation Analyses were performed under this.

5.6 Prediction Modeling

First, the input and the target variables were defined. The dataset was split into training, and test by setting the ratio of 60% - 40%. Then the models were devised using ML techniques, described in Chapter III Section A.

Logistic Regression was performed to interpret a linear classification. Then a Gaussian Naïve Bayes Classification is performed, to interpret a probabilistic classification. Next two bagging algorithms (Decision Tree and Random Forest Classifications) and boosting algorithms (AdaBoost, Gradient and Light Gradient Boosting) were performed. Then an ANN was devised with five hidden layers with specified hidden neurons, and each was added with the ReLU activation function. Adam optimization method is used to increase performance and reduce training time. The regularization technique of randomly dropping neurons during training was used to prevent neurons from co-adapting too much.

Finally, a Stacking ensemble classification was implemented using an algorithm of stacking or Stacked Generalization. The ensemble model was defined by a list of tuples for the four base models which are Random Forest, Gradient Boosting, Ada Boost and Light Gradient Boosting. Then the Logistic Regression was defined as the meta-model combining the predictions from the base models using 5-fold cross-validation.

5.7 Evaluation of Model Performance

The performance of the models developed was evaluated by using the following evaluation metrics. The terms used in classification metrics are introduced.

- TP: True Positive - Both Predicted and Actual are True
- TN: True Negative - Both Predicted and Actual are False
- FP: False Positive - Predicted True but Actual is False
- FN: False Negative - Predicted False but Actual is True

The classification metrics are explained as follows;

A. Accuracy
The accuracy of the model is the total number of correct predictions divided by the total number of predictions.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (4)$$

B. Mean Squared Error (MSE)
MSE is one minus the accuracy score.

$$MSE = 1 - \text{Accuracy Score} \quad (5)$$

C. Precision
The precision of a class defines how reliable the result is when the model predicts that a point belongs to a class.

$$\text{Precision} = TP / TP + FP \quad (6)$$

D. Recall
The recall of a class defines how well the model can predict a class.

$$\text{Recall} = TP / TP + FN \quad (7)$$

E. F1 Score
The F1 score of a class is given by the harmonic mean of precision and recall.

$$\text{F1 Score} = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (8)$$

F. ROC AUC Score
ROC AUC Score is the area under the ROC (Receiver Operating Characteristic) curve obtained by plotting the True Positive Rate against the False Positive Rate. The higher the area covered, the better the model will be.

G. Precision-Recall AUC Score
Precision-Recall AUC Score is the area under the curve of a Precision-Recall curve obtained by plotting Precision against Recall. The higher the area covered, the better the model will be.

6 Experiment Results and Analysis

This section presents the research findings and evaluation of EDA and prediction models developed.

6.1 Findings of EDA

- Distribution of Percentage of Total customers vs. Facility Types
  The top ten facility types which attracted the most customers are presented in Figure 8. Lease drew the greatest number of customers, accounting for 78.07% of the total customer base.

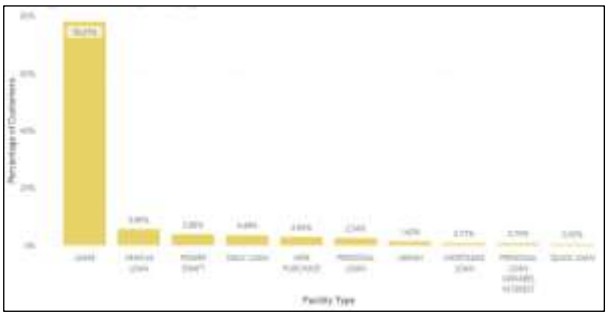


Fig. 8: Percentage of Total Customers vs. Facility Type

- Distribution of Percentage of Total customers vs. Average interest rate and Month
  Figure 9 shows the Loan Customers and average interest rate vs. Month and the highest number of loans has been disbursed in July due to the lowest average interest rate of 23.6%.

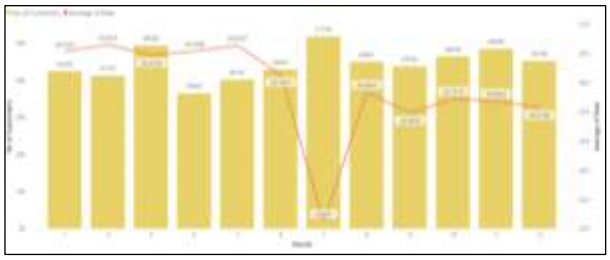


Fig. 9: Percentage of Total customers vs. Average interest rate and Month

- Distribution of Percentage of Total customers vs. Year
  The greatest number of loans were disbursed in 2017 as a result of the targeted promotional activities carried out in that year (Figure 10).

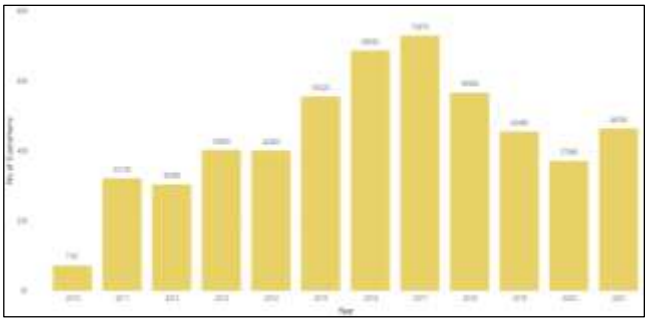


Fig. 10: Percentage of Total Customers vs. Year

- Distribution of Percentage of Total customers vs. Customer Status and Gender
  Figure 11 indicates that 52.68% and 13.67% of total sink customers are males and females respectively. Also, 26.66% and 6.98% of total active customers are males and females respectively.

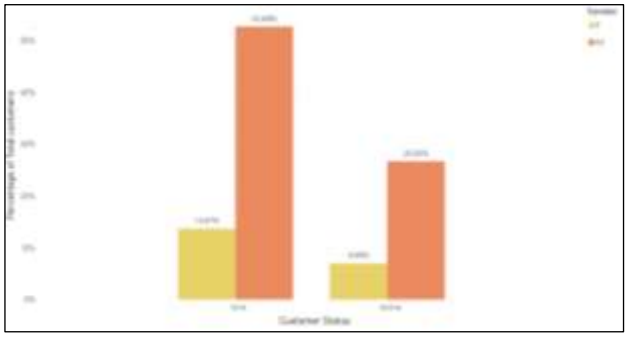


Fig. 11: Percentage of Total Customers vs. Customer Status and Gender

- Distribution of Total customers vs. Gender and Age
  As illustrated in Figure 12, more loans are awarded to males in the 40 – 60 age range.

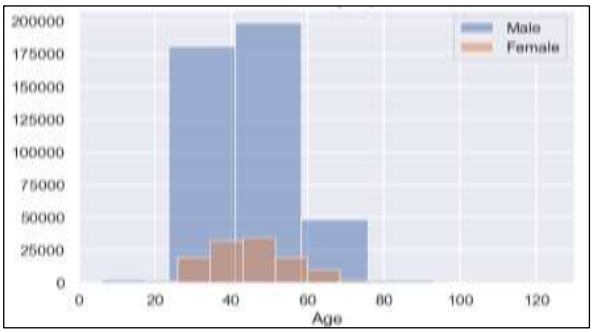


Fig. 12: Distribution of Age of Loan Customers according to Gender

- Distribution of Percentage of Total customers vs. occupation
  As shown in Figure 13, of the total number of loan customers, the proportion of the loan customers working in the service, agriculture, and trade sectors account for 41.16%, 15.75 and 10.7%, respectively.

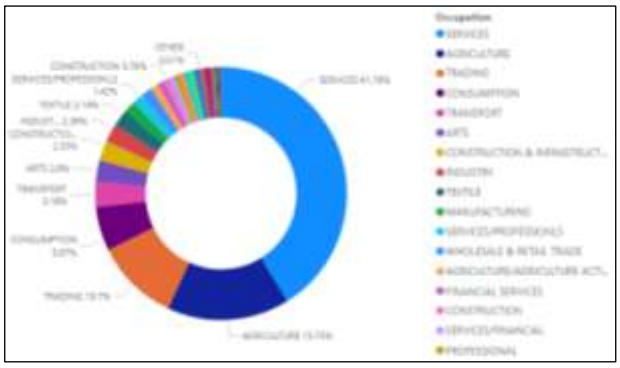


Fig. 13: Number of Loan Customers vs. Occupation

- Distribution of Percentage of Total customers vs. Marital Status
  According to Figure 14, the percentage of married and single sink customers are 54.63% and 8.33%, respectively. Furthermore, of the active customers, 31.09% are married, and 4.50% are single.

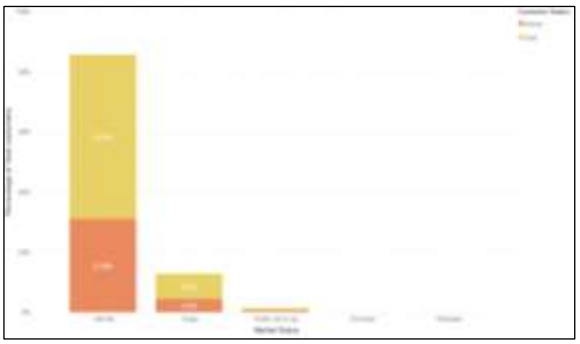


Fig. 14: Percentage of Total Customers vs. Marital Status

- Distribution of Percentage of Total customers vs. District
  According to Figure 15, 19% of loans are disbursed to customers who reside in the Colombo district, with Gampaha (11.98%) and Galle (7.64%) districts following closely behind.

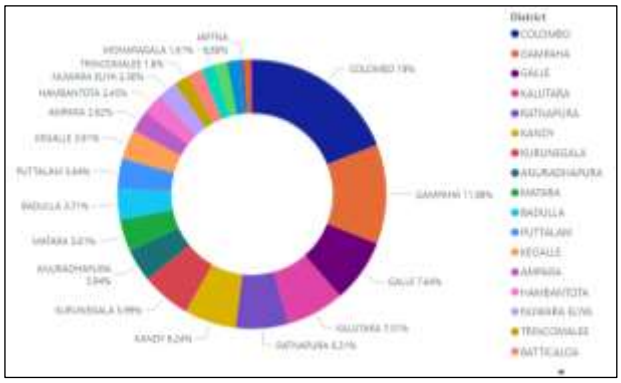


Fig. 15: Percentage of Total Customers vs. District

6.2 Findings of Models Developed

1) Predictions by Logistic Regression Model

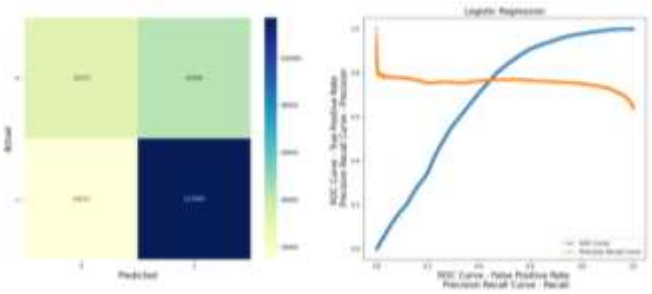
The Logistic Regression performance metrics are presented in Table 1.

Table 1. Model statistics obtained using Logistic Regression

Evaluation Metric	Training set	Test set
Accuracy	0.72	0.72
Mean Squared Error	0.28	0.27
Precision	0.73	0.73
Recall	0.88	0.88
F1-score	0.80	0.80
ROC AUC Score	0.66	0.66
Precision-Recall AUC Score	0.76	0.76

Table 1 shows that the accuracy of Training and Test sets is 0.72, 72% of total predictions are correctly predicted. With a precision of 0.73 for both Training and Test sets, 73% of results can be reliable. Recall of Training and Test sets is 0.88, which means the model is 88% satisfactory in predicting a class. With an F1 score of 0.80 for both the Training and Test sets, the model is interpreted as better quality. ROC AUC score of Training and Test sets is 0.66, which explains that the model is 66% precise in distinguishing between the Active and Sink customers. Precision-Recall AUC Score of Training and Test sets is 0.76, the model exhibits a good balance between precision and recall.

Figure 16 A. shows the Confusion Matrix of the Logistic Regression Model. Type-I Error (FP) is 42898 and Type-II Error (FN) is 14522. Figure 6 B. shows the ROC and Precision Recall curves of the Logistic Regression Model.



A B

Fig. 16: A. Confusion Matrix and B. ROC and Precision-Recall curves of Logistic Regression Model

2) Predictions by Naive Bayes Model

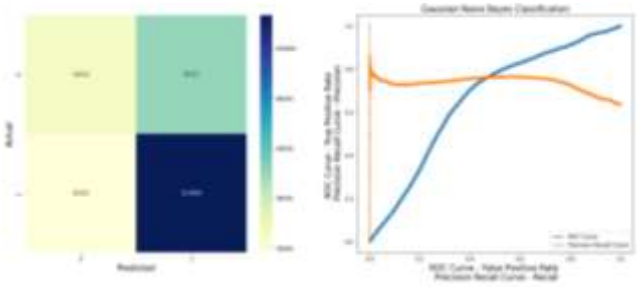
The Naive Bayes classification performance metrics are presented in Table 2.

Table 2. Model statistics obtained using Naïve Bayes Classification

Evaluation Metric	Training set	Test s set
Accuracy	0.66	0.66
Mean Squared Error	0.33	0.33
Precision	0.69	0.69
Recall	0.86	0.86
F1-score	0.76	0.76
ROC AUC Score	0.59	0.59
Precision-Recall AUC Score	0.73	0.73

According to Table 2, an accuracy of 0.66 for both Training and Test sets means 66% of total predictions are correctly predicted. With a precision of 0.69 for both Training and Test sets, 69% of results are reliable. Recall of Training and Test sets is 0.86, the model is 86% satisfactory to predict a class. With an F1 score of 0.76 for both the Training and Test sets, the model is interpreted as better quality. ROC AUC score of Training and Test sets is 0.59, which explains that the model is 59% precise in distinguishing between the Active and Sink customers. Precision-Recall AUC Score of Training and Test sets is 0.73, the model exhibits a good balance between precision and recall.

Figure 17A shows the Confusion Matrix of the Naive Bayes Model. Type-I Error (FP) is 50357 and Type-II Error (FN) is 18182. Figure 17B shows the ROC and Precision Recall curves of the Naive Bayes Model.



A B

Fig. 17: A. Confusion Matrix and B. ROC and Precision-Recall curves of Naive Bayes Model

3) Predictions by Bagging Models

The Bagging model performance metrics are presented in Table 3.

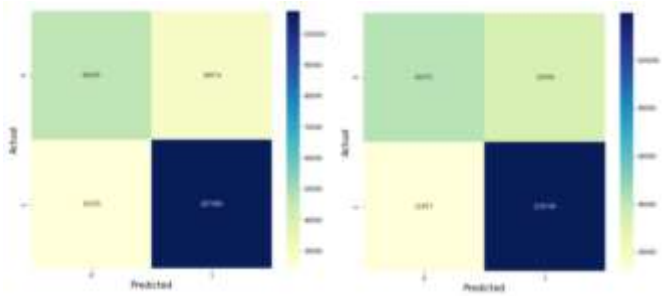
Table 3. Model statistics obtained using Bagging Classification Models

Evaluation Metric	Decision Tree Classification		Random Forest Classification	
	Training set	Test set	Training set	Test set
Accuracy	0.98	0.74	0.98	0.77
Mean Squared Error	0.04	0.25	0.03	0.22
Precision	0.77	0.77	0.77	0.77
Recall	0.81	0.81	0.90	0.90
F1-score	0.80	0.80	0.83	0.83
ROC AUC Score	0.72	0.72	0.73	0.73
Precision-Recall AUC Score	0.83	0.83	0.85	0.85

The accuracy of the training sample for both DT and RF is 0.98, as shown in Table 3. On the other

hand, RF's test accuracy of 0.77 is higher than DT. In both classification models, the Precision of Training and Test sets is 0.77, which means 77% of results are reliable. A higher Recall of 0.90 was obtained by RF, indicating that the model is 90% satisfactory in predicting a class. RF obtained a higher F1 score of 0.83 indicating that the model is of higher quality. RF gained a higher ROC AUC score of 0.73 showing the model is 73% precise in distinguishing between the Active and Sink customers. The RF Precision-Recall AUC Score of 0.85 is greater than the DT score, indicating that the model shows a good balance between precision and recall.

Figure 18 shows the Confusion Matrix of DT and RF Models. In DT, Type-I Error (FP) is lower than RF which was 28874. However, in RF Type-II Error (FN) is lower than DT which was 11857.



A B

Fig. 18: Confusion Matrices of A. Decision Tree and B. Random Forest Models

Figure 19 indicates the ROC curves of Decision Tree and Random Forest Models. According to Figure 19, ROC curves of RF covered a higher area than DT, implying that RF is a superior model to DT.

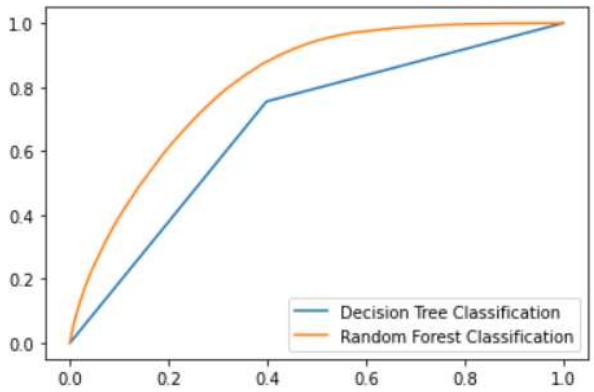


Fig. 19: ROC curves of Decision Tree and Random Forest Models

4) Predictions by Boosting Models

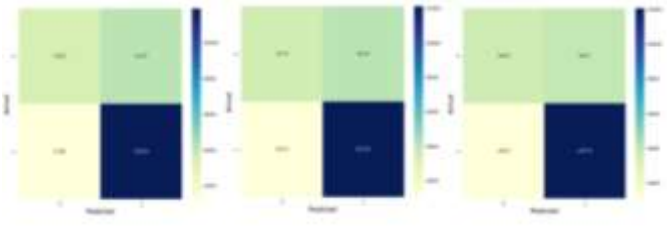
The Bagging model performance metrics are presented in Table 4.

Table 4. Model statistics obtained using Bagging Classification Models

Evaluation Metric	Ada Boost		Gradient Boosting		Light Gradient Boosting	
	Training set	Test set	Training set	Test set	Training set	Test set
Accuracy	0.74	0.74	0.75	0.75	0.76	0.76
Mean Squared Error	0.25	0.25	0.24	0.24	0.23	0.23
Precision	0.74	0.74	0.75	0.75	0.76	0.76
Recall	0.90	0.90	0.91	0.91	0.91	0.91
F1-score	0.81	0.81	0.82	0.82	0.82	0.82
ROC AUC Score	0.68	0.68	0.70	0.70	0.70	0.70
Precision-Recall AUC Score	0.79	0.79	0.82	0.82	0.83	0.83

Higher accuracy and precision of 0.76 were achieved by the LGB model for both the Training and Test sets, as shown in Table 4. A higher Recall of 0.91 was obtained by both Gradient Boosting and LGB, indicating both the models are 91% satisfactory to predict a class. An F1 score of 0.82 was gained by both Gradient Boosting and LGB, interpreting the better quality of both models. A higher ROC AUC score of 0.70 was obtained by both Gradient Boosting and LGB, explaining that the models are 70% precise in distinguishing between the Active and Sink customers. A higher Precision-Recall AUC Score of 0.83 was gained by the LGB model, indicating that the model shows a good balance between precision and recall.

Figure 20 shows the Confusion Matrix of Bagging Classification Models. The lowest Type-I Error (FP) and Type-II Error (FN) were gained by the LGB model, which was 38927 and 10818, respectively.



A B C

Fig. 20: Confusion Matrices of A. Ada Boost B. Gradient Boosting and C. Light Gradient Boosting

Figure 21 indicates the ROC curves of Bagging Classification Models. According to Figure 21, the ROC curves of LGB covered a higher area than the Ada Boost and Gradient Boosting models, implying that LGB is the superior model to others.

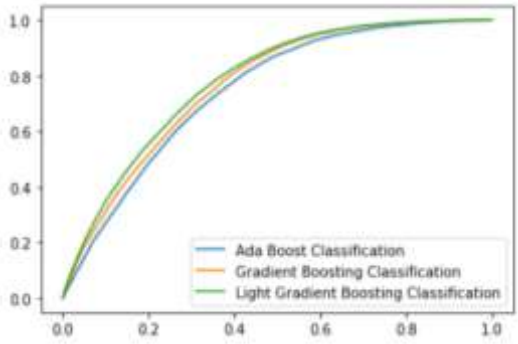


Fig. 21: ROC curves of Bagging Classification Models

5) Predictions by Neural Network Model
A neural network is devised by tuning the hyperparameters such as five hidden layers with specified hidden neurons and each is added with the ReLU activation function. The loss and accuracy values during training are presented in Figure 22.

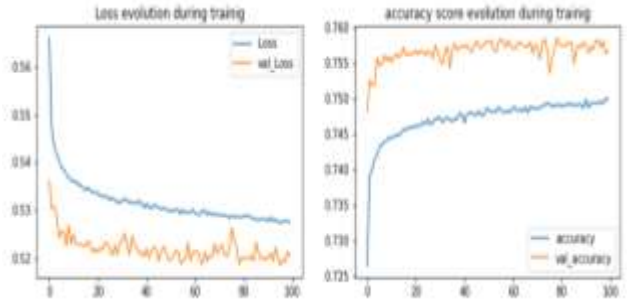


Fig. 22: Loss and Accuracy scores during training

Table 5. Model statistics obtained using the Neural Network Model

Evaluation Metric	Training set	Test set
Accuracy	0.76	0.76
Mean Squared Error	0.22	0.23
Precision	0.75	0.75
Recall	0.91	0.91
F1-score	0.83	0.83
ROC AUC Score	0.74	0.74
Precision-Recall AUC Score	0.84	0.84

According to Table 5, the accuracy of Training and Test sets is 0.76 which means 76% of total predictions are correctly predicted. The precision of Training and Test sets is 0.75, which means 75% of results are reliable. Recall of Training and Test is 0.91, which means the model is 91% satisfactory to predict a class. The F1 score of the Training and Test sets is 0.83, which interprets the better quality of the model. ROC AUC score of Training and Test sets is 0.74, which explains that the model is 74% precise in distinguishing between the Active and Sink customers. Precision-Recall AUC Score of Training and Test sets is 0.84, which means that the model has a good balance between precision and recall.

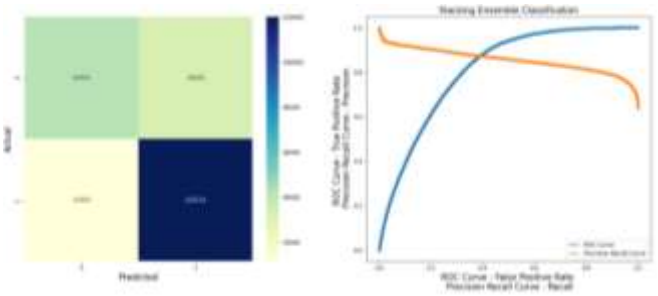
6) Predictions by Stacking Ensemble Model
A Stacking Ensemble model was implemented using an algorithm of stacking or Stacked Generalization. It combines the predictions from multiple well-performing machine learning models. The Stacking Classifier model was first defined by a list of tuples for the four base models which are Random Forest, Ada Boost, Gradient Boosting and Light Gradient Boosting Classifiers, and the meta-model as the Logistic Regression.

Table 6. Model statistics obtained using the Stacking Ensemble Model

Evaluation Metric	Training set	Test set
Accuracy	0.99	0.78
Mean Squared Error	0.01	0.21
Precision	0.7	0.78
Recall	0.92	0.92
F1-score	0.84	0.84
ROC AUC Score	0.75	0.75
Precision-Recall AUC Score	0.86	0.86

According to Table 6, the accuracy and precision of Training and Test sets is 0.78 which means 78% correct predictions of total predictions. Recall of

Training and Test sets is 0.92, which means the model is 92% satisfactory in predicting a class. The F1 score of the Training and Test sets is 0.84, which interprets the better quality of the model. ROC AUC score of Training and Test sets is 0.75, which explains that the model is 75% precise in distinguishing between the Active and Sink customers. Precision-Recall AUC Score of Training and Test sets is 0.86, which means that the model has a good balance between precision and recall.



A B

Fig. 23: A. Confusion Matrix and B. ROC and Precision-Recall curves of Stacking Ensemble Model

The confusion matrix obtained from the Stacking Ensemble model is indicated in Figure 23A. Type-I Error (FP) and Type-II Error (FN) are 33030 and 11459, respectively. ROC Curve obtained from the Stacking Ensemble model as indicated in Figure 23B. The ROC AUC Score, which indicates a True Positive Rate against a False Positive Rate, for training and test sets were 0.75. Precision-Recall AUC Score is the area under the curve of a Precision-Recall curve obtained by plotting Precision against Recall, for training and test sets was 0.86. It covered a greater area, which means it is a better model than others.

6.3 Performance Evaluation of Models Developed

Comparison of Training and Test accuracies of the models developed are shown in Figure 24. It shows that the Stacking Ensemble model outperforms the individual models with statistical significance with a training and test accuracy of 0.99 and 0.78, respectively.

Fig. 24: Training and Test accuracies of models developed

Precision, Recall and F1-score of models developed models are presented in Figure 25. It shows that the Ensemble model outperforms the individual models with statistical significance with Precision, Recall and F1-score of 0.78, 0.92 and 0.84, respectively.

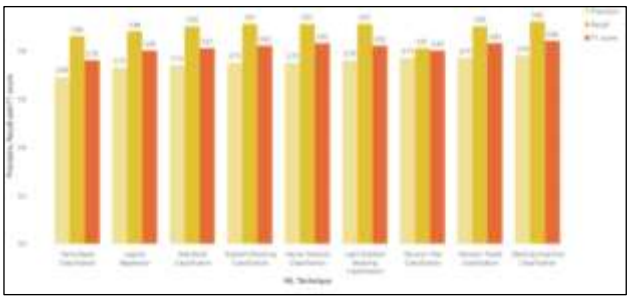


Fig. 25: Precision, Recall and F1-score of models developed

Figure 26 shows the Training and Test Mean Square Error (MSE) of developed models. It shows that the Stacking Ensemble model has the lowest training and test MSE of 0.01 and 0.21, respectively, which indicates that the model is a better fit than the other models.

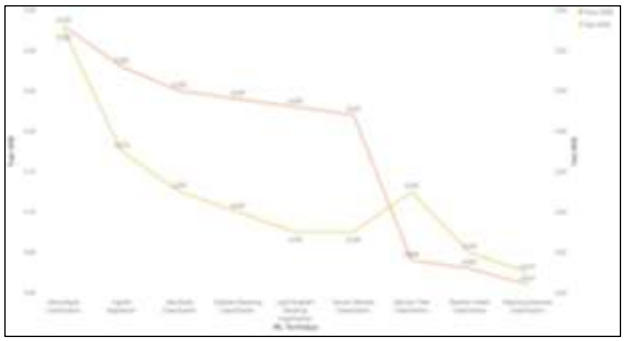


Fig. 26: Training and Test Mean Square Error (MSE) of models developed

ROC curves of developed models are demonstrated in Figure 27. It shows that the Stacking Ensemble model covered a greater area, which means it is a better model for classifying Active and Sink customers than others.

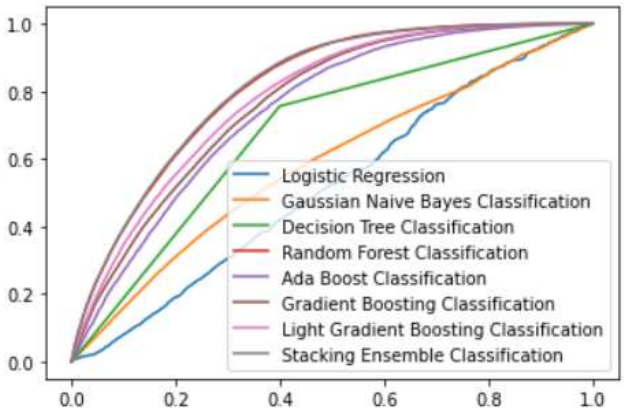


Fig. 27: ROC curves of models developed

ROC AUC Score and Precision-Recall AUC Scores of models developed are presented in Figure 28. Higher ROC AUC Score and Precision-Recall AUC Scores of 0.86 and 0.75 respectively are gained by the Stacking Ensemble model, implying it outperforms the other models.

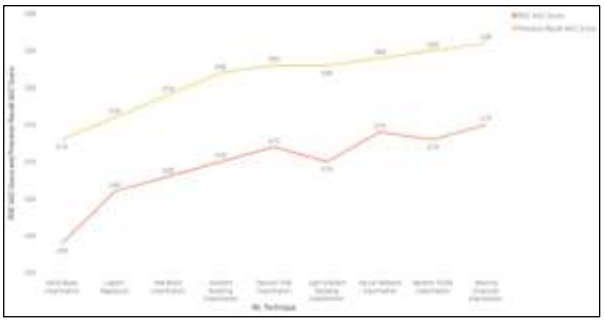


Fig. 28: ROC AUC Score and Precision-Recall AUC Scores of models developed

7 Conclusion

Financial institutions are observed to be increasingly shifting to AI and ML techniques to manage credit risks, financial fraud, money laundering, regulatory risks and customer behavior that can lead to potential revenue losses, etc. Further, it was found that financial institutions could make appropriate decisions or actions in advance regarding these risks, provided the risks associated with loans can be predicted.

Thus, this study proposed a methodology for the evaluation of the credit risks of loans. An extensive data collection was conducted from a leading finance institute in Sri Lanka. Data preprocessing including feature extraction and detection of outliers was performed as vital stages in developing models. An EDA was performed to assess the loan customers for developing marketing strategies and identifying the type of customers who can be approached.

Based on the study findings, it can be indicated that, with 78.07% of the loan portfolio, lease facilities attracted the most customers. The highest number of loans were found to be disbursed to customers in the Colombo district. A higher proportion of loans, 85% of the total loans extended to males between 40-60 years and to married people. Of the total number of loan customers, the proportion of the loan customers working in the service, agriculture and trade sectors account for 41.16%, 15.75 and 10.7%, respectively.

Also, this study focused on a variety of ML techniques to forecast credit risk using Regression, bagging algorithms, boosting algorithms and ANN. Furthermore, the study employed a novel approach, voting-based ensemble learning, which involves several learners trained to forecast the credit risk resulting in a better predictive accuracy than could have been obtained from any of the individual learning models alone. Findings of the comparison of these techniques were used to select the best model that reveals more prominent benefits in the context of predicting credit risk. The model findings suggested that the Stacking Ensemble Classification outperformed the other ML techniques with the highest training and test accuracy of 0.99 and 0.78 respectively, with a lesser MSE of 0.21. The contributions of the study can be used to help financial institutes estimate the credit risk associated with the loan applications they receive to make better decisions regarding loan approval, prevent internal and external fraud, anticipate customer behavior to prevent them from leaving and drawing them with new specially designed loan products etc.

The study may be expanded to a higher level in future by applying further advanced learning algorithms, feature reduction methods and hyperparameter tuning to further improve the model performance. Since the work-study data from only one financial institution was used, it is recommended that further studies be conducted by gathering data from different financial institutions across the country to capture the insights.

References:

[1] F. X. Jency, V. P. Sumathi, and J. S. Sri, "An exploratory data analysis for loan prediction based on nature of the clients," *International Journal of Recent Technology and Engineering*, 2018, vol. 7, no.4.

[2] A. H. Jafar and T.M. Ahmed, "Developing prediction model of loan risk in banks using data mining," *Machine Learning and Applications: An International Journal*, 2016, vol. 3, no. 1, pp. 1–9.

[3] B. Kaur and P. K. Sharma, "Implementation of Customer Segmentation using Integrated Approach," *The International Journal of Innovative Technology and Exploring Engineering*, 2019, vol. 8, no. 6.

[4] S. Vimala and K. C. Sharmili, "Prediction of loan risk using naive Bayes and Support Vector Machine," *International Conference on Advancements in Computing Technologies*, 2018, vol. 4, no. 2, pp. 110-113.

[5] A. Kadam, P. Namde, S. Shirke, S. Nandgaonkar and D. R. Ingle, "Loan Credibility Prediction System using Data Mining Techniques," *International Research Journal of Engineering and Technology*, 2021, vol. 8, no. 5.

[6] P. M. Subia and A. C. Galapon, "Sample model for the prediction of default risk of loan applications using data mining," *International Journal of Scientific &Amp; Technology Research*, 2020, vol. 9, no. 6.

[7] Y. C. Widiyono and S. M. Isa, "Utilization of data mining to predict non-performing loan," *Advances in Science, Technology and Engineering Systems Journal*, 2020, vol. 5, no. 4, pp. 252–256.

[8] K. Arun, G. Ishan, and K. Sanmeet, "Loan approval prediction based on machine learning approach," *IOSR Journal of Computer Engineering*, 2016, vol. 18, no. 3, pp. 79–81.

[9] Y. Wang, J. Xiaomeng and W. Zihan, "Loanliness: Predicting loan repayment ability by using machine learning methods," Stanford.edu, 2021, [Online]. http://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26644913.pdf (Accessed Date: December 12, 2023).

[10] B. E. Chandra and R. Rekha, "Exploring the machine learning algorithm for prediction of the loan sanctioning process," *International Journal of Innovative Technology and Exploring Engineering*, 2019, vol. 9, no. 1, pp. 2714–2719.

[11] S. Fazlollah, M. Houman and S. Stanford, "Classifying a Lending Portfolio of Loans with Dynamic Updates via a Machine Learning Technique," *Mathematics*, 2021, vol. 9, no. 17.

[12] V. S. Kumar, A. Rokade and M. S. Srinath, "Bank loan approval prediction using data mining technique," *International Research Journal of Modernization in Engineering Technology and Science*, 2020, vol. 2, no. 5.

[13] A. Goyal and R. Kaur, "Loan prediction using ensemble technique," *International Journal of Advanced Research in Computer and Communication Engineering*, 2016, vol. 5, no. 3.

[14] M. C. Aniceto, F. Barboza and H. Kimura, "Machine learning predictivity applied to consumer creditworthiness," *Future Business Journal*, 2020, vol. 6, no. 1.

[15] A. Chopra and P. Bhilare, "Application of ensemble models in credit scoring models," *Business Perspectives and Research*, 2018, vol. 6, no. 2, pp. 129–141.

[16] R. P. Kathe, S. D. Panhale and P. P. Avhad, "An approach for prediction of loan approval using machine learning algorithm", *International Journal of Creative Research Thoughts*, 2021, vol. 9, no. 6.

[17] M. Lakhani, B. Dhotre and S. Giri, "Prediction of credit risks in lending bank Loans," *International Research Journal of Engineering and Technology*, 2018, vol. 5, no. 12.

[18] C. Balakrishnan and M. Thiagarajan, "Credit risk modelling for Indian debt securities using machine learning", *Buletin Ekonomi Moneter Dan Perbankan*, 2021, vol. 24, pp. 107–128.

[19] M. Madaan, A. Kumar, C. Keshri, R. Jain and P. Nagrath, "Loan default prediction using decision trees and Random Forest: A comparative study", *IOP Conference Series: Materials Science and Engineering*, 2021, vol. 1022, no. 1, p. 012042.

[20] A. K. I. Hassan, A. Abraham "Modeling Consumer Loan Default Prediction Using Ensemble Neural Networks," *International Conference on Computing, Electrical and Electronic Engineering*, 2013, https://doi.org/10.1109/ICCEEE.2013.6634029.

[21] C. Pascal, S. Ozuomba and C. Kalu, "Application of K-means algorithm for efficient customer segmentation: A strategy for targeted customer services," *International Journal of Advanced Research in Artificial Intelligence*, 2015, vol. 4, no. 10, http://doi.org/10.14569/ijarai.2015.041007.

[22] P. Addo, D. Guegan and B. Hassani, "Credit risk analysis using machine and Deep Learning Models", *Risks*, 2018, vol. 6, no. 2, p. 38.

[23] F. Martinelli, F. Mercaldo, D. Raucci and A. Santone, "Bank credit risk management based on data mining techniques," *Proceedings of the 6th International Conference on Information Systems Security and Privacy*, 2020, DOI: 10.5220/0009371808370843.

[24] G. Anchal and K. Ranpreet, "Accuracy Prediction for Loan Risk Using Machine Learning Models," *International Journal of Computer Science Trends and Technology*, 2016, vol. 4, no. 1.

[25] K. Aditi, S. Nidhi, S. Shreya and G. Archana, "Loan Sanctioning Prediction System," *International Journal of Soft Computing and Engineering*, 2016, vol. 6, no. 4.

[26] S. Mrunal, T. Pooja, S. Priya, S. Swati and P. Sandip, "Data Mining Techniques to Analyses Risk Giving Loan (Bank)", 2016, vol. 2, no. 1.

[27] B. Yogita and A. More, "Comparative analysis of classification based data mining algorithms for credit risk analysis," *International Journal of Engineering & Scientific Research*, 2018, vol. 6, no. 2.

[28] M. S. Sivasree and S. T. Rekha, "Loan credibility prediction system based on decision tree algorithm," *International Journal of Engineering Research & Technology*, 2015, vol. 4, no. 9, pp.825-830.

[29] S. Pidikiti, P. Myneedi and S. Nagarapu, "Loan Prediction by using Machine Learning Models," *International Journal of Engineering and Techniques*, 2019, vol. 5, no. 2.

[30] M. M. Sousa and R. S. Figueiredo, "Credit analysis using data mining: Application in the case of a Credit Union," *Journal of Information Systems and Technology Management*, 2014, vol. 11, no. 2, pp. 379–396.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

- C. L. Perera contributed to all stages of the present study including formulation of the problem, research goals and aims, literature review, design of methodology, data collection, data visualization/presentation, model development, final findings, conclusions and future works.
- S.C. Premaratne provided constant guidance and leadership throughout the planning and execution of the current study.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflict of interest to declare that is relevant to the content of this article.