IMPROVED ACCURACY FOR CREDIT CARD FRAUD DETECTION USING PIPELINING AND ENSEMBLE LEARNING METHODS LOGISTIC REGRESSION COMPARED WITH K-NEAREST NEIGHBOR ALGORITHM

CH. Kiran Kumar[1], S.S.Arumugam[2*]

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

Abstract

Aim: The goal of this study is to provide an improved accuracy for credit card fraud detection using pipelining and ensemble learning methods in logistic regression compared with k-nearest neighbor algorithms to detect credit card fraud and comparing their accuracy. **Materials and Methods:** The sample size for logistic regression (N=10) and for K-nearest neighbor algorithm (N=10) was iterated 20 times to predict credit card fraud. **Results :** logistic regression has significantly better accuracy (98%) compared to k-nearest neighbor (94%)The statistical significance difference 0.00($p<0.05$ independent sample test) value states that the results in the study are significant. **Conclusion:** The results depicted that logistic regression provides good results in detection of credit card fraud over k-nearest neighbor.

Keywords: Credit card fraud detection technique, Novel Classification, En-semble learning, Logistic regression, Random forest, K-nearest neighbor, Support vector machine, Naive bayes, Data mining.

[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: 602015.

[2*]Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India, Pincode: 602015.

1. Introduction

The research of this study is to predict the accuracy percentage of credit card fraud detection (Awoyemi, Adetunmbi, and Oluwadare 2017). As you are moving towards the digital world- cybersecurity is becoming a crucial part of our life. When you talk about security in digital life then the main challenge is to find the abnormal activity (Chertoff 2018). When you make any transaction while purchasing any product online a good amount of people prefer credit cards. The credit cards sometimes help us make purchases even if you don't have the money at that time. but, on the other hand, these features are misused by cyber attackers (Canada and Competition Bureau Canada 2014). To tackle this problem you need a system that can abort the transaction if it finds fishy. Here, comes the need for a system that can track the pattern of all the transactions and if any pattern is abnormal then the transaction should be aborted (White 1976). Today ,you have many machine learning algorithms that can help us classify abnormal transactions (Garg, Chaudhary, and Mishra 2021). The only requirement is the past data and the suitable algorithm that can fit our data in a better form (Brownlee 2018). In this article, I will help you in the complete end-to-end model training process--finally, you will get the best model that can classify the transaction into normal and abnormal types. (Nigrini 2012).

Identifying misinformation of Credit card fraud was implemented by many researchers to bring awareness about credit card fraud detection. Around 20 articles published in IEEE and 200 papers in google scholar. (Awoyemi, Adetunmbi, and Oluwadare 2017) 92% accuracy was obtained with implementation of machine learning models for classifying the fraud detection articles related to credit card fraud detection. (Seeja and Zareapoor 2014) implemented the Logistic Regression machine learning algorithm for predicting financial fraud detection and proved with accuracy of 98%. (*Detecting Credit Card Fraud: An Analysis of Fraud Detection Techniques* 2020) 94% of accuracy obtained for detection of credit card fraud using a machine learning model K-Nearest Neighbor. (Dal Pozzolo et al. 2018) implemented a machine learning algorithm for predicting the accuracy of misinformation about credit card fraud detection and accuracy was 98%. The most cited article was (Garg, Chaudhary, and Mishra 2021) focused on predicting accuracy of misinformation of credit card fraud detection using the Logistic regression machine learning algorithm with an accuracy of 98%(Baesens, Verbeke, and Van Vlasselaer 2015).Our team has extensive knowledge and research experience that has translated into high quality publications(Pandiyan et al. 2022; Yaashikaa, Devi, and Kumar 2022; Venu et al. 2022; Kumar et al. 2022; Nagaraju et al. 2022; Karpagam et al. 2022; Baraneedharan et al. 2022; Whangchai et al. 2022; Nagarajan et al. 2022; Deena et al. 2022)

The research gap identified from the survey is that there are many methods proposed for detecting credit card fraud but most of the methods have less accuracy rate. The main aim of this study is to detect credit card fraud by using logistic regression and k-nearest neighbors to attain better accuracy.

2. Materials and Methods

The research study was done in a Machine learning programming Lab. Saveetha School of Engineering, saveetha Institute of Medical and Technical Science (SIMATS).The number of groups identified for the study are two. The group -1 is logistic regression and group -2 is k-nearest neighbor. Sample size for each group was calculated by using previous study results in credit card fraud detection by keeping g power 80% ,threshold 0.05 and confidence interval as 95% . According to that, the sample size of the logistic regression algorithm (N=10) and k-nearest neighbor algorithm (N=10) were calculated.

The dataset contains the real bank transactions made by European cardholders in the year 2013. As a security concern, the actual variables are not being shared but they have been transformed versions of PCA. Today you have many machine learning algorithms that can help us classify abnormal transactions. The only requirement is the past data and the suitable algorithm that can fit our data in a better form. I will help you in the complete end-to-end model training process. Finally,you will get the best model that can classify the transaction into normal to abnormal types. The dataset collected from the kaggle(http://www.kaggle.com).

Logistic Regression Algorithm

The proposed algorithm is Logistic regression. Logistic regression is one of the most popular machine learning algorithms for novel classification because it is a simple algorithm that performs very well on a wide range of problems. It establishes the relationship between a categorical variable and one or more independent variables. This relationship is used in machine learning to predict the outcome of a categorical variable. It is widely used in many different fields,trading and business and fraud detection and many more.

- Import the dataset from the drive
- Prepare test and trained dataset and complete data preprocessing

- To calculate the logistic function
- To learn the coefficient for a logistic regression model using stochastic gradient descent
- The predictions using a logistic regression model
- Run the code and get accuracy.

K-Nearest Neighbor Algorithm

The proposed algorithm is k-nearest neighbor. K-nearest neighbor is a most popular machine learning algorithm that belongs to the supervised learning technique. The KNN algorithm assumes the similarity between the new case / data and available cases and puts the new case into the category that is most similar to the available categories. KNN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suited category by using KNN algorithm. The knn algorithm can be used for regression as well as for novel classification but mostly it is used for the classification problems. KNN is a non-parametric algorithm , which means it does not make any assumptions on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of novel classification, it performs an action on the dataset. The KNN algorithm at the training phases just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

- Import the dataset from the drive.
- Explore the data in the dataset.
- Pre-process the data.
- It is simple to implement.
- Fitting the k-nn algorithm to the training set.
- Divide the data into training and testing sets.
- Predicting the test result.
- Test accuracy of the result.
- Visualizing the test set result.
- Run the code and get the accuracy.

The software tool used to evaluate the logistic regression and k-nearest neighbor algorithm was google colab with python programming language. The hardware configuration was intel core i3 processor with a RAM size of 4GB.The system type used was 64-bit,os, x64 based processor with HDD of 917 GB. The software configuration includes the windows 8 operating system.

In the proposed model first,perform the data preprocessing on the input images and prepare the data. After that use convolutional neural networks for feature extraction. Later split the data applied it to the novel classification algorithm logistic regression and k-nearest neighbor algorithm by applying 70 percent of data for training and 30 percent for testing next performing evaluation metrics to understand the models.

The analysis was done using IBM SPSS version 21. It is a statistical software tool used for data analysis. For both proposed and existing algorithms 10 iterations were done with a maximum of 10-20 samples and for each iteration the predicted accuracy was noted for analyzing accuracy.

Statistical Analysis

In this research date,time and transaction id are independent variables because they are inputs and remain constant even after changing other parameters, Whereas date,time, transaction id and fraud are dependent variables because they depend on the inputs and vary for every change in the input. The analysis of the research work is done using independent T-Test which is used to compare logistic regression and k-nearest neighbor algorithm to detect credit card fraud.

3. Results

In Table 1, it was observed that LR and KNN algorithms were run at different times in Google colab with a sample size of 20 and accuracy was calculated. The LR algorithm has better accuracy than the RF algorithm.

In Table 2, Independent Sample T-Test was performed to compare the accuracy of LR and KNN and a statistically significant difference was noticed P < 0.00 with 95% confidence level showed that our hypothesis holds good. With respect to changes in the input values (independent variables) the corresponding output values (dependent variables) also changes (Table 2) the mean difference of accuracy was identified as 6.8900.

In Table 3, The statistical analysis of 10 samples was performed. LR obtained .95737 standard deviation with .302765 standard error while KNN obtained 3.02765 standard deviation with .95737 standard error. Accuracy percentage of LR (98) and KNN (94) inferes that LR proves with better accuracy than KNN (Fig. 1). The simple mean Bar graph shows the Standard deviation of LR is better than KNN (Fig. 1).

4. Discussion

In this study the LR and KNN algorithm was analyzed for predicting the accuracy percentage of Credit card fraud detection. It is observed that LR proves with better accuracy (98%) compared to KNN (94%) for predicting Credit card fraud detection. The Novel sigmoid function maps the dataset into higher dimensional space which helps to improve accuracy percentage. The results show the evidence there is a statistically significant difference between the LR and KNN algorithms (Shiny Irene et al. 2021).

This paper (Awoyemi, Adetunmbi, and Oluwadare 2017) shows 80% of accuracy and was implemented using Buzzsumo analytical tool for predicting misinformation of Credit card fraud detection. (Garg, Chaudhary, and Mishra 2021) machine learning techniques were implemented with an accuracy of 71%. (*Detecting Credit Card Fraud: An Analysis of Fraud Detection Techniques* 2020) explains prediction of accuracy using the LR algorithm with an accuracy of 98%. (Dal Pozzolo et al. 2018) Implemented KNN algorithm with an accuracy of 94%. (Canada and Competition Bureau Canada 2014) 98% of accuracy was predicted using the LR algorithm. (White 1976) detecting the fake news using a machine learning model with an accuracy of 94%. (Brownlee 2018) 94% of accuracy was obtained when detecting the credit card fraud detection with machine learning algorithms and compared with the machine learning model. (*Detecting Credit Card Fraud: An Analysis of Fraud Detection Techniques* 2020) implemented machine learning models with an accuracy of 98%.

The attributes that affect accuracy percentage of credit card fraud detection are UserName, ScreenName, Location, Transaction,Time. Original Transaction and Sentiment features are mainly focused to calculate the accuracy percentage of credit card fraud detection. It is proved that the proposed LR has better accuracy compared with previous research articles discussed. It can help the bank to keep track of credit card fraud detection.

The limitation of the proposed work is that the real time dataset with more parameters may give more accurate results of predicting accuracy. In future work, the framework can be extended to include trust information sources such as the " European cardholders" website which could get more parameters related to credit card fraud detection and thus it may result in predicting more accuracy.

5. Conclusion

In this research, a machine learning based model was implemented to detect and classify credit card fraud detection . The proposed model is fully automated , able to extract the features from the images. Based on the obtained results of credit card fraud detection , the accuracy of logistic regression is(98%) and accuracy of K-nearest neighbor is (94%).

Declarations

Conflict of Interest

No conflict of interest in this manuscript.

Author Contributions

Author CHK was involved in data collection ,data analysis, and manuscript writing. Author SSA was involved in conceptualization, guidance and critical review of manuscript.

Acknowledgement

The authors 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 the necessary infrastructure to carry out this work successfully.

Funding

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

- VTech solutions
- Saveetha University
- Saveetha Institute of Medical and Technical Sciences
- Saveetha School of Engineering

6. References

Awoyemi, John O., Adebayo O. Adetunmbi, and Samuel A. Oluwadare. 2017. "Credit Card Fraud Detection Using Machine Learning Techniques: A Comparative Analysis." 2017 International Conference on Computing Networking and Informatics (ICCNI). https://doi.org/10.1109/iccni.2017.8123782.

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.

Baraneedharan, P., Sethumathavan Vadivel, C. A. Anil, S. Beer Mohamed, and Saravanan Rajendran. 2022. "Advances in Preparation, Mechanism and Applications of Various Carbon Materials in Environmental Applications: A Review." Chemosphere. https://doi.org/10.1016/j.chemosphere.2022.134596.

Brownlee, Jason. 2018. Better Deep Learning: Train Faster, Reduce Overfitting, and Make Better Predictions. Machine Learning Mastery.

Canada, Industry, and Competition Bureau Canada. 2014. The Little Black Book of Scams: Your Guide to Protection Against Fraud, The Canadian Edition. Competition Bureau Canada.

Chertoff, Michael. 2018. Exploding Data: Reclaiming Our Cyber Security in the Digital Age. Atlantic Books.

Dal Pozzolo, Andrea, Giacomo Boracchi, Olivier Caelen, Cesare Alippi, and Gianluca Bontempi. 2018. “Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy.” IEEE Transactions on Neural Networks and Learning Systems 29 (8): 3784–97.

Deena, Santhana Raj, A. S. Vickram, S. Manikandan, R. Subbaiya, N. Karmegam, Balasubramani Ravindran, Soon Woong Chang, and Mukesh Kumar Awasthi. 2022. “Enhanced Biogas Production from Food Waste and Activated Sludge Using Advanced Techniques – A Review.” Bioresource Technology. https://doi.org/10.1016/j.biortech.2022.127234.

Detecting Credit Card Fraud: An Analysis of Fraud Detection Techniques. 2020.

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.

Karpagam, M., R. Beaulah Jeyavathana, Sathiya Kumar Chinnappan, K. V. Kanimozhi, and M. Sambath. 2022. “A Novel Face Recognition Model for Fighting against Human Trafficking in Surveillance Videos and Rescuing Victims.” Soft Computing. https://doi.org/10.1007/s00500-022-06931-1.

Kumar, P. Ganesh, P. Ganesh Kumar, Rajendran Prabakaran, D. Sakthivadivel, P. Somasundaram, V. S. Vigneswaran, and Sung Chul Kim. 2022. “Ultrasonication Time Optimization for Multi-Walled Carbon Nanotube Based Therminol-55 Nanofluid: An Experimental Investigation.” Journal of Thermal Analysis and Calorimetry. https://doi.org/10.1007/s10973-022-11298-4.

Nagarajan, Karthik, Arul Rajagopalan, S. Angalaeswari, L. Natrayan, and Wubishet Degife Mammo. 2022. “Combined Economic Emission Dispatch of Microgrid with the Incorporation of Renewable Energy Sources Using Improved Mayfly Optimization Algorithm.” Computational Intelligence and Neuroscience 2022 (April): 6461690.

Nagaraju, V., B. R. Tapas Bapu, P. Bhuvaneswari, R. Anita, P. G. Kuppusamy, and S. Usha. 2022. “Role of Silicon Carbide Nanoparticle on Electromagnetic Interference Shielding Behavior of Carbon Fibre Epoxy Nanocomposites in 3-18GHz Frequency Bands.” Silicon. https://doi.org/10.1007/s12633-022-01825-1.

Nigrini, Mark J. 2012. Benford’s Law: Applications for Forensic Accounting, Auditing, and Fraud Detection. John Wiley & Sons.

Pandiyan, P., R. Sitharthan, S. Saravanan, Natarajan Prabaharan, M. Ramji Tiwari, T. Chinnadurai, T. Yuvaraj, and K. R. Devabalaji. 2022. “A Comprehensive Review of the Prospects for Rural Electrification Using Stand-Alone and Hybrid Energy Technologies.” Sustainable Energy Technologies and Assessments. https://doi.org/10.1016/j.seta.2022.102155.

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

Shiny Irene, D., V. Surya, D. Kavitha, R. Shankar, and S. John Justin Thangaraj. 2021. “An Intellectual Methodology for Secure Health Record Mining and Risk Forecasting Using Clustering and Graph-Based Classification.” Journal of Circuits Systems and Computers 30 (08): 2150135.

Venu, Harish, Ibham Veza, Lokesh Selvam, Prabhu Appavu, V. Dhana Raju, Lingesan Subramani, and Jayashri N. Nair. 2022. “Analysis of Particle Size Diameter (PSD), Mass Fraction Burnt (MFB) and Particulate Number (PN) Emissions in a Diesel Engine Powered by Diesel/biodiesel/n-Amyl Alcohol Blends.” Energy. https://doi.org/10.1016/j.energy.2022.123806.

Whangchai, Niwooti, Daovieng Yaibouathong, Pattranan Junluthin, Deepanraj Balakrishnan, Yuwalee Unpaprom, Rameshprabu Ramaraj, and Tipsukhon Pimpimol. 2022. “Effect of Biogas Sludge Meal Supplement in Feed on Growth Performance Molting Period and Production Cost of Giant Freshwater Prawn Culture.” Chemosphere 301 (August): 134638.

White, Kenneth J. 1976. “The Effect of Bank Credit Cards On the Household Transactions Demand for Money.” Journal of Money, Credit and Banking. https://doi.org/10.2307/1991919.

Yaashikaa, P. R., M. Keerthana Devi, and P. Senthil Kumar. 2022. “Advances in the Application of Immobilized Enzyme for the Remediation of Hazardous Pollutant: A Review.” Chemosphere 299 (July): 134390.

Tables and Figures

Table 1: Predicted Accuracy of CREDIT CARD FRAUD DETECTION (LR algorithm accuracy of 98% and compared with KNN accuracy of 94%)

SL.No	Sample Size	LR algorithm Accuracy (%)	KNN algorithm Accuracy (%)
1	21	98.00	94.00
2	31	97.90	93.50
3	41	97.50	93.00
4	51	97.00	92.50
5	61	96.80	92.00
6	71	96..72	91.50
7	81	96.60	91.00
8	91	96.50	90.50
9	100	96.20	90.00
10	120	96.00	89.00

Table 2: Independent Sample T-test Results with confidence interval of 95% and level of significance of 0.05 (Logistic Regression performs significantly better than K-Nearest neighbor with the value of p=0.000)

	Levene's Test for Equality of Variances		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	11.958	.003	6.862	18	.000	6.89000	1.00415	4.78035	8.99965
Equal variances not assumed			6.862	11.782	.000	6.89000	1.00415	4.67441	9.10559
Loss Equal Variances assumed	11.334	.004	7.805	18	.000	7.29000	1.00854	4.13526	8.44474
Equal variances not assumed			7.805	12.089	.000	7.29000	1.00854	4.09343	8.48657

Table 3: Statistical analysis of LR and KNN. Mean accuracy value, Standard deviation and Standard Error Mean for LR and algorithms as KNN obtained for 10 iterations. It is observed that the LR algorithm performed better than the KNN algorithm.

	Groups	N	MEAN	Std.Deviation	Std.error mean

ACCURACY	LR	10	96.3900	.95737	.30275
	KNN	10	89.5000	3.02765	.95743

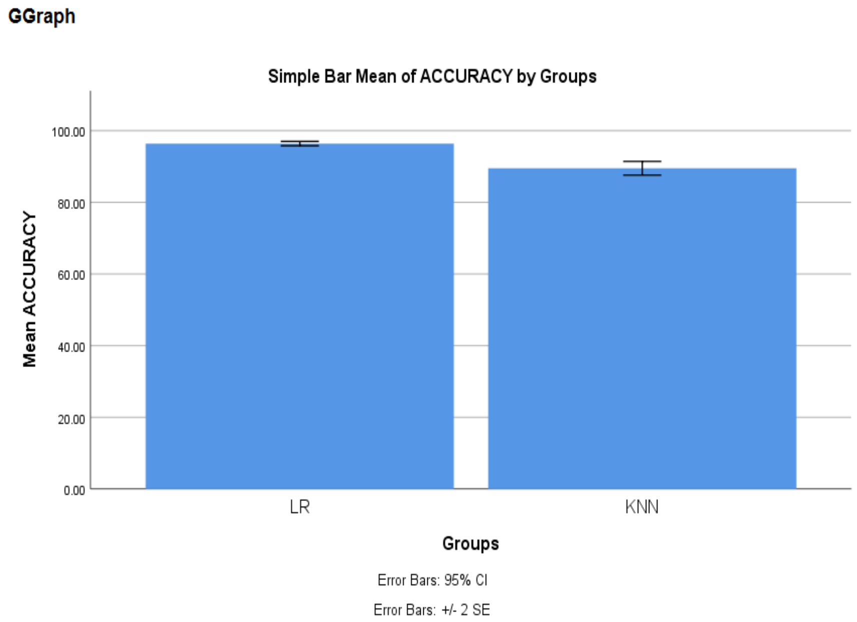

Fig. 1. Comparison of LR algorithm and KNN in terms of mean accuracy. The mean accuracy of LR is better than RF and the standard deviation of LR is slightly better than KNN. X Axis: LR vs KNN Algorithm, Y Axis: Mean accuracy of detection ±1SD.