A Comparative Study between Machine and Deep Learning Models for the Prediction of Bank Credit Recovery

BY

NAZRE IMAM TAHMID
ID: 181-15-10839

NASIMUL HAQUE
ID: 181-15-10989
AND

MD. UMAR FARUQUE
ID: 181-15-10848

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Ahmed Al Marouf
Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Shah Md. Tanvir Siddiquee
Assistant Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

4 JANUARY 2022

APPROVAL

This Project titled "**A Comparative Study between Machine and Deep Learning models for the Prediction of Bank Credit Recovery**", submitted by Nazre Imam Tahmid ID No: 181-15-10839, Nasimul Haque ID No: 181-15-10989 and Md. Umar Faruque ID No: 181-15-10848 to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on – 4 January, 2022.

BOARD OF EXAMINERS

Dr. S.M Aminul Haque (SMAH) **Chairman**
Associate Professor and Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Raja Tariqul Hasan Tusher (THT) **Internal Examiner**
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Md. Sazzadur Ahamed (SZ) **Internal Examiner**
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Shamim H Ripon **External Examiner**
Professor
Department of Computer Science and Engineering
East West University

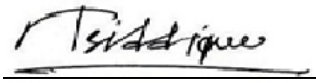
DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Ahmed Al Marouf, Senior Lecturer, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Ahmed Al Marouf (AAM)
Senior Lecturer
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised by:

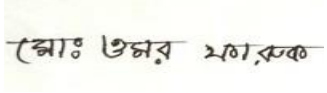


Shah Md. Tanvir Siddiquee (SMTS)
Assistant Professor
Department of Computer Science and Engineering
Daffodil International University

Submitted by:

Nazre Imam Tahmid
ID: 181-15-10839
Department of Computer Science and Engineering
Daffodil International University

Nasimul Haque
ID: 181-15-10989
Department of Computer Science and Engineering
Daffodil International University



Md. Umar Faruque
ID: 181-15-10848
Department of Computer Science and Engineering
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound indebtedness to Supervisor **Ahmed Al Marouf**, **Senior Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of Machine Learning and Deep Learning to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to Dr. Touhid Bhuiyan, Professor and Head, Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Nowadays, technology is advancing rapidly. With the advancement of technology, many institutions are adapting their business with new technologies. Institutions have huge amounts of data about the employers and clients. To handle huge amounts of data, many institutions are applying many data mining techniques to maintain their institutions properly and fast. In financial institutions like banks, to analyze and handle the data about the customer is very necessary. To analyze the credit risk is a primary field in the banking sectors and there are many techniques exist to predict whether a customer is credit worthy or not and the possibility of loan default. In this research, we've used a dataset from a Bangladeshi bank. The dataset is the credit defaulter dataset. We tried to predict the delinquent customers who have the highest possibility of short term credit recovery. We applied some machine and deep learning models to predict the credit recovery. The dataset is imbalanced. First of all we balanced the dataset by using SMOTE technique and then we performed feature scaling, feature selection process on the dataset. Finally, we applied machine and deep learning models. Compared with all of the models, Random Forest (RF) performed better than other models. We applied those models in both Train Test Split and Stratified K-Fold CV methods. In the Train Test Split method, RF gives 93% accuracy and in the Stratified K-Fold CV method, RF gives 94% accuracy. The result of the evaluation and statistical metrics of this model are also good in both of these methods. In the case of deep learning models, the best output comes from Artificial Neural Network (ANN) and Multilayer Perceptron (MLP) with 90% accuracy. Overall RF performed better and can better predict the credit recovery.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Introduction

The banking sector is an important part of a country's economy and plays a vital role in economic development. As a developing country, Bangladesh has improved in many sectors in the recent few years. But the banking system still has some problems. Generally, banks sanction the loan application to a customer by checking the customer's creditworthiness. They use many techniques for it like traditional methods, credit scoring models, machine learning models, etc. If the customer scores are in satisfactory level they sanction loans to them. However, few loans become classified which is also called NPL (Non-performing Loans) for many reasons. There are several reasons behind it like lack of accurate documentation, wrong client selection, weak governance, political force to sanction loans, fake information, violation of rules and regulations, etc. Bangladesh as well as other south Asian countries have faced this problem. It becomes an acute problem for Bangladeshi banks. The percentage of classified loans is 8.18% of the total outstanding loans in Aug 2021 [1]. The ratio of classified loans has fluctuated in the recent few years. Due to the economic downturn situation in COVID-19, the country's NPL in banking sectors have risen by Tk 6,351.29 crore in the first quarter of 2021 as the clients of the bank were unable to repay their loans [2]. The percentage of NPL for several years is illustrated in figure 1.1.1. A higher percentage of classified loans hampered the country's banking sector. The government has taken some steps to reduce it but it is still uprising. So, reducing the classified loans becomes a crying need.

Plenty of researches have been conducted to find customer creditworthiness, classified loan prediction, find credit groups etc. but there is little work to determine the potential customer who can repay the loan and have the significant probability to overcome classified loans in near future. That's why we conduct this research to find the possible customers to recover the credit by using machine and deep learning techniques.

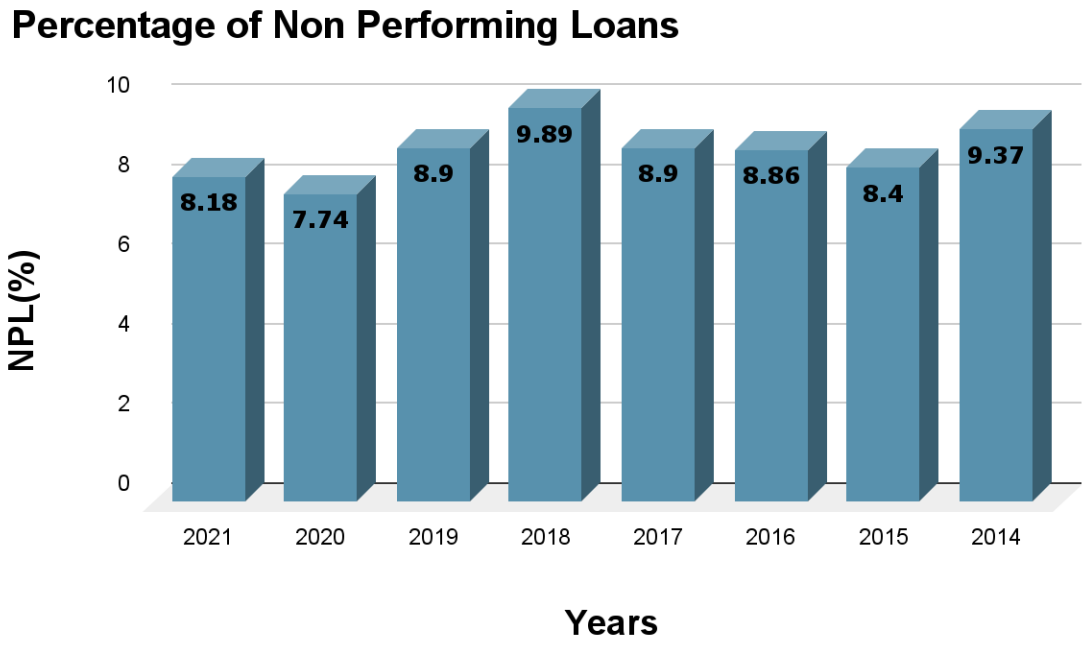


Figure 1.1.1: Percentage of NPL to total gross loans

For this problem, we've used a credit defaulter dataset that is collected from a bank data warehouse. We preprocessed the dataset and find the important features. We also handled the imbalance dataset by using smote oversampling methods. After that, we applied supervised machine and deep learning models to determine the credit recovery possibility for a given customer. We applied both train test split and Stratified K-fold cross-validation methods to see how the models work by using these techniques. Finally, we made a comparison between machine and deep learning models and identified the best-performing model.

1.2 Motivation

The banking system helps to develop the economy of a country. In Bangladesh, the average percentage of classified loan (Non-Performing Loan) is higher than in other countries. This is a severe problem for the Bangladeshi banks and also is an obstacle to economic development. Likewise, other South Asian countries are also experiencing the same problem. If we can identify the possible customers who have higher possibility to repay

the loan in near future then it will be very helpful to reduce the NPL ratio and the bank can be benefited from this. So, it has motivated us to do this project.

1.3 Rationale of the Study

Banks provide finance to many customers or organizations in short terms or long terms. The customer is supposed to repay the money with interest. But due to many reasons, some loans become classified. NPL creates a significant problem in the banking sector for running a smooth banking system. To solve this problem, we can use machine and deep learning techniques to predict that the credit will either be good or bad. This means finding the customers who have a higher possibility to repay the loan. So, using machine and deep learning models we can predict credit recovery.

1.4 Research Questions

Research questions are very important to get the proper guideline for a research paper or thesis. It gives us a proper idea about the work, process and purposes.

RQ1: Can we predict the credit recovery from a loan defaulter dataset?

Yes, we can predict the credit recovery from a loan defaulter dataset using machine and deep learning techniques.

RQ2: Why did we handle the imbalanced dataset?

In our dataset, the distribution of samples of each class are not equal. This is a credit defaulter dataset and has the majority of information about those clients who have bad credit. So, most of the samples belong to the bad credit class and a few samples belong to the good credit class. So, the dataset is imbalanced. If we train a model with this dataset then it might happen in the future that a model with better accuracy will just predict an outcome as bad credit class most of the time and the bank can face problems when sanctioning a new credit to the customer. The model won't be able to predict both classes properly in the future. This is not good for a model and the bank. So, if we handle the imbalance dataset by oversampling the minority class then the model will be able to train itself with same number of samples for each class and it will give a good predictive

outcome and the performance metrics such as Precision, Recall and F1-Score will also be good for both classes. That's why we handled the imbalanced dataset.

RQ3: Why did we perform feature scaling and feature selection for this dataset?

This dataset contains the numeric data. There are a huge number of values in most of the feature columns and the features are in different scale. Models can't learn too well and fast with a huge number of values and different scaled features. So, feature scaling techniques scale the data and make it smaller with the mean 0 and standard deviation of 1. So, all the features come into the same scale and the model performs much better and learns fast.

In this dataset, there are 49 input features and all of the input features are not needed for a model to give a good performance. Models can't learn fast if the model is trained with all of the input features. So, we built a LightGBM model to find the important features and dropped the redundant features. Finally, we trained the other machine and deep learning models with those important features. So, to make the training faster and get a good performance of each model, we performed feature scaling and feature selection process for this dataset.

RQ4: Which algorithm is performing better for the loan defaulter dataset to predict credit recovery?

In this study, we applied the Train Test Split and Stratified K-Fold CV method to each of the machine and deep learning models. Compared with each of the methods and models, we've seen that Random Forest (RF) performed better and gives a better prediction. The result of the performance evaluation and statistical metrics of this model is also good enough compared to each of the models. This model performed better in both of the methods. In the Train Test Split method, RF gives 93% accuracy which we've got accuracy of 94% for this model in the Stratified K-Fold CV method. So, overall Random Forest (RF) algorithm is performing better for this dataset to predict credit recovery.

1.5 Expected Output

We want to apply machine and deep learning models in our dataset and compare the predictive result and also want to identify the best-performing model to predict credit recovery. We also want to make a comparison between Train Test Split and Stratified K-

Fold CV methods to see how each of the machine and deep learning models perform in each of these methods.

1.6 Project Management and Finance

We didn't have to spend any money and didn't have to buy any software or hardware tools to do this project.

The amount of time that we spent on the project activity is given in the following table:

Table 1.6.1: Project Management Timeline

Task	Times
Data Collection and preprocessing	3 months
Literature Review	2 months
Experimental Setup	2 months
Experiments and validation	3 months
Report	2 months
Total	12 months

1.7 Report Layout

Chapter 2 contains the background section which follows terminologies, related works, comparative analysis, the scope of problem and challenges. Research methodology part includes data collection procedure, statistical analysis, proposed methodology and implementation in Chapter 3. The experimental setup, experimental result & analysis and discussion are discussed in Chapter 4. In Chapter 5 we discussed the impact on society, ethical aspects and sustainability of NPL. In Chapter 6, we concluded the thesis with the research summary, conclusion recommendation, and implications for the future.

CHAPTER 2
BACKGROUND

2.1 Preliminaries/Terminologies

For any developing country like Bangladesh, the banking sector is very important. In developing countries, many developing works going on and they also have many growing up industries and companies. For this, developing countries need to be strong in the development of economy. The problem here is that sometimes banks don't get return of the credit they provided to the customer. The bank faces financial loss for this problem and it also resists the economic development of a country. Our research purpose is finding those delinquent customers who have the higher possibility to recover the credit in the short term. Here we used different machine learning models (Logistic regression, Random Forest, K-Nearest Neighbour, Decision Tree, Support Vector Machine, Gaussian Naive Bayes, Bernoulli Naive Bayes, XGBoost, AdaBoost and Linear Discriminant Analysis) & deep learning algorithms (Artificial Neural Network, Multilayer Perceptron, Convolutional Neural Network, Recurrent Neural Network, Long Short Term Memory and Gated Recurrent Unit) on this dataset to predict credit recovery.

2.2 Related Works

Lots of researchers used different machine and deep learning algorithms to predict whether a customer is eligible for taking credit or not. Many of them conducted research on the credit risk of the bank. But there is a little research about the credit recovery of customers. All of the research is quite similar with the process and the applied algorithms. Some of them are described below:

Mohammad Rajib Pradhan et al. worked on performance evaluation of traditional classifiers on prediction of credit recovery. In their research, they used Bangladeshi bank dataset and they performed feature scaling, feature selection and GridSearchCV on their dataset. They found that random forest gives better accuracy (90%) than other classifier models like XgBoost(89%), logistic regression (87%), decision trees (85%), and support vector machines (87%). The model with lowest accuracy is Naive Bayes (24%). [3]

Aquib Abtahi Turjo et al. worked on comparative analysis and implementation of credit risk prediction. The dataset they used in their research was collected from kaggle, then they divided the dataset into 80% for training and 20% for testing. They used Linear regression, Random Forest, Logistic Regression, K-Nearest Neighbor, Gradient Boosting, XGBoost Classifier, Artificial Neural Network, AdaBoost algorithms. Best output found from Gradient Boosting with Accuracy (84.6%), Precision (84%), Recall (99%) and F1 Score (91%). [4]

Mir Ishrak Maheer Dhruba's group et al. worked on application of machine learning in credit risk assessment. They collected dataset from an online free repository controlled by Lending Club. In their dataset, output is "1" and "0". If the output is "1" that means "fully paid" else "changed off". Authors used Logistic Regression, Support Vector Machine. Random Forest, XGB. Finally best output finds from Logistic Regression. [5]

Mehul Madaan et al. worked on loan default prediction using decision trees and random forest. Their research is actually a comparative study between Decision trees and Random forest. In their research, they used Lending Club dataset from Kaggle. After successfully applied these two algorithms they found Random forest with the accuracy level of (80%) and Decision Tree classifier gave 73% accuracy. So, Random Forest algorithm performed better than Decision Tree. [6]

Vishal Singh et al. worked on prediction of modernized loan approval system based on machine learning approach. In their research, they used XGBoost, Random Forest and Decision Tree algorithms. After applied these algorithms on their dataset they found best accuracy from XGBoost with the accuracy of (77%). [7]

Zainab Olalere et al. applied machine learning to predict agricultural loan defaulters among farmers and the algorithms they used are Support Vector Machine, Gradient Boosting, Adaptive Boosting, Logistic Regression. They collected data from financial institution that liaised with farmers in Lavun local government in Nigeria. Comparing all of the algorithms they found the best output from Gradient Boosting with the accuracy (88.57%) and F1-score (90.48%). [8]

Subrata Saha et al. developed the credit risk of bank customers from customer's attribute using neural network. Artificial Neural Networks (ANN) used for pattern recognition, this

model followed three primary components: input layer, hidden layer and output layer. They collected data from Bangladeshi commercial bank. They divided their dataset into two parts: Training and Testing. In their research, they found that correctly predicted percentage in testing are (92.9% & 93.8%) and 10.52% is the overall percentage error data. [9]

Chong Wu et al. developed a hybrid model to explore a credit risk prediction by using deep learning technology. They used hybrid technique by combining DBM and DRBM and they found best accuracy in DBM and DRBM compared with other deep learning methods. [10]

Germanno Teles et al. used Artificial neural network and Bayesian network models for credit risk prediction. By applied Artificial neural network and Bayesian network models on dataset they constructed ANNs is more efficient. [11]

Philip Sarfo-Manu et al. developed an intelligent system for the credit risk in financial institutions. They applied decision tree algorithm to develop an intelligent system and they found the accuracy of 70%. They proposed that their system can be used to predict the clients who are eligible for loans. [12]

Andreas Hild et al. developed the estimating and evaluating probability of default – A using machine learning approaches. Dataset collected from International Financial Research. The algorithms used are LightGBM, XGBoost, Voting Classifier with top 3 models, Voting Classifier with LR NN & LGBM, Deep Neural Network, Random Forest , Logistic Regression, Lasso Logistic Regression, Linear Discriminant Analysis, Decision Tree Classifier, Simple Neural Network, Naive Bayes, Grind Searched Neural Network, K-Nearest Neighbor. In his research, LightGBM (73%) algorithm gave best result than other algorithms. [13]

2.3 Comparative Analysis and Summary

After studying other researchers papers, we've seen that their work is quite relevant to our work. The algorithms they applied to their work and the result they got are listed in the table 2.3.1 below:

Table 2.3.1: Comparative analysis of the relevant paper works

Authors	Year	Applied Models	Result (Accuracy/Evaluation Metrics)
Mohammad Rajib Pradhan, Sima Akter and Ahmed Al Marouf	2020	Logistic Regression	87%
		Naive Bayes	24%
		KNN	87%
		Decision Tree	85%
		Random Forest	90%
		Support Vector Machine	87%
		Multilayer Perceptron	87%
		Adaboost	87%
		XGboost	89%
		Neural Networks	87%
		Linear Discriminant Analysis	87%
Aquib Abtahi Turjo, S.M. Mynul Karim, Tausif Hossain Biswas, Yeaminur Rahman and Ifroim Dewan	2021	Linear Regression	84.42%
		Random Forest	84.55%
		Logistic Regression	84.42%
		K-Nearest Neighbor	83.54%
		Gradient Boosting	84.6%
		XGboost	84.45%
		Artificial Neural Network	84%
		Adaboost	84.58%
Mir Ishrak Maheer Dhruba, Nawab Haider Ghani, Sazzad Hossain and Syed Zamil Hasan Shoumo	2019	Logistic Regression	PCA = 96.3%
			RFECV = 99.6%
			PCA with CV = 95.1%
			RFECV with CV = 99.9%
		Support Vector Machine	PCA = 92.3%
			RFECV = 99.8%
			PCA with CV = 94.5%
			RFECV with CV = 99.9%

		Random Forest	PCA = 87.1%
			RFECV = 99.9%
			PCA with CV = 95.4%
			RFECV with CV = 99.9%
		XGboost	PCA = 88.9%
			RFECV = 99.9%
			PCA with CV = 97.1%
			RFECV with CV = 99.9%
Mehul Madaan, Aniket Kumar, Chirag Keshri, Rachna Jain and Preeti Nagrath	2021	Decision Tree	73%
		Random Forest	80%
Vishal Singh, Ayushman Yadav, Rajat Awasthi and N.Partheeban	2021	Xgboost	77.77%
		Random Forest	76.38%
		Decision Tree	64.58%
Zainab Olalere, Ishaq Oyebisi Oyefolahan and Solomon Adelowo Adepoju	2021	Gradient Boosting	88.57%
		Adaboost	80%
		Random forest	80%
		SVM	80%
		Logistic Regression	82%
Subrata Saha and Sajjad Waheed	2017	ANN	Training = 95.8%
			Testing = 93.3%
Chong Wu, Dekun Gao and Siyuan Xu	2021	DBM + DRBM	88.58%
		SVM	70.89%
Germanno Teles, Joel J. P. C. Rodrigues, Ricardo A. L. Rabelo and Sergei A. Kozlov	2020	Naive Bayes	81.32%
		Neural Network	81.85%

Philip Sarfo-Manu, Gifty Siaw and Peter Appiahene	2019	Decision Tree	70%
Andreas Hild	2021	LightGBM	AUC = 0.73
		XGboost	AUC = 0.72
		Voting classifier with top 3 models	AUC = 0.71
		Voting classifier with LR NN and LGBM	AUC = 0.71
		Deep Neural Network	AUC = 0.71
		Random Forest	AUC = 0.69
		Logistic Regression	AUC = 0.68
		Lasso Logistic Regression	AUC = 0.68
		Decition Tree	AUC = 0.67
		Simple Neural Network	AUC = 0.67
		Naïve Bayes	AUC = 0.64
		Grind Searched Neural Network	AUC = 0.63
		KNN	AUC = 0.55
		LDA	AUC = 0.68

2.4 Scope of the Problem

Usually, the bank does not share its data with the public due to the confidentiality. Thus, data collection was a major problem in this research. This study can help the banks to find those delinquent customers who have the higher possibility to recover the credit in the short term. Bank can identify those customers who have relationship with the bank. So, if a new customer come to the bank to take credit who have recently created an account and are not credit defaulter customer then this study might not be able to identify those customers whether he has the possibility of short term credit recovery or not. So, this can be a problem in sanctioning a credit to new customers.

2.5 Challenges

Every bank keeps customer data secured. So, they usually do not share customer data with others. When we selected this topic for research, we faced the first challenge in data collection. To collect the dataset of credit defaulter is a very challenging task. Here, the majority of information is about the clients who have bad credit. The dataset is imbalanced. To balance the dataset, preprocess the data and apply the deep learning models were also a challenging task and took lots of time.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Research Subject and Instrumentation

This study is about the bank credit recovery of customers. So, the dataset we've used is from a bank. This dataset has raw data and contains the information of the credit defaulter customers. Since the dataset has been collected from the bank, so we didn't need to search the dataset in the internet and didn't need any extra instrumentation for collecting the dataset and doing the research.

3.2 Data Collection Procedure

After selecting this topic, we searched for a dataset that contains the information of the credit defaulter customers. Then we collected the raw data of those customers.

3.3 Statistical Analysis

Table 3.3.1: Statistical properties of the dataset

Property Name	Total
Total number of samples	4600
Total number of features	50
Total number of input features	49
Total number of output classes	2
Total number of class 0 samples	3698
Total number of class 1 samples	902

3.4 Proposed Methodology

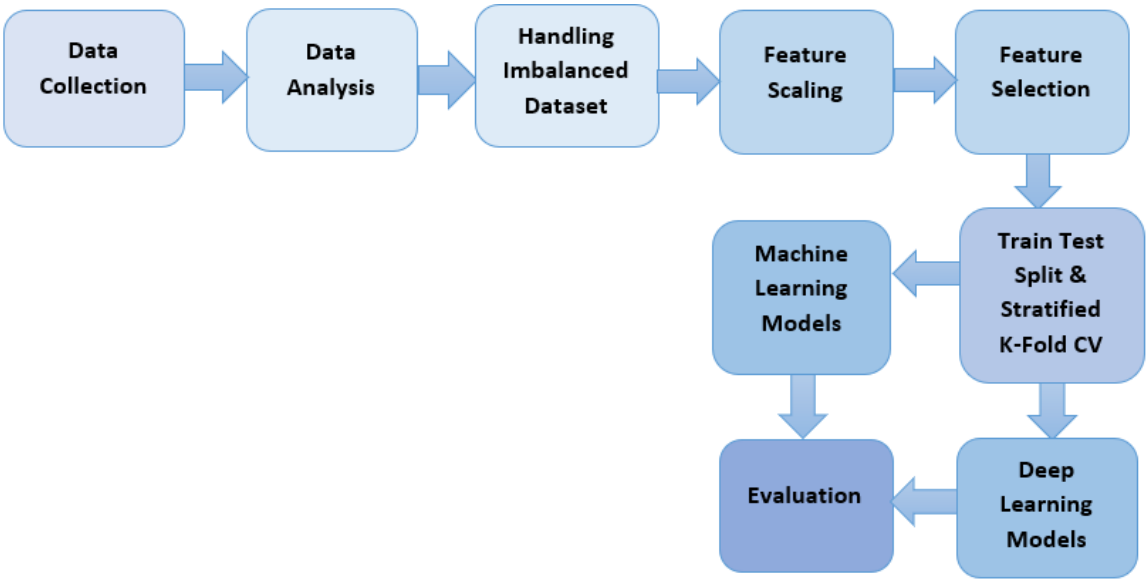


Figure 3.4.1: Proposed Methodology

Data Collection:

In this study, we have collected the dataset which is previously used on "Performance Evaluation of Traditional Classifiers on Prediction of Credit Recovery" this relevant paper [3]. This dataset is about credit defaulter clients or customers. In the dataset, we have observed some data as categorical and some as numerical. The categorical features were labeled and one hot encoded when we got the dataset. This dataset has the majority of information about those clients who have bad credit.

Data Analysis:

Data analysis is the major part to analyze the dataset for checking null values and understanding the relationship between input and output variables. For better accuracy and to get good performance in the machine and deep learning models, a proper dataset is needed. By analyzing the dataset, we can check if there are any missing values in the dataset. Missing values are a huge problem for any models or algorithms. Models return errors if there are any missing values. So, first of all we checked if there are any missing values presented or not. Missing values were not found in this dataset. This part is also called the data preprocessing part in the data mining. After this, we have analyzed the

statistical description of each column in the dataset. Then we have observed how many variables and features are there in the dataset to get the proper idea about the dataset and to visualize the data better. We have seen that 50 features and 4600 samples are there in the dataset. Among the 50 features, input features or attributes are 49 and one is the output. In the output, the total number of classes are two. The class 0 contains 3698 samples and the class 1 contains 902 samples. To get a better understanding, we have plotted the countplot of the total number of samples of each class which have shown in the figure 3.4.2. So, it is necessary to analyze the data before giving it to the models to learn and predict the outcome.

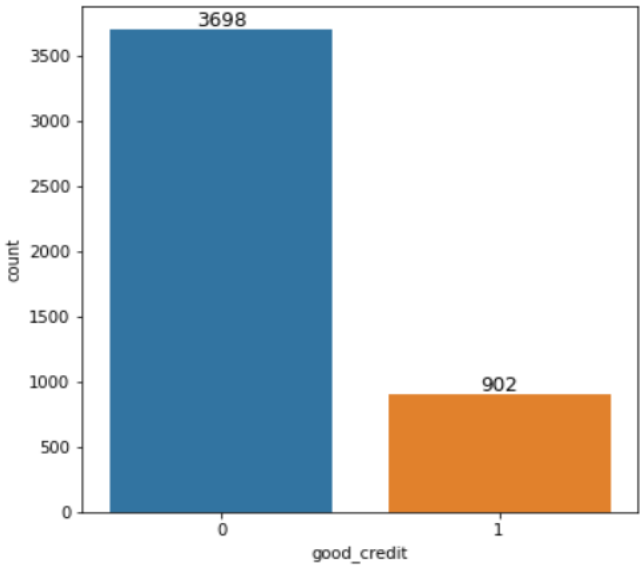


Figure 3.4.2: Countplot of the total number of samples of each class

Handling Imbalanced Dataset:

The dataset used in this study contains 4600 samples. The class 0 contains 3698 samples which is the majority class and is represented as bad credit. The class 1 contains 902 samples which is the minority class and is represented as good credit. Since, the ratio of samples of each class are not equal, the dataset is imbalanced. There are two types of techniques to handle the imbalanced dataset such as Undersampling and Oversampling. Undersampling technique balances the dataset by making the ratio of majority class samples equal to the ratio of minority class samples. So, there is a possibility of dropping many samples from the dataset. Since the dataset has a majority class of 3698

samples, this technique is not good. There are many oversampling techniques. Among those, we have used the Synthetic Minority Oversampling Technique (SMOTE) which takes the minority class and oversamples it. This approach doesn't add any new information, instead it synthesizes the new samples from the existing samples. It is a kind of data augmentation for the minority class. After applying this technique, the ratio of samples of each class are equal and balanced which is shown in the figure 3.4.3.

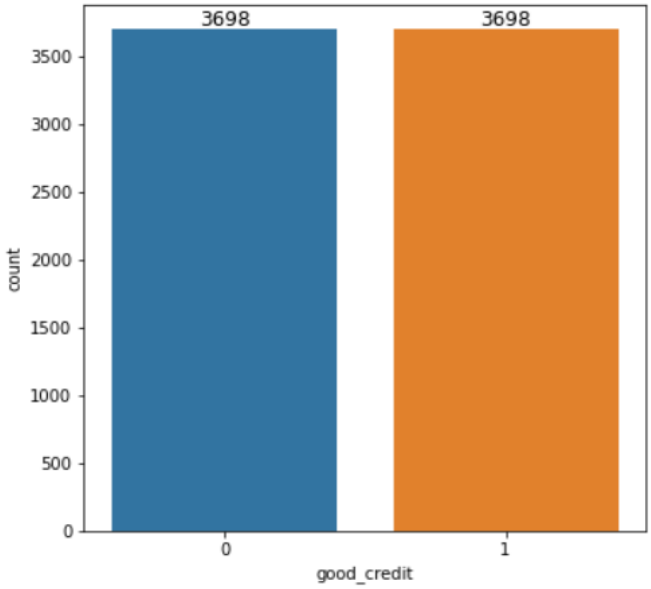


Figure 3.4.3: Countplot of the total number of samples of each class (After applied SMOTE technique)

Feature Scaling:

It is a process of normalizing the range of independent features of the dataset. This process is also referred to as data normalization and also a data preprocessing step. There are some types of feature scaling. Among them, Min-max scaler and standardization techniques are mostly used by people. In this study, we have used the standardization technique to normalize the input or independent features data. It basically scaled and centered the data with mean 0 and standard deviation of 1. It brings different independent features into the same scale. Some models behave and perform much better if the features are into the same scale. It also makes the training of a model faster. In the dataset, there are some huge numbers of values in some features and the features are in different scales which made some machine learning models don't perform well. By using the standardization technique, those huge numbers of values are scaled and the model's training is faster than before and

some models also performed well. Finally, the independent features took the form of normal distribution. Standardization is simply expressed by the following equation:

$$X_{new} = (X - \mu) / \sigma \qquad (i)$$

Here, X is the sample, μ is the sample mean of individual features and σ is the standard deviation.

Feature Selection:

It is the process of dropping redundant features and selecting the necessary features to train a model and get good performance for that model. All of the input features in the dataset are not necessary for a model to learn because there are some redundant features that do not impact so much to the performance of a model. Sometimes, if we train a model by giving all of the input features, the model does not give a good performance and it also takes so much time for a model to learn about the dataset. To obtain the goal outcome and to get the good predictive accuracy of a model all of the input features are not needed. There are some techniques available for feature selection such as XGboost, LightGBM etc. So, to identify which features are important and impacting so much to the performance of a model we have applied the LightGBM model which is a gradient boosting framework to identify the important features and find the performance of each feature. The performance of each feature is illustrated in figure 3.4.4.

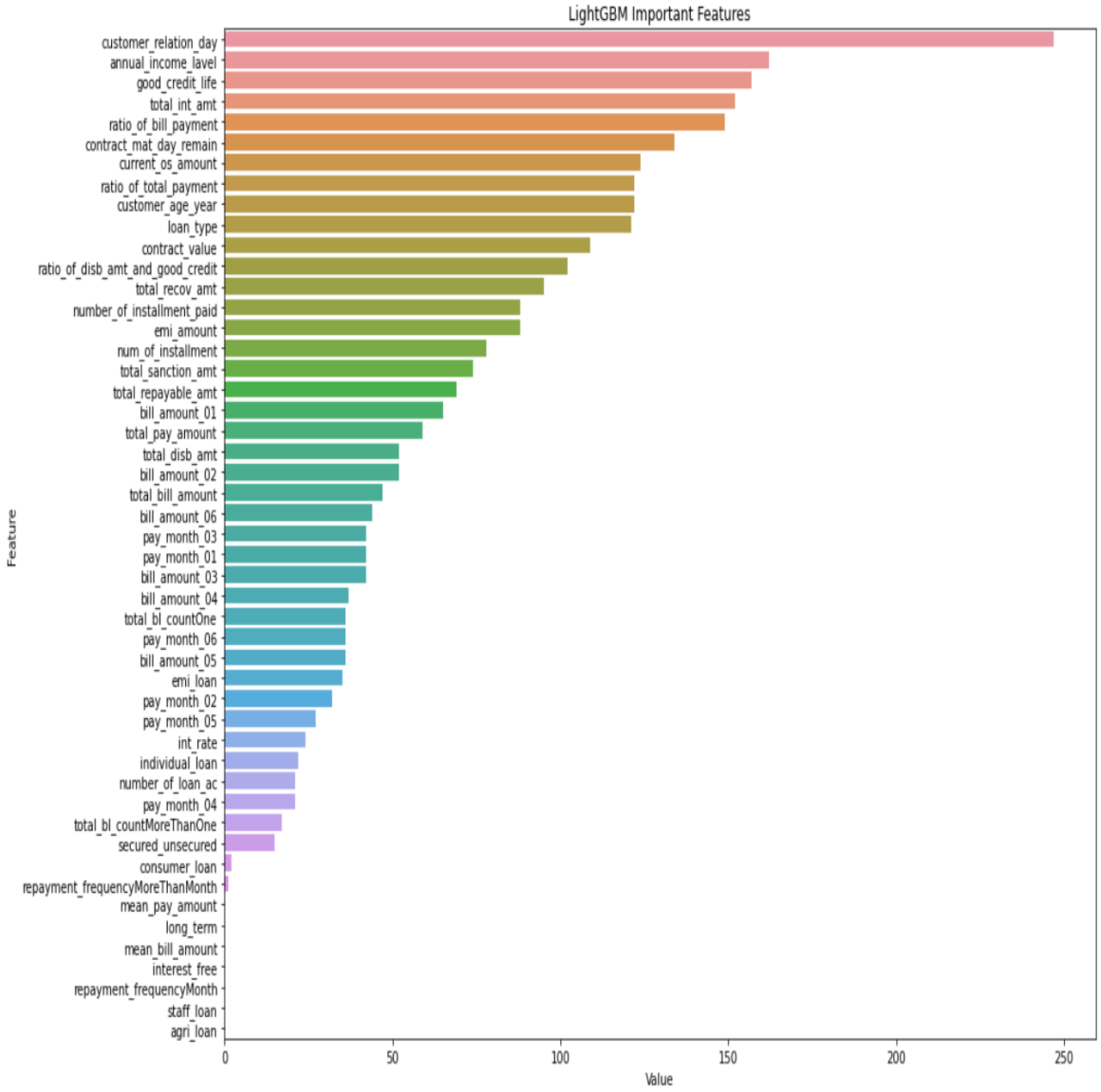


Figure 3.4.4: Performance of each feature by using LightGBM model

From the figure 3.4.4, we see that customer relation day is the most important feature and the last 7 features are not impacting so much to the performance of the model. So, we dropped these seven features and trained the other machine and deep learning models by the important features. After dropping those features, we finally got 42 features as an input. The description of each feature according to the performance is listed in the table 3.4.1.

Table 3.4.1: Description of important features

Feature No.	Description	Feature No.	Description
F1	Customer's relationship days with the bank	F22	2nd month of bill amount
F2	Annual income level	F23	Total bill amount
F3	Life of good credit	F24	6th month of bill amount
F4	Total interest amount	F25	3rd month of payment amount
F5	Ratio of bill payment	F26	1st month of payment amount
F6	Contract maturity remaining day	F27	3rd month of bill amount
F7	Current outstanding amount	F28	4th month of bill amount
F8	Ratio of total payment	F29	Total bad loan count of one
F9	Customer's age	F30	6th month of payment amount
F10	Type of loan	F31	5th month of bill amount
F11	Contract value	F32	Emi loan
F12	Ratio of disbursement amount and good credit	F33	2nd month of payment amount
F13	Total recovery amount	F34	5th month of payment amount
F14	Paid installment	F35	Interest rate
F15	Emi amount	F36	Individual loan
F16	Installment	F37	Number of loan account
F17	Total sanction amount	F38	4th month of payment amount
F18	Total repayable amount	F39	Total bad loan count of more than one
F19	1st month of bill amount	F40	Secured unsecured loan
F20	Total payment amount	F41	Consumer loan
F21	Total disbursement amount	F42	Repayment frequency of more than one month

Train Test Split & Stratified K-Fold Cross Validation:

After analyzing the dataset, doing some data preprocessing steps and selecting the important features, we splitted the dataset into training and testing. Training samples are used to train the model and testing samples are used to evaluate the performance of the model and to see how the model performs and obtain the goal outcome for the unseen data. 80% data for training and 20% for testing has been kept in our study. Since, the dataset is balanced after applied SMOTE technique, we have used stratify as a parameter of the train test split method. This parameter keeps the ratio of samples for each class balanced in training and testing. We have also used random state as a parameter so that each time we divide the dataset into training and testing, it takes the same samples for training and testing every time. We have used the random state as 42. If we change the value of random state,

we get different results for a model. So, this is a problem for a model as the model learn with different train samples every time we give different random state value and the model gives different results. To overcome this issue, we have also applied Stratified K-Fold Cross Validation to each of the machine and deep learning models.

Stratified K-Fold Cross Validation is the extension of K-Fold CV and it keeps the ratio of samples for each class equal in every fold it takes for training and testing. We have used the number of splits as 20. So, it runs 20 number of experiments on the dataset and produce 20 different results for a model. In this method, the model learns with different samples. After finished the total number of splits, we calculate the mean of the results. This method is good to identify and select a best model for the specific dataset. Finally, we made a comparison between train test split and Stratified K-Fold CV to see how those machine and deep learning models performed in each of these. We have also made a comparison between machine and deep learning models.

Machine learning Models:

- Logistic regression (LR)
- Random Forest (RF)
- K-Nearest Neighbour (KNN)
- Decision Tree (DT)
- Support Vector Machine (SVM)
- Gaussian Naive Bayes (GNB)
- Bernoulli Naive Bayes (BNB)
- XGBoost (XGB)
- AdaBoost (AdaB)
- Linear Discriminant Analysis (LDA)

Deep Learning Models:

- Artificial Neural Network (ANN)
- Multilayer Perceptron (MLP)
- Convolutional Neural Network (CNN)
- Recurrent Neural Network (RNN)

- Long Short Term Memory (LSTM)
- Gated Recurrent Unit (GRU)

Evaluation:

We have used some performance evaluation metrics to evaluate the performance of each machine and deep learning models very well. There are many evaluation metrics available to evaluate a model's performance and to get a good understanding about how actually that model performs very well. Among them, we have used Accuracy, Precision, Recall, F1-Score, AUC, MAE, RMSE and MCC in this study. The model which gives good accuracy also has good performance in the other evaluation metrics. We have also shown the ROC curve of each of the models together to understand the ratio between false and true positive rate. Accuracy, Precision, Recall and F1-Score basically comes from the confusion matrix. Confusion matrix is simply a summarization of the prediction results and it displays the relationship between the actual and predicted class. This matrix is very useful in the classification problem. Precision, Recall, F1-Score provide better insights of the model than accuracy. So, only accuracy is not enough to understand the model performance. Other evaluation metrics are also needed. MAE, RMSE, MCC gives us a good statistical evaluation of the model. The performance evaluation and statistical metrics that we have used in this study are briefly described below:

Accuracy: It refers to the ratio of correctly predicted samples to total number of samples. If the dataset contains equal samples for each class, then it works well.

Precision: It refers to the ratio of true positive values to all the predicted positive values.

Recall: It refers to the ratio of true positive values to all the actual positive values.

F1-Score: When there is a need to compare between various models, it is difficult to appoint which one is better with just precision and recall metrics. So, there is a metric that combines both of these and it is F1-Score. It's referred to as the harmonic mean of precision and recall. If the value is higher then the better is the model.

AUC: It is referred to as the Area Under the ROC Curve. It is useful to understand the overall performance of the model. It gives a ratio between false and true positive rates. If the AUC is high then the false positive rate is lower than the true positive rate. A model with higher AUC is the better model.

MAE: It is referred to as Mean Absolute Error and is computed by taking the mean of the absolute difference between actual and predicted values. Lower the value, better the model.

RMSE: It is referred to as Root Mean Squared Error. It is the same as Mean Squared Error which is computed by taking the mean of the square difference between actual and predicted values. In RMSE, root is considered.

MCC: It is referred to as Matthews Correlation Coefficient and is used to measure the quality of classification and generate a value between -1 to +1. The +1 represents perfect, 0 represents average random and -1 represents inverse prediction. So, a model with MCC value closer to +1 gives a good prediction.

The evaluation metrics and the performance of machine and deep learning models are described in section 4.2.

3.5 Implementation Requirements

Hardware Requirement:

Laptop Configuration:

Intel Core i5 8th Gen processor, 12GB RAM, 1TB HDD, 128 GB SSD, 4GB Nvidia Geforce 940MX Graphics Card, Windows 10 64 bit Operating System

Software Requirement:

Google Colab, Python Packages (pandas, matplotlib, numpy, seaborn, imblearn, scikit-learn, tensorflow, keras)

CHAPTER 4
EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Experimental Setup

We've setup the experiment according to the proposed methodology. We have used Google Colab to do the code and used some python packages that are necessary for reading the dataset, analyzing it, handling imbalance data, selecting important features, preprocessing the data, applying machine and deep learning models and visualizing the performance of each model. The laptop that we have used to experiment has an Intel Core i5 8th gen processor with 12GB RAM, 1TB HDD, 128GB SSD, Windows 10 64bit operating system and 4GB Nvidia Geforce 940MX Graphics Card.

4.2 Experimental Results and Analysis

Table 4.2.1: Evaluation and Statistical Metrics of Machine Learning Models by using Train Test Split

Models	Accuracy	Precision	Recall	F1-Score	AUC	MAE	RMSE	MCC
LR	77%	0.80	0.71	0.76	0.83	0.23	0.48	0.54
RF	**93%**	**0.94**	**0.91**	**0.92**	**0.98**	**0.07**	**0.27**	**0.85**
KNN	83%	0.82	0.85	0.84	0.91	0.17	0.41	0.67
DT	86%	0.86	0.86	0.86	0.86	0.14	0.37	0.72
SVM	81%	0.85	0.76	0.80	0.88	0.19	0.43	0.63
GNB	51%	0.51	0.99	0.67	0.69	0.49	0.70	0.08
BNB	64%	0.70	0.51	0.59	0.68	0.36	0.60	0.30
XGB	89%	0.94	0.82	0.88	0.95	0.11	0.34	0.78
AdaB	86%	0.88	0.82	0.85	0.92	0.14	0.38	0.71
LDA	75%	0.80	0.68	0.73	0.82	0.25	0.50	0.51

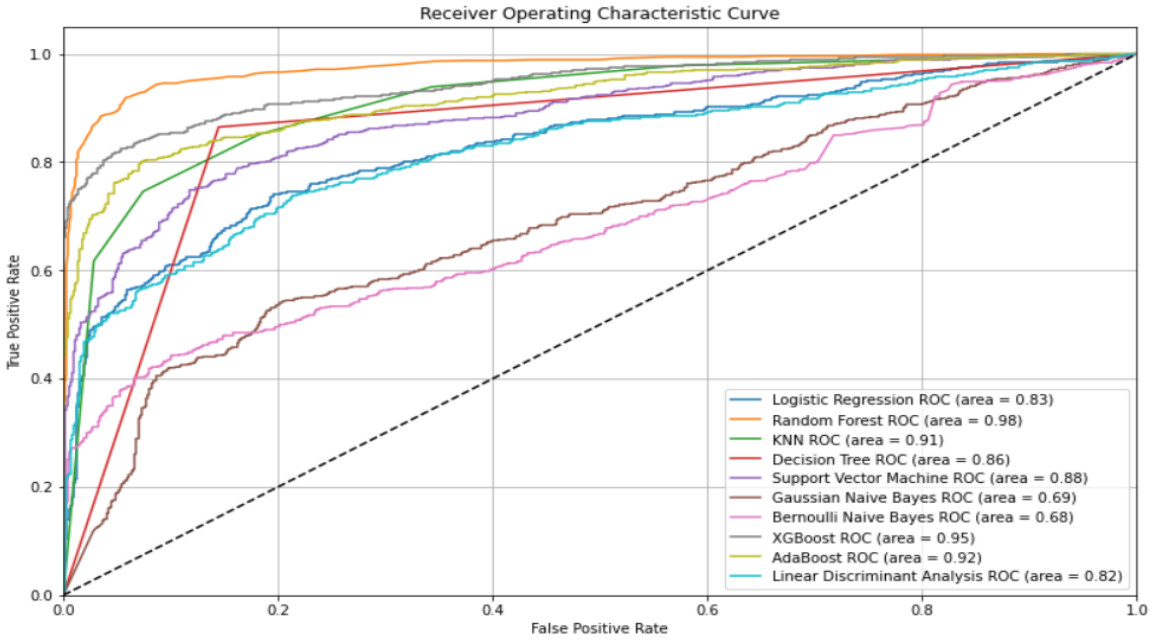


Figure 4.2.1: Roc Curve of Machine Learning Models

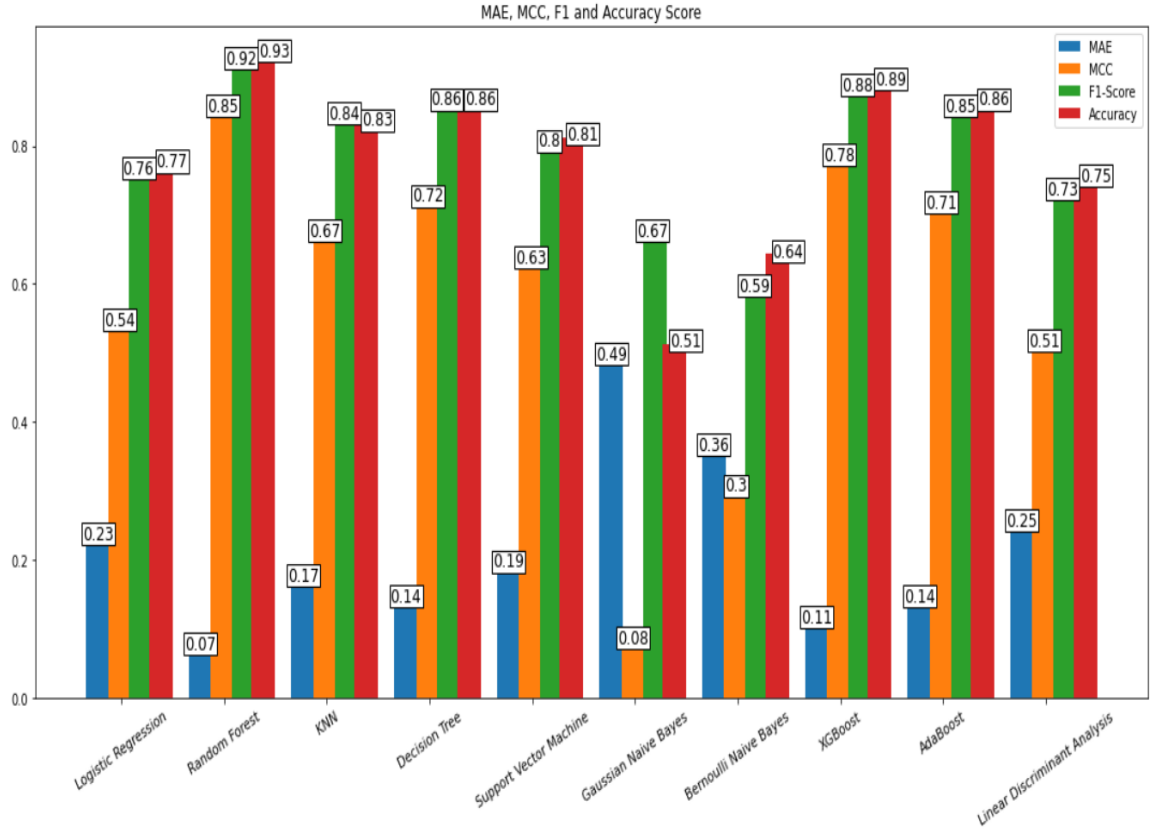


Figure 4.2.2: Bar Chart of MAE, MCC, F1-Score and Accuracy of Machine Learning Models

Table 4.2.2: Evaluation and Statistical Metrics of Deep Learning Models by using Train Test Split

Models	Accuracy	Precision	Recall	F1-Score	AUC	MAE	RMSE	MCC
ANN	**90%**	**0.89**	**0.91**	**0.90**	**0.95**	**0.10**	**0.32**	**0.80**
MLP	89%	0.87	0.90	0.89	0.94	0.11	0.34	0.77
CNN	89%	0.89	0.88	0.89	0.95	0.11	0.33	0.78
RNN	80%	0.82	0.76	0.79	0.87	0.20	0.45	0.59
LSTM	85%	0.89	0.79	0.84	0.91	0.15	0.39	0.70
GRU	86%	0.86	0.85	0.86	0.92	0.14	0.38	0.71

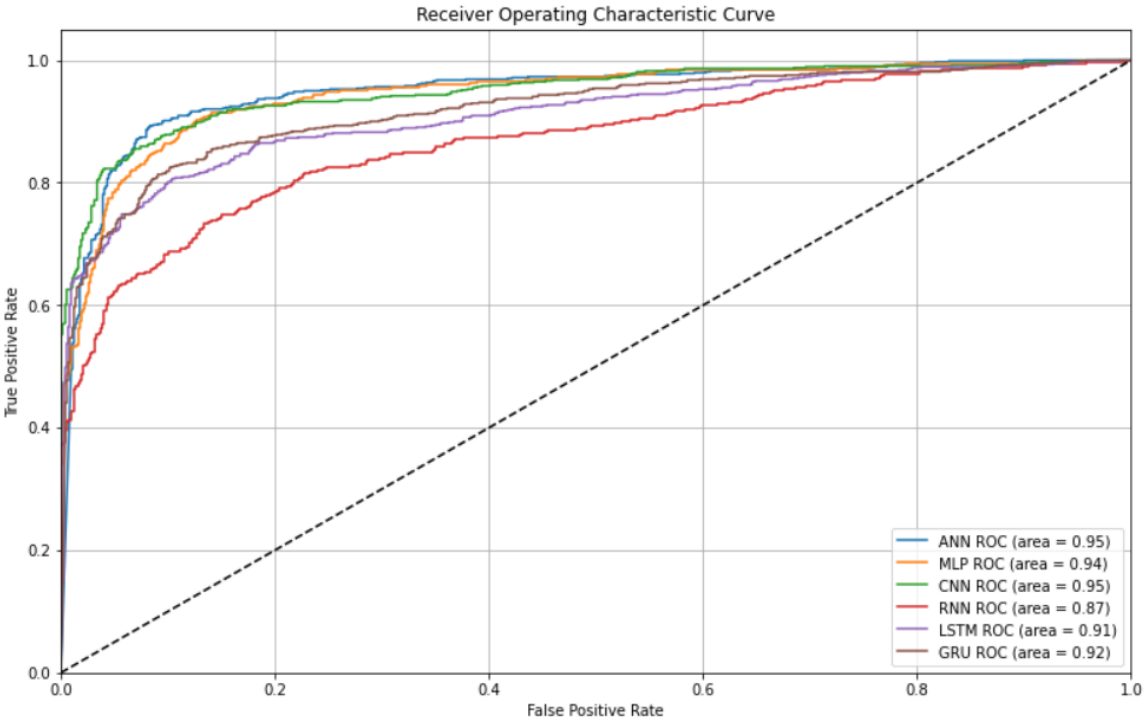


Figure 4.2.3: Roc Curve of Deep Learning Models

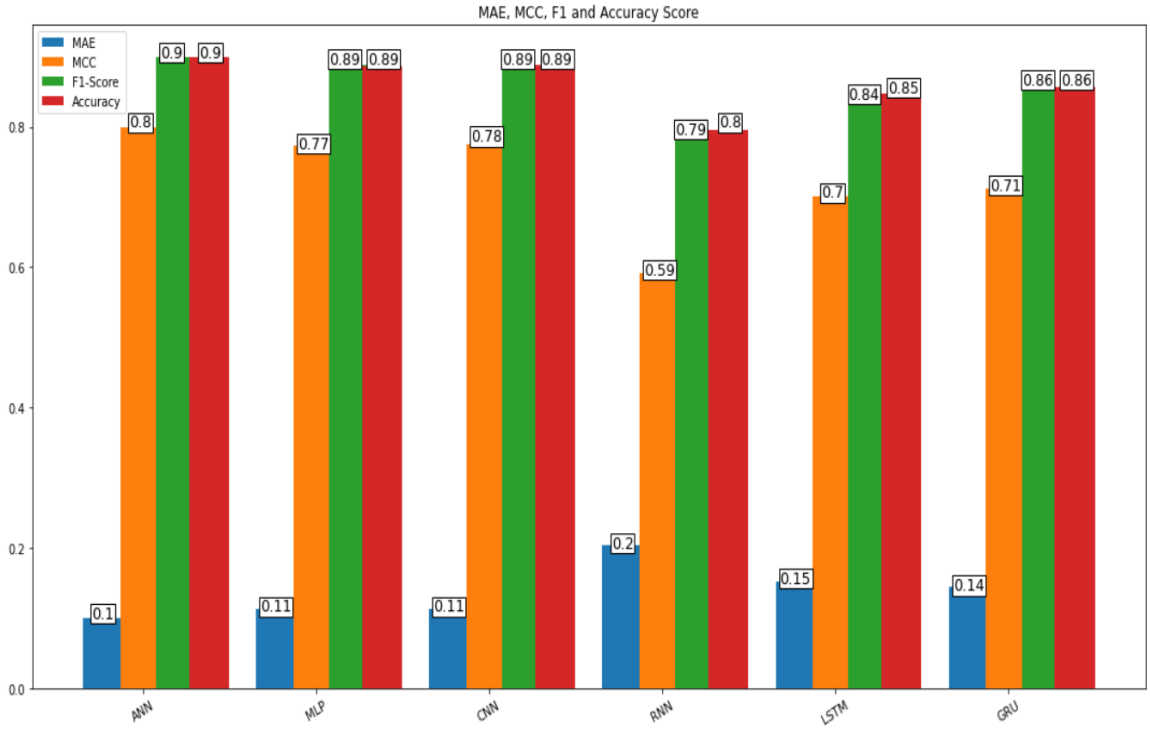


Figure 4.2.4: Bar Chart of MAE, MCC, F1-Score and Accuracy of Deep Learning Models

Table 4.2.3: Evaluation and Statistical Metrics of Machine and Deep Learning Models by using Stratified K-Fold Cross Validation

Models	Accuracy	Precision	Recall	F1-Score	AUC	MAE	RMSE	MCC
LR	75%	0.79	0.69	0.74	0.82	0.25	0.50	0.51
RF	**94%**	**0.96**	**0.92**	**0.94**	**0.98**	**0.06**	**0.25**	**0.88**
KNN	83%	0.83	0.84	0.84	0.91	0.17	0.41	0.67
DT	86%	0.85	0.88	0.86	0.86	0.14	0.37	0.72
SVM	80%	0.85	0.73	0.78	0.88	0.20	0.45	0.60
GNB	53%	0.51	0.96	0.67	0.69	0.47	0.69	0.11
BNB	64%	0.69	0.48	0.57	0.67	0.36	0.60	0.28
XGB	89%	0.95	0.83	0.88	0.95	0.11	0.33	0.79
AdaB	85%	0.88	0.82	0.85	0.92	0.15	0.38	0.71
LDA	74%	0.79	0.65	0.72	0.82	0.26	0.51	0.49
ANN	90%	0.89	0.91	0.90	0.95	0.10	0.32	0.80
MLP	90%	0.89	0.91	0.90	0.95	0.10	0.32	0.79
CNN	88%	0.87	0.89	0.88	0.95	0.12	0.34	0.76

4.3 Discussion

After analyzing the dataset, preprocessing it, selecting the important features and splitting the dataset into training and testing, we've applied the machine and deep learning models to evaluate the performance. First of all, we applied the train test split method and then we applied Stratified K-Fold CV to see how those models perform in each of these and made a comparison between those methods.

By using the train test split method, we find that all the model's results are good enough. Among the machine learning models, Random Forest (RF) gives 93% accuracy, precision 0.94, recall 0.91, F1-score 0.92, AUC 0.98 which performs better than other ML models. It's MAE value 0.07, RMSE value 0.27 which is less than other ML models which means this model showed lower error when predicting the goal outcome. It's MCC value is 0.85 which is higher than the other ML models. XgBoost (XGB) gives 89% accuracy and it's other metrics value is also closer to RF. Gaussian Naive Bayes (GNB) gives the accuracy of 51% which is worse than the other ML models and it's other metrics value is also not so good. Since the AUC value of RF is 0.98, it has lower false positive rate than true positive rate which is clearly understandable in the ROC curve of figure 4.2.1. The bar chart of the performance of each ML model is also shown in the figure 4.2.2 to get a better understanding of the models performance. Overall compared with the ML models, Random Forest (RF) performs better and Gaussian Naive Bayes (GNB) performs worse than the other ML models. RF model gives better prediction and can better identify those customers who have the possibility to recover the credit in the short-term and who have the possibility to recover the credit in the long-term.

Among the deep learning models, ANN gives 90% accuracy and it's other evaluation and statistical metrics result is also good enough. MLP and CNN both give 89% accuracy and their other metrics results are closer to ANN. RNN gives 80% accuracy which is lower than other DL models. The ROC curve of each DL model is also good enough which is shown in the figure 4.2.3 and the bar chart of their performance is shown in the figure 4.2.4. Compared with the DL models, ANN performed better and RNN performed lower than the other DL models.

Compared with all the ML and DL models, the result of all the DL models are good enough than all the ML models but in the case of highest accuracy and the result of evaluation and statistical metrics, RF performed better.

In Stratified K-Fold CV, after the 20 number of splits are completed on the dataset, we've calculated the mean of each metric for each model. We find that all the models performance is almost closer to the train test split method. Here, RF gives 94% accuracy which we got 93% accuracy in the case of the train test split method. The result of the evaluation and statistical metrics of RF is also good here. Since Stratified K-Fold CV took lots of time to produce the final output and we've seen that all the models performance is almost closer to the train test method, hence in the case of DL models we've applied only ANN, MLP and CNN. The performance of these models are also almost closer to the train test split method. Compared with the Stratified K-Fold CV and Train Test Split method, RF performed better and gave a better outcome in Stratified K-Fold CV than the Train Test Split method and GNB performed worse in both of the cases.

CHAPTER 5
IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

A country's development depends on many sectors. The banking sector is one of them. In our society, many people or organizations take credit from the bank for several reasons. But some of them failed to repay the credit in the due time. For this, banks are facing classified loan problem which is also called NPL. Mainly banks earn from the interest charge of the loans. Due to the rise of the NPL ratio, they failed to collect interest payments. So, they have less money to sanction new loans. For this, many entrepreneurs can't get loans from bank. This is hampering the country's economic development. So, if we can predict credit recovery the number of NPL will be decreased. That's why we conduct this thesis to find out the possible customer for credit recovery from the NPLs. It will be very helpful to reduce non-performing loans. The banks can sanction loans to more customers. That will help a lot to reduce the unemployment problem. Many unemployed people can be benefited from this. So, this will have a positive impact on society.

5.2 Impact on Environment

There is no negative impact on environment by this project rather it is helpful for the banking sector.

5.3 Ethical Aspects

The dataset of our study contains the information of credit defaulter customers. Since this is a bank dataset, there is confidentiality about the information of customers. There is a possibility of doing harm to the bank and customers by anyone. To maintain confidentiality, we did not share this dataset to anyone and did not upload this dataset to any kind of internet source rather we just used it for the research purpose. All the customer's information is anonymous. This dataset only contains the raw data and no

picture of the customer has been collected. This study maintains the ethical clauses and we kept all the aspects safe and ethical.

5.4 Sustainability Plan

In this study, if we can predict the delinquent customer who has good credit or has the highest possibility to recover the credit in the short-term, then it will be helpful for a bank to get prior knowledge about the customer before sanctioning him a new credit. Banks will be very benefited from this and it will also reduce the ratio of NPL. That will also help to the economic development of a country.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FURTHER RESEARCH

6.1 Summary of the Study

In this study, we tried to find the delinquent customers who have the highest possibility of short term credit recovery. For this problem, we have used a credit defaulter dataset which is collected from a bank data warehouse. We worked on Google Colab and used some python packages like numpy, pandas, matplotlib, tensorflow, seaborn, keras for doing the work. At first, we pre-processed the data. We have performed a feature scaling technique called standardization in our dataset as the dataset contains many large values and the features are in different scale. We also performed SMOTE oversampling method to balance the imbalanced dataset. We used the LightGBM model for important feature selection. After all the processing we split our dataset using the Train Test Split method and feed it into the machine and deep learning models. We evaluated the models with different evaluation metrics such as Accuracy, Precision, Recall, AUC, F1-Score and statistical metrics such as MAE, RMSE, MCC and got satisfactory results. Then we also applied Stratified K-Fold Cross-Validation to each of the machine and deep learning models. The performance of each model using both Train Test Split and Stratified K-Fold Cross Validation is almost similar. Among all of the models, Random Forest (RF) gives the highest accuracy 94% in the K-Fold Cross Validation technique. We also get the highest 0.96 Precision, 0.98 AUC, 0.94 F1-Score for the same model. Among the deep learning models, ANN and MLP give the highest accuracy 90%. The Gaussian Naive Bayes gives the worst performance which is 53%. Compared to all of the machine learning models with all of the deep learning models, deep learning models give better performance but in the case of highest accuracy, evaluation and statistical metrics, RF performed better.

6.2 Conclusions

NPL occurs when the customers of a bank fail to repay the credit in the due time. It refers to the economic assets where banks don’t get interest payments and it resists the economic development of a country. To make the economic development maintainable, the NPL ratio

needs to be reduced. Hence in this study, we've applied both machine and deep learning models in the bank dataset to predict the delinquent customers who have the highest possibility of short-term credit recovery. The dataset is the credit defaulter dataset and has the majority of information about those customers who have bad credit. We balanced the dataset to make sure that the applied models can predict both good and bad credit classes in the future properly. We performed feature scaling to bring all the features into the same scale. We also performed feature selection to identify which features are important for this dataset and to build an automated system for bank's credit recovery. We made a comparison between machine and deep learning models in both Train Test Split and Stratified K-Fold CV methods. Compared with all of these, the best output has come from the Random Forest (RF) model and Gaussian Naive Bayes (GNB) performed worst. RF model better identifies the delinquent customers who have the highest possibility of short-term credit recovery and will be helpful to reduce the NPL ratio.

6.3 Implication for Further Study

In future, we will try to increase the samples of the dataset and will try to implement more classifier models. We will also try to apply the GridSearchCV method to see how each model performs in this method.

References

[1]Banks see surge in non-performing loans despite policy backup, available at: https://thefinancialexpress.com.bd/economy/banks-see-surge-in-non-performing-loans-despite-policy-backup-1629425220, last accessed on 11th November 2021.

[2] NPLs rise by 6,351C in Q1'21 | Dhaka Tribune, available at: https://www.dhakatribune.com/business/banks/2021/06/15/npls-rise-by-6-351c-in-q1-21, last accessed on 12th November 2021.

[3] Pradhan, M.R., Akter, S. and Al Marouf, A., "Performance Evaluation of Traditional Classifiers on Prediction of Credit Recovery" in Advances in Electrical and Computer Technologies Springer, Singapore, pp. 541-551, 2020.

[4] Turjo, A.A., Karim, S.M., Biswas, T.H., Rahman, Y. and Dewan, I., "Comparative Analysis and Implementation of Credit Risk Prediction Through Distinct Machine Learning Models" in Doctoral dissertation, Brac University, 2021.

[5] Shoumo, S.Z.H., Dhruba, M.I.M., Hossain, S., Ghani, N.H., Arif, H. and Islam, S., "Application of machine learning in credit risk assessment: a prelude to smart banking" in TENCON 2019-2019 IEEE Region 10 Conference (TENCON) IEEE, pp. 2023-2028, October 2019.

[6] Madaan, M., Kumar, A., Keshri, C., Jain, R. and Nagrath, P., "Loan default prediction using decision trees and random forest: A comparative study" in IOP Conference Series: Materials Science and Engineering IOP Publishing, Vol. 1022, No. 1, p. 012042, 2021.

[7] Singh, V., Yadav, A., Awasthi, R. and Partheeban, G.N., "Prediction of Modernized Loan Approval System Based on Machine Learning Approach" in International Conference on Intelligent Technologies (CONIT) IEEE, pp. 1-4, June 2021.

[8] Olalere, Z., "A Comparism of Machine Learning Based Approaches in Predicting Agricultural Loan Defaulters among Farmers in Lavun Local Government Area of Niger State", 2021.

[9] Saha, S. and Waheed, S., "Credit risk of bank customers can be predicted from customer's attribute using neural network" in International Journal of Computer Applications, 161(3), pp.39-43, 2017.

[10] Wu, C., Gao, D. and Xu, S., "A Credit Risk Predicting Hybrid Model Based on Deep Learning Technology" in International Journal of Machine Learning and Computing, 11(3), 2021.

[11] Teles, G., Rodrigues, J.J.P.C., Rabê, R.A. and Kozlov, S.A., "Artificial neural network and Bayesian network models for credit risk prediction" in Journal of Artificial Intelligence and Systems, 2(1), pp.118-132, 2020.

[12] Sarfo-Manu, P., Siaw, G. and Appiahene, P., "Intelligent System for Credit Risk Management in Financial Institutions" in International Journal of Artificial Intelligence and Machine Learning (IJAIML), 9(2), pp.57-67, July-December, 2019.

[13] Hild, A., "Estimating and Evaluating The Probability of Default–A Machine Learning Approach", 2021.

[14] Kaarthik, K., Dharanidharan, G., Navalarasu, R.B. and Sabarinathan, G.,"Machine Learning based Loan Prediction System using Svm and Knn Algorithms" in Turkish Journal of Physiotherapy and Rehabilitation, 32, p.2.

[15] Yontar, M., Dağ, Ö.H.N. and Yanık, S., “Using Support Vector Machine for the Prediction of Unpaid Credit Card Debts” in International Conference on Intelligent and Fuzzy Systems Springer, Cham, (pp. 377-385), July 2019.

[16] Chen, H.Z.,”A new model for bank loan loss-given-default by leveraging time to recovery” in Journal of Credit Risk, 14, pp.1-29, 2018.

[17] Hedblom, E. and Åkerblom, R., “Debt recovery prediction in securitized non-performing loans using machine learning”, 2019.

[18] Lakhani, M., Dhotre, B. and Giri, S.,”Prediction of Credit Risks in Lending Bank Loans” , Methodology, 5(12), 2018.

[19] Lakhani, M., Dhotre, B. and Giri, S., “Prediction of credit risks in lending bank loans using machine learning” in SAARJ Journal on Banking & Insurance Research, 8(1), pp.55-61, 2019.

[20] Ha, S.H. and Krishnan, R., “Predicting repayment of the credit card debt” in Computers & Operations Research, 39(4), pp.765-773, 2012.

[21] Bellotti, A., Brigo, D., Gambetti, P. and Vrins, F., “Forecasting recovery rates on non-performing loans with machine learning” in International Journal of Forecasting, 37(1), pp.428-444, 2021.

[22] Tahmid, N.I., Haque, N., Faruque, U., Keya, M., Khushbu, S.A. and Al Marouf, A., “A Concern of Predicting Credit Recovery on Supervised Machine Learning Approaches” in 12th International Conference on Computing Communication and Networking Technologies (ICCCNT) IEEE, pp. 1-5, 6-8 July, 2021.

Credit Recovery Report

ORIGINALITY REPORT

13%	7%	9%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	"Advances in Electrical and Computer Technologies", Springer Science and Business Media LLC, 2020 Publication	1%
3	Nazre Imam Tahmid, Nasimul Haque, Umar Faruque, Mumenunnessa Keya, Sharun Akter Khushbu, Ahmed Al Marouf. "A Concern of Predicting Credit Recovery on Supervised Machine Learning Approaches", 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021 Publication	1%
4	Mehul Madaan, Aniket Kumar, Chirag Keshri, Rachna Jain, Preeti Nagrath. "Loan default prediction using decision trees and random forest: A comparative study", IOP Conference Series: Materials Science and Engineering, 2021 Publication	1%