Title Page:

Prediction of credit card approval using XGB classifier over Decision tree algorithm

Seeram Balu[1] , Dr.K.Somasundaram[2]

Seeram Balu[1]

Research Scholar,

Department of Computer Science and Engineering,

Saveetha School of Engineering ,

Saveetha Institute of Medical And Technical Sciences,

Saveetha University, Chennai, Tamil Nadu, India, Pin: 602105.

seerambalu18@gmail.com

Dr.K.Somasundaram[2]

Project Guide, Corresponding Author,

Department of Computer Science And Engineering,

Saveetha School of Engineering ,

Saveetha Institute of Medical And Technical Sciences,

Saveetha University, Chennai, Tamil Nadu, India, Pin :602105.

subbiahs.sse@saveetha.com

Keywords: Credit Card approval, Decision Tree Algorithm, XGB classifier, Preprocessing data, Data Manipulation, Hyper parameters, *Reduction*.

ABSTRACT

Aim: With the help of this study, a model that predicts whether a financial institution will be able to approve credit cards for its customers or not. This model can assist a company in making an informed decision regarding whether to accept or refuse a card in order to prevent fraud, which can cause financial institutions to lose money. **Materials and methods:** Classification is performed by XGB classifier(n=10) over Decision tree algorithm (n=10) is for false rate detection. The statistical test difference between g-power 0.08 and alpha value is α=0.05. **Result and Discussion:** The analysis of the result shows that the XGB classifier algorithm has a high accuracy compared to Decision tree algorithm. The mean value is 90.35 and mean accuracy of detection is +/-1SD and significant value is p=0.001, ($p<0.05$) from independent sample T size. This indicates that there is a statistically significant difference between the two algorithms**.** **Conclusion:** The model predicts and examines whether a financial institution will provide a credit card or not to its customers.

Keywords: Credit Card approval, Decision Tree Algorithm, XGB classifier, Preprocessing data, Data Manipulation , Hyper parameters, *Reduction*.

INTRODUCTION

Lending institutions must be able to accurately assess customer credit risk. Credit scoring is a frequently used approach that aids financial firms in determining whether or not to issue credit to a credit application(Anolli, Beccalli, and Giordani 2013). The precise assessment of an applicant's credibility enables banking firms to increase the amount of credit provided while lowering the risk of loss. A reduction in credit score is done if the client is unwilling to repay their loan repayment. In the last few decades, the credit business has seen remarkable expansion(L. Thomas, Crook, and Edelman 2017). Due to the rising number of possible applicants, sophisticated approaches were developed to automate the credit approval process and monitor the borrower's financial health. Because of the enormous number of loan portfolios, even little improvements in scoring accuracy could save financial institutions a lot of money(Yang and Zhou,)(Sharma, Vashistha, and Gupta 2021).A reduction in credit score is done if the client is overdue on their loan repayment.

In the previous four years, Google Scholar has published 17,400 articles on credit card approval using machine learning, with 1,773 articles available on ScienceDirect. Using a credit card is one of the most well-known methods(Bali 2007). A reduction in credit score is done if the client is unwilling to repay their loan repayment. In order to maintain a low credit utilization rate, consider reducition your spending or making periodic bill payments throughout your billing cycle.The majority of people use credit cards because they are a convenient way to pay(Beerbaum and Ahmad). Diverse financial organisations, such as national and private

institutions, use consumer information such as essential information, lifestyles, compensation, term and monthly returns, and current livelihood way of making money to reach a consensus (Roncalli2020). All of this information is analysed before an application is evaluated(Bazarbash 2019).

The purpose of a model of creditworthiness model is to divide credit applicants into two groups: "good credit" applicants who are likely to reimburse their debts, and "poor credit" applicants who should be rebuff credit card due to a high risk of defaulting on their debts(Molnar 2019)(Hamdoun and Rguibi 2019(Bloch 1971)(Weber 2012).

MATERIALS AND METHODS

SIMATS Saveetha School of Engineering's computer science and engineering department carried out the study. Each group was given a total of n=10 iterations in order to increase accuracy . The data set taken from the kaggle website. 438557 rows and 18 columns make up the credit card approval dataset. With 95% confidence and 80% pretest power, the experimental arrangement is maintained.

The dataset which proposed work used in this paper is credit card approval prediction dataset. The Dataset was collected from the open source Kaggle platform. The Hardware configuration were HP i5 processor with a RAM size of 12GB was used.The system type used was 64-bit, OS, x64 based processor with HDD of 917 GB. Windows was used as the operating system, and the Python programming language was employed with the Jupyter Lab tool.

XGB classifier

Extreme Gradient Boosting (XGBoost) is a robust and distributed gradient-boosted decision tree (GBDT) machine learning framework(Misra et al. 2021). It is the primary machine learning application for regression, classification, and ranking problems, and it also supports parallel tree boosting (Gunjan et al. 2020). To comprehend XGBoost, you must first conceive the machine learning concepts and mechanisms that it is based on: supervised machine learning, decision trees, ensemble learning, and gradient boosting are all forms of machine learning techniques.(Brownlee 2016). Before making data into two sections, data manipulation is an important task. Data Manipulation deals with missing values, string data, which parameters should be considered etc(Ranganathan, Chen, and Rocha 2020). Preprocessing data of categorical features by using one hot encoder. Because the model will not accept object data types, preprocessing data of categorical features is mandatory because preprocessing data convert data into integer or float type

Pseudocode for XGB classifier

Input: Dataset for predicting credit card approval.
Step-1: Import and read the dataset.
Step-2: Choose the features at random from the dataset.
Step-3: Data manipulation
Step-4: Preprocessing data of the categorical data
Step-5: Create XGB classifier model.
Step-6: Train the XGB classifier model using the train dataset.
Step-7: Calculate the error between expected output and actual output .
Output: Accuracy in %

Decision Tree Algorithm

The procedure in a decision tree kickoff with the root node of the tree, analogizes the values of copious attributes, and then proceeds to the abut branch until it reaches the culminate leaf node(Chandra and Kuppili 2011). It employs various techniques to examine the split and variables that allow for the most homogeneous population groups(Kim 2003)**.** Before making data into two sections, data manipulation is an important task. Data Manipulation deals with missing values, string data, which parameters should be considered etc. In decision tree algorithm dealing with hyper parameters is tricky because of hyper parameters the model can become overrated or underrated(R. M. Rahman and Hasan 2011). Reduction of decision tree is not allowed. By reduction the algorithm will become overrated or underrated. Maximum tree depth, number of decision trees, minimum number of samples requisite to split, and so on are examples of hyper parameters. After adjusting the Hyper parameters, use testing or new data to discern whether the model is overvalued or underestimated.(Ramkumar and Maheswari 2022). Data visualization plays an important role in explaining how data is distributed. Data visualization gives an idea of the data distribution before and after data manipulation. Data visualization can be done in various forms of graphs like histogram, pie chart, bar chart etc.

Pseudocode for Decision Tree Algorithm

Input: Dataset for predicting credit card approval.
Step-1: Import and read the dataset.
Step-2: Choose the features at random from the dataset.
Step-3: Data manipulation
Step-4: Preprocessing data of the categorical data
Step-5: Dealing with hyper parameters
Step-6: Create a decision tree model with valid hyper parameters .
Step-7: Train the decision tree model using the train dataset.
Step-8: Calculate the error between expected output and actual output .
Output: Accuracy in %

Statistical Analysis

A t-test is a form of presumed steady used to see if there is a significant difference in the means of two groups that are integrated in some way(Fitzmaurice et al. 2008). The output for the grouped statistics was obtained using the spss tool. The t-test is one of many procedures used in statistics for speculation testing(Gomez and Gomez 1984).The IBM SPSS version was the statistical programme employed in this study (Cooksey 2020). For this study, dependent variables include attributes like update and transaction class. In SPSS, the datasets are prepared using sample size as 32 for the XGB classifier and decision tree algorithm(DeCoursey 2003). The Groupid for the XGB classifier is 1 and the Groupid for the decision tree algorithm is 2. The Groupid is a grouping variable and the accuracy is a testing variable (Privitera 2011).

RESULTS:

This data is used for the analysis of the Decision tree algorithm and XGB classifier algorithm. These ten data samples, together with their loss, are also applied to analyze statistical values that can be used for analogies. In Table 1, it is shown that the accuracy of two algorithms, Decision Tree algorithm and XGB classifier algorithm for different N values. The group statistics table depicts the number of samples taken, as well as the mean and standard deviation derived for the precision.
Table 3 represents the outcome of the analysis of the Independent samples test which has been performed for the Decision tree algorithm and XGB classifier algorithm. From Table 3, the significance value for the one tailed test is found to be 0.176, two-tailed is 0.001 and it is found that the Independent samples test has been carried out at Confidence Interval of 95%.

From Table 2, the group statistics values along with the mean, standard deviation and the standard error mean for the two algorithms are also specified. For the data set, the Independent sample T test is used, with the confidence interval set to 95%. Figure 1 shows the comparison of accuracy between Decision Tree algorithm and XGB classifier algorithm. Table 3 shows the independent t sample test calculation for Accuracy and Loss for Decision Tree algorithm and XGB classifier algorithm. Specifies mean difference and standard error difference and the comparative accuracy analysis, mean of loss between the two algorithms are specified. Figure 1 analyzes the mean of accuracy and mean loss among Decision tree algorithm and XGB classifier algorithm.

DISCUSSION:

From the given study the accuracy of the Decision Tree algorithm is 85.1001% when compared to the accuracy of the XGB classifier algorithm is 90.35%. Sample size is given as 10 analyses of

statistics have been done for both the Decision Tree algorithm and the XGB classifier algorithm in order to compare both the algorithms to find the better analysis algorithm in detection of credit card approval. For the given group, accuracy has been calculated. The mean, standard deviation and the standard error mean values for the XGB classifier algorithm are 90.35, 1.58391 and .50088 respectively. Similarly for the Decision Tree algorithm, the mean, standard deviation and the standard error mean values are 85.1001, 1.04044 and .32902 respectively from Table 2.

Compared to previous analysis of Decision Tree algorithm and XGB classifier algorithm for credit card approval,our research has got better accuracy in detecting the credit card approval analysis(Chen, and Rocha 2020). Previously the Decision Tree algorithm got the accuracy of 78.00% and for the XGB classifier algorithm the accuracy was 89.90%. But in our analysis we got the accuracy of the Decision Tree algorithm is 85.1001% and for the XGB classifier algorithm we got the accuracy of 90.35%. Datasets have been expelled from various resources and these datasets may contain several independent and unwanted attributes are there,this should be removed in order to get the best accuracy(L. C. Thomas 1999). Hence,the data is tested and trained to get the best output accuracy. Testing takes a long time. This process necessitates extensive training. Because of the random nature of the dataset, implementation time is likewise lengthy.(Tsolas 2021). The training data can be changed but it is a time consuming process and if data is trained low, the accuracy will be reduced(Beccalli, and Giordani 2013).

Testing and preparation of datasets for both the algorithms Decision Tree algorithm and XGB classifier algorithm is possible by grouping the datasets. Cleaning process of information can be efficiently increased and less time consuming in execution. The course of time utilization in preparing the dataset can be diminished.

CONCLUSION:

The Decision Tree algorithm(85.1001%) and the XGB classifier algorithm (90.35%) were used to create a model which makes predictions to approve credit cards or not. According to the findings of the experiments, the XGB classifier algorithm method outperforms the Decision Tree algorithm.

DECLARATIONS

Conflict of Interests

No Conflict of Interest in this manuscript.

Authors Contributions

Author SB was involved in data collection, data analysis, and manuscript writing. Author KS was involved in conceptualization, data validation, and critical review of the manuscript.

Acknowledgements

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

Funding

We thank the following organizations for providing financial support that enabled us to complete the study.

1. Manac Infotech (P) Limited.
2. Saveetha University.
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering.

REFERENCES

Anolli, M., E. Beccalli, and T. Giordani. 2013. *Retail Credit Risk Management*. Springer.

Bali, Turan G. 2007. "A Generalized Extreme Value Approach to Financial Risk Measurement." *Journal of Money, Credit and Banking*. https://doi.org/10.1111/j.1538-4616.2007.00081.x.

Bazarbash, Majid. 2019. *FinTech in Financial Inclusion: Machine Learning Applications in Assessing Credit Risk*. International Monetary Fund.

Beerbaum, Dirk, and Sammar Ahmad. "Credit Risk According to IFRS 9: Significant Increase in Credit Risk and Implications for Financial Institutions." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2654120.

Bloch, Ernest. 1971. "Two Decades of Evolution of Financial Institutions and Public Policy." *Journal of Money, Credit and Banking*. https://doi.org/10.2307/1991166.

Brownlee, Jason. 2016. *XGBoost With Python: Gradient Boosted Trees with XGBoost and Scikit-Learn*. Machine Learning Mastery.

DeCoursey, William. 2003. *Statistics and Probability for Engineering Applications*. Elsevier.

Fitzmaurice, Garrett, Marie Davidian, Geert Verbeke, and Geert Molenberghs. 2008.

Longitudinal Data Analysis. CRC Press.

Gomez, Kwanchai A., and Arturo A. Gomez. 1984. *Statistical Procedures for Agricultural Research*. John Wiley & Sons.

Gunjan, Vinit Kumar, Sabrina Senatore, Amit Kumar, Xiao-Zhi Gao, and Suresh Merugu. 2020. *Advances in Cybernetics, Cognition, and Machine Learning for Communication Technologies*. Springer Nature.

Hamdoun, Nabila, and Khalid Rguibi. 2019. "Impact of AI and Machine Learning on Financial Industry: Application on Moroccan Credit Risk Scoring." *Journal of Advanced Research in Dynamical and Control Systems*. https://doi.org/10.5373/jardcs/v11sp11/20193134.

Kim, Seong-Jun. 2003. "A Study on the Node Split in Decision Tree with Multivariate Target Variables." *Journal of Korean Institute of Intelligent Systems*. https://doi.org/10.5391/jkiis.2003.13.4.386.

Misra, Rajiv, Rudrapatna K. Shyamasundar, Amrita Chaturvedi, and Rana Omer. 2021. *Machine Learning and Big Data Analytics (Proceedings of International Conference on Machine Learning and Big Data Analytics (ICMLBDA) 2021)*. Springer Nature.

Molnar, Christoph. 2019. *Interpretable Machine Learning*. Lulu.com.

Privitera, Gregory J. 2011. *Statistics for the Behavioral Sciences*. SAGE.

Rahman, Rashedur M., and Fazle Rabbi Md Hasan. 2011. "Implementation of Various Data Processing and Evaluation Techniques on ICDDR,B Surveillance Data to Generate Optimal Decision Tree for Patients Classification." *International Journal of Knowledge Engineering and Soft Data Paradigms*. https://doi.org/10.1504/ijkesdp.2011.045727.

Ramkumar, S., and K. Maheswari. 2022. "Analysis of Error Rate for Various Attributes to Obtain the Optimal Decision Tree." *International Journal of Intelligent Enterprise*. https://doi.org/10.1504/ijie.2022.10048744.

Ranganathan, G., Joy Chen, and Álvaro Rocha. 2020. *Inventive Communication and Computational Technologies: Proceedings of ICICCT 2020*. Springer Nature.

Roncalli, Thierry. 2020. "Credit Scoring Models." *Handbook of Financial Risk Management*. https://doi.org/10.1201/9781315144597-15.

Sharma, Deepika, Ashutosh Vashistha, and Manoj Gupta. 2021. "Review of Credit Risk and Credit Scoring Models Based on Computing Paradigms in Financial Institutions." *The Journal of Credit Risk*. https://doi.org/10.21314/jcr.2021.006.

Thomas, L. C. 1999. *A Survey of Credit and Behavioural Scoring: Forecasting Financial Risk of Lending to Consumers*.

Thomas, Lyn, Jonathan Crook, and David Edelman. 2017. *Credit Scoring and Its Applications, Second Edition*. SIAM.

Tsolas, Ioannis E. 2021. "Firm Credit Scoring: A Series Two-Stage DEA Bootstrapped Approach." *Journal of Risk and Financial Management*. https://doi.org/10.3390/jrfm14050214.

Weber, Olaf. 2012. "Environmental Credit Risk Management in Banks and Financial Service Institutions." *Business Strategy and the Environment*. https://doi.org/10.1002/bse.737.

Yang, Jian, and Yinggang Zhou. "Credit Risk Spillovers among Financial Institutions around the Global Credit Crisis: Firm-Level Evidence." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1783198.

Tables and Figures

Table 1. Comparison between Decision tree algorithm and XGB classifier algorithms with N=10 samples of the dataset with the highest accuracy of respectively 86.58% and 92.782% using the 80% of training and 20% of testing dataset

Size	**Decision Tree accuracy in %**	**XGB Classifier accuracy in%**
1	86.58	92.782
2	86.0	91.99
3	85.7	91.70
4	86.01	91.00
5	85.50	90.5
6	85.10	90
7	84.60	89.45
8	84.31	89.18
9	83.70	88.67
10	83.50	88.01

Table 2. Mean, Standard Deviation and Standard Error mean for XGB classifier and Decision Tree algorithms are given below.

Accuracy	Groups	N	Mean	Std. Deviation	Std. Error Mean
	Decision Tree	10	85.1001	1.04044	.32902
	XGB classifier	10	90.3500	1.58391	.50088

Table 3: We find the mean and variance values by using Levene's test for equality of variance and t-test for equality means. By assuming equal variance and unequal variance values. And accuracy for both of them.

		F	**sig.**	**t**	**df**	**Sig(2-tailed)**	**Mean difference**	**std.Error difference**	**Lower**	**Upper**
Accuracy	Equal variance assumed	1.988	.176	-8.760	18	.001	-5.24990	.59927	-6.50893	-3.99087
	Equal variance not assumed			-8.760	15.548	.001	-5.24990	.59927	-6.52331	-3.97649

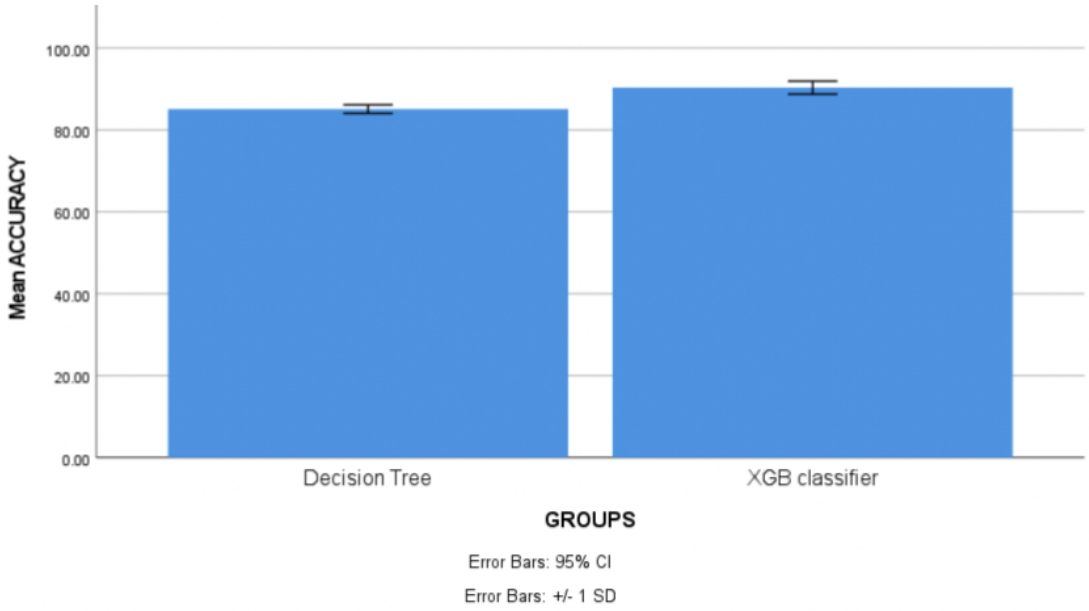

Fig 1. Comparison of XGB classifier algorithm and Decision Tree algorithm in terms of means accuracy. The mean accuracy of the XGB classifier algorithm is better than the Decision Tree algorithm. X-Axis: Decision Tree algorithm Vs : XGB classifier . Y-Axis:Mean Accuracy of Detection +/-1SD