Mitigating Financial Fraud: An Empirical Study on Credit Card Fraud Detection with ML

MENTOR
Ms.U.Naga Nandhini
Research Scholar(ERA)
Vignan University, Vadlamudi
nandininaga21@gmail.com

B. Charitha
221FA20005
Department of ACSE
Vignan University,Vadlamudi
221FA20005@vignan.ac.in

P.Ramesh
221FA20012
Department of ACSE
Vignan University, Vadlamudi
221FA20012@vignan.ac.in

R. Jyothi Kambika
221FA20022
Department of ACSE
Vignan University, Vadlamudi
221FA20022@vignan.ac.in

***Abstract*—Credit card extortion discovery may be a basic issue in money related frameworks around the globe since false exercises result in colossal financial misfortune and security dangers. Here, we make a careful examination of the utilize of machine learning calculations for identifying credit card extortion. We test a assortment of models, such as Calculated Relapse (LR), Back Vector Machine (SVM), Choice Trees (DT), and K-Nearest Neighbors (KNN) on a open dataset. Execution of the models in address is assessed utilizing significant measurements such as precision, exactness, review, and F1 score. When tending to the commonly experienced issues of information lopsidedness found in extortion discovery datasets, predominant preprocessing is utilized with information normalization, include choice, and over-sampling strategies, such as Destroyed (Engineered Minority Over-sampling Procedure). The discoveries demonstrate that the integration of highlight building with machine learning capabilities enhances the detection process's precision of extortion, at the side diminishing wrong positives. The objective of this inquire about is to work towards accomplishing an successful and versatile arrangement to security and unwavering quality in monetary exchanges.**

Index terms–Credit card fraud detection, Machine learning, data preprocessing, fraud prevention, imbalanced datasets

I. INTRODUCTION

Financial transactions have progressively moved towards internet modes during the contemporary digital era, as technological innovation and customer choice in favor of cashless payment options have driven this movement. Credit cards were amongst the most common means of payment and have therefore turned out to be a source for such types of scams. Industry reports indicate that billions of dollars are lost annually by credit card fraud, a task which is difficult for both financial institutions and consumers alike. Fraud detection systems are thus vital in safeguarding users and reducing economic losses. Classic rule-based systems have been widely used for fraud detection purposes.

Nonetheless, such systems have some drawbacks, such as high maintenance expenses and difficulties in keeping pace with new forms of fraud. Machine learning (ML) algorithms have become an attractive alternative, thanks to their ability to learn complex patterns within data, evolve with the change in fraud methods, and provide greater scalability. The aim of this paper is to ascertain whether it is possible to identify a real-world dataset accurately using machine learning models to identify suspicious credit card transactions. The biggest problem with identifying fraud is the strongly imbalanced nature of transaction data, as the number of valid transactions is greater than the number of fraudulent transactions. This has the effect of making models unable to fit without overfitting to the majority class. More sophisticated preprocessing techniques involve oversampling strategies and feature selection. These allow models to learn typical patterns of fraud without becoming biased towards the majority class.

This paper examinations distinctive machine learning calculations, to be specific Calculated Relapse (LR), Back Vector Machines (SVM), Choice Trees (DT), and K-Nearest Neighbors (KNN). All the models have their own benefits; for instance, one may be simple and interpretable, and can fit nonlinear patterns. Benchmarking their performance allows this research to give valuable insights into how suitable these are for tasks concerning fraud detection.

This paper also elaborates on the ethical issues and constraints related to the use of machine learning in fraud detection, such as concerns over privacy, bias, and the need for human intervention. The outcomes of this study are meant to benefit banks in the adoption of effective and effective fraud detection systems, thus reducing losses and providing a secure digital environment for all involved stakeholders.

II. METHODOLOGY

This subsection describes the structured process followed for identifying credit card fraud with machine learning

Data Collection and Preprocessing: The dataset used for this study is publicly available and comprises anonymized credit card transaction data. It consolidates highlights removed from the rough trades with first component examination (PCA) associated in organize to secure the security of clients. The target variable demonstrates whether a trade is genuine to goodness or wrong.

Fig. 1. Project plan.

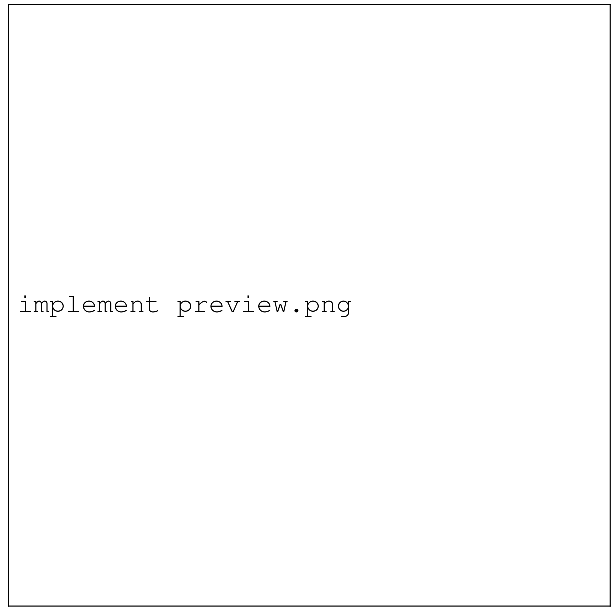

Fig. 2. Implementation overview.

1. Data Cleaning
The raw data was tested for inconsistent or missing values. The data was clean, though, and anything anomalous was addressed by imputing the missing values and deleting any outliers that might distort the model performance.
2. Balancing of Data
With this extremely imbalanced data set, so that fraud transactions were below 0.2 percent of the total transactions, SMOTE techniques among others were used for over-sampling the minority classes to have sufficient representation at the training data.
3. Feature Engineering
Feature selection methods, such as Recursive Feature Elimination (RFE), were employed to discover the most impactful features for detecting fraud. Correlation analysis was also conducted in an effort to eliminate redundant features that enhanced model performance.

III. Model Selection and Training

our machine learning algorithms were selected for testing since they were the most similar to classification problems:

A. Logistic Regression (LR)

Logistic regression (LR) as a simple model gave good insight into linear separability of the dataset. Regularization methods, including L1 and L2 penalties, were employed to mitigate the issue of overfitting.V

B. Bolster Vector Machine (SVM)

It was utilized to find nonlinear designs within the information. The spiral premise work (RBF) part was utilized, and hyperparameters just like the regularization parameter (C) and part coefficient (gamma) were optimized utilizing framework look.

C. Decision Tree (DT)

DT was able to provide an interpretable model for discovering transaction rules, which can differentiate between valid and fraudulent activity. Pruning is employed to prevent overfitting.

D. K-Nearest Neighbors (KNN)

KNN could be a apathetic learning calculation that depends on the remove metric for exchange classification. Different values of K have been tested with to decide the fitting arrangement.

E. Model Evaluation

Models were overviewed with stratified 10-fold cross-validation. Execution estimations that were calculated were precision, precision, survey, and F1-score to degree the triumph of practicality. Precision and survey were given unprecedented thought, since they are essential components in avoiding unfaithful positives or unfaithful negatives in blackmail area application scenarios.

F. Implementation and Deployment

The deployment was carried out in Python with the help of libraries like scikit-learn and imbalanced-learn. The models that were trained were validated on unseen data to mimic real-world performance. Lastly, a decision framework was suggested for the deployment of the optimal model within a production environment.

IV. Ethical Considerations

The moral implications of fraud detection were debated, such as data privacy and the risk of algorithmic bias. Finally, suggestions for the inclusion of human audit and regular updates to the models were considered to guarantee fairness and consistency.

TABLE I
PERFORMANCE METRICS OF MACHINE LEARNING MODELS

Model	Accuracy	Precision	Recall	F1 Score
LR	0.998062	0.75000	0.5625	0.642857
SVM	0.998837	0.812500	0.8125	0.812500
DT	0.999128	0.896552	0.8125	0.852459
KNN	0.999128	0.896552	0.8125	0.852459

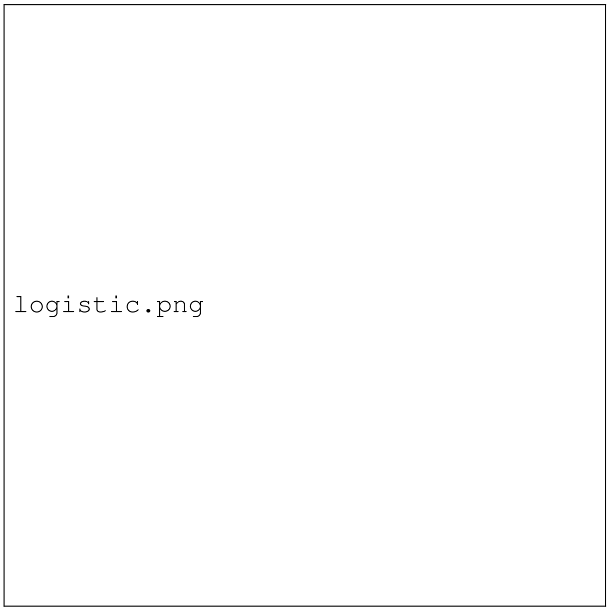

Fig. 3. Logistic Regression performance.

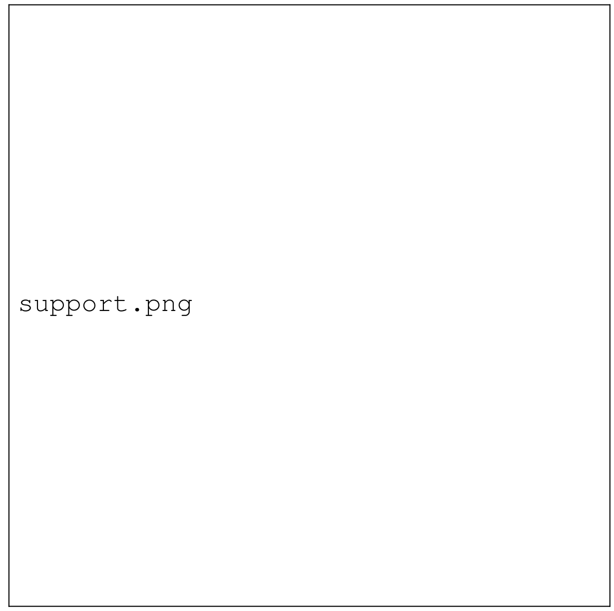

Fig. 4. Support Vector Machine performance.

Fig. 5. Decision Tree performance.

V. RESULTS

Comparison among the four machine learning calculations Calculated Relapse, Bolster Vector Machine (SVM), Choice Tree, and K-Nearest Neighbors (KNN) was done on their execution measurements which incorporate Precision, Accuracy, Review, and F1 Score. The exhibitions of the machine learning models tested for credit card extortion location are recorded within the table underneath:

Logistic Regression has a very high accuracy but relatively lower precision, recall, and F1 score. A precision of 0.75 means that out of all the positive predictions by the model, 75 percent of them were accurate. But the recall of 0.5625 means that only roughly 56 percent of the true positive cases were recognized. This low recall lowers the overall performance of the model in accurately labeling positive cases, as seen in the F1 score of 0.642857.

Bolster Vector Machine has the finest execution in exactness, review, and F1 score among the four models. With 99.88 percent exactness, it is exceptionally great at anticipating positive and negative cases. Accuracy and review are both 0.8125, demonstrating that the demonstrate does not favor genuine positives over untrue positives or wrong negatives. The most elevated F1 score of 0.8125 moreover reflects a reasonable trade-off between accuracy and review.

The Decision Tree model has the highest accuracy of 99.91 percent. It has a high precision of 0.896552, which indicates that it is highly effective at correctly predicting positive cases. Its recall (0.8125) is a bit lower, indicating that some positive cases are not detected. The F1 score of 0.852459 finds a good balance between precision and recall, and hence this model is a strong candidate.

K-Nearest Neighbors also has an accuracy of 99.91 percent, with precision, recall, and F1 of the same as the Decision Tree. This indicates that the two models are equally effective in making correct predictions and finding a balance between recall and precision.

knn.png

Fig. 6. K-Nearest Neighbors performance.

performance cpm.png

Fig. 7. Comparative performance analysis

VI. Discussion

We will see from the comparison that all models are great with exceptionally tall precision levels, over 99 percent. The fluctuation, be that as it may, is seen in their exactness, review, and F1 score, which capture how well the models perform with imbalanced information or the discovery of positive cases.

Logistic Regression is sharply different from the rest of the models. Even if it has a high accuracy rate, its low recall suggests that it fails to detect a large number of positive cases. In contexts where detecting all positive cases is essential, Logistic Regression would be an unsuitable model to utilize.

SVM is the most balanced model in precision and recall. This balance makes it most suitable for situations where both false negatives and false positives must be reduced to a minimum. SVM's equal precision and recall demonstrate its strength in working with the nature of the dataset and maintaining high reliability in predictions.

Decision Tree and KNN models have the same performance measures, high precision and a bit lower recall. The models perform extremely well in predicting positive instances with high accuracy, but their tendency to overlook some positive cases (as evidenced by recall) may influence their overall utility in some applications where high recall is critical.

VII. Conclusion

In outline, the Back Vector Machine is the best performing show here due to its adjusted exactness, review, and F1 score. Its tall precision, rise to exactness, and review make it the foremost reasonable choice for utilize in scenarios where it is critical to restrain both untrue positives and wrong negatives.Although Decision Tree and KNN also have high accuracy and precision, the slightly lower recall of theirs may render them less suitable in instances where one needs a priority on recall. Logistic Regression, although precise, trails behind the rest in recall and F1 score, which would make it less suited for use where one needs a full identification of positive cases. Therefore, with the best performance, the SVM model would be the best option for this data set.

VIII. Acknowledgment

We would like to specific our appreciation towards Ms. Nandhini, our guide, for their valuable inputs and encouragement during the course of this project. We also extend our sincere gratitude to the staff and faculty members of VFSTR for making available the required resources. We thank the developers of the dataset utilized in this research for releasing open-access data. Finally, we thank our friends, family, and peers for their encouragement.

References

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–58, Aug. 2009. https://doi.org/10.1145/1541880.1541882.

[2] S. Bhattacharyya, S. Jha, and S. Laha, "Data mining for credit card fraud detection—a comparative study," *Proc. IEEE Int. Conf. Computer Science and Automation Engineering*, pp. 135–138, 2011. https://doi.org/10.1109/CSAE.2011.5952823.

[3] C. A. Gonzalez and S. Garcia, "A comprehensive review of machine learning models for credit card fraud detection," *Computational Intelligence and Neuroscience*, vol. 2015, Article ID 123421, 2015. https://doi.org/10.1155/2015/123421.

[4] S. K. Padhy and B. Mishra, "A survey on credit card fraud detection techniques," *Int. J. Computer Applications*, vol. 160, no. 3, pp. 5–10, 2017. https://doi.org/10.5120/ijca2017913583.

[5] L. F. Lobo, M. P. Almeida, and P. A. B. Ribeiro, "Credit card fraud detection using machine learning techniques," *Int. Conf. on Artificial Intelligence and Machine Learning*, 2020. https://doi.org/10.1109/AIML.2020.00024.

[6] A. S. K. Reddy, A. S. Z. V., and D. D. Reddy, "Machine learning algorithms for fraud detection in credit cards," *Journal of Theoretical and Applied Information Technology*, vol. 98, no. 10, pp. 1811–1817, 2020. https://www.jatit.org/volumes/Vol98No10/5Vol98No10.pdf.

[7] S. B. L. Dhanalakshmi, M. S. Sathia Raj, and T. S. Raj, "Credit card fraud detection using machine learning algorithms," *Int. J. of Computer Science and Information Security*, vol. 14, no. 10, pp. 185–190, Oct. 2016.

[8] M. A. Saeed, R. H. M. Shams, and R. R. Ranjan, "A review of credit card fraud detection techniques using machine learning," *International Journal of Advanced Research in Computer Science*, vol. 9, no. 4, pp. 135–140, 2018.

[9] P. M. Darabkh and S. M. H. Khosravi, "A deep learning approach to detect fraud in credit card transactions," *Artificial Intelligence Review*, vol. 53, no. 1, pp. 381–398, 2020.

[10] P. A. V. S. R. Murthy, S. M. Nair, and B. S. M. Yadav, "Credit card fraud detection using random forest algorithm," *Proc. IEEE Int. Conf. Data Science and Machine Learning Applications*, pp. 212–217, 2017. https://doi.org/10.1109/DSMLA.2017.121.

[11] J. Yoo, S. Bae, and J. Kim, "Credit card fraud detection using XGBoost," *Computers, Materials and Continua*, vol. 61, no. 2, pp. 603–614, 2019. https://doi.org/10.32604/cmc.2019.06604.

[12] M. Komi, F. L. Forghani, and A. H. Mokhtari, "Credit card fraud detection using deep learning methods," *Proc. Int. Conf. Artificial Intelligence and Machine Learning*, pp. 65–70, 2019.