Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| **RESEARCH ARTICLE**

Credit Card Client's Payment Prediction for Next Month Using Machine Learning Algorithms

Faria Rahman Annasha[1] ✉ **Sabbir Hossen**[2], **Monoara Sultana Morzina**[3], **Md. Solaiman Kabir**[4] and **Md. Showrov Hossen**[5]

[12345]*Lecturer, Department of CSE, City University, Dhaka and Bangladesh*

Corresponding Author: Faria Rahman Annasha, **E-mail**: rahmanfaria98@gmail.com

| ABSTRACT

With the quick growth of the credit card system, there is a rising number of misconduct rates on credit card loans, which creates a financial risk for commercial banks. Thus, successful resolutions of the risks are significant for the sound advancement of the industry in the long term. Numerous financial banks and organizations become more and more attentive to the issue of credit card default because it brings about a high probability of financial risks. Credit risk plays a significant part in the financial business. One of the main functions of a bank is to issue loans, credit cards, investment mortgages, and other credit. One of the most popular financial services offered by banks in recent years has been the credit card. With its constant rise in risk factors, the banking industry is perhaps the most fragile and volatile in the world. Credit risk remains a crucial element for financial institutions that have experienced losses amounting to hundreds of millions of dollars as a result of their incapacity to retrieve the funds disbursed to clients. In the banking industry, it is now vital to forecast whether a borrower will be able to repay the loan. In this paper, we applied different machine learning classifiers, including Random Forest, K Nearest Neighbor, Logistic Regression, Decision Tree, Decision Tree with AdaBoosting, and Random Forest with AdaBoosting, to build a credit default prediction model. The results show that the AdaBoosting model achieved better accuracy than the other machine learning algorithms. Our proposed technique can support financial organizations in controlling, identifying, and monitoring credit risk, and it can identify credit card clients who pay the loan in the next month.

| KEYWORDS

Credit card, loans, payment, machine learning, random forest, decision tree, adaboosting

| ARTICLE INFORMATION

ACCEPTED: 01 August 2024 **PUBLISHED:** 07 August 2024 **DOI:** 10.32996/jcsts.2024.6.3.8

1. Introduction

With its constantly rising risk factors, the banking industry is among the most vulnerable and unstable in the world. Financial organizations have lost hundreds of millions of dollars as a result of their inability to retrieve the money that was supplied to customers. Credit risk has continued to play a significant role in these losses. In the financial industry, it has become critical to anticipate or estimate whether a person will be able to repay the loan. The decision to issue a loan is based on a number of variables, such as the applicant's prior credit history, the total amount requested, the state of the economy, the applicant's age, occupation, and so on. Additionally, the borrower must be able to repay the loan in the allotted period. The clients of credit cards are the main subject of this study among the different loans provided to individuals. Since the amount due on a credit card varies each month depending on the user's spending habits, credit cards are considered revolving credits. A pre-approved credit limit is provided to the customer based on his credit history and several other factors (Torvekar, 2019).

The banking sector has undergone a revolutionized transformation to the big data paradigm, which has altered how financial organizations function. People are now in better shape financially and in terms of job chances, thanks to the gradual recovery from the effects of the recent financial crisis. Over the last several years, there has been a sharp rise in the demand for financial services, which has resulted in an enormous amount of data in terms of volume, accuracy, and diversity (Sayjadah, 2018). Financial institutions are becoming more significant, and this has led to pressure on them to offer a variety of services, including credit facilities, investments, mortgages, and retail banking. Banks are thus figuring out how to use their current data sets to their advantage in order to deal with the issue of emerging data. Every time a consumer transacts with a bank, vast amounts of data are collected, including everything from demographic information to online history (Hussein, 2020).

Data analytics has changed the scenario for banks in terms of tackling their various difficulties and minimizing their market flaws. With the use of analytics, banks are now better equipped to handle the real-time data provided by their consumers (Xu, 2017). Banks may now explore a completely new horizon thanks to predictive and prescriptive analytics, which was not conceivable with their prior descriptive methodology. In order to better forecast trends, the banking industry is currently renowned for its effective application of machine learning techniques that combine numerous categorization approaches to separate its consumers (Juneja, 2020). One of the key forecasts about credit card default that worries banks is credit scoring, which helps them understand why consumers are more likely to fail. In order to more easily predict their customers' default, they must pay close attention to every tiny detail about them and watch the installment information that is put into the credit rating writing. The primary objective of this work is to develop a credit default prediction model that benefits the banking industry by utilizing several machine learning techniques, including Logistic Regression, K Nearest Neighbor, Decision Tree, Random Forest, Decision Tree with AdaBoosting, and Random Forest with AdaBoosting.

1.1 Motivation

The banking sector has faced huge losses in recent times due to borrowers' unwillingness to repay their money, which has led to bad debts. One of the essential purposes behind this is the improper independent direction and giving of advances to un-qualified candidates. Risk executives and the ID of credit chance of clients require treatment with significant measures of information from a wide assortment of perspectives. Bank directors today, in this manner, experience a critical test in the ID of expected defaulters. Giving Credit cards to inadequate candidates has prompted tremendous misfortunes for the banks. Besides clients with reimbursing capacity, however, aggregating weighty credit and overutilization of it as much as possible can also prompt serious misfortunes for banks. Visas are spinning credits, and the sum to be paid changes consistently. Thus, compelling checking of these angles can assist the banks in following the records that have the likelihood of default and, in this manner, make the essential move. Consolidating AI methods for the expectation of defaulters can be valuable for banks in planning preventive measures and accordingly staying away from misfortunes.

2. Literature Review

Credit card default prediction using machine learning algorithms is a popular research area. In assessing the likelihood of defaulting to credit card clients, Yeh et al. [2009] found that artificial neural networks performed better than machine learning techniques. Neural networks scored better than various conventional scoring models in terms of type II errors and predicted accuracy, according to Aktan et al. [2009]. Using machine learning approaches, Khandani et al. [2010] developed forecasting models for consumers' default guesses in 2010. In order to significantly increase default and customer misbehavior categorization rates, they used consumer credit transactions and credit agency information, with linear regression R2s of forecasted misbehaviors. Support vector machines and generic programming are better models for categorizing loan applicants, according to Singh and Aggarwal's research [2011].

Research by Aktan et al. [2009], which was published in 2009, revealed that neural networks performed better than certain conventional scoring models in terms of type II errors and prediction accuracy. Models of credit card borrower default were presented by Bellotti et al. [2013] and used both macroeconomic factors and social data on credit card customers. In general, it was discovered that models with macroeconomic and social data were statistically more significant than other models. Kraus et al.'s [2010] empirical study of credit risk made use of client personal information and data from a German bank. When it came to forecasting credit card defaulters, the accuracy rate of the Random Forest Ensemble approach was the greatest, according to a study done in 2016 by Jacob et al. [2016]. In 2017, Pasha et al. [2017] looked at the possibility of credit card default. In this work, six distinct data mining methodologies are modeled using datasets. The results of this study show that the neural network has the highest accuracy and performs the best in predicting credit card default.

A comprehensive survey of the literature on the application of machine learning techniques in the context of credit risk evaluation was carried out by Abbas Keramati et al. [2011]. Their objective is to fully assess each approach and pinpoint its limitations in order to create a model with more power. They presented a hybrid model with improved capabilities after carefully examining ten data mining approaches. They found that hybrid SVM models yield superior results. Vasilica Oprea et al. [2022] calculated the likelihood of default and looked at the accuracy of various data extraction and processing techniques. A similar problem was attempted to be solved using clustering techniques by TUDOR et al. [2012]. Customer groups were created using the kNN clustering technique based on attribute similarity. A machine-learning approach has been suggested by Khandani et al. [2018] that projects future dangers three to twelve months ahead of time. Their predictions have prevented losses ranging from 6% to 23%. However, using this technique for long-term investments carries risk. In order to estimate the likelihood of a loan default, Addo et al. recently developed binary classifiers using actual data from deep learning and machine learning models [2020]. He observed that the accuracy provided by neural networks is not always superior to that of earlier machine learning systems. Additionally, he demonstrated that the performance of tree-based models is higher when additional hyperparameters are used. He would have had far more accuracy, though, had the writers included a feature extraction phase.

A straightforward default discriminant model is constructed in the literature [2019] using the Bayes discriminant principle. In addition, the experiment on the dataset demonstrates whether or not enterprises default can use the key quantitative financial variables to discriminate, especially when enterprises' qualitative data are lacking. The author analyzes all short-term loan enterprise financial data from a state-owned commercial bank. The authors also discovered that some issues with the firm financial statements' data quality persisted even after a number of sample washing and processing steps.

It is concluded that due to imbalanced data sets, the discriminant model's accuracy is not optimal. In conclusion, scholarly scholars have made some progress toward interpretable models; nonetheless, the accuracy of the models that are now in use is low because of unbalanced data sets. A neural network application for evaluating credit risks has been reported by Angelini et al. [2020]. She created two models: an original feed-forward NN model and a modified one with a unique feed-forward NN architecture. He came to the conclusion that since both had good accuracy, neural networks might also be used to forecast credit fraud, but he was unable to provide the model's accuracy measure. Later, on a different dataset [2023], Khashman et al. applied the same model that Angelini et al. [2018] had suggested. For the training dataset, he discovered an accuracy of 99.25%, and for the validation dataset, 73.17%. With an overall accuracy rate of 83.6%, moneylenders should not be able to trust their clients.

Neural networks outperformed classification methods, according to Soltan et al. [2020]. After using three supervised neural network layers to train three models, he discovered that the maximum weighted voting learning method yielded the best accuracy of 91.44%, which was significantly higher than that of the classification models. The effectiveness of least-squares support vector machines for classification and logistic regression for prediction was compared by Trustorff et al. [2022]. He employed a small amount of training data and an input dataset with a large variation to demonstrate how much better SVM performed on these datasets than logistic regression. We found that AdaBoosting approaches can improve classification accuracy and reduce false alarm rates after examining various relevant research studies that focused on machine learning techniques for credit card default prediction.

3. Methodology

The proposed methodology flowchart followed by experiments outlined in Fig. 1. Credit card client's dataset is used for our experiments. We collected our dataset from Kaggle open-source dataset. This dataset contains a total of 30000 instances and 25 features. First of all, we pre-processed the dataset; then, we applied different classifiers to classify the credit card clients as default and non-default. Finally, we evaluated the results of different classifiers.

3.1 Pre-processing of the data

The methods by which we pre-process our dataset are described in data pre-processing. It was necessary to preprocess the dataset before using machine learning classification methods.

3.1.1 Handling the missing value

A machine learning classification model may encounter issues if a dataset has any missing values. We utilized the Python Scikit-learn module to address the missing values in our dataset, which includes some missing values.

3.1.2 Encoding the Categorical Data

The majority of machine learning models cannot be constructed using categorical data if any feature is included. Numerical numbers must be used to encode categorical data.

3.1.3 Normalization Feature Scaling

One way to normalize the independent variable inside a range is to use feature scaling with normalization. Using (1), we scaled the values of all independent features from 0 to 1 by normalizing feature scaling. Because normalization prevents the higher value feature from superseding the lower value feature during classification, we employed it.

$$X(norm) = \frac{X - MIN(X)}{MAX(X) - MIN(X)} \qquad (1)$$

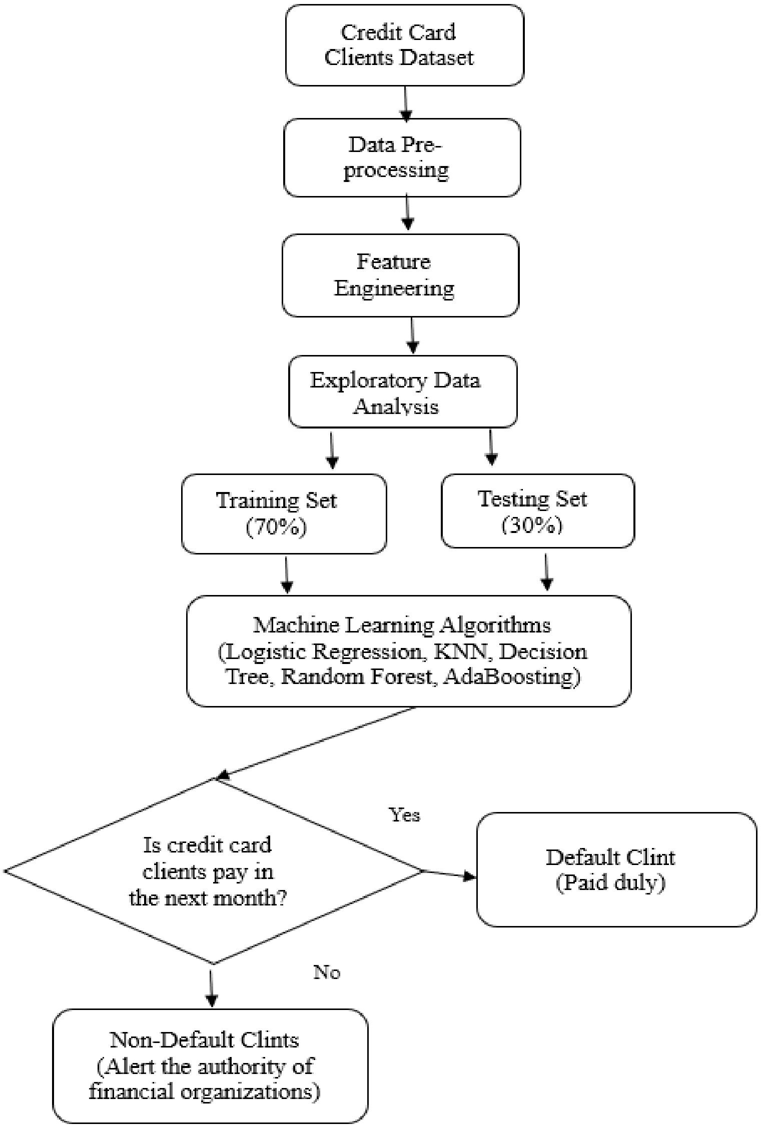
Credit Card Clients Dataset
Data Pre-processing
Feature Engineering
Exploratory Data Analysis
Training Set (70%)
Testing Set (30%)
Machine Learning Algorithms (Logistic Regression, KNN, Decision Tree, Random Forest, AdaBoosting)
Is credit card clients pay in the next month?
Yes
Default Clint (Paid duly)
No
Non-Default Clints (Alert the authority of financial organizations)

Fig. 1. Methodology flowchart for credit card default client's prediction

3.2 Exploratory Data Analysis

We have carried out EDA to get some insight into the dataset.

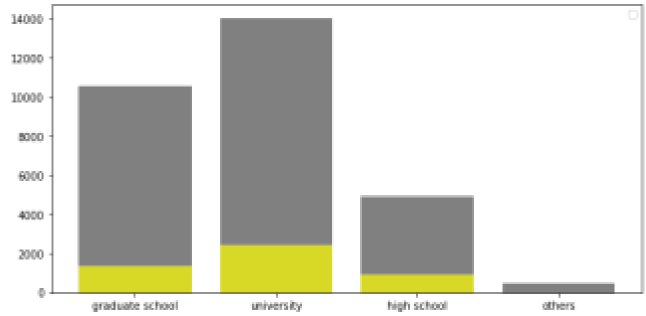

Fig. 2. Education distribution plot for the default clients

Fig. 2 shows the education distribution plot for the default clients where the education total is indicated by grey color and default education is indicated by yellow. The higher the education, the lower the probability of defaulting the next month.

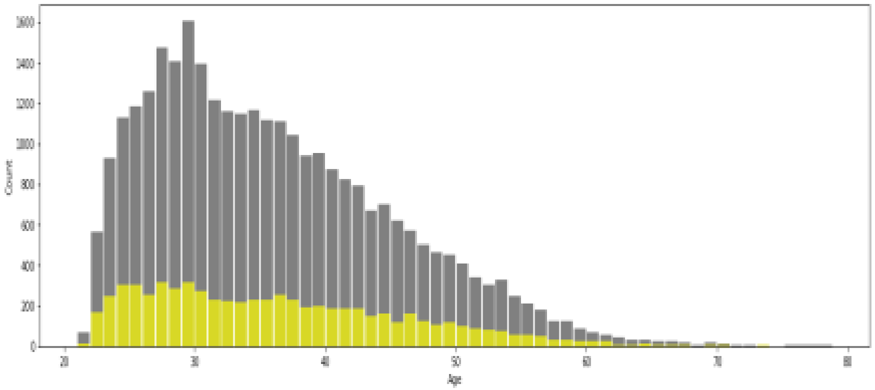

Fig. 3. Age distribution plot for the default clients

Fig. 3 shows the age distribution plot for the default clients where the age total is indicated by grey color and the default age is indicated by yellow.

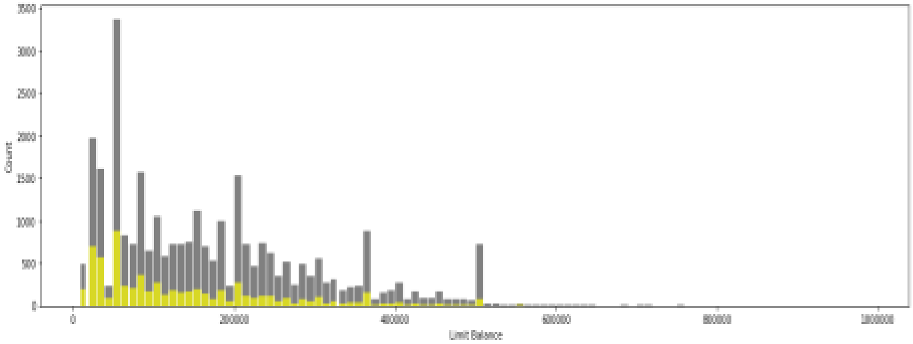

Fig. 4. Limit Balance distribution plot for the default clients

Fig. 4 shows the limit balance distribution plot for the default clients where the total limit balance is indicated in grey color, and the default limit balance is indicated in yellow. From the above histogram plot, we can say that the greater the age of the customer, the lesser the chances for that customer to be a defaulter. Also, we can infer that this credit card company has a maximum of customers aged 25 to 35 years.

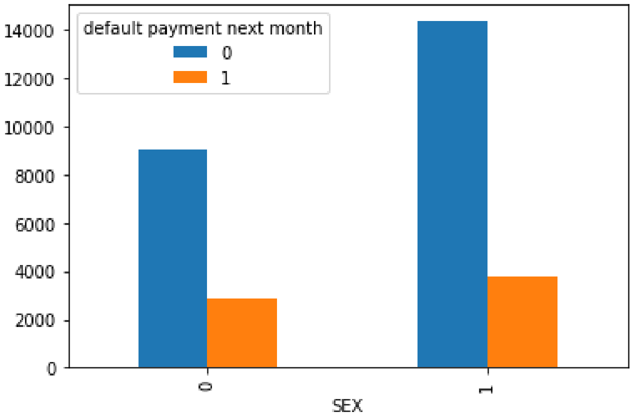

Fig. 5. Gender distribution plot for the default clients (0=male, 1=female)

Fig. 5 shows the gender distribution plot for the default clients. From the histogram plot, we can infer that the bank has more male customers than female customers.

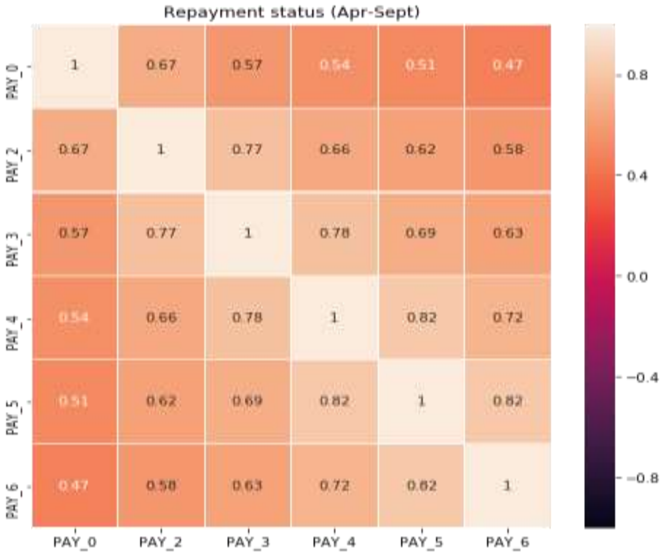

Fig. 6. Correlation matrix for the repayment status (Apr-Sept)

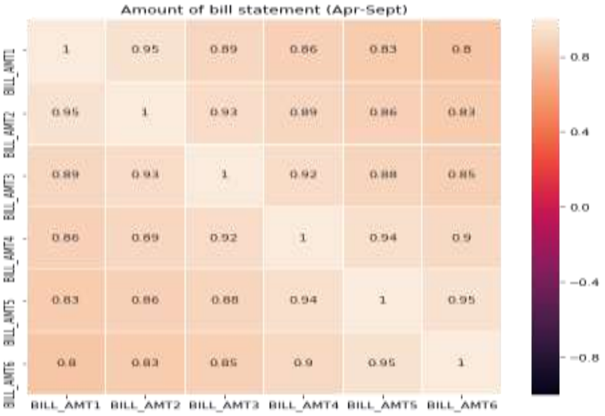

Fig. 7. Correlation matrix for Amount of bill statement (Apr-Sept)

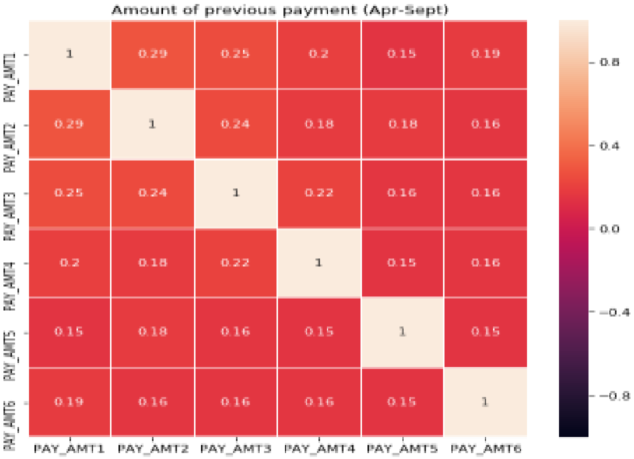

Fig. 8. Correlation matrix for Amount of previous statement (Apr-Sept)

Fig. 6 shows the correlation matrix for the repayment status for the April to September months of Pay_0 to Pay_6 of the credit card client's dataset. Fig. 7 shows the correlation matrix for the Number of bill statements for the April to September months of

Bill_ATM_1 to Bill_ATM_6 of the credit card client's dataset. Fig. 8 shows the correlation matrix for the Amount of the previous statement for the April to September months of Pay_ATM_1 to Pay_ATM_6 of the credit card client's dataset.

TABLE I. CREDIT CARD CLIENTS DATASET FEATURES

Name	Description
LIMIT_BAL	Amount of the given credit
SEX	Gender
EDUCATION	Education of the Clients
MARRIAGE	Marital Status
AGE	Age
PAY_0 to PAY_6	Repayment status (Apr-Sep)
BILL_ATM1 to BILL_ATM_6	Amount of Bill Statement (Apr-Sep)
PAY_ATM_1 to PAY_ATM_6	Amount paid (Apr-Sep)
DPNM	Default Payment Next Month (yes=1, no=0)

3.3 Machine Learning Classifiers

For credit card default client's prediction, we are applying some machine learning classifiers. In this section, we discussed these classifiers.

3.3.1 K Nearest Neighbor

Under supervised learning, the K Nearest Neighbor (KNN) classifier operates. In order to classify individual data points and members of the instance-based learning family, proximity is used. By calculating the distance between the test data point and each training set data point, KNN finds the N nearest neighbors to the test data point. Simple voting is used among the N data points in our credit card client's dataset to decide whether to classify the traffic as default or non-default.

3.3.2 Logistic Regression

Logistic regression is an approach for data classification that forecasts the likelihood of a binary class (Yes/No). A logistic function that is statistically used to predict a particular class probability is called the sigmoid function.

3.3.3 Decision Tree

A decision tree is a supervised learning classification method that divides data into leaves and decision nodes according to a preset parameter. The decision node is where the data is being split, and the ultimate result is represented by the leaves. To categorize the credit card customers, we employed the decision tree as a binary classification tree.

3.3.4 Random Forest

One supervised learning method for handling classification issues is Random Forest. It is a collection of algorithms for decision trees. Random Forest uses the Bootstrap Aggregation or Bagging Ensemble technique to select a random sample from the dataset. The credit card clients were categorized using the Random Forest ensemble technique. Based on the majority vote of several decision trees, Random Forest makes the ultimate determination.

3.3.5 AdaBoosting

AdaBoost algorithm, short for Adaptive Boosting, is **a Boosting technique used as an Ensemble Method in Machine Learning.** It is called Adaptive Boosting, as the weights are re-assigned to each instance, with higher weights assigned to incorrectly classified instances.

4. Results and Discussion

This section presents the classifier's performance evaluation. The performance measures that are frequently employed for any classification problem are recall, f1 score, accuracy, and precision. Based on the following, we computed the accuracy, precision, f1 score, and recall values to assess the performance of the classifier.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \qquad (2)$$

$$Precision = \frac{TP}{TP+FP} \qquad (3)$$

$$Recall = \frac{TP}{TP+FN} \qquad (4)$$

$$F1\ Score = \frac{2*Recall*Precision}{Recall+Precision} \qquad (5)$$

where,

TP = No. of 'Default' data points correctly classified as 'Default'.

FP = No. of 'Default' data points incorrectly classified as 'Non-default'.

TN = No. of 'Non-default'' data points correctly classified as 'Non-default''.

FN = No. of 'Non-default' data points incorrectly classified as 'Default'.

The results were evaluated for credit card clients default prediction based on the performance metrics for binary classification of two classes as 'default' and 'non-default': True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN).

Classification reports of different classifiers for default and non-default client classes are shown in Table II.

TABLE II. COMPARISON OF THE CLASSIFICATION REPORTS

Classifiers	**Classes**	**Precision**	**Recall**	**F1 Score**
KNN	Default	0.887	0.886	0.886
	Non-default	0.886	0.885	0.886
Logistic Regression	Default	0.863	0.842	0.852
	Non-default	0.852	0.874	0.853
Decision Tree	Default	0.893	0.892	0.892
	Non-default	0.892	0.892	0.891
Random Forest	Default	0.895	0.893	0.893
	Non-default	0.893	0.893	0.892
Decision Tree with AdaBoosting	Default	0.926	0.915	0.946
	Non-default	0.935	0.905	0.926
Random Forest with AdaBoosting	Default	0.936	0.915	0.926
	Non-default	0.945	0.935	0.946

It is observed from Table II that Decision Tree with AdaBoosting and Random Forest with AdaBoosting perform better than the other classifiers. We showed the comparison of different classifiers' accuracy in Table III. From Table III, it is observed that Decision Tree with AdaBoosting and Random Forest with AdaBoosting classifier obtained the highest accuracy of all the other classifiers. We executed the entire experiment in the Python Jupyter Notebook environment.

TABLE III. COMPARISON OF THE CLASSIFIERS ACCURACY

Classifiers	Classes	Accuracy
KNN	Default	88.82%
	Non-default	88.79%
Logistic Regression	Default	86.43%
	Non-default	86.37%
Decision Tree	Default	84.20%
	Non-default	84.17%
Random Forest	Default	89.41%
	Non-default	89.38%
Decision Tree with AdaBoosting	Default	92.64%
	Non-default	93.61%
Random Forest with AdaBoosting	Default	93.77%
	Non-default	94.74%

5. Conclusion

The purpose of this research is to anticipate credit card default in banks by using machine learning techniques. AdaBoosting has a prediction accuracy of over 90% based on the outcomes analysis. Before issuing a customer a credit card, banks can use machine learning to evaluate their credit risk. The primary goal of banks is to provide their customers with worthwhile goods and services, and in order to remain competitive, they must continue to be inventive and creative. Banking institutions may now reach their clientele in a more tailored way thanks to machine learning technology. Banks can gain from using analytics in the workplace in a number of ways.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Affiliation H.I and Affiliation, B.A (2009). A comparison of data mining techniques for credit scoring in banking: A managerial perspective, *Journal of Business Economics and Management*. 233-240, 2009.

[2] Bellotti T and Crook, J. (2013). Forecasting and stress testing credit card default using dynamic models, *International Journal of Forecasting*. 563-574, 2013.

[3] Dastile X and Celik, T. (2010). Statistical and machine learning models in credit scoring: A systematic literature survey, *Applied Soft Computing*

[4] Hussein A.M and Gheni, H.Q. (2020). PREDICTION OF CREDIT CARD PAYMENT NEXT MONTH THROUGH TREE NET DATA MINING TECHNIQUES, *International Journal of Computing*. 97-105, 2020.

[5] Juneja, S. (2020). Defaulter Prediction for Assessment of Credit Risks using Machine Learning Algorithms, in *4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India.

[6] Jamal K and Hu, S. (2019). Enhanced Recurrent Neural Network for Combining Static and Dynamic Features for Credit Card Default Prediction, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, UK, 2019.

[7] Khan, F.N and Israt, L. (2020). Credit Card Fraud Prediction and Classification using Deep Neural Network and Ensemble Learning, 2020 IEEE Region 10 Symposium (TENSYMP), Dhaka, Bangladesh, 2020, doi: 10.1109/TENSYMP50017.2020.9231001.

[8] Khandani A.E and Kim, A.J. (2010). Consumer credit-risk models via machine-learning algorithms, *Journal of Banking & Finance,* Journal of Banking & Finance.

[9] Keramati A and Yousefi, N. (2011). A Proposed Classification of Data Mining Techniques in Credit Scoring, in *Proceedings of the 2011 International Conference on Industrial Engineering and Operations Management*, Kuala Lumpur, Malaysia.

[10] Khan, F.N and Israt, L. (2020). Credit Card Fraud Prediction and Classification using Deep Neural Network and Ensemble Learning, *2020 IEEE Region 10 Symposium (TENSYMP)*, Dhaka, Bangladesh, 2020, doi: 10.1109/TENSYMP50017.2020.9231001.

[11] Mahboob T and Shaukat, K. (2018). An Investigation of Credit Card Default Prediction in the Imbalanced Datasets, *IEEE Access,* 12. 23-34

[12] Pasha D and Shahzad, F. (2017). Performance Comparison of Data Mining Algorithms for the Predictive Accuracy of Credit Card Defaulters, *Computer Science*.

[13] Sayjadah Y and Alotaibi, F. (2018). Credit Card Default Prediction using Machine Learning Techniques, in *Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)*, Subang Jaya, Malaysia.

[14] Sayjadah, Y and Kasmiran, K.A. (2018). Credit Card Default Prediction using Machine Learning Techniques, 2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA), Subang Jaya, Malaysia, 2018, pp. 1-4, doi: 10.1109/ICACCAF.2018.8776802.

[15] Said I and Qu, Y. (2022). A Study on the Performance Comparison of Five Popular Machine Learning Models Applied for Loan Risk Prediction, 2022 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, doi: 10.1109/CSCI58124.2022.00123.

[16] Singh, R. and Aggarwal, R. (2011). Comparative Evaluation of Predictive Modeling Techniques on Credit Card Data," *International Journal of Computer Theory and Engineering,* vol, 598-603

[17] TUDOR A.I and BÂRA, A. (2012). Clustering Analysis for Credit Default Probabilities in a Retail Bank Portfolio, *Database Systems Journal Academy of Economic Studies*. 23-30, 2012.

[18] Torvekar N and P. S. (2019). Game, "Predictive Analysis of Credit Score for Credit Card Defaulters," *International Journal of Recent Technology and Engineering (IJRTE)*. 283-286.

[19] Venkatesh A and Jacob, S.G. (2016). Prediction of Credit-Card Defaulters: A Comparative Study on Performance of Classifiers, *International Journal of Computer Applications* 36-41, 2016.

[20] Xu P and Ding, Z. (2017). An improved credit card users default prediction model based on RIPPER, in *13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, Guilin, China.

[21] Yeh, I.C and Lien, C.H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients, *Expert Systems with Applications*. 2473-2480, 2009.

[22] Yang M and Yi L. (2022). Applying machine learning algorithms to predict default probability in the online credit market: Evidence from China, *International Review of Financial Analysis*.

[23] Yu, Y. (2020). The Application of Machine Learning Algorithms in Credit Card Default Prediction, in *International Conference on Computing and Data Science (CDS)*, USA

[24] Zhang, L. and Zhou, C. (2023). Adaptive Feature Cross-Compression for Credit Default Prediction, in IEEE Access. 94322-94334, 2023, doi: 10.1109/ACCESS.2023.3309834.