IMPROVED ACCURACY IN CREDIT CARD FRAUD DETECTION USING NOVEL LOGISTIC REGRESSION OVER GRADIENT BOOSTING ENSEMBLE CLASSIFIER

P. Atchaya[1], K.Somasundaram[2*]

Article History: Received: 12.12.2022 **Revised:** 29.01.2023 **Accepted:** 15.03.2023

Abstract

Aim: To enhance the accuracy in credit card fraud detection using Novel Logistic Regression and Gradient Boosting Ensemble Classifier.
Materials and Methods: This study contains Novel Logistic Regression and Gradient Boosting Ensemble Classifier. Each algorithm consists of a sample size of 70 and the study parameters include alpha value 0.05, beta value 0.2 and the power value 0.8. Their accuracies are compared with each other using different sample sizes also.
Results: The Novel Logistic Regression is 93.59% more accurate than Gradient Boosting Ensemble Classifier of 92.70% in detecting fraudulent transactions. Significance value for accuracy and loss is 0.030 ($p<0.05$).
Conclusion: The Novel Logistic Regression model is significantly better than Gradient Boosting Ensemble Classifier in detecting fraudulent transactions. It can be also considered as a better option for credit card fraud detection.

Keywords: Novel Logistic Regression, Gradient Boosting Ensemble Classifier, Fraudulent transactions, Credit Card, Accuracy, Machine Learning.

[1]Research Scholar, Department of Information Technology, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India. Pincode: 602105.
[2*]Department of Information Technology, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India. Pincode: 602105.

1. Introduction

Fraud in credit card transactions is unauthorized and unwanted usage of an account by someone other than the owner of that account (Dighe, Patil, and Kokate 2018). Fraud detection involves monitoring the activities of populations of users in order to estimate, perceive or avoid objectionable behavior, which consist of fraud, intrusion, and defaulting. The biggest advantage is its quick access to credit (Thennakoon et al. 2019). Credit cards function on a credit basis, which suggests you get to use your card now and buy your purchases later. The money used doesn't leave your account, thus not denting your bank balance whenever you swipe. Machine learning algorithms are employed to analyze all the authorized transactions and report the suspicious ones (Mishra and Ghorpade 2018). This paper aims to conduct comparative analyses of identification of fraudulent activity on credit card utilizing Gradient Boosting Ensemble Classifier, Novel Logistic Regression techniques to explore the most accurate method of classifying a credit card transaction as fraudulent or non-fraudulent by which algorithm and combination of factors are considered. Its applications are Machine learning, statistical analysis, and behavior monitoring are utilized in fraud detection to uncover the patterns and techniques employed by criminals to conduct fraud. When fraud precursors are discovered, the system can intervene before any damage is caused (Baesens et al. 2015).

In the last 5 years,there have been 247 articles in google scholar and 295 in IEEE xplore. Gradient boosting is a widely-used machine learning algorithm, due to its efficiency, accuracy and interpretability. Financial institutions have continuously improved their fraud system. Aggregation strategy to create a new set of features based on analyzing periodic behavior of the time of transaction (Raj et al. 2011). A neural network based fraud detection trained on large samples of labeled credit card account transactions and tested on a hold out dataset. To detect accuracy and early detection (Maniraj et al. 2019). Grow in numbers,taking larger shares in the payment system. Improved fraud detection to maintain viability of the payment system (Tinio and California State Polytechnic). Future direction to improve both techniques and results Credit card fraud detection using bayesian and neural networks (Lamba 2020).

Our institution is passionate about high quality evidence based research and has excelled in various domains (Vickram et al. 2022; Bharathiraja et al. 2022; Kale et al. 2022; Sumathy et al. 2022; Thanigaivel et al. 2022; Ram et al. 2022; Jothi et al. 2022; Anupong et al. 2022; Yaashikaa et al. 2022; Palanisamy et al. 2022). Some datasets are aimed at theoretical research rather than processing it as per the real life application. Therefore identifying fraudulent transactions is challenging. Most of the existing standard feature extraction processes are for short-term analysis, so researchers have made their own feature set. Finally a paper is proposed assuming all the limitations. This paper solely focuses on fraudulent credit card transactions to increase the accuracy of prediction.

2. Materials and Methods

The work is carried out in the Soft Computing Lab, Department of Information Technology, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai. In this study, Novel Logistic Regression and Gradient Boosting Ensemble Classifier are compared. The study consists of two sample groups i.e Each group consists of 10 samples with a pretest power of 0.18. The sample size was set at 0.05, with an enrollment ratio of 1, a G power of 80%, a confidence interval of 95%, and G power of 80%. The dataset for categorization came from Kaggle Inc. Database, an open-source data repository for credit card fraud detection using numerous machine learning techniques.

Data Preparation

The input dataset is collected from Kaggle for this study (https://www.kaggle.com/mlg-ulb/creditcardfraud). The dataset contains 31 attributes. The Time attribute represents the date and time of transactions. Transactions represented from V1 to V28 which are used to represent the transactions done and Time attribute represents transactions done at a particular time interval. The amount attribute represents the amount of money transacted from one account to another. The dataset contains 2,84,808 transactions. The study's independent variable is transactions, time, amount and its values. The dependent attributes are accuracy and precision. The dataset is separated into training and testing sets with a test size of 10.

NOVEL LOGISTIC REGRESSION

Novel Logistic Regression is one among the foremost popular classification algorithms in machine learning. The Novel Logistic Regression model describes relationships between predictors which will be continuous, binary, and categorical. Dependent variables can be binary. Based on some predictors we predict whether something will be found or not. We estimate the probability of belonging to every category for a given set of predictors. It is basically a statistical model which makes use of a logistic function to model a binary dependent variable. This model is mainly used where there is a chance of occurrence of a binary classification issue. It works well on linearly separable classes. Pseudocode and Accuracy Values for the regression model is mentioned in Table 1 and Table 3. The odds ratio is one concept using which we can also define the logit function given in equation (1) and (2) below,

Odds Ratio — p/(l — p) (1)

Logit (P) = log p/l-p (2)

GRADIENT BOOSTING ENSEMBLE CLASSIFIER

Gradient boosting re-defines boosting as a numerical optimization problem where the target is to minimize the loss function of the model by adding weak learners using gradient descent. Gradient may be a first-order in iterative optimization algorithm for locating an area of minimum differentiable function given in equation (3) below, As gradient boosting is predicated on minimizing a loss function, differing types of loss functions are often used leading to a versatile technique which will be applied to regression, multi-class classification, etc. Pseudocode and Accuracy Values for the regression model are mentioned in Table 2 and Table 4.

y(pred) = y1 + (eta * r1) + (eta * r2) + ... + (eta * rN) (3)

The minimum requirement to run the softwares used here are intel core I3 dual core CPU@3.2 GHz , 4GB RAM , 64 bit OS, 1TB Hard disk Space Personal Computer and Software specification includes Windows 8 , 10 , 11 , Python 3.8 , and MS-Office. Statistical Package for the Social Sciences Version 26 software tool was used for statistical analysis. An independent sample T-test was conducted for accuracy. Standard deviation, standard mean errors were also calculated using the SPSS Software tool. The significance values of proposed and existing algorithms contain group statistical values of proposed and existing algorithms. The independent variables are transactions, time, V1 to V28, amount and the dependent variable is accuracy and precision.

3. Results

The group statistical analysis on the two groups shows Novel Logistic Regression (93.59%) has more mean accuracy than Gradient Boosting Ensemble Classifier (92.70%) and the standard error mean is slightly less than Novel Logistic Regression. The accuracies are recorded by testing the algorithms with 10 different sample sizes and the average accuracy is calculated for each algorithm. Table 6 shows the group statistic analysis, representing logistic regression and Gradient boosting ensemble classifier. Figure.1 shows the comparison of Novel Logistic Regression and Gradient Boosting Classifier in terms of mean accuracy and loss.

4. Discussion

From the results of this study, Novel Logistic Regression is proved to have better accuracy. Novel Logistic Regression has an accuracy of 93.59% whereas Gradient Boosting Ensemble Classifier has an accuracy of 92.70% is shown in Fig. 1. Accuracy and loss values for two algorithms Novel Logistic Regression and gradient boosting are denoted as given in Table 5. Group statistics table shows a number of samples that are collected. Mean and standard deviation obtained and accuracies are calculated and entered.

It's critical for credit card firms to be able to spot fraudulent credit card transactions so that customers cannot be charged for things they didn't purchase. Modeling prior credit card transactions with data from those that turned out to be fraudulent is part of the Credit Card Fraud Detection Problem (Dornadula and Geetha 2019). Gradient Boosting is one of the most extensively used classification techniques. Gradient Boosting is a combination of many decision trees. Many weak learners are integrated to improve the overall classification performance of the model. A set of classification and regression trees makes up the tree ensemble model (Varmedja et al. 2019). Individual classifiers in machine learning are not always capable of providing the maximum potential accuracy. As a result, many classifiers are utilized to obtain the highest level of accuracy and resilience (Adepoju et al. 2019). Credit card usage has grown exceedingly popular in today's economic climate. These cards allow users to make large-scale payments without having to carry significant amounts of cash. They've altered the way people make cashless payments and made any type of payment more convenient for buyers (Bianchini et al. 2009).

Applying Novel Logistic Regression for machine learning isn't a difficult task. However, it comes with its own limitations. The Novel Logistic Regression won't be ready to handle an outsized number of categorical features. In the example we reduced the number of features to a very large extent. Table 7, shows independent ‘t’ sample tests for algorithms. The comparative accuracy analysis, mean of loss between two algorithms are specified and shown in Table 6. It's future scope is the financial services firm tapped into vast databases of customer information culled from a variety of sources, including phone devices, IP addresses, in-person and online purchasing behavior, credit card preferences, geospatial locations, web analytics, and a variety of other sources.

5. Conclusion

Based on this study, the mean accuracy of Novel Logistic Regression is 93.59% compared to Gradient Boosting Ensemble Classifier which has a mean accuracy of 92.70%. Hence it is inferred that Novel Logistic Regression can predict fraudulent transactions more significantly than Gradient Boosting Ensemble Classifier. It can be used in predicting Credit card fraud detection in the future.

Declarations

Conflicts of Interest

No conflicts of interest in this manuscript.

Authors Contributions

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

Acknowledgement

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. BestEnlist company.
2. Saveetha University.
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering.

6. References

Dornadula, Vaishnavi Nath, and S. Geetha. 2019. "Credit Card Fraud Detection Using Machine Learning Algorithms." Procedia Computer Science. https://doi.org/10.1016/j.procs.2020.01.057.

Mishra, Ankit, and Chaitanya Ghorpade. 2018. "Credit Card Fraud Detection on the Skewed Data Using Various Classification and Ensemble Techniques." 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS). https://doi.org/10.1109/sceecs.2018.8546939.

Baesens, Bart, Wouter Verbeke, and Veronique Van Vlasselaer. 2015. Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection. John Wiley & Sons.

Varmedja, Dejan, Mirjana Karanovic, Srdjan Sladojevic, Marko Arsenovic, and Andras Anderla. 2019. "Credit Card Fraud Detection - Machine Learning Methods." 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH). https://doi.org/10.1109/infoteh.2019.8717766.

Jain, Vinod, Mayank Agrawal, and Anuj Kumar. 2020. "Performance Analysis of Machine Learning Algorithms in Credit Cards Fraud Detection." 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO). https://doi.org/10.1109/icrito48877.2020.9197762.

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.

Zheng, Lutao, Guanjun Liu, Chungang Yan, Changjun Jiang, Mengchu Zhou, and Maozhen Li. 2020. "Improved TrAdaBoost and Its Application to Transaction Fraud Detection." IEEE Transactions on Computational Social Systems. https://doi.org/10.1109/tcss.2020.3017013.

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.

Lamba, Harshit. 2020. Credit Card Fraud Detection In Real-Time.

Maniraj, S. P., Aditya Saini, Shadab Ahmed, Swarna Deep Sarkar, SRM Institute of Science and Technology, and INDIA. 2019. "Credit Card Fraud Detection Using Machine Learning and Data Science." International Journal of Engineering Research and. https://doi.org/10.17577/ijertv8is090031.

Aydogan, Murat, and Ali Karci. 2018. "Spam Mail

Detection Using Naive Bayes Method with Apache Spark.” 2018 International Conference on Artificial Intelligence and Data Processing (IDAP). https://doi.org/10.1109/idap.2018.8620737.

Kaur, Darshan, and Shubhpreet Kaur. “Machine Learning Approach for Credit Card Fraud Detection (KNN & Naïve Bayes).” SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3564040.

Jain, Rajni, Bhupesh Gour, and Surendra Dubey. 2016. “A Hybrid Approach for Credit Card Fraud Detection Using Rough Set and Decision Tree Technique.” International Journal of Computer Applications. https://doi.org/10.5120/ijca2016909325.

Khine, Aye Aye, and Hint Wint Khin. 2020. “Credit Card Fraud Detection Using Online Boosting with Extremely Fast Decision Tree.” 2020 IEEE Conference on Computer Applications(ICCA). https://doi.org/10.1109/icca49400.2020.9022843.

Adepoju, Olawale, Julius Wosowei, Shiwani Lawte, and Hemaint Jaiman. 2019. “Comparative Evaluation of Credit Card Fraud Detection Using Machine Learning Techniques.” 2019 Global Conference for Advancement in Technology (GCAT). https://doi.org/10.1109/gcat47503.2019.8978372.

TABLES AND FIGURES

Table 1. Pseudocode for Novel Logistic Regression

//I : Input dataset records
Import required packages.
Convert data sets into numerical values after the extraction feature.
Assign data to X train, Y train, X test and Y test variables.
Using train_test_split()function, pass training and testing variables.
Give test_size and the random_state as parameters for splitting the data.
Adding Regression, Dense Layer to the model.
Compiling model using matrices as accuracy.
Calculate accuracy of model.
OUTPUT//Accuracy

Table 2. Pseudocode for Gradient Boosting Ensemble Classifier

//I : Input dataset records
Import required packages.
Convert data sets into numerical values after the extraction feature.
Assign data to X train, Y train, X test and Y test variables.
Using train_test_split()function, pass training and testing variables.
Given test_size and ‘n_estimaors’ : model = GradientBoostingClassifier().
Compiling model using matrices as accuracy.

Calculate accuracy of model.
OUTPUT//Accuracy

Table 3. Accuracy of Fraud Detection using Novel Logistic Regression

Test size	**Accuracy**
Test 1	93.01
Test 2	93.11
Test 3	93.60
Test 4	93.75
Test 5	94.30
Test 6	94.09
Test 7	93.39
Test 8	93.08
Test 9	93.23
Test 10	94.41

Table 4. Accuracy of Fraud Detection using Gradient Boosting Ensemble Classifier

Test size	**Accuracy**
Test 1	94.05
Test 2	93.80
Test 3	92.48
Test 4	92.18
Test 5	91.46
Test 6	92.67
Test 7	91.18
Test 8	92.62
Test 9	93.23
Test 10	93.40

Table 5. Group , Accuracy , Loss value uses 8 Columns with 8 width data for Fraud Detection in Credit card.

S.NO	**Name**	**Type**	**Width**	**Decimal**	**Columns**	**Measure**	**Role**
1	Group	Numeric	8	2	8	Nominal	Input
2	Accuracy	Numeric	8	2	8	Scale	Input
3	Loss	Numeric	8	2	8	Scale	Input

Table 6. Group Statistical Analysis of Novel Logistic Regression and Gradient Boosting Ensemble Classifier. Mean, Standard Deviation and Standard Error Mean are obtained for 10 samples. Novel Logistic Regression has higher mean accuracy and lower mean loss when compared to Gradient Boosting Ensemble Classifier.

	GROUP	**ALGORITHM**	**N**	**MEAN**	**Std. Deviation**	**Std.Error Mean**
Accuracy	1	Logistic Regression	10	93.5970	.52156	.16493
	2	Gradient Boosting Ensemble Classifier	10	92.7070	.94145	.29771
Loss	1	Logistic Regression	10	6.3940	.51097	.16158
	2	Gradient Boosting Ensemble Classifier	10	7.1930	.98900	.31275

Table 7. Independent Sample T-test: Confidence interval as 95% and level of significance as 0.05. Novel Logistic Regression is insignificantly better than Gradient Boosting Ensemble Classifier with p value 0.030 ($p<0.05$)

Accuracy		**Levene's Test for Equality of Variance**		**t-test for Equality of Means**						
		F	**Sig.**	**t**	**df**	**Sig. (2-tailed)**	**Mean Difference**	**Std. Error Difference**	**95% Confidence Interval of the Difference Lower Upper**	
Accuracy	**Equal variances assumed**	2.476	.030	2.615	18	.018	.89000	.34035	.17496	1.60504
	Equal variances not assumed			2.615	14.049	.020	.89000	.34035	.16027	1.61973

Error	**Equal variances assumed**	5.538	.030	-2.270	18	.036	-.79900	.35202	-1.53857	-.05943
	Equal variances not assumed			-2.270	13.485	.040	-.79900	.35202	-1.55673	-.04127

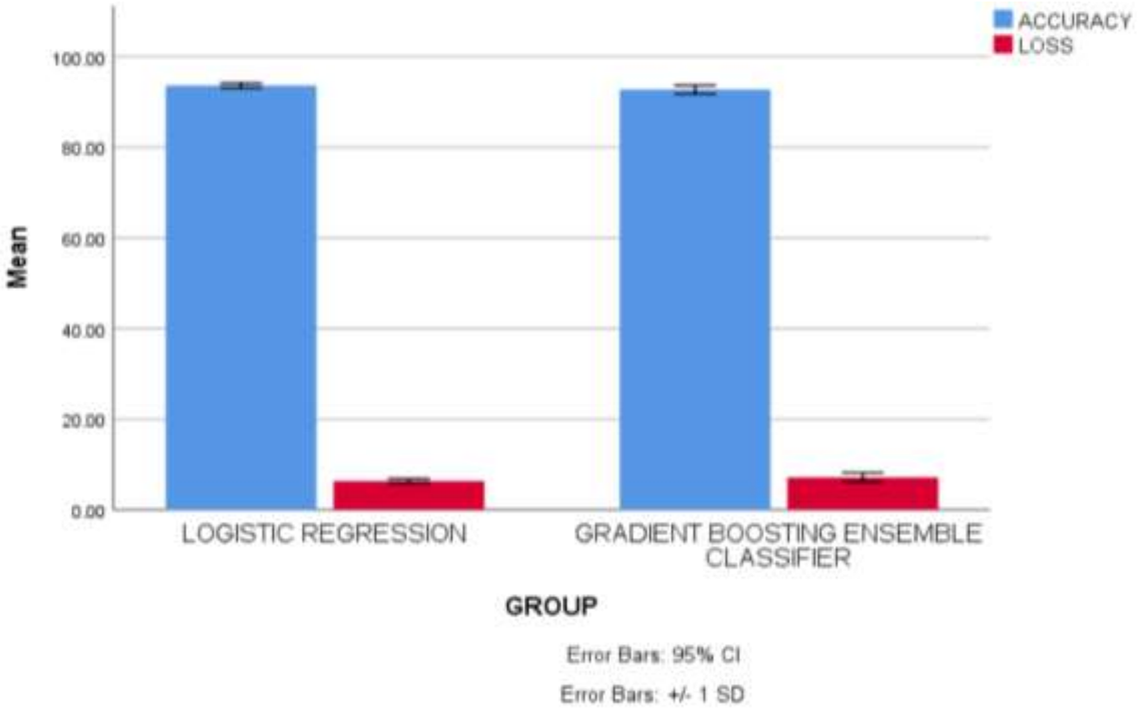

Fig. 1. Comparison of Novel Logistic Regression and Gradient Boosting Classifier in terms of mean accuracy and loss. The mean accuracy of Novel Logistic Regression is better than Gradient Boosting Classifier; Standard deviation of Novel Logistic Regression is slightly better than Gradient Boosting Ensemble Classifier. X Axis: Novel Logistic Regression vs Gradient Boosting Ensemble Classifier and Y Axis: Mean accuracy of detection ± 1 SD.