TEESSIDE
UNIVERSITY

THE USE OF MACHINE LEARNING TO PREDICT CREDIT CARD DEFAULT PAYMENTS

A Report of Machine Learning ICA Project

TSEGHA MADONNA (Student)

B1215326

Contents

Abstract:

A financial institution's main responsibility is to offer credit lending services. Using technological advancements to expedite the process is advised to improve financial inclusion and fulfil the growing needs at both the macro and micro levels.Using technological advancements to expedite the process is advised to improve financial inclusion and fulfil the growing needs at both the macro and micro levels.

Financially speaking, a client's credit is granted only after they satisfy specific requirements that guarantee they will repay the principal plus interest within the allotted period.

Credit experts have taken notice of the problem of defaulting on credit obligations. In recent years, there has been evidence of the effectiveness of using machine learning algorithms to evaluate a potential client's creditworthiness. The current analysis provides insightful information that can help financial institutions reduce the credit risk they incur when giving credit to their customers. These observations are quite valuable for improving the credit procedure and reducing the probability of credit risk.

The use of machine learning models to forecast credit card defaults is described in this paper. The dataset includes 30,000 distinct cases and 24 characteristics. Eight distinct machine learning models were utilised in order to predict the likelihood of a credit card holder going into default. The correctness of the models was the main emphasis of the evaluation process. The study's conclusions showed that the XGB Classifier had the greatest accuracy rate, 90.52%, followed by Random Forest and KNN, with 89.54% and 86.15%, respectively. This document includes an examination and feature selection of the data, an extensive assessment of the pertinent literature, a set of experiments, and a discussion of the results.

***Keywords**: credit card default prediction, machine learning, XGB Classifier, Long-Short Term Memory (LSTM), Random Forest, KNN, Adabooster, Decision Tree, SVM, Logistic Regression, Gaussian Naive Bayes, accuracy, cross-validation, hyperparameter tuning, data pre-processing, feature scaling, data exploration, feature selection, imbalance, SMOTEENN, Kaggle*

Introduction:

When a debtor is unable to satisfy their credit card liabilities' payment requirements, a credit card default occurs. This is a common problem in the credit industry, and the consequences could result in significant financial losses for the creditor. As such, it is necessary to predict the likelihood of credit card default for the creditor to take appropriate action to mitigate the risk.

It has been shown that machine learning is effective at predicting credit card default.

By analysing historical data, machine learning algorithms can identify patterns and predict the likelihood of nonpayment. The current work explores the use of machine learning models to predict credit card defaults. Because of its excellent accuracy and efficiency, the banking industry has made extensive use of XGBoost, a gradient-boosting algorithm that has become quite popular. Research projects that use XGBoost and Long-Short Term Memory (LSTM) models to predict credit risk have increased dramatically in the last few years (Gao et al., 2021).

These studies' main goals have been to improve the models' accuracy and deal with the problem of unbalanced data.

Importance of The problem

The importance of credit card default rests in its ability to result in significant financial losses for the creditor, which could lead to lower income and higher expenses if legal action is necessary to recoup the unpaid balance. It is critical to foresee the possibility of credit card default to reduce this risk. By identifying borrowers who have a high possibility of defaulting, suitable steps can be implemented to manage risk exposure, such as lowering loan limits or raising interest rates.

Review of Relevant Literature

The precision of machine learning algorithms in predicting credit card default has been the subject of numerous studies. (2018) Husejinovic et al.

Seven machine learning methods were compared by (Neema and Soibam, 2017): random forest, k-NN, neural

networks, naive Bayes, decision trees, linear discriminant analysis, and SVM. The accuracy with which different algorithms predicted credit card defaults was compared in this study. The Random Forest algorithm in the study yielded the greatest results in terms of cost-accuracy balance. Review of Relevant Literature the precision of machine learning

algorithms in predicting credit card default has been the subject of numerous studies. (2018) Husejinovic et al.

Different techniques for ensemble learnin; naive Bayes, SVM, decision trees, k-NN, and logistic regression. For evaluating performance, accuracy, sensitivity, and specificity were employed. The best credit card client default detection rate (71.3%) was achieved by the naive Bayes model, while the logistic regression model had the highest overall accuracy (0.820).

Seven machine learning methods were compared by (Neema and Soibam, 2017): random forest, k-NN, neural networks, naive Bayes, decision trees, linear discriminant analysis, and SVM. The accuracy with which different algorithms predicted credit card defaults was compared in this study. The Random Forest algorithm in the study yielded the greatest results in terms of cost-accuracy balance.

(Yu, 2020) used machine learning techniques to forecast credit card defaults. The SVM method demonstrated the greatest accuracy rate of 81.1%, according to the data. In terms of precision, recall, and F1-score, the Random Forest method outperformed the other algorithms, according to the study. The predictive power of logistic regression, decision trees, neural networks, and support vector machines for credit card default has been investigated in other research.

An online learning system based on adaptive boosting was presented by (Lu et al., 2017) for the real-time detection of credit card defaults. The researchers represented application data, such as payment history, credit scores, and education, using a feature vector. Decision tree learning was improved using the AdaBoost technique by utilising original examples. Reweighting was applied to fresh data to stabilise changes in stochastic and dynamic processes.

(Gao et al., 2021) evaluated XGBoost-LSTM models for credit card user default prediction using transaction flow data. The study found that the default prediction performance of the XGBoost-LSTM model is good. As a result, the research supports deep learning developers of financial applications.

In 2021, Zhang and Chen employed the XGBoost algorithm to forecast defaults by Chinese bond issuers. The Synthetic Minority Over-sampling Technique (SMOTE) was applied since the dataset was unbalanced. The study found that the XGBoost algorithm performs effectively in skewed data. SMOTE was also effective in balancing out samples. Predicting bond defaults in emerging economies can be aided by this technique.

Ensemble approaches use many machine learning models' predictions to improve credit card default prediction. (Wang et al., 2011) compared bagging, boosting, and stacking to assess credit scores using four base learners: logistic regression, decision tree, artificial neural network, and support vector machine. Across all three credit datasets, bagging outperformed the other ensemble learning methods.

Machine learning techniques have the potential to enhance credit card default prediction when compared to conventional statistical methods.

(Yu, 2020) used machine learning techniques to forecast credit card defaults. The SVM method demonstrated the greatest accuracy rate of 81.1%, according to the data. In terms of precision, recall, and F1-score, the Random Forest method outperformed the other algorithms, according to the study. The predictive power of logistic regression,

decision trees, neural networks, and support vector machines for credit card default has been investigated in other research.

An online learning system based on adaptive boosting was presented by (Lu et al., 2017) for the real-time detection of credit card defaults. The researchers represented application data, such as payment history, credit scores, and education, using a feature vector. Decision tree learning was improved using the AdaBoost technique by utilising original examples. Reweighting was applied to fresh data to stabilise changes in stochastic and dynamic processes.

(Gao et al., 2021) evaluated XGBoost-LSTM models for credit card user default prediction using transaction flow data. The study found that the default prediction performance of the XGBoost-LSTM model is good. As a result, the research supports deep learning developers of financial applications.

In 2021, Zhang and Chen employed the XGBoost algorithm to forecast defaults by Chinese bond issuers. The Synthetic Minority Over-sampling Technique (SMOTE) was applied since the dataset was unbalanced. The study found that the XGBoost algorithm performs effectively in skewed data. SMOTE was also effective in balancing out samples. Predicting bond defaults in emerging economies can be aided by this technique.

Ensemble approaches use many machine learning models' predictions to improve credit card default prediction. (Wang et al., 2011) compared bagging, boosting, and stacking to assess credit scores using four base learners: logistic regression, decision tree, artificial neural network, and support vector machine. Across all three credit datasets, bagging outperformed the other ensemble learning methods.

Machine learning techniques have the potential to enhance credit card default prediction when compared to conventional statistical methods.

Methodology

Data exploration and Feature selection

The UCI Machine Learning Repository Credit Card Dataset provided the dataset used in this study to predict credit card default. The dataset consists of 30,000 observations and 24 attributes, each of which represents a credit, card-holding individual. The dataset includes a variety of demographic and credit-related variables that were gathered from Taiwanese client credit card records between April 2005 and September 2005, such as age, gender, educational attainment, marital status, credit limit, payment history, and outstanding balance. The credit card holder's payment default is shown by the dependent variable, which is a binary variable.

	Column	Description
0	LIMIT_BAL	Amount of the given credit (NT dollar): it inc...
1	SEX	Gender (1 = male; 2 = female).
2	EDUCATION	Education (1 = graduate school; 2 = university...
3	MARRIAGE	Marital status (1 = married; 2 = single; 3 = o...
4	AGE	Age (year).
5	PAY_0	Repayment status in September, 2005 (-1=pay du...
6	PAY_2	The repayment status in August, 2005 (scale sa...
7	PAY_3	The repayment status in July, 2005 (scale same...
8	PAY_4	The repayment status in June, 2005 (scale same...
9	PAY_5	The repayment status in May, 2005 (scale same ...
10	PAY_6	The repayment status in April, 2005 (scale sam...
11	BILL_AMT1	Amount of bill statement in September, 2005 (N...
12	BILL_AMT2	Amount of bill statement in August, 2005 (NT d...
13	BILL_AMT3	Amount of bill statement in July, 2005 (NT dol...
14	BILL_AMT4	Amount of bill statement in June, 2005 (NT dol...
15	BILL_AMT5	Amount of bill statement in May, 2005 (NT dollar)
16	BILL_AMT6	Amount of bill statement in April, 2005 (NT do...
17	PAY_AMT1	Amount paid in September, 2005 (NT dollar)
18	PAY_AMT2	Amount paid in August, 2005 (NT dollar)
19	PAY_AMT3	Amount paid in July, 2005 (NT dollar)
20	PAY_AMT4	Amount paid in June, 2005 (NT dollar)
21	PAY_AMT5	Amount paid in May, 2005 (NT dollar)
22	PAY_AMT6	Amount paid in April, 2005 (NT dollar)
23	default payment next month	Default payment (1=Yes, 0=No)

Table 1: Dataset Feature Description.

Data Pre-processing:

This pre-processing stage started by renaming a column. The purpose of this metric was to make sure that the column titles were consistent and easy to understand. The describe function was used to create the dataset's summary statistics. This improved understanding of the distribution of individual characteristics was made possible, and the data types for the variables were verified (Figure 1).

```
LIMIT_BAL                   int64
SEX                         int64
EDUCATION                   int64
MARRIAGE                    int64
AGE                         int64
PAY_0                       int64
PAY_2                       int64
PAY_3                       int64
PAY_4                       int64
PAY_5                       int64
PAY_6                       int64
BILL_AMT1                   int64
BILL_AMT2                   int64
BILL_AMT3                   int64
BILL_AMT4                   int64
BILL_AMT5                   int64
BILL_AMT6                   int64
PAY_AMT1                    int64
PAY_AMT2                    int64
PAY_AMT3                    int64
PAY_AMT4                    int64
PAY_AMT5                    int64
PAY_AMT6                    int64
default payment next month  int64
dtype: object
```

Figure 1: Checking Data Types.

Exploratory Data Analysis (EDA):

Finding patterns in the data and understanding the relationships between the variables depend on the EDA phase. First, a histogram plot was created to show the target variable's distribution graphically (see Figure 2). The target variable's distribution was found to be unbalanced, with the non-default class accounting for the majority of cases, according to the research.

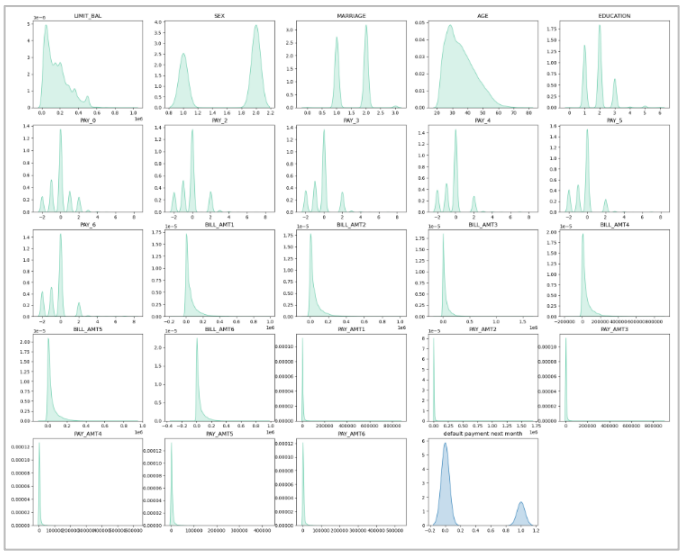

Figure 2: Density Plot

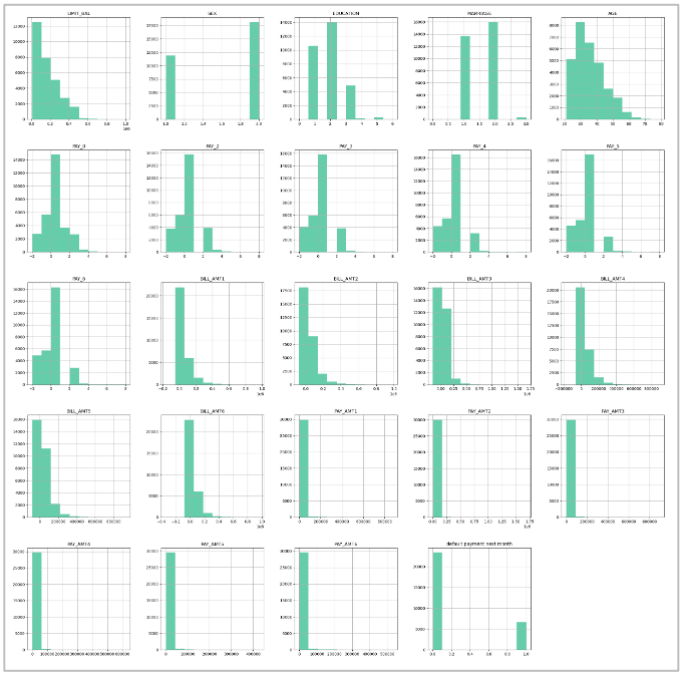

Figure 3: Histogram

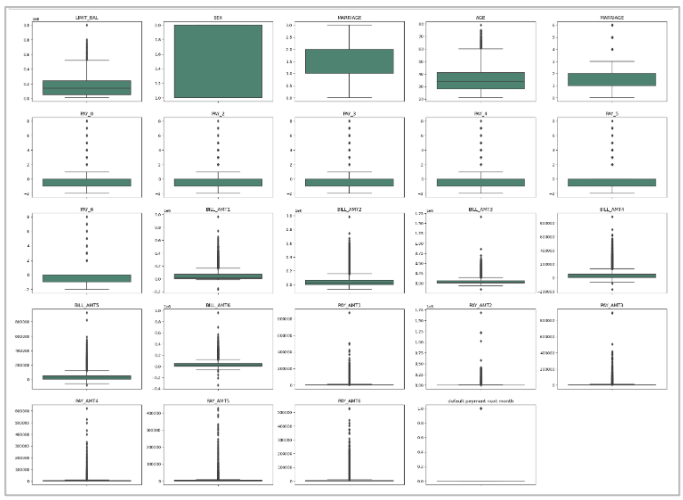

Figure 4: Box Plots

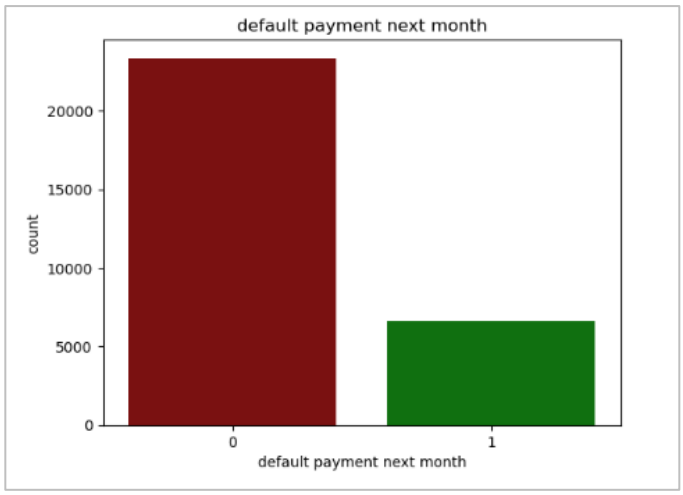

Figure 5: Bar chart showing unbalanced dataset.

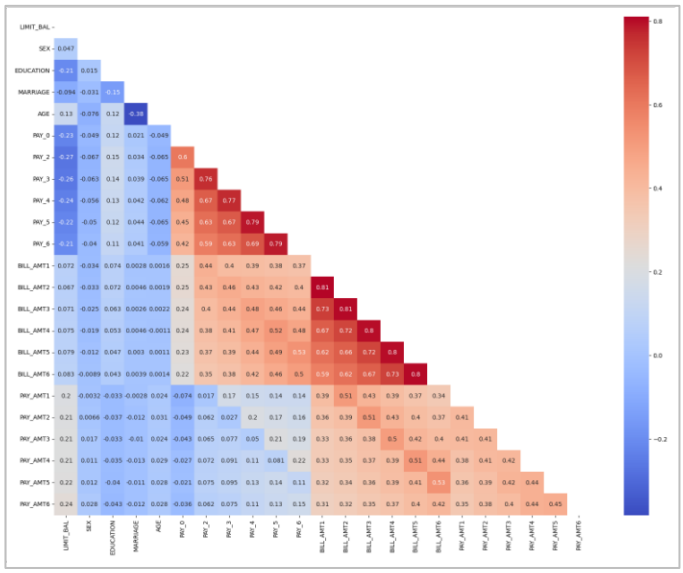

Figure 6: Correlation Matrix before balancing.

Balancing the Dataset:

The dataset's class imbalance problem was fixed using the SMOTEENN approach. By undersampling the majority class and oversampling the minority class, the SMOTEENN technique is a hybrid methodology that addresses the problem of imbalanced data. It does this by integrating the Synthetic Minority Over-sampling Technique (SMOTE) and Edited Nearest Neighbour (ENN). By using this process, the model's effectiveness was increased and the dependent variable's balance was made easier.

A correlation matrix was created to graphically depict the correlations between the features in the dataset after the target variable was balanced (Figure 8). High levels of correlation between some features were found by analysing the correlation matrix, which raised the possibility of information redundancy about the model.

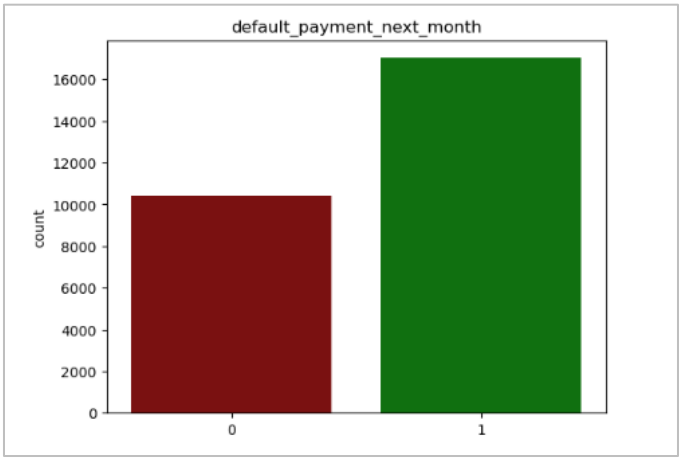

Figure 7: Bar chart showing balanced dataset.

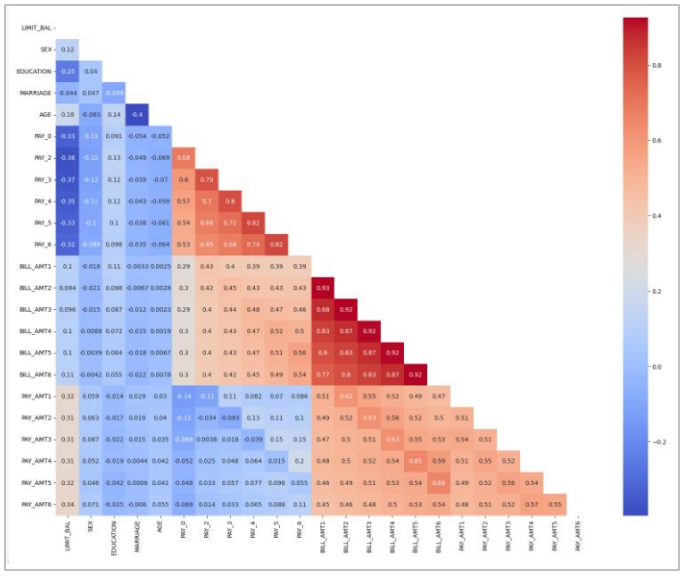

Figure 8: Correlation Matrix after balancing

Future research projects might look into different feature selection techniques to improve the model's effectiveness.

Model Training and Evaluation:

Eight machine learning models in total were used to forecast the likelihood of credit card defaults. A 70:30 ratio was used to divide the dataset into training and testing sets respectively.

Models Used

Eight machine learning models were evaluated based on how well they could predict credit card defaults. Models that were selected were:

1. XGB Classifier

2. The Forgotten Forest

3. Kendall 3. Adabooster

4. Decision Tree

5. The Logistic Regression Model

6. Naive Bayes with Gaussian

These models were chosen because of their popularity and performance in earlier studies on credit card delinquency predictions.

Evaluation Metrics

Accuracy was the evaluation metric used for the models. The accuracy metric concerns the percentage of cases that have been accurately classified relative to the total number of cases.

Cross Validation

A five-fold cross-validation method was used to evaluate the models' performance and reduce the risk of overfitting. Five separate folds were created from the dataset; four of these folds were used for training the models, while the fifth fold was used for testing. Every fold was used as the test set once during the five iterations of the process. The average accuracy throughout the five folds was used to calculate each model's final accuracy score.

Hyperparameter Tuning

To optimise the models' performance, a grid search strategy was used to find each model's ideal hyperparameters. Hyperparameters are those that are predefined and not derived from the data and are set before the model is trained. Grid search involves analysing several combinations of hyperparameters to find the best setup that provides the best performance. For every hyperparameter, a range of values was used, and the set of parameters that produced the highest accuracy score was selected.

The models' memory, accuracy, precision, and F1 score were evaluated. The results obtained are shown in (Table 2) below.

Model	Accuracy	Precision	Recall	F1-score
Random Forest	89.7%	91.0%	93.0%	92.0%
KNN	86.15%	86.0%	92.0%	89.0%
XGB Classifier	86.12%	88.0%	85.0%	89.0%
Adabooster	80.99%	84.0%	87.0%	85.0%
SVM	75.0%	73.0%	94.0%	83.0%
Decision Tree	74.4%	76.0%	87.0%	81.0%
Logistic Regression	72.84%	73.0%	90.0%	81.0%
Gaussian Naive Bayes	68.31%	67.0%	96.0%	79.0%

Table 2: Performance Metrics of the Models.

The results showed that, with a score of 90.52%, the XGBoost Classifier had the highest degree of accuracy. The Random Forest algorithm and the K-Nearest Neighbours (KNN) approach came next, with 89.54% and 86.15%, respectively, of the points. The Random Forest model performed better overall as seen by its highest precision, recall, and F1-score. Support Vector Machine and Gaussian Naive Bayes models demonstrated lower performance metrics, which indicates that they are less effective in predicting credit card defaults.

Results:

Our research's conclusions imply that applying machine learning models to the field of credit card default prediction can be successful. The pre-processing of the data was accomplished by using feature scaling. The target variable was balanced through the application of the SMOTEENN approach. Accuracy was the evaluation metric used for the models. Using a value of k = 5, k-fold cross-validation was used to evaluate the models' effectiveness and reduce the risk of overfitting. To find each model's ideal hyperparameters, a grid search was used.

The precision metrics of the eight models are shown in the following table both before and after grid search and cross-validation were applied.

Model	Accuracy (before CV)	Accuracy (after CV and grid search)
XGB Classifier	86.12%	90.52%
Random Forest	89.72%	89.54%
KNN	86.15%	85.37%
Adabooster	80.99%	82.5%
Decision Tree	81.95%	81.22%
SVM	75.00%	78.29%
Logistic Regression	72.84%	72.85%
Gaussian Naive Bayes	68.0%	65.94%

Table 3: Models' Acuracy.

After using grid search and cross-validation, the XGB Classifier and Random Forest models achieved the most notable accuracy scores of 90.52% and 89.54%, respectively, as shown in the table. The results indicate that machine learning models can be used to predict credit card defaults with a high degree of accuracy. The models' precision can be further increased by applying cross-validation and grid search techniques.

Further investigation could be carried out to improve the effectiveness of the models by adding more relevant factors or using different approaches, such as group approaches.

Discussion:

Our research's conclusions suggest that using machine learning models can be a useful tool for predicting credit card defaults. The models with the best accuracy scores

were the Random Forest and XGB Classifier models, indicating that they are the most effective models for the task at hand.

Interestingly, it can be observed that the models' accuracy might be improved by adding more features or by utilising more advanced techniques like deep learning. Moreover, while accuracy is an important evaluation statistic, using other metrics like precision, recall, and F1-score could provide a more comprehensive assessment and evaluation of the models' functionality.

It is also imperative to take into account any potential biases that can be present in the models and data. When new data is added to the models, they cannot perform well at generalisation if the dataset is not representative of the population. Furthermore, the models can reinforce or even worsen preexisting biases if they are trained on incomplete data. Therefore, it is essential to carefully review and address everything potential biases in the models and data.

In the end, it is imperative to take into account the ethical implications linked to the application of machine learning algorithms in the credit card default prediction domain. Inequalities may be maintained or even worse if the models are applied to deny credit to particular people or groups. As such, it is critical to ensure that models are used fairly and ethically and to carefully consider the potential ethical and social consequences of their use.

Conclusion

In conclusion, the current study examined the use of machine learning algorithms for credit card default prediction. The dataset consists of several demographic and credit-related characteristics, such as age, gender, marital status, educational attainment, credit limit, payment history, and outstanding balance. The target variable was a binary value that represented the possibility of a credit card holder payment default.

Eight machine learning models were evaluated based on how well they could forecast credit card defaults. The XGB Classifier, Random Forest, KNN, Adabooster, Decision Tree, SVM, Logistic Regression, and Gaussian Naive Bayes were the models used in the investigation. Accuracy was the evaluation metric used for the models, and k-fold cross-validation with k = 5 was performed to reduce the risk of overfitting. Grid search was used to find each model's ideal set of hyperparameters.

The results of grid search and cross-validation revealed that the Random Forest and XGB Classifier models had the highest accuracy scores, 89.54% and 90.52%, respectively. According to the results, machine learning models may be used to forecast credit card defaults with a high degree of accuracy. This prediction accuracy can be further increased by applying cross-validation and grid search techniques.

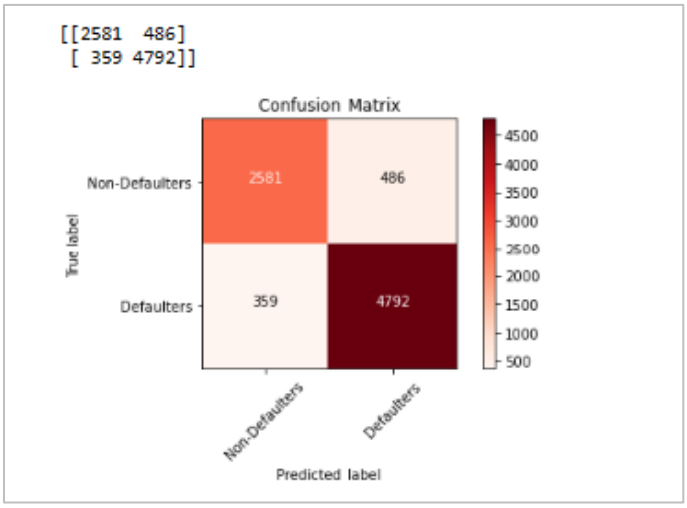

Figure 9: Model Confusion matrix

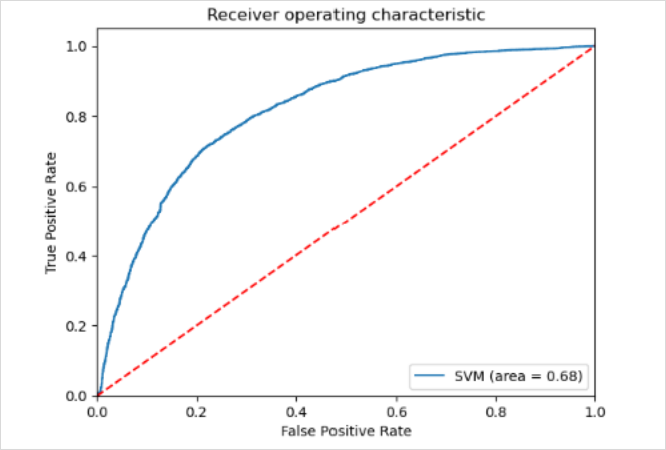

Figure 10: SVM ROC

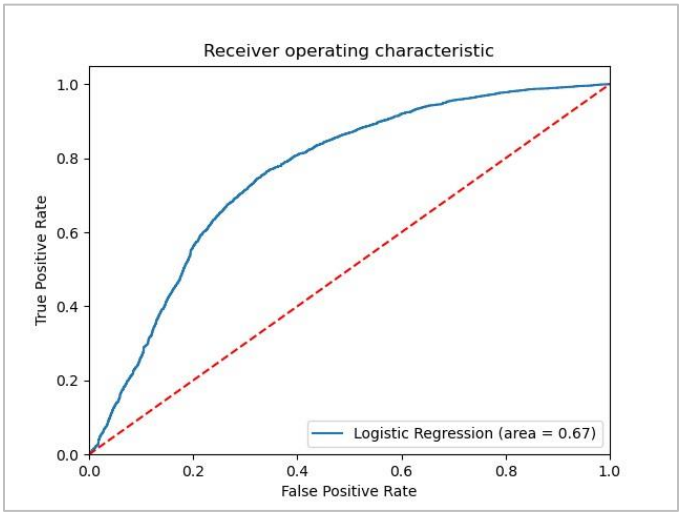

Figure 11: Logistic Regression ROC

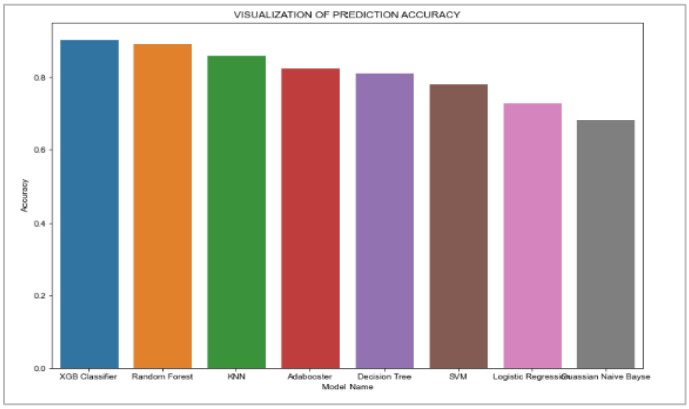

Figure 11: Comparative analysis of the models

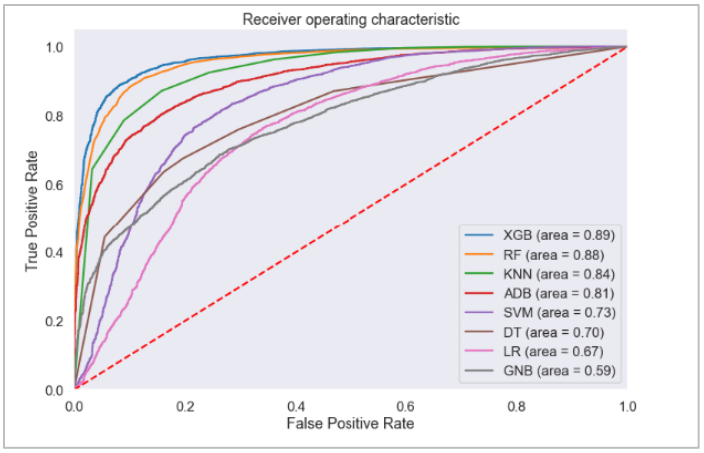

Figure 12: Consolidated ROC Curves of the models

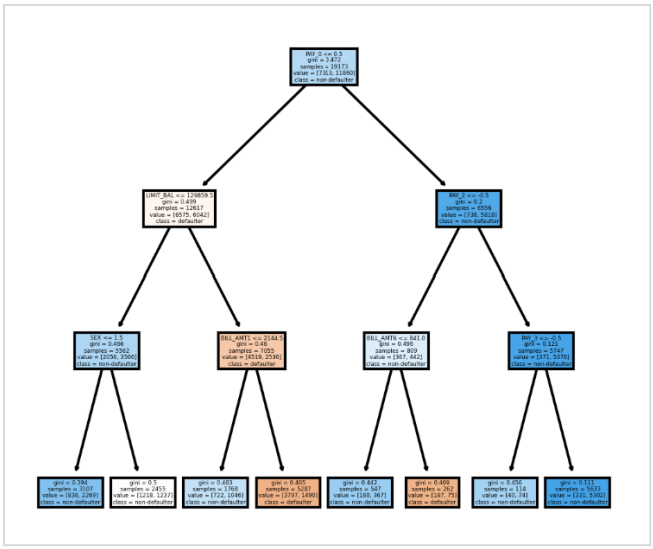

Figure 13: Decision Tree Model

Future work

Although they can be improved, our research demonstrates that machine learning systems are capable of predicting credit card defaults. Future study areas could look into:

Extra Characteristics/Features: Even with the credit-related and demographic characteristics in the dataset, extra variables could increase model accuracy. The employment status, income, and other financial obligations of the credit card holder may provide insight into their ability to make payments.

Other models: While the Random Forest and XGB Classifier models did well in our investigation, there is a chance that other machine learning models would do better. Deep learning methods can be used by neural networks to predict credit card defaults.

Time series data were included in the study using a cross-sectional dataset. Credit card users' payment habits can shift as a result of unforeseen spending or changes in income. By collecting dynamic components, time series data may increase the precision of the model.

Generally speaking, however, there are lots of chances to enhance machine learning models that anticipate credit card default.

References

GAO, J., SUN, W. & SUI, X. 2021. Research on Default Prediction for Credit Card Users Based on XGBoost-LSTM Model. *Discrete Dynamics in Nature and Society,* 2021**,** 1-13.

HUSEJINOVIC, A., KEČO, D. & MASETIC, Z. 2018. Application of Machine Learning Algorithms in Credit Card Default Payment Prediction. *A Husejinovic, D Keco, Z Masetic, International Journal of Scientific Research,* 7**,** 425-426.

LU, H., WANG, H. & YOON, S. W. Real time credit card default classification using an adaptive boosting-based online learning algorithm. IIE Annual Conference. Proceedings, 2017. Institute of Industrial and Systems Engineers (IISE), 422-427.

NEEMA, S. & SOIBAM, B. 2017. The comparison of machine learning methods to achieve the most cost-effective prediction for credit card default. *Journal of Management Science and Business Intelligence,* 2**,** 36-41.

WANG, G., HAO, J., MA, J. & JIANG, H. 2011. A comparative assessment of ensemble learning for credit scoring. *Expert systems with applications*, 38**,** 223-230.

YU, Y. The application of machine learning algorithms in credit card default prediction. 2020 International Conference on Computing and Data Science (CDS), 2020. IEEE, 212-218.

ZHANG, Y. & CHEN, L. 2021. A study on forecasting the default risk of a bond based on xgboost algorithm and over- sampling method. *Theoretical economics letters,* 11**,** 258.

Appendixes:

Codes written for the project will be shared in a separate document.