ANALYTICAL APPROACH FOR DETECTION OF CREDIT CARD FRAUD USING LOGISTIC REGRESSION COMPARED WITH NOVAL RANDOM FOREST

B. Sri Sai Chowdary[1], J. Chenni Kumaran[2*]

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

Abstract

Aim: The main aim of the research is to detect Credit Card Fraud using Logistic Regression (LR)compared with the Novel Random Forest (RF).
Materials and Methods: When implementing an accurate prediction model it might not be sufficient to just consider one or two parameters. This analysis will be fed to the prediction model. Following logistic Regression Algorithm, Novel Random Forest Algorithm Based on the Previous Collected Datasets can Predict the Upcoming credit card fraud With Calculations.
Result: Comparison is done by using SPSS Software.The Logistic Regression algorithm produces 83.5% whereas Random forest algorithm produces 94.89% accuracy while detecting credit card fraud on a data set ($p > 0.05$). Hence Random forest is better than Logistic Regression.
Conclusion: After Using iterations get that by using logistic Regression algorithm get 83.5% (0.83) and novel Random Forest algorithm get 94.89% (0.94).So can say that By using the novel Random Forest Algorithm get more Accuracy than logistic Regression Algorithm.

Keywords: Credit Card Fraud Detection, Classification, Logistic Regression, Machine Learning, Novel Random Forest, legitimate.

[1]Research Scholar, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and technical Science, Saveetha University, Chennai, Tamil Nadu, India. Pincode: 602105.
[2*]Project Guide, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India. Pincode: 602105.

1. Introduction

The cards are subsequently used fraudulently by impostors who have intercepted them. Card fraud transactions stored by financial issuers are very small compared to legitimate transactions, which results in a high imbalance credit card dataset (Mqadi, Naicker, and Adeliyi 2021). The manual method is estimated by different fraud investigators who check the separate transaction and generate binary feedback on every fraud transaction. Fraud cases in the transaction are the primary barrier while enhancing e-commerce and also cause a massive loss in the economy (Mehbodniya et al. 2021). This problem is costly for financial institutions, whether it is money or reputation, which is why they look for various solutions to prevent fraud. However, fraudsters also use technology to evolve and destroy these protective measures (Peng, Hao, and Pal 2021). Card fraud can happen with the theft of the physical card as well as with the compromise of the card, including skimming, breach, and account takeover, which would otherwise look like a legitimate transaction (Gao et al. 2019). Credit card fraud occurs whenever a credit card is used without the consent of its legitimate owner with the aim of either making purchases or stealing money (Gao et al. 2019; Buonaguidi et al. 2021). Credit Card Fraud is one of the biggest threats to business establishments today. Credit card fraud begins either with the theft of the physical card or with the important data associated with the account, such as card account number or other information that necessarily be available to a merchant during a permissible fraud transaction that is not legitimate (Hyder John 2019).

In the last five years, Google scholar identified almost 7100 research articles on Credit Card Fraud Detection using one of the machine learning algorithms, logistic regression compared with the novel Random Forest. Identification in illegal credit card dealings is an enormously intricate issue, as aspects are rarely beneficial if considered independently (Morine et al. 2017). The next approach employs supervised learning schemes to teach appropriate classifiers on the collection of trades involving genuine and fraudulent cases. The directed learning scheme works by removing out the swindle features from the deceitful trade (Venkata Suryanarayana, Balaji, and Venkateswara Rao 2018). Financial fraud has posed a serious menace that are far reaching consequences for individuals, corporate organizations, the government, and the finance industry(Gao et al. 2019). Creditcard span is an extensive term for deceits committed, including an imbursement card such as a credit card as a degraded source of money in dealings ("Credit Card Fraud Detection System Based on Operational & Transaction Features Using SVM and Random Forest Classifiers" n.d.). Classifiers are typically employed to analyze all the authorized transactions and alert the most suspicious ones. Alerts are then inspected by professional investigators who contact the cardholders to determine the true nature (either genuine or fraudulent) of each alerted fraud transaction (V and Gokula 2019). In order to minimize costs of detection it is important to use expert rules and statistical-based models (e.g. Machine Learning) to make the first screen between genuine and potential fraud and ask the investigators to review only the cases with high risk (Venkata Suryanarayana, Balaji, and Venkateswara Rao 2018).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) (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 existing system shows poor accuracy. The study is to improve the accuracy of Classification(Mehbodniya et al. 2021) by incorporating logistic Regression and comparing performance with the novel Random Forest. The proposed model improves classifiers to achieve more accuracy for Credit Card Fraud Detection.

2. Materials and Methods

This study setting was done in the Soft Computing Laboratory, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences. The number of required samples in research is two in which group 1 is logistic Regression compared with group 2 is novel Random Forest Algorithm. The samples were taken from the device and iterated 10 times to get the desired accuracy with G power. A dataset consisting of a collection of stocks was downloaded from the Kaggle repository (salihfurkansaglam 2021)

Algorithm for logistic regression

Logistic Regression is a Machine Learning algorithm that is used for classification problems, it is a predictive analysis algorithm and based on the concept of probability.

Steps

Stage 1: Import python Libraries.
Stage 2: Investigate and Clean the Information.
Stage 3: Change the out Factors by making Faker Factors.
Stage 4: Split Preparing Information and Test Datasets.
Stage 5:Change the Mathematical Factors: Scaling.
Stage 6: Fit the Strategic Relapse Model.

```
from sklearn.linear_model import LogisticRegression
classifier = LogisticRegression()
classifier.fit(X_train, y_train)
pred = classifier.predict(X_train)
print(classifier.score(X_train,y_train))
```

Stage 7: Assess the Model.
Stage 8: Decipher the Outcomes

Algorithm for Random Forest

Novel Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

Steps

Step 1: Load the important libraries
Step 2: pick N irregular records from the informational index.
Step 3: Build a decision tree based on these N records.
Step 4:Pick the quantity in the Random variable algorithm and rehash steps 2 and 3.

```
# importing rfc classifier
from sklearn.ensemble import RandomForestClassifier
RFC = RandomForestClassifier()
RFC.fit(xtrain, ytrain)
ypred = RFC.predict(xtest)
```

Step 5: Fitting the Random forest model

```
n_outliers = len(fraud_)
n_errors = (ypred != ytest).sum()
print("Random Forest Classifier")
```

Step 6: Evaluating model's performance

Recall that the testing setup includes both hardware and software configuration choices. The laptop has an Intel Core i5 5th generation CPU with 8GB of RAM, an x86-based processor, a 64-bit operating system, and a hard drive. Currently, the software runs on Windows 10 and is programmed in Python. Once the program is finished, the accuracy value will appear. Procedure: Wi-Fi laptop connected. Chrome to Google Collaboratory search Write the code in Python. Run the code. To save the file, upload it to the disc, and create a folder for it. Log in using the ID from the message. Run the code to output the accuracy and graph.

Statistical Analysis

SPSS is a software tool used for statistics analysis. The proposed system utilized 10 iterations for each group with predicted accuracy noted and analyzed. Independent samples t-test was done to obtain significance between two groups.

3. Results

Table 1 shows the accuracy value of iteration of LR and novel RF. Table 2 represents the Group statistics results which depict LR with mean accuracy of 83.5% and standard deviation is 1.84. Novel RF has a mean accuracy of 94.89% and standard deviation is 1.82. Proposed novel RF algorithm provides better performance compared to the LR algorithm. Table 3 shows the independent samples T-test value for LR and novel RF with Mean difference as -11.39, std Error Difference as 0.82. Significance value is observed as 0.96 ($p > 0.05$).

Figure 1 shows the bar graph comparison of mean of accuracy on LR and novel RF algorithm. Mean accuracy of LR is 83.5% and novel RF is 94.89%.

4. Discussion

In this study, the iteration based on the previous historical datasets are considered for getting the accurate value of Credit Card Fraud Detection using logistic Regression Algorithm and novel Random Forest algorithm on the basis of anaconda navigator (Jupituer notebook) and SPSS. After getting many iterations, compare all the Selective algorithms and decide to get The accurate values of crop yield prediction.

The implementation of the machine became to learn about Credit Card Fraud Detection. Have a look at makes a speciality of the credit card datasets obtained from diverse portals belonging to some districts of Karnataka inside the country. Datasets ordered in a properly based totally way. The logistic Regression set of guidelines is used for the prediction model and its accuracy is received. The destiny is outstanding for the implementation of system studying algorithms inside the vicinity of Credit Card Fraud Detection and are hoping to put in force greater advanced algorithms in order that the system becomes more efficient. Hoping to make tool prediction extra sturdy and attain excessive accuracy with the assistance of greater datasets and advanced algorithms.

5. Conclusion

In this study, Credit Card Fraud Detection using a novel RF algorithm provides better accuracy than LR algorithm.

Declaration

Conflict of Interests

No conflict of interests in this manuscript

Authors Contribution

Author SS Chowdary was involved in data collection, data analysis, manuscript writing. Author J Chenni Kumaran was involved in conceptualization, data validation, 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 Sciences (Formerly known as Saveetha University) for providing the necessary Infrastructure to carry out this work successfully.

Funding:

Thanks for the following organizations for providing financial support that enabled us to complete the study.

1. Qbec Infosol
2. Saveetha School of Engineering.
3. Saveetha University
4. Saveetha Institute of Medical and Technical Sciences

6. References

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.

Buonaguidi, Bruno, Antonietta Mira, Herbert Bucheli, and Viton Vitanis. 2021. "Bayesian Quickest Detection of Credit Card Fraud." *Bayesian Analysis* -1 (-1). https://doi.org/10.1214/20-ba1254.

"Credit Card Fraud Detection System Based on Operational & Transaction Features Using SVM and Random Forest Classifiers." n.d. Accessed February 9, 2022. https://doi.org/10.1109/ICCAKM50778.2021.9357709.

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.

Gao, Jiaxin, Zirui Zhou, Jiangshan Ai, Bingxin Xia, and Stephen Coggeshall. 2019. "Predicting Credit Card Transaction Fraud Using Machine Learning Algorithms." *Journal of Intelligent Learning Systems and Applications* 11 (3): 33–63.

Hyder John, Sameena Naaz. 2019. "Credit Card Fraud Detection Using Local Outlier Factor and Isolation Forest" 7 (4): 1060–64.

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.

Mehbodniya, Abolfazl, Izhar Alam, Sagar Pande, Rahul Neware, Kantilal Pitambar Rane, Mohammad Shabaz, and Mangena Venu Madhavan. 2021. "Financial Fraud Detection in Healthcare Using Machine Learning and Deep Learning Techniques." *Security and Communication Networks* 2021 (September). https://doi.org/10.1155/2021/9293877.

Morine, Kevin J., Xiaoying Qiao, Vikram Paruchuri, Mark J. Aronovitz, Emily E. Mackey, Lyanne Buiten, Jonathan Levine, et al. 2017. "Reduced Activin Receptor-like Kinase 1 Activity Promotes Cardiac Fibrosis in Heart Failure." *Cardiovascular Pathology: The Official Journal of the Society for Cardiovascular Pathology* 31 (November): 26–33.

Mqadi, Nhlakanipho Michael, Nalindren Naicker, and Timothy Adeliyi. 2021. "Solving Misclassification of the Credit Card Imbalance Problem Using Near Miss." *Mathematical Problems in Engineering* 2021 (July). https://doi.org/10.1155/2021/7194728.

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.

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.

Peng, Sheng-Lung, Rong-Xia Hao, and Souvik Pal. 2021. *Proceedings of First International Conference on Mathematical Modeling and Computational Science: ICMMCS 2020*. Springer Nature.

salihfurkansaglam. 2021. "FraudDetection-Accuracy%99 Recall%99 Precision%99." Kaggle. December 24, 2021. https://kaggle.com/salihfurkansaglam/frauddetection-accuracy-99-recall-99-precision-99.

Venkata Suryanarayana, S., G. N. Balaji, and G. Venkateswara Rao. 2018. "Machine Learning Approaches for Credit Card Fraud Detection." *International Journal of Engineering & Technology* 7 (2): 917–20.

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.

V, Gokula Krishnan, and Krishnan V. Gokula. 2019. "Credit Card Fraud Detection Using Random Forest Algorithm." *International Journal for Research in Applied Science and Engineering Technology*. https://doi.org/10.22214/ijraset.2019.3215.

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.

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. Accuracy Values for LR and novel RF

S.NO	LR	RF
1	85.00	92.00
2	82.00	93.50
3	84.00	94.68
4	85.00	96.80
5	83.00	95.00
6	80.00	96.00
7	86.00	92.00
8	85.00	95.90
9	82.00	96.25

10	83.00	96.80

Table 2. Group Statistics Results-LR has an mean accuracy (83.5%), std.deviation (1.84), whereas for novel RF has mean accuracy (94.89%), std.deviation (1.82).

Group Statistics					
Accuracy	Groups	N	Mean	Std deviation	Std. Error Mean
	LR	10	83.50	1.840	.58214
	RF	10	94.89	1.828	.57822

Table 3. Independent Samples T-test -novel RF seems to be significantly better than LR (p=0.99)

Accuracy	**Independent Samples Test**								
	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**
Equal variances assumed	.003	.960	-13.885	18	.155	-11.39400	.82070	-13.12691	-9.66910
Equal variances not assumed			-13.885	17.999	.155	11.39400	.82070	-13.12692	-9.66919

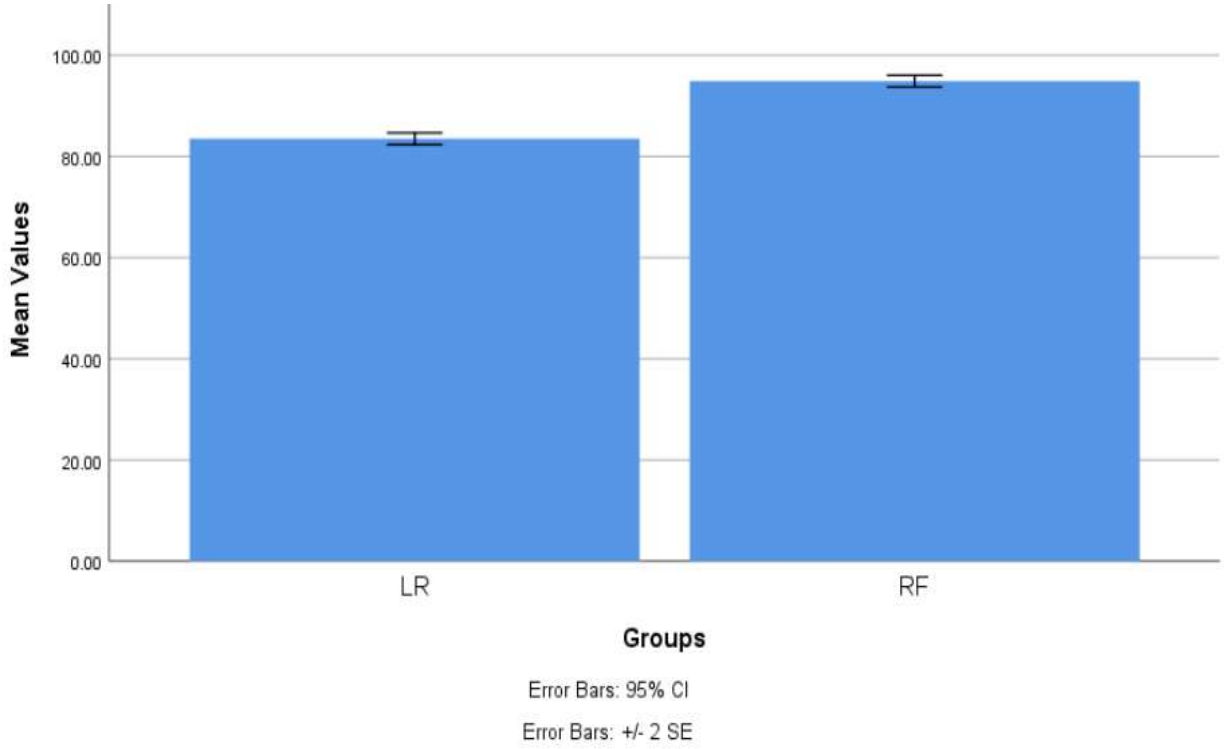

Fig. 1. Bar Graph Comparison on mean accuracy of LR (83.5%) and novel RF (94.89%). X-axis: LR, novel RF, Y-axis: Mean Accuracy with $\pm$2 SE.