Navigating the Labyrinth of Credit Card Fraud: A Novel Perspective with Decision Tree Algorithms  
Albert, Muhammad Raffi Hakim, Hanif Muhamamd Sangga Buana, Jordan Rubin Kurnia

**Abstract** The Importance of Credit Card Transactions has influenced the economy of the worldwide for the better future, yet saifd systems is still riddled with fraudulent interactions which harms the security and safety of the users who wanted to utilize said benefits. We have designed a system that is capable of detecting any accounts linked to fraudulent attempts within credit card transactions. we demonstrate the power of AdaBoost and RandomForest, two robust decision tree-based algorithms, in navigating this intricate labyrinth. Our research showcases that these algorithms, when applied to the detection of fraudulent transactions, achieve a remarkable accuracy of over 95%. This paper underscores the effectiveness of AdaBoost and RandomForest in handling high-dimensional and imbalanced data, common in credit card fraud detection. Our findings suggest that the integration of these algorithms can significantly enhance the accuracy and reliability of credit card fraud detection systems, thereby illuminating the path towards more secure financial transactions.

*Index Terms*—AdaBoost, RandomForest, Decision Tree, Credit Card, Classification, Fraud, Data Mining

# INTRODUCTION[[1]](#footnote-2)

I

. n today’s fast-paced digital society, credit cards have become an indispensable tool for financial transactions. They offer convenience, enable cashless transactions, and are widely accepted both domestically and internationally. The advent of online shopping and digital services has further amplified their importance, making them a cornerstone of modern economic activity. However, this widespread use of credit cards has also made them a prime target for fraudulent activities. Fraudsters employ sophisticated techniques to deceive cardholders and financial institutions, leading to unauthorized transactions. These fraudulent activities not only result in significant financial losses but also breach the trust of consumers and undermine the integrity of financial systems.

The risk associated with credit card transactions is further exacerbated by the challenges in detecting and preventing fraud. Factors such as the enormous volume of transactions, the adaptive nature of fraud techniques, and the imbalanced nature of fraud occurrences make detection a complex task. Despite this, there`s opportunity to establish a well-defined system to not only determine & detect fraudulent actions, but also remember identity associated with fraudulent activity to prevent further actions by bad faith actors and streamline the processes.

One of the ways of detecting fraudulent actions is by utilizing Decision Tree algorithms well-establishedby communities. Several famous ones such as RandomForest & AdaBooster have been applicated in various real-life scenarios and have helped improve the quality & performance of the system. We expect said technology also influences the fraudulent detection in positive & improving ways.

RandomForest[1] is a decision tree algorithm Random Forest, introduced by Breiman in 2001, is a popular machine learning method that uses an ensemble of decision trees for prediction. It offers higher accuracy than a single decision tree, especially for complex datasets with large dimensions and intricate variable interactions. Each tree in the Random Forest is constructed using randomly selected training datasets and predictor variables. The final prediction is an aggregation of the results from each tree. A major benefit of using random forest for prediction modeling is the ability to handle datasets with a large number of predictor variables

Another algorithm will be tested is AdaBooster[2] which is a versatile algorithm capable of various attachment with other types of algorithms. However when paired with strong ones such as Decision Tree AdaBoost can be proven strong as it was fast, efficient and works well with high-bias/low variance datasets.

This paper focuses on overview of RandomForest & Adaboost algorithms, it`s applications in credit card fraud detection and examining the performance between two algorithms.

# Related Works

A study conducted by Tiwari etc. [3] tested Decision Tree Algorithm with various other algorithms with different discipline. Decision trees are popular due to their ease of use and flexibility in handling different data types. However RandomForest are used to address the instability and sensitivity of single trees, random forests are ensembles of independent decision trees, offering better computational efficiency. They use two sources of randomness: bootstrapped samples and a random subset of data attributes for each tree. This results in variance among trees, making the model easy to use and effective.

Another Study by Bagga & Others [4] also iterated & explored a bit on usage of RandomForest & AdaBoost as a way of not only improve accuracy but also cover the weaknesses of various Decision-Tree related algorithm by combining multiple poorly performing classifiers to build up a strong classifier which gives high accuracy. Weights set of each classifier and training the sample data in each iteration so as to ensure accurate predictions of unusual behavior.

Also important to be noticed is works by Cherif etc. [5] somewhat discusses the Isolation Forest algorithm, an anomaly detection method developed by Fei Tony Liu. This algorithm determines the isolation of data, i.e., how distant a given data point is from the remaining data points. It isolates anomalies using binary trees, providing a faster anomaly detector that identifies anomalies directly without profiling all the regular instances.

# METHODOLOGY

To Fully utilize the concept of Credit Card Fraudulent Detection, one must establish various algorithms for the model. It`s also very important to process the data and value inside for data mining purposes.

## Data Selection & Preprocessing

Datasets that will be used will be “Synthetic Financial Datasets For Fraud Detection” by Edgar Lopez-Rojaz that can be collected in Kaggle. The dataset contains ~6.36m variables with 11 attrbiutes for each label. The main attribute of said dataset is categorial as fraud/not fraud. However, there are two problems that must be resolved before the datasets can be used to the fullest.

1. The datasets are simply huge, with each process that can take extremely long times.
2. Also said datasets were simply lopsided in terms of fraud classification. It risks overfitting issues.

As such necessary steps must be taken to resolve both issues

1. Problems regarding the huge datasets can be resolved by shrinking the datasets into manageable size. Risks associated with this approach is that we might misses out on variables with important value.
2. Overfitting can be resolved with Bootstrapping method, a resampling technique used to estimate the uncertainty or variability of a statistic or model parameter. It involves creating multiple random samples, called bootstrap samples, by sampling with replacement from the original dataset.

The final result of Datasets is a very manageable and balanced datasets with size of ~16.5k variables with balanced size of fraudulent & not-fraudlent sizes.

## Data Transformation

Random Forest is a widely-used machine learning algorithm developed by Leo Breiman and Adele Cutler, which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems. Most common steps of RandomForest are displayed as such:

* Step 1: In the Randomforest model, a subset of data points and a subset of features is selected for constructing each decision tree. Simply put, n random records and m features are taken from the data set having k number of records.
* Step 2: Individual decision trees are constructed for each sample.
* Step 3: Each decision tree will generate an output.
* Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression, respectively.

|  |
| --- |
|  |
| https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/ |

The second algorithm to be considered is AdaBoost Classifier with Decision Tree as Strong foundation for algorithm. AdaBoost algorithm, short for Adaptive Boosting, is a Boosting technique used as an Ensemble Method in Machine Learning. It is called Adaptive Boosting as the weights are re-assigned to each instance, with higher weights assigned to incorrectly classified instances.

|  |
| --- |
| Step 1: Calculate sample Weight |
|  |
| Step 2: Sample Classification based ona vaialble attributes |
| Step3: Calculate the Influence towards the sample attribute category |
|  |
| Step 4: Calculate Total Error & Perfomance |
|  |
| Step 5: Decrease Error by recalibrating the error/deviation |
| Step 6: Reclassify each value of dataset based on given classification |
| Step 7+: Reiterate all previous steps by recalculating sample weights and normalizing the value. How many iterations depends on how accurate we want. |

Integration of the code will use python 3.12.0 versions with several libraries installed and used for this experiment purposes:

* streamlit: Streamlit is an open-source Python library that allows you to create custom web apps for machine learning and data science.
* pandas: Pandas is a powerful Python library for data analysis and manipulation.
* Randomforest & adabooster: respective algorithms used for algorithm purposes
* StandardScaler: Standardize features by removing the mean and scaling to unit variance to allows machine for an easier computation whilst maintaining the original value of the datasets
* Train\_test\_split: Splits the datasets for testing & validation purposes.

All those libraries will be essential into establishing and activating necessary algorithm for testing & applications purposes.

# DATASET & TESTING METHODS

Datasets that will be used will be “Synthetic Financial Datasets For Fraud Detection” by Edgar Lopez-Rojaz that can be collected in Kaggle. The dataset contains ~6.36m variables with 11 attrbiutes for each label.

|  |
| --- |
|  |
|  |

From the dataset we can determine the correlation matrix for each label.

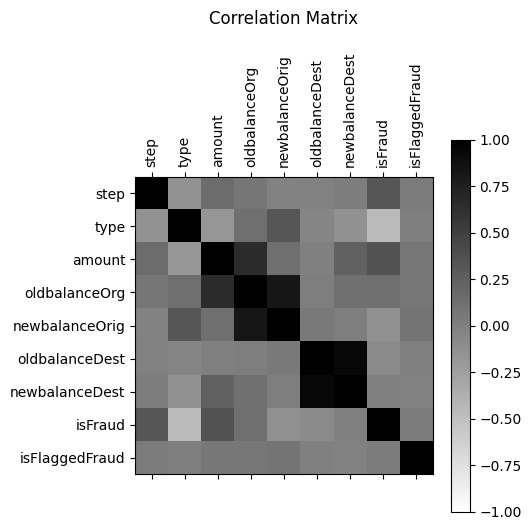


Figure 1. Correlation Matrix

Users must input various Data into machine before they can determine whenever said interaction constitutes as fraudulent action.

* step - maps a unit of time in the real world. In this case 1 step is 1 hour of time. Total steps 744 (30 days simulation). []
* type - CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER.
* amount - amount of the transaction in local currency.
* nameOrig - customer who started the transaction
* oldbalanceOrg - initial balance before the transaction
* newbalanceOrig - new balance after the transaction.
* nameDest - customer who is the recipient of the transaction
* oldbalanceDest - initial balance recipient before the transaction. Note that there is not information for customers that start with M (Merchants).
* newbalanceDest - new balance recipient after the transaction. Note that there is not information for customers that start with M (Merchants).
* isFraud – Whenever said transation is fraudulent action.
* isFlaggedFraud – Whenever said transaction is considered as fraudulent.

|  |
| --- |
|  |

After inputting various necessary value for classification, user can choose which kind of algorithms that wanted to be used.

|  |
| --- |
|  |

After confirming the value & algorithms to be used, system first needs to check whevener inserted ID, both origin & destination were involved in fraudulent processes, if both were found to have fraudulent record the system immediately consider both as fraudulent and skips the process. Otherwise, the system will apply the trained model into assigned value to determine whenever the system were considered as fraudulent actions. Users will be notified whenever this transaction is considered as fraudulent and stores implicated IDs of both Origin & Destination for further classifications.

# Result, Disucssion, & Implications

After we are testing the method, now we can determinate the confusion matrix for each Algorithm.

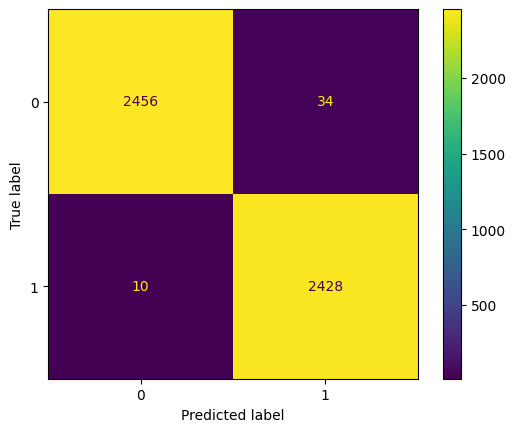


Figure 2. RandomForest Confusion matrix

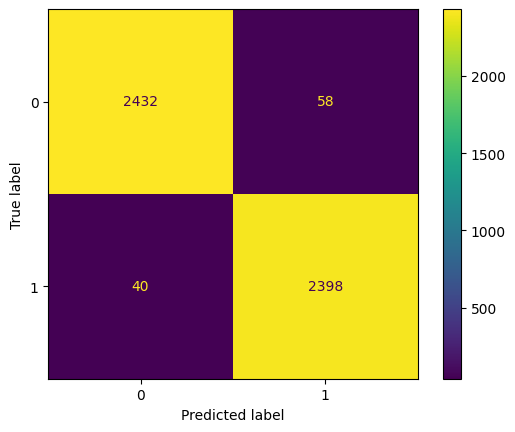


Figure 3. AdaBooster Confusion matrix

After we determinate the confusion matrix, we can calculate the Accuracy, Recall, and Precision for each algorithm.

|  |  |  |
| --- | --- | --- |
| Algorithm | RandomForest | AdaBooster |
| Accuracy | 99% | 98% |
| Recall | 99% | 98% |
| Precision | 99% | 98% |

Based on this test, we have determined several findings, regarding about how

1. RandomForest & AdaBoosterbarely closes in eachother, however RandomForest is slightly better based than AdaBooster on the tests.
2. RandomForest Capability of handling overfitting or large data definitely helps with processing and classification
3. AdaBoost capability for learning from previous adjustment helps improve the calculation & classification and less prone to overfitting compared to other weak learners

The Further Implication of this founding is that it`s generally possible to create a relatively simple system that can detect and recognize and also classify said transaction with reported value as fraudulent actions. Said system also demonstrates the capability & flexibility if Decision Tree algorithm as algorithm easy to apply and be taken advantage of in various scenarios.

# Conclusion

Our study underscores the critical importance of addressing fraudulent interactions within credit card transactions to ensure the security and safety of users, thereby fortifying the worldwide economy. Through the implementation of AdaBoost and RandomForest, two powerful decision tree-based algorithms, we have demonstrated their efficacy in navigating the complex landscape of credit card fraud detection.

Furthermore, our exploration into the capabilities of AdaBoost highlights its strength in learning from previous adjustments, enhancing both calculation and classification. Notably, AdaBoost proves to be less prone to overfitting compared to other weak learners, contributing to its reliability in the detection of fraudulent activities.

The practical implication of our findings suggests the feasibility of developing a straightforward yet effective system for detecting, recognizing, and classifying fraudulent transactions. The flexibility and ease of application of decision tree algorithms, exemplified by RandomForest and AdaBoost, position them as valuable tools in enhancing the accuracy and reliability of credit card fraud detection systems.

In summary, our research not only contributes to the advancement of credit card fraud detection methodologies but also underscores the potential of decision tree algorithms in creating robust and adaptable systems for secure financial transactions. These insights pave the way for the development of more sophisticated and reliable fraud detection mechanisms, ensuring a safer and more secure future for credit card transactions globally.

References

1. G. Jaime Lynn Speiser, Michael E. Miller, Janet Tooze, Edward Ip, A comparison of random forest variable selection methods for classification prediction modeling, Expert Systems with Applications, Volume 134, 2019, Pages 93-101, ISSN 0957-4174, <https://doi.org/10.1016/j.eswa.2019.05.028>.
2. Sarah Azmi, S., & Baliga, S. (2020). An Overview of Boosting Decision Tree Algorithms utilizing AdaBoost and XGBoost Boosting strategies. International Research Journal of Engineering and Technology, 7(5), 6867–6870. <https://www.semanticscholar.org/paper/An-Overview-of-Boosting-Decision-Tree-Algorithms-Azmi-Baliga/3e454a7286fd1ab0c4d5b2d9a8d8e819928be6f4>
3. Tiwari, Pooja & Mehta, Simran & Sakhuja, Nishtha & Kumar, Jitendra & Singh, Ashutosh. (2021). Credit Card Fraud Detection using Machine Learning: A Study.
4. Siddhant Bagga, Anish Goyal, Namita Gupta, Arvind Goyal, Credit Card Fraud Detection using Pipeling and Ensemble Learning, Procedia Computer Science, Volume 173, 2020, Pages 104-112, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2020.06.014>.
5. Cherif, A., Badhib, A., Ammar, H., Alshehri, S., Kalkatawi, M., & Imine, A. (2023). Credit card fraud detection in the era of disruptive technologies: A systematic review. Journal of King Saud University - Computer and Information Sciences, 35(1), 145–174. <https://doi.org/10.1016/j.jksuci.2022.11.008>
6. Chen, J. I., & Lai, K. (2021). Deep Convolution Neural Network Model for Credit-Card Fraud Detection and Alert. Journal of Artificial Intelligence and Copsule Networks, 3(2), 101–112. <https://doi.org/10.36548/jaicn.2021.2.003>
7. Probst, P., Wright, M. N., & Boulesteix, A. (2019). Hyperparameters and tuning strategies for random forest. WIREs Data Mining and Knowledge Discovery, 9(3). <https://doi.org/10.1002/widm.1301>
8. Ileberi, E., Sun, Y., & Wang, Z. (2022). A machine learning based credit card fraud detection using the GA algorithm for feature selection. Journal of Big Data, 9(1). <https://doi.org/10.1186/s40537-022-00573-8>
9. Shahraki, A., Abbasi, M., & Haugen, Ø. (2020). Boosting algorithms for network intrusion detection: A comparative evaluation of Real AdaBoost, Gentle AdaBoost and Modest AdaBoost. Engineering Applications of Artificial Intelligence, 94, 103770. <https://doi.org/10.1016/j.engappai.2020.103770>
10. E. Ileberi, Y. Sun and Z. Wang, "Performance Evaluation of Machine Learning Methods for Credit Card Fraud Detection Using SMOTE and AdaBoost," in IEEE Access, vol. 9, pp. 165286-165294, 2021, doi: 10.1109/ACCESS.2021.3134330.

1. . [↑](#footnote-ref-2)