Original Article

Comparison of Novel Optimized Random Forest Technique and Logistic Regression for Credit Card Fraud Detection with Improved Precision

M.Shahid Saif Ali Baig[1], K.Jaisharma[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, Tamilnadu, India. Pincode - 602 105.

Abstract

Aim: Objective is to improve precision for credit card fraud detection by using Novel Optimized Random Forest Technique (NORFT) and comparison with Logistic Regression (LT). **Materials and Methods:** In NORFT, it uses multiple Decision Trees to detect the credit card fraud by culminating the maximum attained probability values. The groups consist of NORFT and LR for comparison analysis. The sample size was estimated by using Clinicalc online tool, which is determined as N=2500 for each group with g-power value as 80% and datasets are collected from various web sources with recent study findings and threshold 0.05%, confidence interval 95% mean and standard deviation.. **Results:** The implementation resulted in precision as such NORFT (92.52%) and LR (71.60%). The statistical significance was performed using Independent Sample T-test between the groups, the study has a significance value of ($p>0.05$) i.e. $p=0.649$ and states does not have any difference in research. **Conclusion:** Novel Optimized Random Forest Algorithm (NORFT) has significantly finer precision than Logistic Regression Algorithm (LR).

Keywords: Machine Learning, Novel Optimized Random Forest, Fraud Detection, Logistic Regression, Credit bureau, Credit Card Holders.

DOI: 10.47750/pnr.2022.13.S03.082

INTRODUCTION

In recent times, credit bureau fraud is grabbing attention from many people across the world. This also leads to the increase in fraud that takes place during the mislead of information of users by manual or by physical thief of card (Goyal and Sharma 2020). This represents the importance of Credit card fraud detection to control the innocent people to get affected. There exists many algorithms that have more precision value by comparing to other algorithms and the existing model used in this logistic regression (Nieto 2009). The model uses machine learning algorithms for getting better precision techniques to prevent the people from hackers or attackers by Credit Card holders. The importance is to reduce fraud and to save money from hackers and new technologies (Singh and Mahrishi 2020). The Master Card fraud detection can be used in most of the credit bureaus involving applications such as warehouse stores, online shopping, e-commerce business where credit card holders use for digital transactions (Garg, Chaudhary, and Mishra 2021).

This study mainly discusses credit card fraud detection. To understand the existence of current trends, past 5 years papers were reviewed, 30 related articles published in IEEE Xplore and 98 in ScienceDirect were considered. Based on the high citations and impact factors, 4 best papers were taken for final consideration in this research study. A comparison of machine learning algorithms guide for detecting fake visa card transactions discussed by authors (Dhankhad, Mohammed, and Far 2018). In this study, the use of credit card fraudulent transactions using dataset to compare oversampling and undersampling methods for imbalance classification (Shamsudin et al. 2020)**.** The article discusses various Machine learning algorithms that are often used in fintech to detect suspicious transactions (AbdulSattar and Hammad 2020). The endmost paper published was a demonstration of recognizing Mastercard extortion transactions by using decision trees and logistic regression (Kumar 2021). After analyzing these papers, the best study found a study of machine learning methods for detecting fraudulent credit card holders transactions.

Our team has extensive knowledge and research experience that has translate into high quality publications(Bhansali et al. 2021; Jayanth et al. 2021; Sudhakar, Ravel, and Perumal 2021; Sathiyamoorthi et al. 2021; Deepanraj et al. 2021; Raju et al. 2021; Arun Prakash et al. 2020; Kamath et al. 2020; Shanmugam et al. 2021; Rajasekaran et al. 2020; Adhinarayanan et al. 2020; Rajesh et al. 2020; Aurtherson et al. 2021). The lacunae in the current exploration algorithms has the maximum precision of 71.60%. The proposed model aim is to improve the precision above the existence of 71.60% using NORFT and compare the result with LR.

Materials and Methods

The research had been performed in the Data Analytics Laboratory of the CSE department in Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences. The experiment was designed to have two groups namely, group 1 as NORFT and group 2 as LR. The sample size was calculated using online sample calculator to derive the sufficient comparison result, based on the evaluation 2500 samples were used for each group using G-power of 80% with a confidence interval at 95%, alpha value as 0.05 and beta value as 0.2 (Kumar 2021).

Novel Optimized Random Forest

Random Forest (RF) is the most popular Machine Learning algorithm used widely for classification prediction. The Random Forest algorithm creates many decision tree structures to find the least weighted value for classification. Based on Random Forest (RF), supervised learning approach the proposed algorithm uses a technique of cumulating the branch weight to get maximized probability values. Using the maximized probability values, the precision rate drastically increases to categories the credit card fraud detection with less Mean Squared Error (MSE) given in Equation (1) (Kumar 2021).

$$MSE = 1/N\ I = 1N\ (fi-Yi)\text{^}2 \quad (1)$$

where, N is defined as the number of data points.

MSE is defined as Mean squared error.
fi is the worth returned by the model.
Yi is the real incentive for information point I.

Logistic Regression Algorithm

The machine learning algorithm named Logistic Regression goes under the family of supervised learning. It is used to predict the values of dependent variables in the form of yes or no or in binary form of 1 or 0 as categories (Singh and Mahrishi 2020).

Pseudocode for Logistic Regression Algorithm

Require: Dataset {X, y}, the parameters λ, $\lambda 1$ and $\lambda 2$ are picked by 10-overlap cross-approval
Ensure: β
Initialize all $\beta_j(m) = 0(j = 1,2,3,...,p), m = 0$;
repeat
Compute Z(m) and W(m) and the misfortune work Equation in view of $\beta(m)$;
for j=1: ρ do
Compute $Z_i^{(j)}(m) \leftarrow \sum_{\square \neq \square} x_{ik}\beta_k(m)$;
$w_j(m) \leftarrow \sum_{\square=\square}^{\square} W_i(m)x_{ij}(Z_i(m)-Z_i^{(j)}(m))$;
Update $\beta_j(m)$ using Equation ;
end for
Let m ←m+1,β(m+l) ← $\beta(m)$;
until $\sum_{\square=1}^{\square} (|\beta\square(\square + 1)|-|\beta\square(\square)|) < 10^{-8}$

The model was built with the hardware configuration components consisting of HP i5 processor, RAM 8GB, HDD 1TB and software components consisting of Windows 10 OS, Google Collab, Chrome Browser, MS Excel, Standard browser and spreadsheet editors. The model is trained with 70% of data from Credit Card Fraud dataset downloaded in Kaggle. The dataset contains 284807 entries in imbalance information cooled from 1994, there are 31 attributes available for each entries namely time, v1 to v8. Before training the model, the data underwent preprocessing, cleaning and feature selection processes. Once the training was completed, the model was tested using the rest of 30% data. The validations are performed to ensure the correctness of the derived output from the model.

Statistical Analysis

IBM SPSS version 22 software is used for statistical analysis between Novel Optimized Random Forest Technique Algorithms (NORFT) and Logistic Regression Algorithms (LR). The independent variables are Time, User identities (Kiruthika et al. 2020), Amount, Sensitive features (V1 TO V28) and dependent variables are

prediction categories 1-fraud, 0-Otherwise. The Independent Sample T-Test was conducted by setting the confidence interval 95% and significance value $p<0.05$ between the NORF and LR (Kumar 2021).

Results

Table 1 depicts the result comparison using group statistics of Novel Optimized Random Forest Technique Algorithms (NORFT) and Logistic Regression (LR). For the sample size of 2500, the NORFT has Mean of 92.5200, Std. Deviation of .72296 and Std.Error Mean as .16166. The LR has the same sample size of 2500, Mean of 71.6020, Std. Deviation of .716020 and Std.Error Mean as .18114.

Table 2 depicts the Independent sample T-test performed, to identify the significance relationship between Novel Optimized Random Forest Technique (NORFT) and Logistic Regression (LR). In this T-test preset confidence interval as 95% with significance value $p>0.05$. From the table, the significance obtained value of 0.649 and the mean difference of precision for both Equal variances not assumed mean difference is 20.91800.

Figure 1 represents the bar chart diagramatic the graphical representation of NORTH and LR. In X-axis both the algorithms were taken and on top of it error bars are drawn, Y-axis has the mean precision scaling for the confidence interval 95% and error bars at +/-1SD.

Discussion

The comparison of two algorithms describes that, the NORFT has the precision percentages of 92.52% and LR has the precision of 71.60%. The proposed model has achieved better precision compared to LR, but the significance has $p=0.649$ ($p>0.05$) which means there does not exist any difference between the two groups results due to dataset inconsistency.

There are many articles similar to study of proposed models where findings are Machine learning is being used to detect Master card fraud in real time (Thennakoon et al. 2019). In this study the author tells about Machine learning and deep learning algorithms are used to detect credit bureau card fraud (Azhan and Meraj 2020). Main scope in this research is about Charge card Fraud Detection Using Genetic Algorithm Feature Selection on Naive Bayes, Random Forest, and Svm (Saheed et al. 2020). This study shows that Fake brain organization and calculated relapse is utilized to detect credit card holder fraud (Sahin and Duman 2011). The opposite algorithm for the proposed study is fraudulent credit card detection using a neural network (Georgieva, Markova, and Pavlov 2019) which has better precision of 72% by comparison to norft having 92.52% of precision (Ren, Ye, and Li 2017).

The limitations of the proposed work would be, due to the huge data sensitive information like privacy and security reasons the feature selection becomes more complex and because of that the privacy is getting lacking. In future, feature selection algorithms can be optimized more effectively for the sensitive information containing problems, which may improve the classification compatibility in near further applications.

Conclusion

The precision rate of Novel Optimized Random Forest Technique Algorithms (NORFT) has been improved to 92.52% compared to Logistic Regression (LR), which has 71.60%. This also proved using statistical significance value attained as $p>0.05$, and there does not pertain any significance in the final result.

DECLARATIONS

Conflict of Interest

No conflict of interest in this manuscript.

Authors Contribution

Author MSSAB was involved in dataset collection, algorithm development, data analytics laboratory, and manuscript writing. Author KJS was involved in validation and review of the manuscript.

Acknowledgements

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

Fundings

We would like to thank the following organization for providing financial support that enabled us to complete this study.

1. Biozone Pvt.Ltd, Chennai.
2. Saveetha University.
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering.

References

1. AbdulSattar, Khadija, and Mustafa Hammad. 2020. "Fraudulent Transaction Detection in FinTech Using Machine Learning Algorithms." *2020 International Conference on Innovation and Intelligence for Informatics, Computing and Technologies (3ICT)*. https://doi.org/10.1109/3ict51146.2020.9312025.
2. Azhan, Mohammed, and Shazli Meraj. 2020. "Credit Card Fraud Detection Using Machine Learning and Deep Learning Techniques." *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*. https://doi.org/10.1109/iciss49785.2020.9316002.
3. Dhankhad, Sahil, Emad Mohammed, and Behrouz Far. 2018. "Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study." *2018 IEEE International Conference on Information Reuse and Integration (IRI)*. https://doi.org/10.1109/iri.2018.00025.
4. Garg, Vaishali, Sarika Chaudhary, and Anil Mishra. 2021. "ANALYSING AUTO ML MODEL FOR CREDIT CARD FRAUD DETECTION." *International Journal of Innovative Research in Computer Science & Technology*. https://doi.org/10.21276/ijircst.2021.9.3.5.
5. Georgieva, Sevdalina, Maya Markova, and Velizar Pavlov. 2019. "Using Neural Network for Credit Card Fraud Detection." *RENEWABLE ENERGY SOURCES AND TECHNOLOGIES*. https://doi.org/10.1063/1.5127478.
6. Goyal, Yogita, and Anand Sharma. 2020. *Credit Card Fraud Detection and Analysis Through Machine Learning*.
7. Kiruthika, Usha, S. Kanaga Suba Raja, C. J. Raman, and V. Balaji. 2020. "A Novel Fraud Detection Scheme for Credit Card Usage Employing Random Forest Algorithm Combined with Feedback Mechanism." *2020 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS)*. https://doi.org/10.1109/icpects49113.2020.9337045.
8. Kumar, Tulika. 2021. "Comparison of Logistic Regression and Decision Tree Method for Credit Card Fraud Detection." *International Journal for Research in Applied Science and Engineering Technology*. https://doi.org/10.22214/ijraset.2021.34241.
9. New Mexico. Attorney General's Office. 2001. *Preventing Credit Card Fraud*.
10. Nieto, Jennifer. 2009. "Financial Advantages: Misleading Information in the Credit Card-Processing Industry." *Texas Dental Journal* 126 (3): 274–75.
11. Ren, Huorong, Zhixing Ye, and Zhiwu Li. 2017. "Anomaly Detection Based on a Dynamic Markov Model." *Information Sciences* 411 (October): 52–65.
12. Saheed, Yakub K., Moshood A. Hambali, Micheal O. Arowolo, and Yinusa A. Olasupo. 2020. "Application of GA Feature Selection on Naive Bayes, Random Forest and SVM for Credit Card Fraud Detection." *2020 International Conference on Decision Aid Sciences and Application (DASA)*. https://doi.org/10.1109/dasa51403.2020.9317228.
13. Sahin, Y., and E. Duman. 2011. "Detecting Credit Card Fraud by ANN and Logistic Regression." *2011 International Symposium on Innovations in Intelligent Systems and Applications*. https://doi.org/10.1109/inista.2011.5946108.
14. Shamsudin, Haziqah, Umi Kalsom Yusof, Andal Jayalakshmi, and Mohd Nor Akmal Khalid. 2020. "Combining Oversampling and Undersampling Techniques for Imbalanced Classification: A Comparative Study Using Credit Card Fraudulent Transaction Dataset." *2020 IEEE 16th International Conference on Control & Automation (ICCA)*. https://doi.org/10.1109/icca51439.2020.9264517.
15. Singh, Bhupendra, and Mehul Mahrishi. 2020. "Comparing Different Models for Credit Card Fraud Detection." *SKIT Research Journal*. https://doi.org/10.47904/ ijskit.10.2.2020.8-12.
16. Thennakoon, Anuruddha, Chee Bhagyani, Sasitha Premadasa, Shalitha Mihiranga, and Nuwan Kuruwitaarachchi. 2019. "Real-Time Credit Card Fraud Detection Using Machine Learning." *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. https://doi.org/10.1109/confluence.2019.8776942.

TABLES AND FIGURES

Table 1. The Novel Optimized Random Forest Algorithm (NORFT) (92.5200) method and grouped statistics were compared using group statistics for recorded data from simulation for 2500 iterations (71.6020). In comparison, The Novel Optimized Random Forest Algorithm (NORFT) has a high level of Precision.

	Algorithm	**N**	**Mean**	**Std. Deviation**	**Std. Error Mean**
Precision	NORFT	2500	92.5200	.72296	.16166
	LR	2500	71.6020	.81007	.18114

Table 2. For logged data from simulation, an independent sample test was performed for 3600 iterations to set the confidence interval to 92.5200%. The results yielded p=0.649 (p>0.05) insignificant probability.

	Levene's Test for Equality of	**T-test for Equality of Means**

		Variance								
		f	Sig	t	df	Sig.(2-tailed)	Mean Difference	Std.Error Difference	95% Confidence of the Differences	
									Lower	Upper
Precision	Equal variances assumed	.211	.649	86.159	38	.000	20.9180	.24279	20.426	21.4
	Equal variances not assumed			86.159	37.51	.000	20.9180	.24278	20.426	21.4

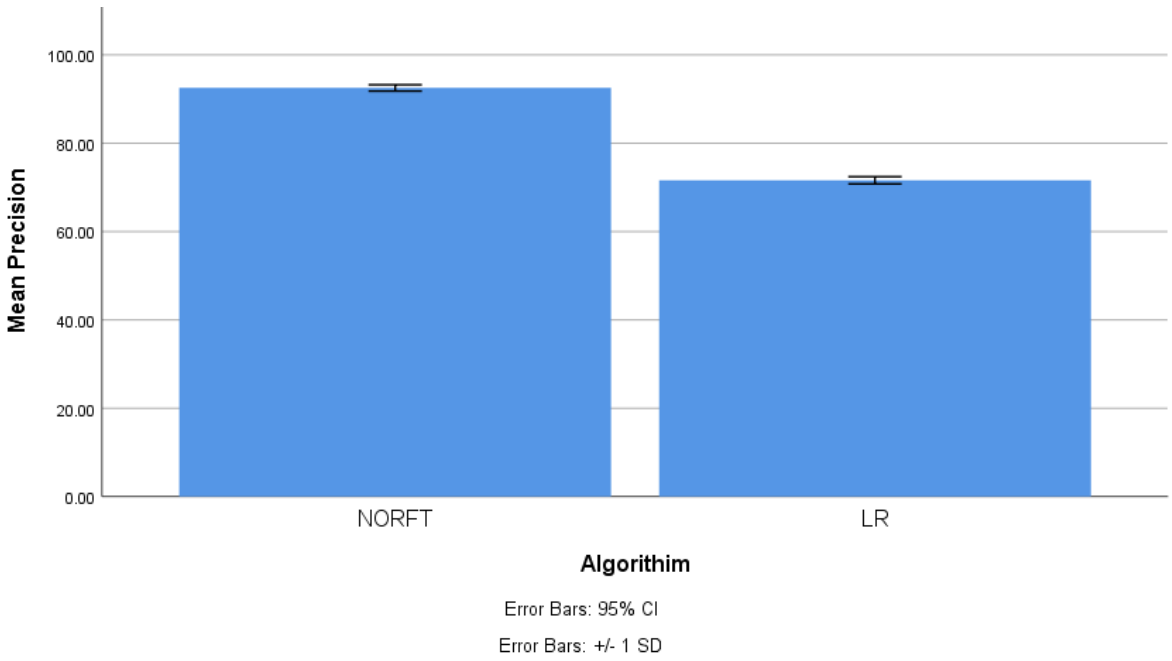

Fig. 1. Performance comparison between Novel Optimized Random Forest Algorithm (NORFT) (92.52%) and Logistic Regression Algorithms (LR) (71.60%). X axis Novel Optimized Random Forest Algorithm (NORFT) Vs Logistic Regression Algorithms (LR) Y axis Mean precision. Error Bar +/-1 SD.