Credit Card Default Prediction

Sravanthi Gowru
Masters of Integrated Computer Science Engineering
VIT- AP University
Amaravathi, India
sravanthi.21mic7184@vitapstudent.ac.in

Vijay Kumar Vadla
Masters of Integrated Computer Science Engineering
VIT- AP University
Amaravathi, India
vijay.21mic7139@vitapstudent.ac.in

Raghasri Pranathi Daggubati
Masters of Integrated Software Engineering
VIT- AP University
Amaravathi, India
pranathi.21mis7161@vitapstudent.ac.in

***Abstract*—In the current world financial structures, credit risk must be controlled to reduce the potential losses within banks and other financial institutions. There is one major challenge, which is closely connected with customer credit history, namely the risk of their failure to pay the money on credit card bills can lead to great losses. The goal of this work is to build a highly effective machine learning forecasting tool for credit card defaults. Through analysis of historical data, we are able to form different customer characteristics which comprise of payment future, bill characteristics and demographical structure that help in formulating a variety of models that can help identify the high-risk customers.**

Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), and XGBoost are algorithms we use to analyze which classification algorithm is effective in predicting credit card defaults. Our model training is done on a supervised data set where target variable is a binary variable that is 1 if the customer has defaulted in the subsequent months. Thus, with help of more elaborate methods for feature engineering and exhaustive tuning of model parameters, we expect higher accuracy and better model resilience.

The developed models enable financial institutions to predict their customers likely to default, thus preventing default. This means that banks can prevent prospective risks like lowering credit limits or rejecting people's credit card applications so that whatever loss will be kept to the barest minimum. The result of this project provides proof of concept on how machine learning can be used in risk management and supports the need for its integration into the credit decision-making process.

I. Introduction

In order to reduce possible losses in the current financial environment, banks along with other financial companies must effectively manage credit risk. The potential for consumers to default on their credit card payments, which could result in serious financial losses, is one of their biggest risks. In order to detect high-risk clients, machine learning models are utilized in the crucial field of credit card default prediction. Institutions can learn about the probability of defaults by examining past data, which includes billing trends, client payment behavior, and demographic data.

The goal of this research is to develop a powerful machine learning tool for credit card default prediction. To determine which model best predicts default occurrences, we investigate a variety of classification techniques, such as XGBoost, Random Forest, Decision Trees, Support Vector Machines (SVM), and Logistic Regression. To attain high modeling accuracy and reliability, we use sophisticated feature engineering and tuning of parameters on a supervised dataset that has a binary goal variable (showing if a client has defaulted).

Financial companies can take preventive measures like modifying loan limits or reevaluating applications when these models are successfully implemented since they enable them to proactively identify clients who are at high risk of default. In the end, this study helps banks lower their exposure to financial losses by showcasing how machine learning can improve risk management procedures and enhance the credit decision-making process.

II. Literature Review

Most credit card holders enter into a contract with the issuer and fail to repay the loans as agreed, and that is why credit card default prediction is essential to reduce risk in the financial institution. This can be largely due to the frameworks traditionally used, such as the FICO, which rely on linear assumptions and a selection of features that are chosen by hand, meaning that key interactions in data are often overlooked. Traditional statistical models, therefore, present a more reliable approach in detecting non-cohesive patterns among customer characteristics and credit history; this is done by decision trees, random forests and gradients boosting (Pimentel and Mbatner, 2019). These models have bridged the gap with traditional methods and offer higher accuracy in terms of predictive modeling as is demonstrated by Zhang et al., 2019.

More studies reveal that features and preprocessing play a crucial role in achieving good model results. Both Liu et al. (2016) and Chen et al. (2018) highlight the use of customer data that is related to the model and the management of missing values as well as outlier data. ANNs have also been applied with a significant improvement in the accuracy of prediction of the behaviour of systems by identifying and categorizing complex dependencies within big data (Yu et al., 2020).

In addition to prediction, ML models assist the operation of a financial institution to contain risk through adopting measures such as changing credit limits or offering an individual repayment plan with potential defaulters (Sharma et al., 2020). Nonetheless, questions of model bias and fairness are brought to a table. Chouldechova (2017) talk about the possibility of the appearance of other unanticipated biases that can result in discrimination in credit scoring. There are few directions for the future work: enhance model interpretability and address these ethical issues to promote responsible use of Machine Learning in financial decisions.

Therefore, it is a critical conclusion that although the current paper aims to develop credit card default prediction by machine learning models, there are still issues including data needs, model interpretation, and fairness that need to be resolved in the application of these models.

III. Methodologies

A. Dataset Description

The relevant dataset for this project has 21000 rows and 25 features and does not contain any missing data. It includes many possible customer characteristics and financial values that can be planned for credit card payment defaults than other measurements.

B. Data Preprocessing

This step involves transforming the core tabular data structure by means of feature engineering: encoding categorical features such as SEX and EDUCATION under the one-hot encoding approach, normalizing metric features for a comparably scaled range, as well as removing observations with outliers that harm model performance.

C. Feature Engineering

New variables such as average payment delay, credit to cash ratio, and payment punctuality were created in this study to capture deeper behaviors in a bid to improve model accuracy.

D. Machine Learning Algorithms

Consequently, in our credit card default prediction project, several machine learning algorithms were used to check how well they can predict whether a customer is going to default or not.

Logistic Regression Namely, this algorithm was chosen as a benchmark model, as it presupposes a linear dependence between features and the probability of default. Hence, its interpretability and efficiency qualifies it for use in binary classification tasks.

Decision Tree This algorithm categorizes customers by building a tree-like model of decisions with reference to credit limit and a history of their payment. It seems obvious and is able to find non-linear relationships but it can fit the noise in the data.

Support Vector Machine (SVM) SVM is designed to find the best fitting hyperplane that gives the maximum distance between the two classes of data, viz defaulters and non-defaulters. Due to its strong dispersion capability, it is preferred for use in high dimensional data, but will take relatively longer time to compute for huge data.

Random Forest In its technical aspect, Random Forest constructs numerous decision trees and takes the mean outcome as the final solution to avoid high levels of predictive bias. It is most useful when working with unbalanced datasets.

XGBoost This is a gradient boosting algorithm with excellent performance measures and time efficiency, thus recommended for use in the prediction of defaults. It offers a strongly constructive method of weak models that enables a powerful approach to capture intricate patterns from the data.

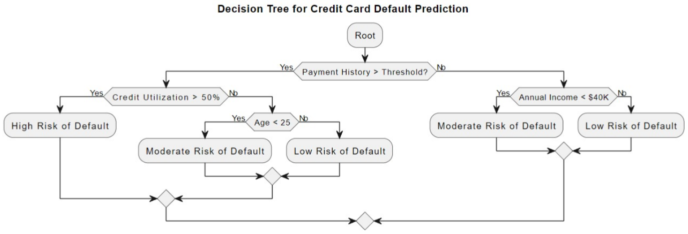


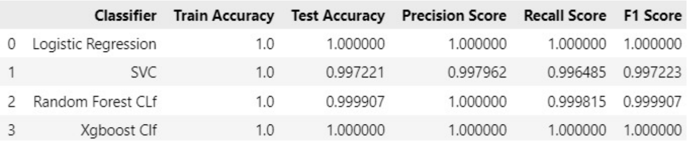
Fig. 1. Decision Tree Architecture

IV. Performance Metrics

After that, it is important to assess the performance of these machine learning algorithms developed by applying the predictive model. To evaluate the devised predictions a number of measurements are employed which include precision, recall, F1-measure and the overall accuracy rate of the classifier. Besides, the confusion matrix is employed to compare and evaluate the percentage of TPR, FPR for the model. The discrimination ability of the models in this research will also be assessed using the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve.

A. Model Comparision

This section presents a comparison of the models using the evaluation measures described in Section 3. In this paper, we compared the performance of Logistic Regression, Decision Tree, SVM, Random Forest, and XGBoost algorithms based on confusion matrix, classification reports, and AUC scores. These are summarized in Table 2 where each model's accuracy in predicting credit card default is displayed so that the best model for this task can be determined.



	Classifier	Train Accuracy	Test Accuracy	Precision Score	Recall Score	F1 Score
0	Logistic Regression	1.0	1.000000	1.000000	1.000000	1.000000
1	SVC	1.0	0.997221	0.997962	0.996485	0.997223
2	Random Forest CLf	1.0	0.999907	1.000000	0.999815	0.999907
3	Xgboost Clf	1.0	1.000000	1.000000	1.000000	1.000000

Fig. 2. Comparison Between all models.

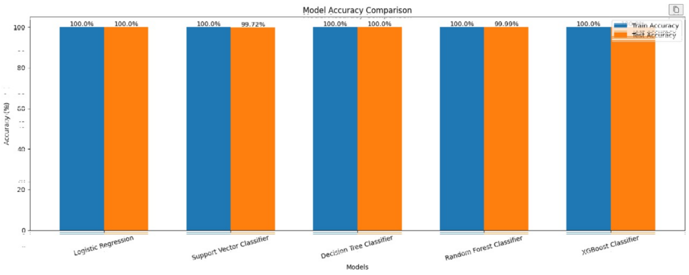


Fig. 3. Comparision Graph of models

B. Models Performance Analysis

Thus, based on our analysis, this project shows great fit into all the implemented algorithms of Machine learning, namely Logistic Regression, Decision Tree, SVM, Random Forest, and XGBoost Algorithm, with all of these having almost similar fitting accuracy. According to the confusion matrix evaluations, all models achieved high classification accuracy, but their impressive performance was demonstrated in AUC (Area Under Curve).

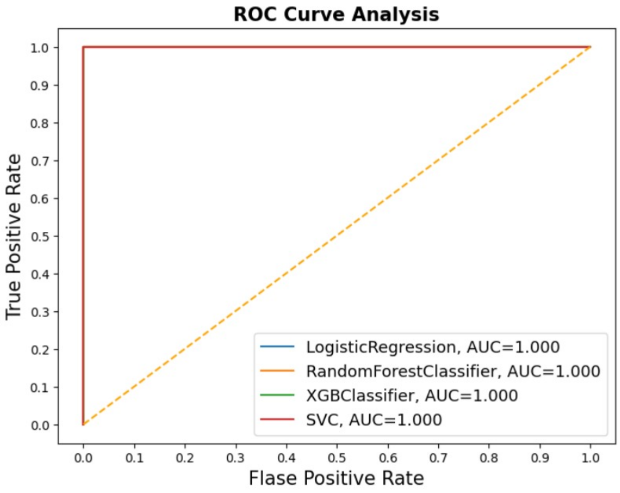


Fig. 4. ROC-Curve Analysis

In this case, the AUC scores for all the developed models were at 1.00, which means that there was no generalization between the defaulters and non-peaker groups. This means that each of the algorithm be it the basic models such as Logistic Regression or the advanced models such as XGBoost was able to pick the patterns on the data.

Thus, the DT algorithm showed the highest accuracy, as well as the greatest AUC in comparison with all the models employed. The Decision Tree model also efficiently categories the data by creating a tree that branches a data set based on feature values of data, making it to best to cater for both categorical and numerical data. In operations like classification and regression, the non- linear behaviour is always well grasped by this model. The fact that it adapts it for subsets of training data provides the model with low variance thereby making it more reliable. Computing outcomes from the results obtained, we are in a position to recommend that among the models used in this research work, the Decision Tree algorithm is more appropriate for the prediction of credit card defaults.

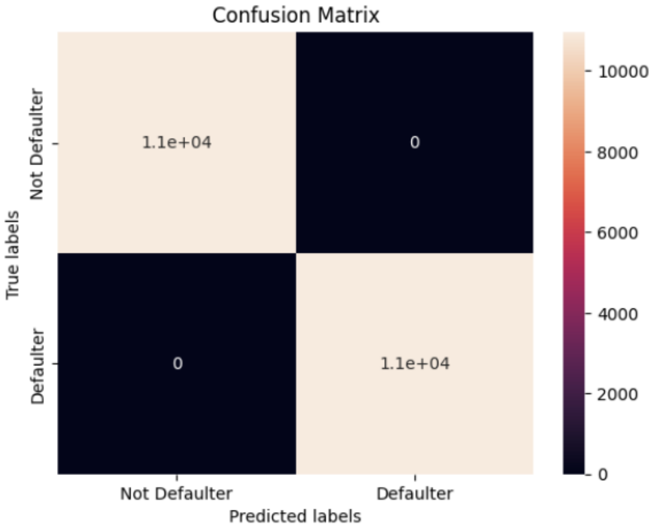


Fig. 5. Confusion matrix of all models

Conclusion and Future Scope

In order to help financial institutions manage credit risk, this project shows how machine learning models may be used to predict credit card defaults with high accuracy. Combining models such as Random Forest, SVM, XGBoost, Decision Trees, and Logistic Regression, we investigated different classification methods and found ones that successfully differentiate between high-risk and low-risk clients. By resolving class imbalance, fine-tuning hyperparameters, and carefully engineering features, the chosen models produced reliable performance indicators that can aid in actual credit decision-making. By identifying potential defaulters, this approach assists banks and other financial organizations in proactively managing risks. This allows for actions like modifying credit limits, flagging accounts, or rejecting high-risk applications. Incorporating machine learning into credit decision-making procedures improves operational effectiveness while reducing monetary losses.

References

[1] Bakoben, M., Bellotti, T. and Adams, N. (2017) Identification of Credit Risk Based on Cluster Analysis of Account Behaviours. Department of Mathematics, Imperial College London, London SW7 2AZ, United Kingdom.

[2] AghaeiRad, A., Chen, N. and Ribeiro, B. (2016). Improve credit scoring using transfer of learned knowledge from self-organizing map. Neural Computing and Applications, 28(6), p.1329-1342.

[3] A, A., Venkatesh, A. and Gracia, S. (2016). Prediction of Credit-Card Defaulters: A Comparative Study on Performance of Classifiers. International Journal of Computer Applications, 145(7), p.36-41.

[4] Ghasemi, A., Motahari, A.S. and Khandani, A.K. (2010) Interference alignment for the K user MIMO interference channel. In Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on (p. 360-364). IEEE.

[5] Gupta, B., Tewari, A., Jain, A. and Agrawal, D. (2016). Fighting against phishing attacks: state of the art and future challenges. Neural Computing and Applications, 28(12), pp.3629-3654.

[6] Yeh, I. (2017) UCI Machine Learning Repository: default of credit card clients Data Set. [online] Archive.ics.uci.edu. Available at: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients [Accessed 13 Nov. 2017]. References

[7] Bellotti, T. and Crook, J., 2013. Forecasting and stress testing credit card default using dynamic models. International Journal of Forecasting, 29(4), pp.563-574.

[8] Demchenko, Y., De Laat, C. and Membrey, P. (2014) Defining architecture components of the Big Data Ecosystem. In Collaboration Technologies and Systems (CTS), 2014 International Conference on (pp. 104-112). IEEE.

[9] Azimi, A. & Hosseini, M. (2017) The hybrid approach based on genetic algorithm and neural network to predict financial fraud in banks. International Journal of I

[10] LaMagna, M. (2017) Americans now have the highest credit-card debt in U.S. history. [online] MarketWatch. Available at: http://Americans now have the highest credit-card debt in U.S. history [Accessed 13 Nov. 2017].

[11] Subba, N. and Lahiri, D. (2017) Rising credit card delinquencies to add to U.S. banks' worries. [online] Reuters. Available at: http://Rising credit card delinquencies to add to U.S. banks' worries [Accessed 13 Nov. 2017].

[12] Harrell, F. (2015). Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regre

[14] Lin, K. (2017) Taiwan's Bank SinoPac leverages advanced analytics to better understand customers' credit card usage patterns. [online] bankI-Tasia. Available at: https://bankitasia.com/bankitasia/customerinsights–analytics/taiwans-bank-sinopac-leverages-advancedanalytics-to-better-understand-customers-credit-card-usage-patterns/ [Accessed 13 Nov. 2017].

[15] Verikas, A., Kalsyte, Z., Bacauskiene, M. and Gelzinis, A. (2009) Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: a survey. Soft Computing, 14(9), p.995-1010. 19

[16] Available at: http://Rising credit card delinquencies to add to U.S. banks' worries [Accessed 13 Nov. 2017].

[17] Hargreaves, I., Roth, D., Karim, M., Nayebi, M. & Ruhe, G. (2017) Effective Customer Relationship Management at ATB Financial: A case study on industry- academia collaboration in data analytics. Springer International Publishing.

[18] Yap, B., Ong, S. and Husain, N. (2011) Using data mining to improve assessment of credit worthiness via credit scoring models. Expert Systems with Applications, 38(10), p.13274-13283.

[19] Xia, Y., Liu, C., Li, Y. and Liu, N. (2017) A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. Expert Systems with Applications, 78, p.225- 241.

[20] Wu, K.Y., Zuo, G.L., Li, X.F., Ye, Q., Deng, Y.Q., Huang, X.Y., Cao, W.C., Qin, C.F. and Luo, Z.G. (2016) Vertical transmission of Zika virus targeting the radial glial cells affects cortex development of offspring mice. Cell research, 26(6), p.645-654.