Original Article

Analysis of Accuracy for Approving Bank Loan using Numerical Data of Customer by Comparing Decision Tree over Logistic Regression Algorithm

Ch.Venkata Sandeep[1], Dr.T. Devi[2*]

[1]Research Scholar, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India, Pincode: 602 105.

[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, Pincode: 602105.

Abstract

Aim: To predict the loan of the person using a Novel Decision Tree (DT) over the Logistic Regression algorithm (LR). **Materials and Methods**: The existing model uses a Logistic Regression algorithm and in the proposed Novel Decision Tree. The 20 sample values are used to find out the mean. Std. deviation, std. error means. The sigmoid function is calculated using Levene's Test for Equality of Variances both assumed and non-assumed. Decision trees use different algorithms to decide to split a node into two or more sub-nodes. The purity of the node depends upon the target variable. The prediction of the loan can be done only after knowing all the factors. The sample size was measured as 40 for both groups using G power (80%). **Results**: The graph explains the comparison of the mean accuracy value with algorithms Novel Decision Tree and Logistic Regression where the mean accuracy is about 88% and 70%. The statistical significant value obtained is 0.78 ($p>0.05$), it shows insignificance between the groups based on an independent sample T-Test. **Conclusion**: The mean accuracy rate of the Novel Decision tree algorithm has been improved to 88% compared to Logistic Regression which is having around 70% mean accuracy.

Keywords: Logistic Regression, Novel Decision Tree, Sigmoid Function, Prediction, Mean Accuracy, Machine Learning.

DOI: 10.47750/pnr.2022.13.S04.209

INTRODUCTION

Loans are crucial for all the sectors starting right with individuals for study purposes to small and mid-sized businesses. Yet, the banks can't collect back all the loans. Defaults are unavoidable due to numerous factors such as bankruptcy, domestic problems, several dependents, etc. Analyzing the patterns and behaviors based on real-time data gives an edge over the loan approval process. This research explains (Maimon and Lior 2014) the importance of employing analytical algorithms (Molnar 2019) to analyze and implement predictive methods to classify the data as defaulters and otherwise. This research gives a clear-cut insight into applying accurate (Harrington 2012) analysis of numerical data of customers (Siddig, Ibrahim, and Elkatatny 2021) by using the Novel Decision tree over the logistic regression algorithm. The applications are building an effective predictive model in banking and computation of accuracy for the loan prediction and calculation of the classification accuracy for loans (Taraba 2021).

There are about 9400 articles published since 2017 which are relevant to the topic on IEEE Xplore and about 82 articles on ScienceDirect.The existing algorithm, Logistic Regression, a popular statistical machine learning algorithm, is used to classify data based on the outcome variable of extreme ends using predictive analysis where a logarithmic line separates the ends (Xu, Lu, and Xie 2021). For studying and analyzing, the data is collected from Kaggle and used for predictions. Since the output is targeted to be a binary, the sigmoid function

is used to develop the model. Multiple parameters such as personal attributes, like age, credit history, duration, amount, scores, account details, Business Values, Customer Assets, etc are taken into account to find the probability of default per person. The algorithm primarily ensures the data is (Ferreira et al. 1993)cleaned to elude missing values in the numerical data set. After which the training of the model with appropriate data sets starts to happen. Since pre-processing is a major area in the model, it becomes time-consuming and expensive. The model also lacks realism as it does not address the class imbalance and studying nonstationary environments is impossible. The proposed algorithm, Novel Decision Tree, a collection of nodes, creates decisions on features connected to certain classes. Random Forests, a supervised learning algorithm, is built based on Novel Decision trees that start by selecting random features, say k out of m. Then, the best split criteria are used to find the root node from the randomly selected k features. Daughter nodes are then calculated using the same criteria, a function like gin Index () or Infogain (), gives the fittest splitting attributes. A tree is formed with a root node and the target being the leaf node (Ross Quinlan 1993). Analyzing the customer data provides a scope for the bank to minimize its under-performing assets. This process is iterated to give rise to multiple randomly created forests. The number of trees and the random variables are important parameters in this algorithm. Information Gain (G) is used to split the data in a particular tree into daughter nodes. This algorithm reduces the total cost and processing time and can achieve higher precision (Taraba 2021).

Our institution is passionate about high quality evidence based research and has excelled in various fields (Parakh et al. 2020; Pham et al. 2021; Perumal, Antony, and Muthuramalingam 2021; Sathiyamoorthi et al. 2021; Devarajan et al. 2021; Dhanraj and Rajeshkumar 2021; Uganya, Radhika, and Vijayaraj 2021; Tesfaye Jule et al. 2021; Nandhini, Ezhilarasan, and Rajeshkumar 2020; Kamath et al. 2020).The existing system has issues and major parts (Zhu et al. 2021).It is quite apparent when looking at the rising need for new algorithms to solve the drawbacks of existing models. Since the Novel Decision Tree algorithm has a growing edge over the logistic regression algorithm which confines to non-stationary environments. The study aims to determine the better performing algorithm out of the Novel Decision tree and logistic regression models for accurately analyzing real-time numerical data of customers. The comparison is done using a mean accuracy graph.

MATERIALS AND METHODS

The setup of the research has been performed in the Data Analytics Laboratory, Department of CSE in Saveetha School Engineering, Saveetha Institute of Medical and Technical Sciences. The study uses a credit card dataset downloaded from Kaggle. The sample size was measured as 40 for both groups using Gpower (80%) with an alpha value of 0.05 and a beta value of 0.95("Decision Tree Algorithm Based on OLAP Multidimensional Data Set System" 2016).

In sample preparation of group 1 which represents the proposed system ensembles the decision tree algorithm which splits into different modules before finalizing the result. The proposed Decision tree algorithm uses 20 sample values where various statistical metrics are evaluated to get to a mean accuracy. This value is used to find the comparison between the existing and proposed models. In the Decision tree algorithm, the data is continuously divided for a certain parameter. It uses the tree structure for solving the problem by dividing them into nodes and classes. The 2-tailed Sigmoid function determines the mean of the algorithm.

The group 2 represents the existing model using the logistic regression algorithm. The 20 sample values are used to find out the mean, Std.deviation, std.error mean. The sigmoid function is calculated using Levene's Test for Equality of Variances both assumed and non-assumed (Kelleher, Namee, and D'Arcy 2020; Ross Quinlan 1993). The 2-tailed Sigmoid function is calculated using the T-test for equality of means.The dataset used for the existing model has been imported by Kaggle by downloading the dataset which has records of around 20 (Harrington 2012) sample values and different attributes related to the output of the data. The logistic regression algorithm is one of the most popular machine learning algorithms. It is used to predict the categorical dependent variable using a given set of variables.

This study was implemented using Jupyter lab, and the hardware configuration required is an intel i3 processor, 50 GB HDD, 4GB RAM, and the software configuration required is a Windows OS. The project is mainly a comparison of two algorithms one being Novel Decision Tree and the other being Logistic Regression , where a data set named DataSet of customers 1 containing 20 rows is used to find the mean accuracy of both algorithms. All the datasets have been collected from the kaggle. Different datasets are compared individually (Schmid 2020).

Statistical Analysis

The statistical software which is used for analyzing IBM SPSS version 22 (64 bit) is an analysis software, which is done by uploading a dataset to the software which gives the output as independent variables N, means, std. deviation, std. error means, with the mean accuracy as the output for the given models Novel Decision tree and logistic regression. The dependent variables are output accuracy and cross-validation. The independent variables are the time period of experience values (S.Vijayarani et al. 2011).

RESULT

Table 1 represents the pseudocode for supporting the Decision Tree by focusing on the data frames and prediction amount. Initially, all the variables are taken into consideration which supports the algorithm.
Table 2 represents the pseudocode for Linear Regression which is an optimization algorithm for finding the local minimum of a differentiable function. Gradient descent is used to find values of function parameters.
Table 3 represents the group statistics of DT and LR algorithms. For the taken samples N=20 the mean accuracy of DT (88.1%) and LR (70.2%).
Table 4 shows the independent variables which define the equality of the variances and equality of means with the significance p=0.78 obtained. The frequency is 0.076 at time 12.7 diff. frequency obtained 18 for the given samples.
Figure 1 is a bar graph created by comparing annual income with the saving amount done by a person. It shows the variation among the people by representing them individually. It also shows the amount of interest that is implemented for the amount taken.
Figure 2 shows the comparison of both the DT and LR with the X-axis of the Decision tree vs Logistic Regression Algorithm and the Y-axis of Mean accuracy of detection 1 SD. It compares all the values obtained from the data sets and shows the plot.

DISCUSSION

Based on the results obtained by independent T-test analysis, the significance value is determined. The significance value has 0.78, where the $p>0.05$ and an accuracy of 88% which is higher than the 70% mean accuracy with insignificance between the groups.
The analysis of both algorithms has been done with Table 3 representing the group statistics and Table 4 representing the independent variables and a bar graph which represents the comparison of the two algorithms with the accuracy (Lee et al. 2018) percentages of 88% for Novel Decision tree and logistic regression (Conway and White 2012) with an accuracy of 70%.In uncertain times like today, defaulters are pretty commonplace in the banking sector. An accurate assessment of the real-time data aggregated from customers (Williams, n.d.; S.Vijayarani et al. 2011) is vital to approve loans. The paper explains how the Novel Decision Tree is comparatively accurate over the Logistic Regression Algorithm which is used for classifying the data sets into finding the right customers (Rokach 2008) for approving loans. The table explains the independent variables which define the Equality of the Variances and Equality of Means with the Sig. (2-tailed) =0.000 for both assumed and non-assumed variances (Rokach 2008; Molnar 2019) and a mean difference of 18.20000 for both assumed and non-assumed variances and 95% of confidence value for lower and upper intervals are 15.20747 (both assumed and non-assumed) and 21.19253 (both assumed and non-assumed) respectively. The graph table displays the Input parameters such as the name of the active data set, filters, weight, split files, and the number of rows in the working data file (Ross Quinlan 1993).
Factors affecting the research work are the predictive models that specify the comparison of two models with the best performance and accuracy. Although the results of the study are better in both experimental and statistical analysis, there are certain limitations in the work. The evaluation of accuracy cannot provide a better outcome on larger data sets. However, the work can be enhanced by applying optimization algorithm techniques, to achieve better accuracy. Feature selection algorithms can be used before classification to improve the accuracy of classifiers. The future scope of the study explains how it will be useful in the future for many applications with improved accuracy than other algorithms that don't take into account the necessary number of variables by carefully observing the credit risk while evaluating the credit score or to be precise, the approval score

CONCLUSION

The mean accuracy rate of the Novel Decision tree algorithm has been improved to 88% compared to Logistic Regression which is having around 70% mean accuracy. This suggests the proposed system provides an accurate improvement for loan approval.

DECLARATIONS

Conflicts of interests

No conflicts of interests in the manuscript.

Authors Contribution

Author CHVS was involved in data collection, data analysis, and manuscript writing. Author SPC was involved in conceptualization, data validation and critical review of 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. Saveetha University
2. Saveetha Institute of Medical and Technical Sciences
3. Saveetha School of Engineering
4. Sree Vidya High School, Nandigama

REFERENCES

1. Conway, Drew, and John Myles White. 2012. *Machine Learning for Hackers: Case Studies and Algorithms to Get You Started*. "O'Reilly Media, Inc."
2. "Decision Tree Algorithm Based on OLAP Multidimensional Data Set System." 2016. *Rev. Téc. Ing. Univ. Zulia*. https://doi.org/10.21311/001.39.6.22.
3. Devarajan, Yuvarajan, Beemkumar Nagappan, Gautam Choubey, Suresh Vellaiyan, and Kulmani Mehar. 2021. "Renewable Pathway and Twin Fueling Approach on Ignition Analysis of a Dual-Fuelled Compression Ignition Engine." *Energy & Fuels: An American Chemical Society Journal* 35 (12): 9930–36.
4. Dhanraj, Ganapathy, and Shanmugam Rajeshkumar. 2021. "Anticariogenic Effect of Selenium Nanoparticles Synthesized Using Brassica Oleracea." *Journal of Nanomaterials* 2021 (July). https://doi.org/10.1155/2021/8115585.
5. Ferreira, José, Joaquim Correia, Thomas Jamet, and Ernesto Costa. 1993. "An Application of Machine Learning in the Domain of Loan Analysis." *Machine Learning: ECML-93*. https://doi.org/10.1007/3-540-56602-3_160.
6. Harrington, Peter. 2012. *Machine Learning in Action*. Simon and Schuster.
7. Kamath, S. Manjunath, K. Sridhar, D. Jaison, V. Gopinath, B. K. Mohamed Ibrahim, Nilkantha Gupta, A. Sundaram, P. Sivaperumal, S. Padmapriya, and S. Shantanu Patil. 2020. "Fabrication of Tri-Layered Electrospun Polycaprolactone Mats with Improved Sustained Drug Release Profile." *Scientific Reports* 10 (1): 18179.
8. Kelleher, John D., Brian Mac Namee, and Aoife D'Arcy. 2020. *Fundamentals of Machine Learning for Predictive Data Analytics, Second Edition: Algorithms, Worked Examples, and Case Studies*. MIT Press.
9. Lee, Shin-Jye, Zhaozhao Xu, Tong Li, and Yun Yang. 2018. "A Novel Bagging C4.5 Algorithm Based on Wrapper Feature Selection for Supporting Wise Clinical Decision Making." *Journal of Biomedical Informatics* 78 (February): 144–55.
10. Maimon, Oded Z., and Rokach Lior. 2014. *Data Mining With Decision Trees: Theory And Applications (2nd Edition)*. World Scientific.
11. Molnar, Christoph. 2019. *Interpretable Machine Learning*. Lulu.com.
12. Nandhini, Joseph T., Devaraj Ezhilarasan, and Shanmugam Rajeshkumar. 2020. "An Ecofriendly Synthesized Gold Nanoparticles Induces Cytotoxicity via Apoptosis in HepG2 Cells." *Environmental Toxicology*, August. https://doi.org/10.1002/tox.23007.
13. Parakh, Mayank K., Shriraam Ulaganambi, Nisha Ashifa, Reshma Premkumar, and Amit L. Jain. 2020. "Oral Potentially Malignant Disorders: Clinical Diagnosis and Current Screening Aids: A Narrative Review." *European Journal of Cancer Prevention: The Official Journal of the European Cancer Prevention Organisation* 29 (1): 65–72.
14. Perumal, Karthikeyan, Joseph Antony, and Subagunasekar Muthuramalingam. 2021. "Heavy Metal Pollutants and Their Spatial Distribution in Surface Sediments from Thondi Coast, Palk Bay, South India." *Environmental Sciences Europe* 33 (1). https://doi.org/10.1186/s12302-021-00501-2.
15. Pham, Quoc Hoa, Supat Chupradit, Gunawan Widjaja, Muataz S. Alhassan, Rustem Magizov, Yasser Fakri Mustafa, Aravindhan Surendar, Amirzhan Kassenov, Zeinab Arzehgar, and Wanich Suksatan. 2021. "The Effects of Ni or Nb Additions on the Relaxation Behavior of Zr55Cu35Al10 Metallic Glass." *Materials Today Communications* 29 (December): 102909.
16. Rokach, Lior. 2008. *Data Mining with Decision Trees: Theory and Applications*. World Scientific.
17. Ross Quinlan, J. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann.
18. Sathiyamoorthi, Ramalingam, Gomathinayakam Sankaranarayanan, Dinesh Babu Munuswamy, and Yuvarajan Devarajan. 2021. "Experimental Study of Spray Analysis for Palmarosa Biodiesel-diesel Blends in a Constant Volume Chamber." *Environmental Progress & Sustainable Energy* 40 (6). https://doi.org/10.1002/ep.13696.
19. Schmid, Bernhard. 2020. "Decision Letter: A Remote Sensing Derived Data Set of 100 Million Individual Tree Crowns for the National Ecological Observatory Network." https://doi.org/10.7554/elife.62922.sa1.
20. Siddig, Osama, Ahmed Farid Ibrahim, and Salaheldin Elkatatny. 2021. "Application of Various Machine Learning Techniques in Predicting Total Organic Carbon from Well Logs." *Computational Intelligence and Neuroscience* 2021 (August): 7390055.
21. S.Vijayarani, Dr Dr, S. Vijayarani Dr Vijayarani, Assistant Professor, School of Computer Science and Engineering, Bharathiar University, Coimbatore, and M. Sangeetha M. Sangeetha. 2011. "A Novel Privacy Preserving Approach for Decision Tree Learning." *Indian Journal of Applied Research*. https://doi.org/10.15373/2249555x/mar2014/40.
22. Taraba, Peter. 2021. "Linear Regression on a Set of Selected Templates from a Pool of Randomly Generated Templates."

Machine Learning with Applications. https://doi.org/10.1016/j.mlwa.2021.100126.
23. Tesfaye Jule, Leta, Krishnaraj Ramaswamy, Nagaraj Nagaprasad, Vigneshwaran Shanmugam, and Venkataraman Vignesh. 2021. "Design and Analysis of Serial Drilled Hole in Composite Material." *Materials Today: Proceedings* 45 (January): 5759–63.
24. Uganya, G., Radhika, and N. Vijayaraj. 2021. "A Survey on Internet of Things: Applications, Recent Issues, Attacks, and Security Mechanisms." *Journal of Circuits Systems and Computers* 30 (05): 2130006.
25. Williams, Gareth. n.d. "Novel Antibiotics from a 'White Box' 2D Structural Fingerprint Decision Tree." https://doi.org/10.26434/chemrxiv.14387885.
26. Xu, Junhui, Zekai Lu, and Ying Xie. 2021. "Loan Default Prediction of Chinese P2P Market: A Machine Learning Methodology." *Scientific Reports* 11 (1): 18759.
27. Zhu, Siyao, Cassandra Mitsinikos, Lisa Poirier, Takeru Igusa, and Joel Gittelsohn. 2021. "Development of a System Dynamics Model to Guide Retail Food Store Policies in Baltimore City." *Nutrients* 13 (9). https://doi.org/10.3390/nu13093055.
28. EPRA International Journal of Research & Development (IJRD). https://doi.org/10.36713/epra4323.

TABLES AND FIGURES

Table 1. Pseudocode for supporting Decision Tree over Logistic Regression Algorithm based on the initialize,compute,update. Sum of vectors compares the values.

Input: Credit score $x_{i,}$ labels y_i
Output: Sum of vectors ,a array, b and DT
Procedure 1: Initialize:$a_{i,}$=0,f_i = -y 2:Compute:Continuous.remove('Loan Status') Category.remove('Loan Status') 3:Update:Loan status 4:Compute:list(find dataframe.columns) 5: Until dataframe.loc[dataframe['Credit Score']>850 6:Update the threshold b 7:Store the status value 8:Update the data entry 9: Determine the prediction amount.

Table 2. Pseudocode for Linear Regression which is an optimization algorithm for finding the local minimum of a differentiable function.Gradient descent is used to find values of a function parameters.

Input: D: Expenditure amount T: Unique terms in all documents
Output: Accuracy
Procedure: for dataframe.loc[dataframe['Months since last delinquent'].isnull() for dataframe[list(dataframe.columns)].isnull().sum() wij= accuracy of approval Often t_i in document d_j End for document End for of term

Table 3. Statistical calculations for independent samples tested between DT and LR Algorithm. The mean accuracy of DT is 88.100 and LR is 70.200. Standard deviation of DT is 3.02765 and LR Algorithm is 3.33500. T-Test for comparison for DT (0.95743) and LR (1.05462).

	Algorithm	**N**	**Mean**	**Std.Dev**	**Std.Error Mean**

Accuracy	DT	20	88.100	3.02765	0.95743
	LR	20	70.200	3.33500	1.05462

Table 4. Comparison between significance level for Decision Tree over Logistic Regression, the statistical significant value obtained is 0.78 in terms of accuracy with a 95% confidence interval.

		Levene's test for equality variance		**T-test for Equality of Means**						
				T-test for Equality of Means						
		F	**sig**	**t**	**df**	**Sig** (2-tailed)	**Mean Difference**	**std**	**95.5% confidence interval of the**	
									Lower	**Upper**
Accuracy	Equal variance assumed	0.076	0.78	12.7	18	0.000	-10.8	1.42	15.207	21.19253
	Equal variance not assumed			12.7	17.83	0.000	-10.8	1.42	15.205	21.19453

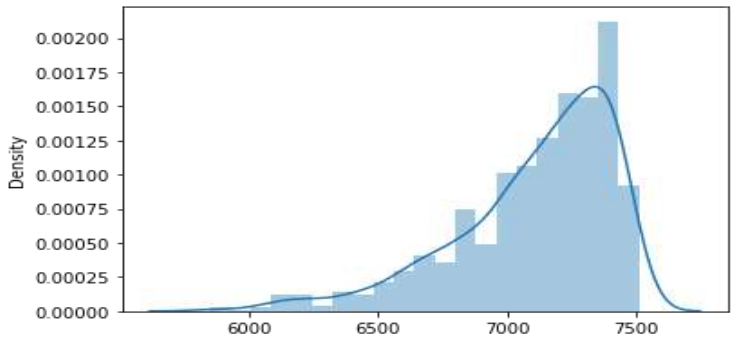

Fig. 1. The density of the graph describes the expenditure of the people compared with their annual income. The saving amount shows the graph variation of the individual persons. The total amount to be paid to the bank depends on the amount and time period.

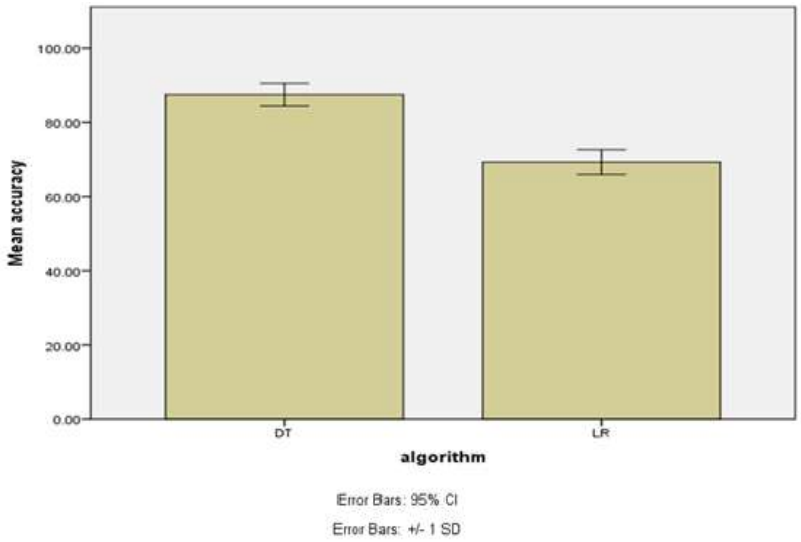

Fig. 2. The graph represents the comparison of the mean accuracy value with algorithms Decision Tree and Logistic Regression where the mean accuracy of the Novel Decision tree is about 88% and the mean accuracy value of the Logistic Regression is about 70%. X axis: Decision tree vs Logistic Regression Algorithm, Y axis: Mean accuracy of detection ± 1 SD.