Title Page:

Prediction of credit card approval using Decision Tree algorithm over Logistic Regression

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, Hyper parameters, Decision Tree algorithm, Logistic Regression, Data visualization, Data Manipulation,*Reduction.*

ABSTRACT

Aim: The purpose of this research is to develop a model that can forecast if a financial institution will be able to approve credit cards for its clients. To prevent fraud, which could result in financial institutions losing money, this approach can help an organisation decide precisely whether to approve or reject a credit card. **Materials and methods:** Classification is performed by Decision tree algorithm (n=10) over Logistic regression (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 Decision Tree algorithm has a high accuracy compared to Logistic Regression. The mean value is 85.1001 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 serves as both a predictor and an analyst to decide if a financial institution will issue a credit card to the customer. According to our observations, prior default, years worked, credit score, and debt are the four most crucial factors that any financial business host would take into account when deciding whether to issue a credit card.

Keywords: Credit Card approval, Hyper parameters, Decision Tree algorithm ,Logistic Regression, Data visualization, Data Manipulation,*Reduction*.

INTRODUCTION

In today's world, every procedure has been digital. Because of sophisticated information processing technology, money transfers from one account holder to another have become possible in seconds.(National Research Council et al. 1997). Not only that, but every company in every industry has gone digital, including railways, insurance, health care, fashion technology, education, sales and commerce, and advertising(Schwab 2017). One such industry is banking, where each person's financial situation is taken into account when applying for a loan, credit card, or other financial product. If the loan applicant's credit score is good, banks will be happy to lend to him, but he or she can choose any bank(S. Rahman 2014). A reduction in credit score is done if the client is overdue on their loan repayment.

Credit card churn forecast is not immune to this circumstance. As a result, banks should take proactive steps to prevent churn among existing credit card customers. Withholding existing clients helps a company's overall revenue and keeps its excellent name in a competitive industry. As a result, utilising customer management models, every company takes important steps to keep existing clients or reduction of clients. Because customer retention saves time, money, and resources when compared to acquiring new customers, it is a critical responsibility(National Research 2000)(Mertz 2021).

Over the last four years, 17,400 articles on credit card approval using machine learning have been published in Google Scholar, with 1,773 articles available in ScienceDirect.A credit card is one of the most well-known choices(Porter 2008). The majority of consumers use credit cards to conduct their transactions since they are a convenient way to pay. 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 (IMF and World Bank 2019). All of this data is analysed before an application is considered(*New York Magazine* 1992)(Liu and Motoda 2013).

The datasets in previous studies had been oversampled. The prior study used in reduction erroneous datasets as its input data set(Fox et al. 2007). Positive classes are heavily sampled. After cleaning the dataset, the number of tuples obtained is relatively low(Rouquette et al. 2014). As a result, the input dataset is extremely small(Shmueli, Patel, and Bruce 2006). The current research has a very low degree of accuracy. The study's goal is to enhance the algorithms' reliability while also increasing the size of the input dataset.(Luengo et al. 2020)(Raschka 2015)

MATERIALS AND METHODS

SIMATS Saveetha School of Engineering's computer science and engineering department carried by 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. The tool utilised was Jupyter Lab using the Python programming language, and the operating system was Windows.

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).Reduction of decision tree is not allowed. By reduction the algorithm will become overrated or underrated. 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 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 visualization before data manipulation
Step-4: Data manipulation
Step-5: Create decision tree model with valid hyper parameters
Step-6: Model training with decision tree samples
Step-7: Calculate the error between expected output and actual output .
Step-8: Data visualization after getting output
Output: Accuracy in %

Logistic Regression

Model the situation For forecasting continuous (numeric) variables, regression models can be effective. Approved, on the other hand, has a binary target value that can only be 1 or 0. The applicant will either be granted or denied a credit card; no partial credit cards will be granted. This could apply linear regression to predict the approval decision using a threshold, with anything below 0 and anything above 1 as a result. Regrettably, the anticipated values may fall significantly outside of the intended 0 to 1 range. As a result, neither linear nor multivariate regression can accurately predict the values(Thomas 1999). Rather, logistic regression will be more effective because it will generate a chance that the goal value is 1. Probabilities are always a fraction of a percent(Irizarry 2019). Before making data into two sections, data manipulation is an important task because data manipulation deals with missing values, string data, which parameters should be considered etc(Navarro, n). 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(Verzani 2018). Data visualization can be done in various forms of graphs like histogram, pie chart, bar chart etc.

Pseudocode for Logistic Regression 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 visualization before data manipulation
Step-4: Data manipulation
Step-5: Create logistic regression model.

Step-6: Train the logistic regression model using the train dataset.
Step-7: Calculate the error between expected output and actual output .
Step-8: Data visualization after getting 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(Daniel and Cross 2018). The output for the grouped statistics was obtained using the spss tool (Bhattacherjee 2012). The t-test is one of many procedures used in statistics for speculation testing.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 10 for the decision tree algorithm and logistic regression algorithm(DeCoursey 2003). The Groupid for the decision tree algorithm is 1 and the Groupid for the logistic regression algorithm is 2. The Groupid is a grouping variable and the accuracy is a testing variable (Privitera 2011)(Boslaugh 2012).

RESULTS:

This data is used for the analysis of the Decision tree algorithm and Logistic regression algorithm. These ten data samples, together with their loss, are also applied to analyze statistical values that can be used for analogies.(Gyeera.)(Abdulhafedh 2022). In Table 1, it is shown that the accuracy of two algorithms, decision tree algorithm and logistic regression algorithm for different n values.(G and Manoj 2021)(Harrell 2013). The group statistics table depicts the number of samples taken, as well as the mean and standard deviation derived for the precision.(Holmes, Illowsky, and Dean 2018).

Table 3 represents the outcome of the analysis of the Independent samples test which has been performed for the Decision Tree algorithm and the Logistic regression algorithm. From Table 3, the significance value for the one tailed test is found to be .048, two-tailed is 0.000 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%. The graph in fig 1 compares the accuracy of the logistic regression algorithm and the decision tree algorithm. Table 3 shows the independent t sample test calculation for Accuracy and Loss for decision tree algorithm and logistic regression 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 Logistic regression algorithm

DISCUSSION:

From the given study the accuracy of the Decision tree algorithm is 85.1001% when compared to the accuracy of the logistic regression is 55.5070%. Statistics have been done for both the Decision tree algorithm and the logistic regression algorithms 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 Decision tree algorithm are 85.1001, 1.04044 and .32902 respectively. Similarly for logistic regression, the mean, standard deviation and the standard error mean values are 55.5070, 1.44201 and .45600 respectively from Table 2.

Compared to previous analysis of Decision tree algorithm and logistic regression algorithm for credit card approval,our research has got better accuracy in detecting the credit card approval analysis. Previously, the Decision tree method achieved an accuracy of 82.69%, while the logistic regression approach achieved an accuracy of 65.0%. However, according to our exploration, the Decision tree method has an accuracy of 85.1001%, whereas the logistic regression technique has an accuracy of 55.5070%.(Thomas 1999). 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(Navarro). Hence,the data is tested and trained to get the best output accuracy. Testing takes a long time. This process necessitates extensive training. Reduction because of the random nature of the dataset, implementation time is likewise lengthy.(Daniel and Cross 2018). The training data can be changed but it is a time consuming process and if data is trained low, the accuracy will be reduced(Fitzmaurice et al. 2008 and Lavrakas 2008) .

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

CONCLUSION:

The Decision Tree Algorithm(85.1001%) and the Logistic Regression Algorithm(55.5070%) were used to create a model which makes predictions to approve credit cards or not. According to the findings of the experiments, the Decision Tree Algorithm method outperforms the Logistic Regression 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

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

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

REFERENCES

Abdulhafedh, Azad. 2022. "Comparison between Common Statistical Modeling Techniques Used in Research, Including: Discriminant Analysis vs Logistic Regression, Ridge Regression vs LASSO, and Decision Tree vs Random Forest." *OALib*. https://doi.org/10.4236/oalib.1108414.

Bhattacherjee, Anol. 2012. *Social Science Research: Principles, Methods, and Practices*. CreateSpace.

Boslaugh, Sarah. 2012. *Statistics in a Nutshell*. "O'Reilly Media, Inc."

Chandra, B., and Venkatanaresh Babu Kuppili. 2011. "Heterogeneous Node Split Measure for Decision Tree Construction." *2011 IEEE International Conference on Systems, Man, and Cybernetics*. https://doi.org/10.1109/icsmc.2011.6083761.

Cooksey, Ray W. 2020. *Illustrating Statistical Procedures: Finding Meaning in Quantitative Data*. Springer Nature.

Daniel, Wayne W., and Chad L. Cross. 2018. *Biostatistics: A Foundation for Analysis in the Health Sciences*. Wiley.

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.

Fox, Kathleen M., Sanjay K. Gandhi, Robert L. Ohsfeldt, James W. Blasetto, and Harold E. Bays. 2007. "Effectiveness of Rosuvastatin in Low-Density Lipoprotein Cholesterol Lowering and National Cholesterol Education Program Adult Treatment Panel Guideline III LDL‐C Goal Attainment Compared to Other Statins among Diabetes Mellitus Patients: A Retrospective Study Using an Electronic Medical Records Dataset in the United States." *Current Medical Research and Opinion*. https://doi.org/10.1185/030079907x219580.

G, Manoj Kumar, and Kumar G. Manoj. 2021. "Accuracy Analysis for Logistic Regression Algorithm and Random Forest Algorithm to Detect Frauds in Mobile Money Transaction." *Revista Gestão Inovação E Tecnologias*. https://doi.org/10.47059/revistageintec.v11i4.2182.

Gyeera, Thomas Weripuo. "Regression Analysis of Predictions and Forecasts of Cloud Data Centre KPIs Using the Boosted Decision Tree Algorithm." https://doi.org/10.36227/techrxiv.14538486.

Harrell, Frank E. 2013. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer Science & Business Media.

Holmes, Alexander, Barbara Illowsky, and Susan Dean. 2018. *Introductory Business Statistics*.

International Monetary Fund, and World Bank. 2019. *Fintech: The Experience So Far*. International Monetary Fund.

Irizarry, Rafael A. 2019. *Introduction to Data Science: Data Analysis and Prediction Algorithms with R*. CRC Press.

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.

Lavrakas, Paul J. 2008. *Encyclopedia of Survey Research Methods*. SAGE Publications.

Liu, Huan, and Hiroshi Motoda. 2013. *Instance Selection and Construction for Data Mining*. Springer Science & Business Media.

Luengo, Julián, Diego García-Gil, Sergio Ramírez-Gallego, Salvador García, and Francisco Herrera. 2020. *Big Data Preprocessing: Enabling Smart Data*. Springer Nature.

Mertz, David. 2021. *Cleaning Data for Effective Data Science: Doing the Other 80% of the Work with Python, R, and Command-Line Tools*. Packt Publishing Ltd.

National Research Council, Computer Science and Telecommunications Board, and Committee on Intellectual Property Rights in the Emerging Information Infrastructure. 2000. *The Digital Dilemma: Intellectual Property in the Information Age*. National Academies Press.

National Research Council, Division on Engineering and Physical Sciences, Computer Science and Telecommunications Board, Commission on Physical Sciences, Mathematics, and Applications, and Committee on Maintaining Privacy and Security in Health Care Applications of the National Information Infrastructure. 1997. *For the Record: Protecting Electronic Health Information*. National Academies Press.

Navarro, Daniel. *Learning Statistics with R*. Lulu.com.

New York Magazine. 1992.

Porter, Michael E. 2008. *Competitive Advantage: Creating and Sustaining Superior Performance*. Simon and Schuster.

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.

Rahman, Saimunur. 2014. *Introduction to E-Commerce Technology in Business*.

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.

Raschka, Sebastian. 2015. *Python Machine Learning*. Packt Publishing Ltd.

Rouquette, A., S. M. Côté, J. Hardouin, and B. Falissard. 2014. "Item Response Theory (IRT) Used to Enhance Accuracy of Data Analyses in Longitudinal Studies of Child Development: A Simulation Study." *PsycEXTRA Dataset*. https://doi.org/10.1037/e500122015-141.

Schwab, Klaus. 2017. *The Fourth Industrial Revolution*. Penguin UK.

Shmueli, Galit, Nitin R. Patel, and Peter C. Bruce. 2006. *Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner*. John Wiley & Sons.

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

Verzani, John. 2018. *Using R for Introductory Statistics*. CRC Press.

Tables and Figures

Table 1. Comparison between Decision tree algorithm and Logistic regression algorithms with N=10 samples of the dataset with the highest accuracy of respectively 86.58% and 57.07% using the 80% of training and 20% of testing dataset

Size	**Decision Tree accuracy in %**	**Logistic Regression accuracy in %**
1	86.58	57.07
2	86.0	57.00
3	85.7	56.80
4	86.01	56.70
5	85.50	56.50
6	85.10	55.00
7	84.60	54.5
8	84.31	54
9	83.70	54.10
10	83.50	53.40

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

Accuracy	Groups	N	Mean	Std. Deviation	Std. Error Mean
	Logistic Regression	10	55.5070	1.44201	.45600
	Decision Tree	10	85.1001	1.04044	.32902

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 algorithms

		F	**sig.**	**t**	**df**	**Sig(2-tailed**	**Mean difference**	**std.Error difference**	**Lower**	**Upper**
Accuracy	Equal variance assumed	4.521	.048	-52.628	18	<.001	-29.59310	.56231	-30.77446	-28.4117
	Equal variance not assumed			-52.628	16.373	<.001	-29.59310	.56231	-30.78294	-28.4032

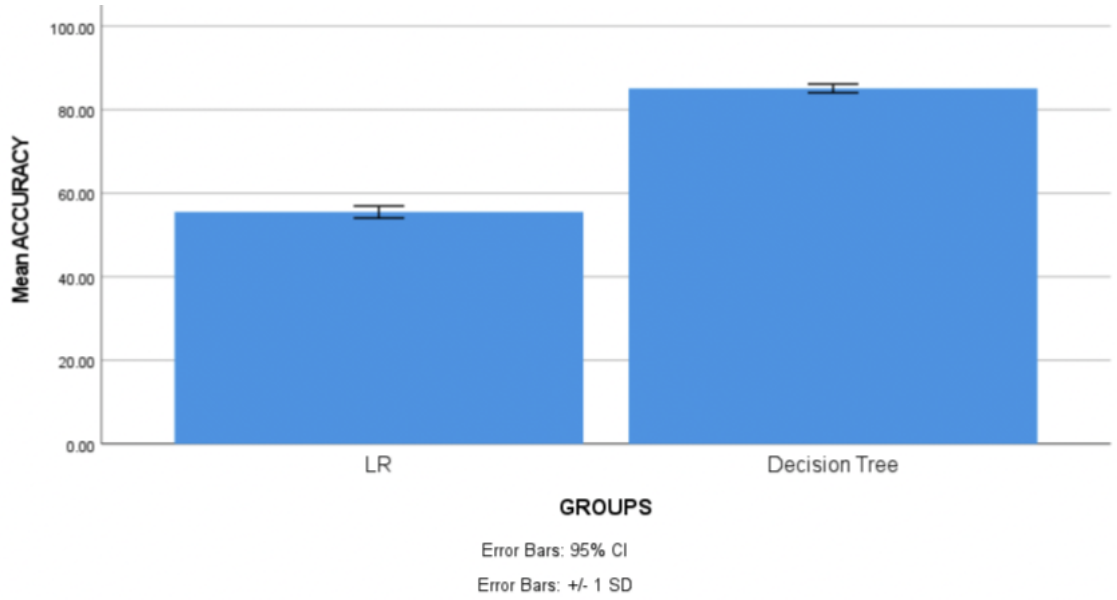

Fig 1. Comparison of Logistic Regression algorithm and Decision Tree algorithm in terms of means accuracy. The mean accuracy of the Decision Tree algorithm is better than Logistic Regression. X-Axis:Decision Tree algorithm Vs :Logistic Regression. Y-Axis:Mean Accuracy of Detection +/-1SD.