Novel Logistic Regression over Naive Bayes Improves Accuracy in Credit Card Fraud Detection

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

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

[2*]Project Guide, Corresponding Author, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India. Pincode: 602105.

ABSTRACT

Aim: To enhance the accuracy in credit card fraud detection using Naive Bayes and Novel Logistic Regression. **Materials and Methods:** This study contains Naive Bayes and Novel Logistic Regression. 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 Naive Bayes of 85.88% in detecting fraudulent transactions. Significance value for accuracy and loss is 0.255 ($p>0.05$). **Conclusion:** The Novel Logistic Regression model is significantly better than Naive Bayes in detecting fraudulent transactions. It can be also considered as a better option for credit card fraud detection.

Keywords: Novel Logistic Regression, Naive Bayes, Fraudulent Transactions, Credit Card, Accuracy, Machine Learning.

INTRODUCTION

Credit card fraud is defined as the unauthorized and undesired use of a credit card account by someone who is not the owner of the account (Hussain et al. 2021). The study of large groups of people's behavior in order to foresee, identify, or prevent undesired behavior such as fraud, intrusion, or default is known as fraud detection (Kelleher et al. 2020). The most major benefit is the ease with which credit can be obtained. It’s drawbacks are the transactions contain a limited amount of data, such as transaction amount, merchant category code (MCC), acquirer number, date and time, and the merchant's address. (Novakovic and Markovic 2020). To analyze all authorized transactions and report the ones that are questionable, machine learning techniques are used. This paper employs Naive Bayes and Novel Logistic Regression techniques to conduct comparative analyses of credit card fraud detection in order to determine the most accurate method of classifying a credit card transaction as fraudulent or non-fraudulent, taking into account a variety of factors and algorithms. The application of any credit card fraud detection system is to detect suspicious events and report them to an analyst while enabling normal transactions to proceed without interruption (Serra et al. 2021).

Google Scholar has indexed 247 papers in the last five years, compared to 295 for IEEE Xplore. Gradient boosting is a popular machine learning method due to its efficiency, accuracy, and interpretability. The fraud detection systems used by financial organizations are regularly updated (Duboue 2020;). An aggregation technique was utilized to produce a new set of characteristics based on the periodic behavior of transaction time. A fraud detection neural network was trained and tested on a holdout dataset using a large sample of labeled credit card account transactions. To evaluate the accuracy and early detection of credit card fraud using a neural network (Kaur and Kumar 2020). They are gaining a larger share of the payment system as a result of their expanding numbers. To keep the payment system working, fraud detection has been improved (Nandi et al. 2022). Future ideas for improving both methodology and outcomes (Jahanbakhshi et al. 2021).Our team has extensive knowledge and research experience that has translate into high quality publications(Bhavikatti et al. 2021; Karobari et al. 2021; Shanmugam et al. 2021; Sawant et al. 2021; Muthukrishnan 2021; Preethi et al. 2021; Karthigadevi et al. 2021; Bhanu Teja et al. 2021; Veerasimman et al. 2021; Baskar et al. 2021)

The main objective of this paper is to investigate the various credit card fraud detection strategies in order to understand their benefits and drawbacks and to collect and segment data is the first step in the machine learning fraud detection process then the collected data is used to find whether the transaction is fraudulent or not. Future scope based on the comparison of the various credit card fraud detection approaches presented above, it is evident that logistic regression performs the best in this scenario. However, machine learning cannot accurately recognise the names of fraud and fraud transactions for the supplied dataset, which is one of the paper's flaws. We can overcome this obstacle for the project's future development using a variety of ways.

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 Naive Bayes 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 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 0.2.

NOVEL LOGISTIC REGRESSION

Novel Logistic Regression is one of the most widely used classification algorithms in machine learning. The Novel Logistic Regression model describes the relationships between continuous, binary, and categorical predictors. Binary dependent variables are a possibility. We can anticipate whether something will be found or not based on a few features. For a given collection of predictors, we determine the likelihood of belonging to each category. In this statistical model, a logistic function is employed to model a binary dependent variable. This model is utilized when there is a possibility of a binary classification problem. It's best for classes that can be split in a linear fashion. It's best for classes that can be split in a linear fashion. One idea that can be used to define the logit function is the odds ratio. Pseudocode and Accuracy Values for the regression model is mentioned in Table 1 and Table 3.

NAIVE BAYES

The Naive Bayes technique is a supervised learning algorithm that uses the Bayes theorem to solve classification problems. It is mostly used in text classification problems that necessitate a large training dataset. The Naive Bayes Classifier is a simple and effective classification method for developing fast machine learning models that can make quick predictions shown in Fig. 1. It's a probabilistic classifier, meaning it makes predictions based on the probability of an object. The Naive Bayes Algorithm is commonly used for spam filtration, sentiment analysis, and article classification. The Bayes theorem, often known as Bayes' rule or Bayes' law, is a mathematical formula for determining the probability of a hypothesis based on prior data. This is determined by conditional probability. Pseudocode and Accuracy Values for the regression model are mentioned in Table 2 and Table 4.

STATISTICAL ANALYSIS

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 10, 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.

RESULT

The group statistical analysis on the two groups shows Novel Logistic Regression (93.59%) has more mean accuracy than Naive Bayes (85.88%) 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. Here table 1 and table 2 shows the pseudocode for logistic regression and convolutional neural networks, table 3 and table 4 shows the accuracy of the algorithms. Table 5 shows the Group, Accuracy, Loss values. The Graph for mean of Accuracy and mean of Loss by Groups is given in Fig. 2. Table 6 consists of the statistical analysis of Novel Logistic Regression and Convolutional Neural Network. Mean,Standard Deviation. Table 7 consists of an Independent Sample T-test.

DISCUSSION

From the results of this study, Novel Logistic Regression is proved to be having better accuracy. Novel Logistic Regression has an accuracy of 93.59% whereas Naive Bayes has an accuracy of 85.88%. Table 5, shows that group, accuracy and loss values for two algorithms Novel Logistic Regression and Naive bayes are denoted. Group statistics table shows a number of samples that are collected. Mean and standard deviation obtained and accuracies are calculated and entered.

To reduce their losses, all credit card issuing banks must implement effective fraud detection systems (Raj et al. 2011). Many people have created and executed various approaches, tactics, and procedures in order to provide a solution to avoid credit card theft (Kaur and Kumar 2020). Every day, over a million transactions take place, all of which must be verified for validity. To accomplish this, the system can be trained to distinguish between fraudulent and non-fraudulent transactions (Baesens et al. 2015). The machine learning models' live transactions, expected results, and other critical data were stored in a Data Warehouse (Thennakoon et al. 2019).

It's simple to use Novel Logistic Regression for machine learning. It does, however, come with some drawbacks such as Novel Logistic Regression will be overwhelmed by a large number of categorical features. In this example, we drastically reduced the amount of attributes. Novel Logistic Regression would fail to provide us with an accurate prognosis. Table 7, shows independent t sample tests for algorithms. The comparative accuracy analysis, mean of loss between two algorithms are specified.

CONCLUSION

Based on this study, the mean accuracy of Novel Logistic Regression is 93.59% compared to Naive Bayes which has a mean accuracy of 85.88%. Hence it is inferred that Novel Logistic Regression can predict fraudulent transactions more significantly than Naive Bayes. It can be used in predicting Credit card fraud detection in the future.

DECLARATIONS

Conflicts of Interest

No conflicts of interest in this manuscript.

Author’s 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, Chennai.
2. Saveetha University.
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering.

REFERENCES

1. 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.
2. 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.
3. 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.
4. 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.
5. Lamba, Harshit. 2020. *Credit Card Fraud Detection In Real-Time*.
6. 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.
7. 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.
8. 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.
9. 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.

10. 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.
11. 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.
12. 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.
13. 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.
14. 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.
15. 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.

TABLES AND FIGURES

Table 1. Pseudocode for Novel Logistic Regression

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

Table 2. Pseudocode for Naive Bayes

```
//I : Input dataset records
    1. Import required packages.
    2. Convert data sets into numerical values after the extraction feature.
    3. Assign data to X train, Y train, X test and Y test variables.
    4. Using train_test_split()function, pass training and testing variables.
    5. Formula for Naive Bayes
           $P(A|B)=P(B|A)P(A)P(B)$
    6. Compiling model using matrices as accuracy.
    7. 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 Naive Bayes

Test size	Accuracy
Test 1	86.36
Test 2	85.86

Test 3	86.09
Test 4	86.32
Test 5	86.17
Test 6	86.07
Test 7	85.46
Test 8	85.25
Test 9	85.50
Test 10	85.78

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 Naive Bayes. 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 Naive Bayes.

	GROUP	ALGORITHM	N	MEAN	Std. Deviation	Std.Error Mean
Accuracy	1	Logistic Regression	10	93.5970	.52156	.16493
	2	Naive Bayes	10	85.8860	.38240	.12092
Loss	1	Logistic Regression	10	6.4030	.52156	.16493
	2	Naive Bayes	10	14.1140	.38240	.12092

Table 7. Independent Sample T-test: Confidence interval as 95% and level of significance as 0.05. Novel Logistic Regression is insignificantly better than Naive Bayes with p value 0.255 (p>0.05).

	Levene's	**t-test for Equality of Means**

Accuracy		**Test for Equality of Variance**								
		F	**Sig.**	**t**	**df**	**Sig. (2-tailed)**	**Mean Difference**	**Std. Error Difference**	**95% Confidence Interval of the Difference Lower**	**Upper**
Accuracy	**Equal variances assumed**	1.383	.255	37.704	18	.000	7.71100	.20451	7.28134	8.14066
	Equal variances not assumed			37.704	16.507	.000	7.71100	.20451	7.27853	8.14347
Error	**Equal variances assumed**	1.383	.255	-37.704	18	.000	-7.71100	.20451	-8.14066	-7.28134
	Equal variances not assumed			-37.704	16.507	.000	-7.71100	.20451	-8.14347	-7.27853

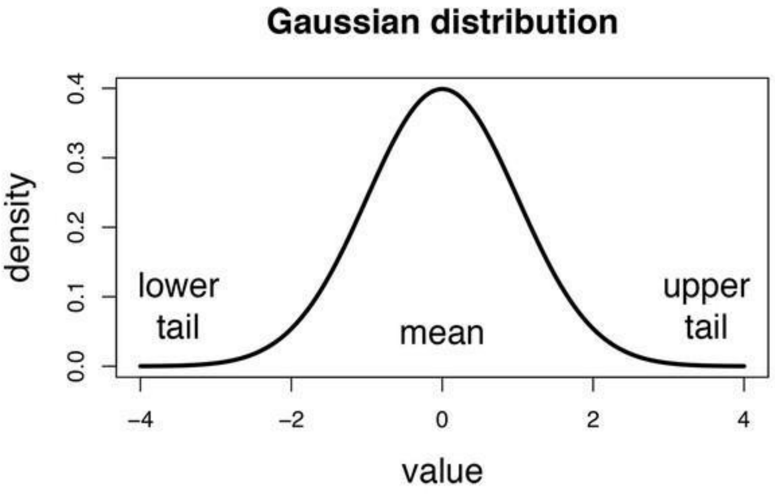

Fig. 1. Gaussian Distribution

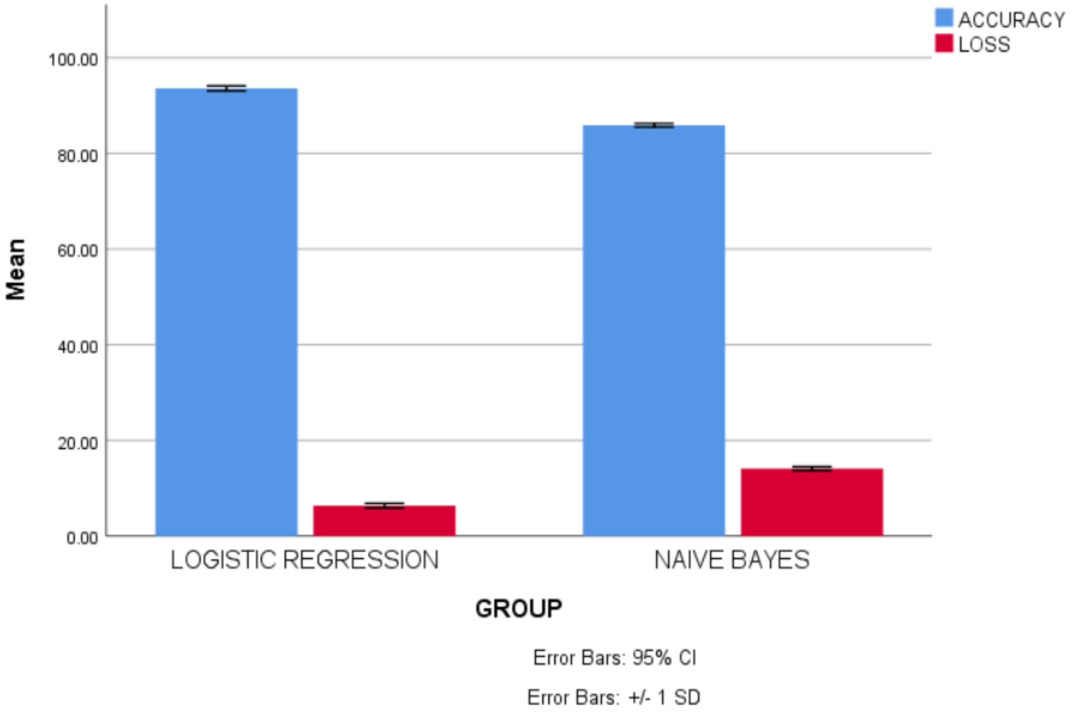

Fig. 2. Comparison of Novel Logistic Regression and Naive Bayes. in terms of mean accuracy and loss. The mean accuracy of Novel Logistic Regression is better than Naive Bayes.; Standard deviation of Novel Logistic Regression is slightly better than Naive Bayes. X Axis:Novel Logistic Regression vs Naive Bayes and Y Axis: Mean accuracy of detection ± 1 SD.