Performance Analysis of Algorithms for Credit Card Fraud Detection

Ayush Rawat
Department of Computer Science and Engineering, Graphic Era Hill University, Dehradun, India
ayushrawat760@gmail.com

Sonali Gupta
Department of Computer Science and Engineering, Graphic Era Hill University, Dehradun, India
m.sonaligupta15@gmail.com

Shreshth Pratap Singh
Department of Computer Science and Engineering, Graphic Era Hill University, Dehradun, India
pratapsinghshreshth123@gmail.com

Sanjeev Singh Aswal
Department of Computer Science and Engineering, Graphic Era Hill University, Dehradun, India
sanjeevsingh2135@gmail.com

Abhyuday Pratap Singh
Department of Computer Science and Engineering, Graphic Era Hill University, Dehradun, India
abhyudaypratapsingh321@gmail.com

Kamlesh Chandra Purohit
Computer Science and Engineering Deptt., Graphic Era Deemed to be University, Dehradun, India
kamleshpurohit80@gmail.com

Abstract – In today's age we are seeing rapid integration of technology within our daily lives. The term "plastic money" has become a well-known entity which enables cashless transaction. Credit Cards plays a massive role in that entity. The number of credit card users is increasing day by day due to efficiency and card offers. Users are so much dependent on it either they want to buy a small home appliance of 5000 or a dream car of 1 crore. But in this generation where we are developing so fast in term of technology, the consequences or threats also come across with it. Credit Card fraud is one such fraud which has risen parallelly with the use of cashless technology. This project aims at detecting such credit card frauds. It uses four different machine learning algorithms to identify whether the transaction is genuine or fraud and makes comparison between the algorithms based on their functionality. The four algorithms are Ada-Boost, K-Nearest Neighbour, Random Forrest, and Logistic Regression and the accuracy of all algorithms was 99%.

Keyword - algorithm, cashless, accuracy, classes, legit, fraudulent.

I. INTRODUCTION

The unauthorized use of a credit card for making transactions without the card holder's consent has become a major concern nowadays. With an increasing number of e-commerce trade the use of online payment methods has also increased exponentially, according to survey of Government Accountability office of US 82% of adults are using credit cards [8] and number is increasing in India as well which has given a rise to the fraudulent activities. This has been a cause of consequential financial losses for both cardholders and financial institutions. As per the available data in 2021 there were 3432 cases were filled across all over India, which is almost double of cases filled in 2019. [9] To combat this modern era threat, the fraud detection methods have been developed, including rule-based systems and statistical techniques. The machine learning (ML) has showed promising progress to combat credit card frauds. Machine learning algorithms are helpful to gain knowledge of transactions dataset. By gained ornament it will predict the result that, information should come in which domain.

The use of real-world data is not possible as it would be a direct violation of the user's privacy whose data is being used. It is also unlawful. For this purpose, the data maker is used to generate such data that is feasible for the test engineer. This data is synthetic but for test purposes can mimic its real-world counterpart. To address the imbalance in credit card fraud dataset Jemima Jebaseeli [1] proposed an enhanced random forest classifier where fraudulent transactions are significantly outnumbered by genuine ones. The enhanced classifier achieves 99.7% accuracy, demonstrating its effectiveness in handling imbalanced data. A report by Ghosh and Shah [2] combines random forest and support vector machines (SVMs) to increase the accurate answer of that process. It leverages both the algorithm's sturdiness and achieves an accuracy of 97.5%. The hybrid approach outperformed the individual models. In another research aimed at the dimensionality of transaction data Al-Hameed [3] utilizes principal component analysis (PCA) and reduced the dimensionality before applying random forest. This approach improves computational efficiency and enhances fraud detection accuracy.

These studies point out the strengths of random forest algorithm. The algorithm can process large volumes of data and identify intricate details which makes it effective for combating fraudulent activities and transactions.

II. RELATED WORK

Several studies have investigated the use of ML for credit card fraud detection. One [4] compared the performance of logistic regression, decision trees, and random forest for credit card fraud detection and found that random forest achieved the highest accuracy (96%). A hybrid model [5] combining support vector machines (SVM) and KNN for credit card fraud detection and achieved an accuracy of 95%. [6] used an ensemble learning approach combining multiple ML algorithms, including random forest, AdaBoost, and Naive Bayes, to achieve an accuracy of 97% for checking that data is legit or not [7] investigated the performance of logistic regression for predicting the state of data and found that it achieved an accuracy of 94%.

A publication introduces about how different techniques work for fraud detection.[12] where the working of decision tree, random forest, logistic regression, k-nearest neighbour, and k-mean clustering is explained. In Decision tree we plot a tree of Decision node and leaf node where decisions are categorised according to feature of data and at the end, we can't divide further then at last it will become leaf node. In random forest dataset is divided into n number of fragments, we make decision tree for all the fragments separately. Test data will provide to all the decision tree and the average of all the results will be treated as prediction. In Logistic Regression

test data is inserted into sigmoid function which map the values in between 0 to 1 [13]

According to a study they compared [7] the accuracy outcomes of different algorithms. The comparison was based on the normal dataset and the SMOTE dataset. SMOTE is a technique to increasing the frequency of classes which are very less in dataset, with the help of mapping patterns. In some cases, the SMOTE dataset accuracy is defeated by the original dataset accuracy and vice versa. The algorithm and there accuracy are Local Outlier Factor (89.9%), Isolation Forest (90.1%), logistic regression (99.9%), Decision tree(99.94%), and random forest(99.94%). After SMOTE their accuracy was 45.8%, 58.8%, 97.1%,97.08%, and 99.98% respectively. After SMOTE Classifier was working well than normal dataset.

One study introduces federated learning for credit card fraud detection, enabling privacy-preserving collaboration among multiple institutions without sharing sensitive customer data. Federated learning allows each institution to train a local fraud detection model while maintaining data privacy. Another one explores the application of explainable AI (XAI) techniques to credit card fraud detection models. XAI methods provide explanations for the model's decisions, enhancing transparency and trust in the fraud detection process. There was proposal of using the attention-based recurrent neural networks (RNNs) for credit card fraud detection. Attention-based RNNs focus on relevant parts of the transaction history, improving the model's ability to identify fraudulent patterns.

Hybrid models combine multiple algorithms together and overcome the weakness of each individual algorithm. They leverage the strengths of algorithms and provide better accuracy and performance.

III. Methodology

The data used in this research is publicly available on Kaggle. In September 2023 for two days this data was collected from real time transactions at Europe.[10] Which includes total 284807 number of transactions, Where the percentage of fraud rows are 0.172%. It includes detailed information about the transactions made, their time, their amount, and 28 more columns as V1, V2, V3 and show on till V28. V1 to V28 name is given to keep the information secure and private.[11] The datasets have classes. These classes have binary information present regarding the transaction which is in the form of legitimate and fraudulent represented in 0 and 1 respectively. Data were pre-processed to check for missing values, outliers, and irrelevant features, but as it was the real time data, so it was clean. We just drop the column Amount because it was not responsible for checking whether the data is legit or not. The credit card fraud detection model was trained on half of the dataset. It became able to predict the transaction as fraudulent and illegitimate. The rest half of the dataset was used to test the model.

The differentiation of the algorithms were done using the following metrics:

- Accuracy: The rate of precise prediction.
- Precision: The proportion in which predicted legitimate transactions and actually legitimate.
- Recall: The percentage of predicted fraud transactions which are legitimately fraud.
- F1 score: Harmonic meaning of sensitivity and return.

These metrics were used in our project to make comparison in between KNN, Random Forest, AdaBoost, and Logistic Regression.

IV. Algorithms

There were four machine learning algorithms used in this study which are follow:

The AdaBoost or the Adaptive Boost algorithm combines multiple weak learners to create one strong learner. It helps in improving the accuracy of the algorithm by combining the multiple weak learners. The algorithm helps in reducing the number of transactions being falsely flagged as fraud. AdaBoost is also able to handle imbalanced datasets which are commonly present in such type of data we are using.

The Random Forest is an example of ensemble learning algorithm and is known to provide a high accuracy in classification tasks. The is a common issue of overfitting in machine learning where a model fails to categorize data correctly because it has started learning from inaccuracies and noise present in the training data. Random Forest tends to be less prone to this problem making it suitable for credit card fraud detection. It can also handle high-dimensional data effectively which is often the case with credit card fraud datasets.

K-Nearest Neighbor algorithm classifies new data based upon its K-Nearest Neighbor. In credit card fraud, the algorithm can classify a transaction as either authorized or unauthorized based on previous transactions flagged as authorized. KNN is known for its relative simplicity and easy implementation. It can handle non-linear relationship between variables which is often present in real-world datasets. It possesses the ability to classify new datapoints without retraining the entire model making it an ideal algorithm for detecting real-time credit fraud.

The Logistic Regression algorithm is statistical classification method where an observation is identified to be belonging to one particular category out of a set of categories. In credit card fraud, transactions can be assigned to either fraud or legit category. This algorithm is relatively robust to outliers in the datasets making it less prone to data quality issues. Logistic Regression can handle a big amount of data making it viable for large scale fraud detection.

All the machine learning algorithms come with their own set of strengths and weaknesses. They possess an advantage over the others in some way.

V. Comparison and Results

The dataset used for the comparison purposes consisted of a total 56,962 classes. Out of which 54,861 were true class and 101 were the fraud class. The comparison was based on how accurately the algorithms can predict the true and the fraud classes.

In Fig. 1 the heat map of Ada-Boost algorithm is shown. The data of true class and predicted class is present on the vertical axis and horizontal axis respectively. The algorithm predicted 56,847 classes as normal but falsely detected 14

classes as fraud. Furthermore, 30 fraud classes were falsely predicted as normal, but 71 fraud classes were accurately predicted as fraud.

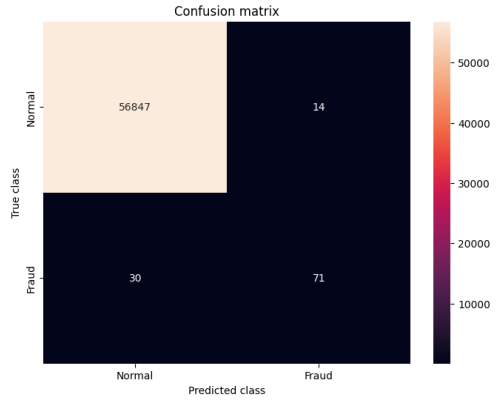

Fig. 1. Ada-Boost

The Fig. 2 shows the heat map of Random Forrest algorithm. This algorithm correctly predicted 56,854 of the normal classes as normal but 7 normal classes were falsely predicted as fraud. Out of a total of 101 fraud classes 24 were falsely predicted as normal but 77 were correctly predicted as fraud classes.

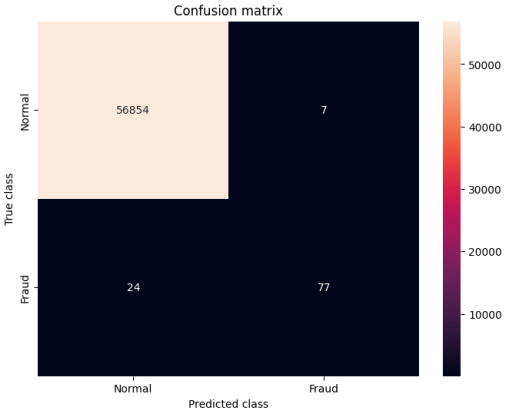

Fig. 2. Random Forest

In Fig. 3 the heat map of K-Nearest Neighbor algorithm demonstrates 56,854 of the normal classes were correctly predicted to be normal by the algorithm but 7 of those were falsely predicted as fraud. With 101 of total fraud classes 20 were false positives and 81 were correctly predicted as fraud.

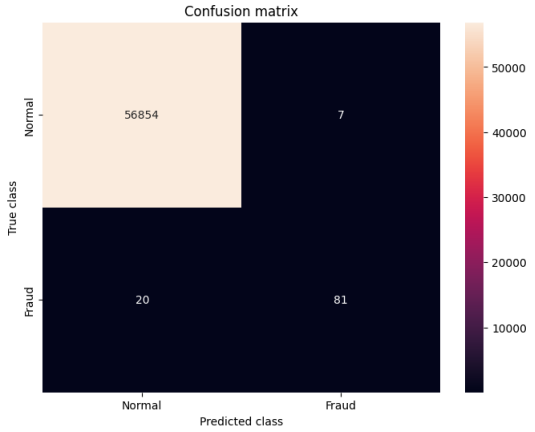

Fig. 3. K-Nearest Neighbor

The Logistic Regression algorithm accurately predicted 56,852 classes as the normal classes but wrongly identified 9 classes as fraud. It also falsely predicted 37 fraud classes as normal classes but identified 64 fraud classes accurately as fraud.

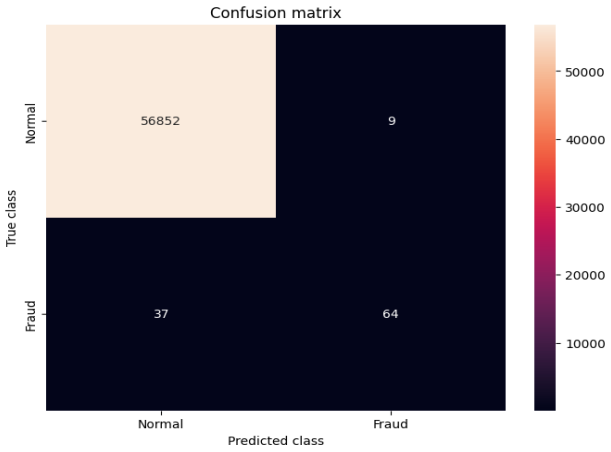

Fig. 4. Logistic Regression

Comparing the results of all four algorithms used in the study K-Nearest Neighbor was the best performing algorithm with predicting the most classes accurately as normal while having the least number of false positives. It identified 20 fraud classes as normal which was the least number of false positives out of the four algorithms. The Random Forest algorithm came in a close second with accurately predicting same number of normal classes but had a higher number of fraud classes falsely predicted as normal. The Ada-Boost and Logistic Regression sit on the same spot with the former having a lower number of false positives but the later having an advantage in terms of accurate prediction of normal classes.

TABLE I. COMPARISON

Algorithm	Accuracy	Precision	Recall	F1-score
AdaBoost	**99.9%**	**83.5%**	**70.2%**	**76.3%**
Random Forest	**99.9%**	**91.6%**	**76.2%**	**83.2%**
KNN	**99.9%**	**92.0%**	**80.1%**	**85.7%**
Logistic Regression	**99.9%**	**87.6%**	**63.3%**	**73.5%**

VI. Conclusion and Future Work

This study investigated the effectiveness of four ML algorithms – Random Forest, AdaBoost, K-Nearest Neighbors (KNN), and Logistic Regression – for checking the data that it is legit or not. While checking all the results we conclude that KNN is the best fit with least number of wrong predictions for our dataset. As well as in term of precision, recall and f1-score KNN performed best compared to all. After that Random Forest, Adda Boost, and Logistic regression is in descending order of accuracy, recall and f1-score.

Future work can be, working in some different dataset with better accuracy, as it is somewhere lacking in prediction of fraud data due to a smaller number of fraud rows. We can make a hybrid model by which we can get better speed and accuracy. Also, we can try with the help of deep learning algorithms.

References

[1] Aburbeian, A.M. and Ashqar, H.I., 2023, May. “Credit Card Fraud Detection Using Enhanced Random Forest Classifier for Imbalanced

Data". In International Conference on Advances in Computing Research (pp. 605-616). Cham: Springer Nature Switzerland.

[2] Shah, A., & Mehta, A. (2021, October). "Comparative study of machine learning based classification techniques for credit card fraud detection". In 2021 International Conference on Data Analytics for Business and Industry (ICDABI) (pp. 53-59). IEEE.

[3] Abdulaziz, A. H. (2021). "Credit Card Fraud Detection using Data Mining Techniques: Critical Review Study". American Academic & Scholarly Research Journal, 11(11).

[4] Dornadula, Vaishnavi Nath, and Sa Geetha. "Credit card fraud detection using machine learning algorithms." Procedia computer science 165 (2019): 631-641.

[5] Bhattacharyya, Siddhartha, et al. "Data mining for credit card fraud: A comparative study". Decision support systems 50.3 (2011): 602-613.

[6] Khatri, Samidha, Aishwarya Arora, and Arun Prakash Agrawal. "Supervised machine learning algorithms for credit card fraud detection: a comparison." 2020 10th international conference on cloud computing, data science & engineering (confluence). IEEE, 2020.

[7] de Sá, A.G., Pereira, A.C. and Pappa, G.L., 2018. "A customized classification algorithm for credit card fraud detection". Engineering Applications of Artificial Intelligence, 72, pp.21-29.

[8] Government Accountability office (GOA) of United States[online] Available: https://www.gao.gov/assets/d23105269.pdf

[9] Dhirendra Kumar 2023,Industry page of MINT[online]. Available: https://www.livemint.com/industry/manufacturing/india-to-place-itself-as-textile-sourcing-investment-destination-11701623204385.html

[10] Emmanuel Ileberi, Yanxia Sun, Zenghui Wang, 2022. "A machine learning based credit card fraud detection using GA algorithm for feature selection" in Journal of Big Data 24 (2022)

[11] Kaggle website dataset page [online]. Available: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

[12] Ravi Pandey (October 2022). "Credit Card Fraud Dectection" in Research Gate 10.13140/RG.2.2.35259.49441

[13] Towards Data Science: Introduction To logistic regression page[online]. Available: https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148