Title Page:

Credit Card Fraud Detection using Logistic Regression Compared with Naive Bayes to Improve Accuracy

Bhargav Chowdari [1,], S.Parthiban[2]

Bhargav Chowdari[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: 602105.
bhargavchowdarig18@saveetha.com

S.Parthiban[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: 602105.
parthibans.sse@saveetha.com

Keywords: Naive Bayes, Logistic Regression, Fraud detection, Supervised Learning, Machine learning.

ABSTRACT

Aim: The principle objective of this article is to improve the accuracy of Credit card fraud detection using Novel Logistic Regression compared with the Naive Bayes. **Materials and Methods**: The categorizing is performed by adopting a sample size of n = 10 in Novel Logistic Regression and sample size n = 10 in Naive Bayes with a sample size = 10, obtained using the G-power value 80%. **Results:** The analysis of the results shows that the Novel Logistic Regression has a high accuracy of (99.89) in comparison with the Naive Bayes(66.17). There is a statistically significant difference between the study groups with ($p< 0.05$). **Conclusion**: For Credit card fraud detection it shows that the Novel Logistic Regression appears to generate more accuracy than the Naive Bayes.

Keywords: Naive Bayes, Novel Logistic Regression, Fraud detection, Supervised Learning, Machine learning.

INTRODUCTION

In today's environment, Using a credit card is a common event in today's world. It's widely utilized for online transactions and payments. Credit cards can be used in a variety of ways. The use of credit cards has expanded tenfold, increasing the chances of fraud in such purchases. Credit card fraud costs the global economy billions of dollars. Fraud is defined as a deception to make illegal gains with someone else's money. Credit card fraud can be done in many different ways. (Nandi et al. 2022) By using lost or stolen cards, making fake or counterfeit cards, duplicating the original website, removing or altering the magnetic strip on the card that carries the user's information, by phishing by skimming or stealing data from a merchant's end. (Nandi et al. 2022) One of the techniques of purchasing products or services is by using a credit card. Fraud detection is the process of distinguishing between fraudulent and non-fraudulent transactions so that customers can enjoy their shopping or other transactions without delay. Many detections, such as the evolutionary algorithm, itemset mining, and migrating birds' algorithm, have been used to solve this problem(Namanda 2016).

Credit card fraud detection system cases in and around the world are sequenced by comparing machine learning algorithms 1385 journals from IEEE Xplore digital library 894 articles from ScienceDirect, 450 articles from Google Scholar 385 articles from collaborative algorithm and frequent pattern mining algorithm are two highly correlated parts in detection of college. Among all the articles and journals the most cited papers are(Kaur et al. 2019)). Many proposed a feature selection algorithm with Novel Logistic Regression for designing a high-level intelligent system for credit card detection classification(Seeja and Zareapoor 2014).In my opinion,(Jurgovsky 2019)Several works have demonstrated that the performance of Novel Logistic Regression is high and provides high accuracy in the prediction of credit card fraud detection. A study by(Reddy, Vijaya Kumar Reddy, and Ravi Babu 2018) compares the accuracy

of various mining detecting algorithms in detecting credit card fraud. It is important to analyze and compare the various classification algorithms that provide better accuracy(Bonney 1992; Nandi et al. 2022). Hence the aim of this research study is to use the Novel Logistic Regression algorithm inorder detect Credit card fraud by achieving better accuracy.

.

The Existing System, suffers With data Scalability, that impacts user based collaborative filtering algorithms whose performance falls with the growth in the number of users grows in admission system.The credit card fraud detection methods that were available traditionally for detecting fraud lack certain factors such as size of data along with the huge amount of data used for training purposes.When the number of data keeps on increasing the time taken for training them and the accuracy of the outcome gets affected.Hence the aim of this research study is to use the Novel Logistic Regression algorithm inorder detect Credit card fraud by achieving better accuracy.

MATERIALS AND METHODS

The research work was carried out in the machine learning lab, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences where the laboratory facilitates a high configuration system to obtain the experimental results. The number of groups identified for the study were two with the sample size (52) of per group. The computation is performed using G-power with 80% with alpha value 0.05 and beta value is 0.95 with a confidence interval at 95%.

The sample preparation group 1 is done for Novel logistic regression Algorithm, logistic regression is a family of algorithms in which comparable people or things are identified in numerous ways and ratings are calculated based on the ratings of similar users.

The sample preparation group 2 is done for Naive Bayes. By detecting regular patterns, we may group elements that are highly connected together and quickly uncover shared properties and correlations. By doing Naive Bayes, it is possible to perform further machine learning activities such as clustering, classification, and other machine learning tasks.

MATLAB is a high performance language for technical computing. It integrates computation, visualization and programming in an easy to use environment where problems and solutions are expressed in familiar mathematical notation. The r2014a version of matlab is used for improving the algorithms and minimum 8GB RAM 1 TB HDD storage is required to run the output.

The data collection is taken from the open source access website IEEE-dataport.org and kaggle that is used for credit card fraud detection using novel logistic regression and Support vector machine algorithm technique.The open data set contains 614 entries.

Statistical Analysis

Statistical software used is IBM SPSS with version 26.0 to find standard deviation, mean, standard error mean, mean difference, sig and F value. Independent variables are segment size and data size. Dependent variables are dataset size and accuracy of outcome. Independent T-Test analysis is carried out in this research work.

NOVEL LOGISTIC REGRESSION

It is one of the most popular Machine Learning algorithms under the Supervised Learning technique. It is used for predictive analysis of the categorical dependent variable using a given set of independent variables. Novel Logistic Regression predictive analysis of the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. The following algorithm steps follow

Step 1: Start.

Step 2: Load datasets path.

Step 3: Extract feature values.

Step 4: Apply the Novel Logistic Regression

Step 5: Separates the dominant players

Step 6: Identity the null values

Step 7: Process to recover from null values

Step 8: Stop

NAIVE BAYES

Naive Bayes is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. Naive Bayes classification algorithm is a probabilistic classifier. It is based on probability models that incorporate strong independence assumptions.Therefore they are considered as naive. The following algorithm steps follow

Step 1: Start.

Step 2: Load datasets path.

Step 3: Extract feature values.

Step 4: Apply the Naive Bayes

Step 5: Separates the dominant players

Step 6: Identify the null values

Step 7: Process to recover from null values

Step 8: Stop

RESULTS

Table1 represents the comparison of accuracy in credit card fraud detection by using Novel Logistic Regression and Naive bayes, by iterating in credit card fraud for various numbers of times.

Table 2 the statistical analysis of the Novel Logistic Regression algorithm and the Naive Bayes. Mean accuracy value, Standard deviation is observed that the Novel Logistic Regression algorithm performed better than the Naive Bayes algorithm. The Novel Logistic Regression algorithm obtained 1.55 standard deviations with 0.49 standard error while the Naive Bayes algorithm obtained 1.48 standard deviations with 0.46 standard error. Also the independent sample t-test was used to compare the accuracy of two algorithms and a statistically significant difference was noticed with the Novel Logistic Regression model obtained 99.89% accuracy and Naive Bayes having 66.17% of accuracy. When compared with the other algorithm's performance, the proposed Novel Logistic Regression classifier is significantly better than the Naive Bayes algorithm.

Table 3 represents the significance of the data and standard error difference, where the significance of Novel Logistic Regression and Naive Bayes with the confidence interval is 99.89% and the level of significance of 0.05.

Figure 1 shows the Bar Graph for Comparison of Accuracy. Mean accuracy of Novel Logistic Regression is better than Naive Bayes and standard deviation of ant colony optimization is slightly better than Naive Bayes. X axis:Novel Logistic Regression vs Naive Bayes. Y axis: Mean accuracy of detection ±1 SD .

DISCUSSION

The work shows that Novel Logistic Regression is better than Naive Bayes and the at detecting credit card fraud in terms of accuracy and cross-validation. (Baesens, Verbeke, and Van Vlasselaer 2015)However, the average error of Naive Bayes appears to be higher than logistic. From the experimental results performed in Jupyter, the accuracy and cross-validations of Novel Logistic Regression are 99.89% and 66.17%. Novel Logistic Regression has better significance ($p<0.05$) than SVM method and while using the independent sample t-tests.This shows that Novel Logistic Regression is better than Naive Bayes. The different parameters such as TP rate, FP rate, cross-validations are also (Hadi, Hafidudhin, and Tanjung 2018)compared. According to the SPSS plot, the proposed Novel Logistic Regression classifier performs better in terms of accuracy (99.89%) and cross-validations (66.17%) than the Naive Bayes algorithm.The most important aspect in forecasting credit card fraud detection is accuracy and cross-validation(Lawi and Aziz 2018). A machine learning-based diagnostic system for Credit card fraud detection was proposed using a creditcard.csv dataset. Popular machine learning algorithms, three feature selection algorithms, the cross-validation method, and seven classifier performance metrics such as classification accuracy, specificity, sensitivity, Matthews correlation coefficient, and delay execution were used by the study(Bonney 1992; Jurgovsky 2019).

In the study by, (Hadi, Hafidudhin, and Tanjung 2018)a scalable solution was proposed to enable credit card fraud detection. Naive Bayes algorithm was used on a Spark framework to predict and showed that even with a dataset of 300 documents, the study achieved a higher(Li et al. 2021).In the study of(Akgun and Mei 2020), attribute filtering, frequent element extraction, and various data mining techniques such as decision tree, support vector machine, and KNN classifications are used to predict credit card fraud detection. When it comes to predicting Credit card fraud detection, the cross validations of the Novel Tree Specific Random Forest algorithm was better than other data mining algorithms(Seera et al. 2021). The accuracy of the Novel Logistic Regression algorithm depends on the size of the training and testing data set. In our study, the accuracy and cross validations appear to be better than Naive Bayes. However, the average error appears to be higher in our proposed work which should be minimized(Jurgovsky 2019).

Although the results of the study are better in both experimental and statistical analysis, there are some limitations in the work. Accuracy assessment cannot provide a better result on larger data sets. In addition, in logistic regression, the average error seems to be higher than Naive Bayes. It would be preferable if the average error could be considerably reduced (Nandi et al. 2022). However, the work can be improved by applying optimization algorithm techniques, to achieve better cross validations and lower mean error(Dorronsoro et al. 1997). Feature selection algorithms can be used prior to classification to improve the classification accuracy of classifiers. Therefore, thanks to the Data mining algorithms, we can reduce the computation time and improve the accuracy of the classification of classifiers. (Jurgovsky 2019)

CONCLUSION

The Novel Logistic Regression is a classification technique that uses the mean to improve accuracy and cross validation. The work shows that the prediction of accuracy and cross validations for credit card fraud detection using logistic regression(99.89%) is better than the Naive Bayes(66.17%) at accurately detecting fraud, but the average error is slightly greater than logistic. Therefore, it is concluded that the Novel Logistic Regression provides acceptable accuracy and cross validations compared to Naive Bayes data mining algorithms

DECLARATION

Conflict of interest

No conflict of interest in this manuscript.

Author Contribution

Author GBC was involved in data collection, data analysis, algorithm framing, implementation and manuscript writing. Author SP was involved in designing the work flow, guidance and 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 research work successfully

Funding: We thank the following organizations for providing financial support that enabled us to complete the study.

1. Manac Infotech Pvt Ltd, HYD

2.Saveetha University

3.Saveetha Institute of Medical And Technical Sciences

4.Saveetha School of Engineering

REFERENCES

Akgun, O. C., and J. Mei. 2020. "An Energy Efficient Time-Mode Digit Classification Neural Network Implementation." *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences* 378 (2164): 20190163.

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.

Bonney, R. 1992. *Preveventing Credit Card Fraud*.

Dorronsoro, J. R., F. Ginel, C. Sgnchez, and C. S. Cruz. 1997. "Neural Fraud Detection in Credit Card Operations." *IEEE Transactions on Neural Networks / a Publication of the IEEE Neural Networks Council* 8 (4): 827–34.

Hadi, Sholikul, Didin Hafidudhin, and Hendri Tanjung. 2018. "Comparison of Conventional Systems Credit Card and Credit Card Shariah as Alternative Construction Credit Card on Banking System." *Jurnal Manajemen*. https://doi.org/10.32832/jm-uika.v8i1.733.

Jurgovsky, Johannes. 2019. *Context-Aware Credit Card Fraud Detection*.

Kaur, Paramjit, Kewal Krishan, Suresh K. Sharma, and Tanuj Kanchan. 2019. "ATM Card Cloning and Ethical Considerations." *Science and Engineering Ethics* 25 (5): 1311–20.

Lawi, Armin, and Firman Aziz. 2018. "Classification of Credit Card Default Clients Using LS-SVM Ensemble." *2018 Third International Conference on Informatics and Computing (ICIC)*. https://doi.org/10.1109/iac.2018.8780427.

Li, Yuening, Zhengzhang Chen, Daochen Zha, Kaixiong Zhou, Haifeng Jin, Haifeng Chen, and Xia Hu. 2021. "Automated Anomaly Detection via Curiosity-Guided Search and Self-Imitation Learning." *IEEE Transactions on Neural Networks and Learning Systems* PP (September). https://doi.org/10.1109/TNNLS.2021.3105636.

Namanda, Marvin. 2016. *Future Issues with Credit Card Fraud Detection Techniques*. GRIN Verlag.

Nandi, Asoke K., Kuldeep Kaur Randhawa, Hong Siang Chua, Manjeevan Seera, and Chee Peng Lim. 2022. "Credit Card Fraud Detection Using a Hierarchical Behavior-Knowledge Space Model." *PloS One* 17 (1): e0260579.

Reddy, R. Vijaya Kumar, R. Vijaya Kumar Reddy, and U. Ravi Babu. 2018. "Efficient Handwritten Digit Classification Using User-Defined Classification Algorithm." *International Journal on Advanced Science, Engineering and Information Technology*. https://doi.org/10.18517/ijaseit.8.3.5397.

Seeja, K. R., and Masoumeh Zareapoor. 2014. "FraudMiner: A Novel Credit Card Fraud Detection Model Based on Frequent Itemset Mining." *TheScientificWorldJournal* 2014 (September): 252797.

Seera, Manjeevan, Chee Peng Lim, Ajay Kumar, Lalitha Dhamotharan, and Kim Hua Tan. 2021. "An Intelligent Payment Card Fraud Detection System." *Annals of Operations Research*, June, 1–23.

Table 1. Comparison between Novel Logistic Regression and Naive Bayes algorithms with N= 10 samples of the dataset with the highest accuracy of respectively 99.89% and 66.17% in sample 1 (when N=1) using the dataset size = 614 and the 70% of training and 30% of testing data.

ITERATIONS	LOGISTIC REGRESSION	Naive bayes
1	99.89	66.17
2	98.80	65.60
3	97.82	64.84
4	97.80	64.20
5	96.64	63.33
6	95.72	62.89
7	95.06	61.66
8	94.54	60.33
9	93.56	59.62
10	93.02	58.01

Table 2. Statistical results of Novel Logistic Regression and Naive Bayes algorithms. Mean accuracy value, Standard deviation and Standard Error Mean For LR and Naive Bayes algorithms are obtained for 10 iterations. It is observed that the LR (99.89%) performed better than the Naive Bayes (66.17%) algorithm

	Algorithms	Sample(N)	Mean	Std Deviation	Std Error Mean
Accuracy	LOGISTIC REGRESSION	10	96.28	2.28	0.723
	Naive Bayes	10	62.66	2.70	0.855

Table 3. The Independent sample t-test of the significance level Novel Logistic Regression and Naive Bayes algorithms results with two tailed significant values (p=0.001). Therefore both the LR and the Naive Bayes algorithms have a significance level less than 0.05 with a 95 % confidence interval.

	Lavene's test for equality of variances	**T-Test Equality of Means**	**95% Confidence interval of the difference**

		F	sig	t	df	sig(2 tailed)	Mean diff	Std.error	Lower	Upper
Accuracy	**Equal variance Assumed**	0.29	.593	30.013	18	<0.001	33.620	1.120	31.266	35.973
	Equal variances not assumed			30.013	17.513	<0.001	33.620	1.120	31.266	35.973

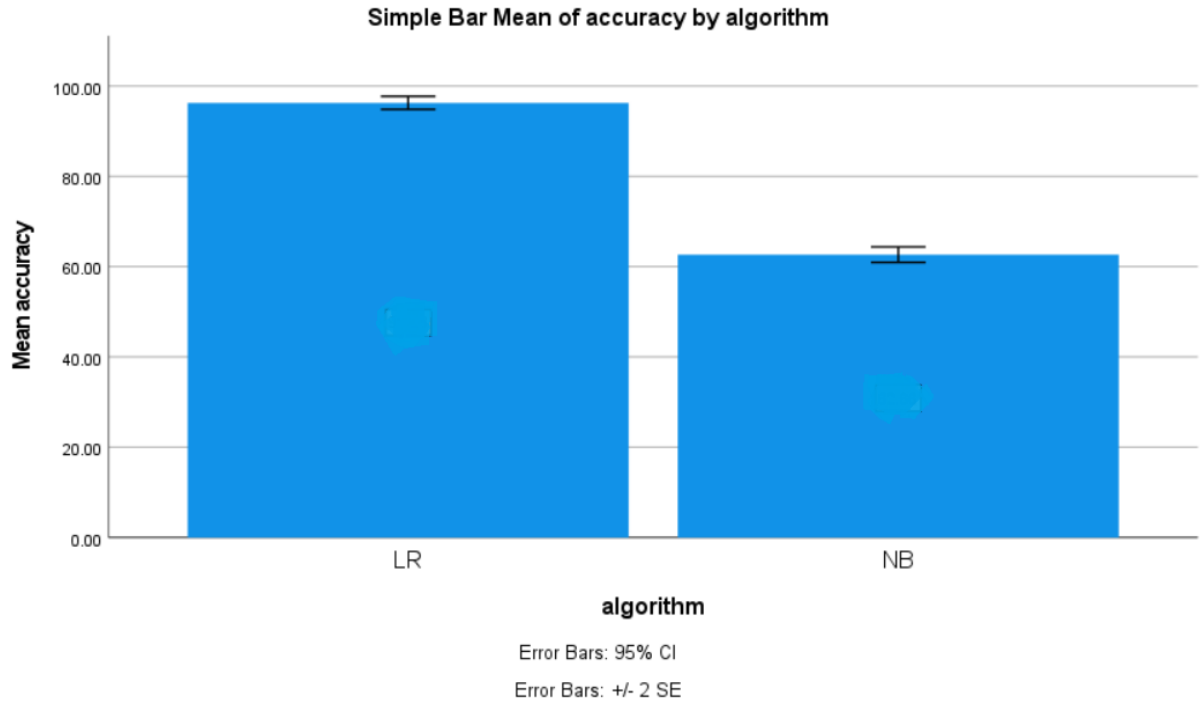

Fig. 1. Comparison of Novel Logistic Regression algorithm and Naive Bayes in terms of accuracy. The mean accuracy of the LR algorithm is better than Naive Bayes and standard deviation of LR is slightly better than Naive Bayes algorithm. X Axis : LR, Naive Bayes Y Axis: Mean accuracy of detection ± 1 SD .