IEEE-CIS Fraud Detection

Siddharth Kushawaha
Department of Mathematical Sciences
Stevens Institute of Technology
Hoboken, United States
skushawa@stevens.edu

Sheetal Shivaraj Kubsad
Department of Mathematical Sciences
Stevens Institute Of Technology
Hoboken, United States
skubsad@stevens.edu

***Abstract*—Fraud detection is a significant challenge across numerous industries, such as e-commerce and banking services. The upsurge in online transactions has come along with an unexpected increase in fraud. In this paper, we propose a fraud detection system conducted on IEEE-CIS Fraud Detection dataset developed by VESTA corporation. Our model employs the Balanced Random Forest Classifier with the Adaboost algorithm to increase the convergence and treat the imbalanced dataset as well. We used different classifiers such as Logistic Regression, CatBoost, Random Forest Classifier, and Gaussian NB. On Comparing the different models, we observed the highest AUC Score of 0.94 with the Balanced Random Forest Classifier.**

***Index Terms*—Banking Services, Fraudulent, Balanced Random Forest Classifier,Adaboost, CatBoost**

I. Introduction

In the current digital era, frauds are more common because of the growing reliance on technology across various sectors. Fraudulent actions like credit card fraud, identity theft, and online scams are extremely hazardous for financial institutions, enterprises, and individuals in general. Ensuring the integrity and security of digital systems now heavily depends on identifying and stopping these fraudulent acts. There have been numerous implementations of fraud prevention systems to protect consumers and financial institutions from fraudulent transactions. These systems, while effective in many cases, often result in false positives that can be both inconvenient and embarrassing for customers, since they need the whole process to be secure and non-cumbersome. This project aims to address the need for improving the accuracy and efficiency of fraud prevention systems in collaboration with the IEEE Computational Intelligence Society (IEEE-CIS) and Vesta Corporation, a leading payment service company that created the dataset for “Research Prediction Competition”

Significant time, money, and effort have been invested into developing a fraud-detection system to stop financial loss. A range of machine-learning [1] algorithms have been used to analyze large amounts of data. These involve standard methods like logistic regression [1], support vector machines [2], RandomForest Classifier [3] as well as state-of-the-art techniques like gradient boosting trees [3] and deep learning [4]. The most promising of them are gradient boosting trees deep learning and Random Forest Classifier.

Our proposed system, aims to use various machine learning techniques on the transaction data which contains both genuine and fraudulent transactions, giving an insight into the mode of transaction and entire transaction details, to accurately classify transactions and provide speed, efficiency, and accuracy. In order to pave the way for more effective and flexible fraud prevention tactics, we explore innovative techniques for anomaly detection, model optimization, and feature extraction. The goal is to minimize false positives and enhance the customer experience by developing innovative and advanced fraud detection methods. The main contribution of the project is a model of Random Forest Classifier and Adaboost to achieve maximum AUC Score when compared to the other models we used.

In the following sections, we delve into related work in the field, EDA , Data Preprocessing, and Implementation, presenting our approach in detail and discussing the experimental results. Through this project, we not only aim to enhance the current understanding of fraud detection but also contribute valuable insights to the broader field of machine learning, demonstrating the potential of intelligent algorithms in combating fraudulent activities.

II. Related Work

This section explores the related research in the field of fraud detection and prevention.

In[1], Wenkai Deng proposed a data mining-based system for fraud detection by using logistic Regression, support vector machine (SVM), and Random Forest Regressor. Columns with maximum null values were deleted and recursive feature elimination was used to eliminate each feature iteratively. The research was concluded by getting maximum accuracy with the use of the Random Forest Regression Model. Mainly AUC curve and accuracy were used to perform model evaluation.

Neghia Nguyen in[2] proposed a research model for fraud detection using the features of CatBoost and Deep Neural Network (DNN) to exploit both new and historic customer data. The authors used the SMOTE method to treat the imbalanced dataset and utilized the property of Catboost for the automatic handling of categorical features and properties of PCA [5] for feature selection. The authors used the AUC Score and Precision-Recall Curve as evaluation metrics for the model.

Lijie Chen in [3] presents an approach to detect frauds using StackNet which allows the use of multiple models such as Catboost, LightGBM, and RandomForest which have been used in this research project. Feature Engineering is implemented by deleting columns with high correlation and data padding is also performed to ensure missing values are handled accurately. The authors implemented Stacknet in levels in which LightGBM Regressor, CatBoost Regressor, and GradientBoosting regressor comprised the first level, whereas RandomForestRegressor comprised the second level. The model is evaluated by the AUC score.

Hassan Najadat in [4], proposed a system using a deep learning model called BiLSTM-MaxPooling-BiGRU-MaxPooling(BiLSTM and BiGRU) which is used to extract the most important features, and other machine learning models are also used on the dataset. The authors solved the problem of the imbalanced dataset by performing Random-over Sampling, Random-Under Sampling, and Synthetic Minority Oversampling Technique (SMOTE) [6]. The model implemented takes two inputs, categorical and numerical which are fed to the dropout layer, and the output is combined, which is then used in the final layer with the sigmoid activation function for the last stage prediction.

III. Methodology

A. Description Of Dataset

Feature	**Description**
TransactionDT	Timedelta from a given reference datetime (not an actual timestamp)
TransactionAMT	Transaction payment amount in USD
ProductCD	the code for each transaction
Card1 - Card6	payment card information such as card type, card category, issue bank, country
addr	address
dist	distance
P and (R) emaildomain	purchaser and recipient email domain
C1-C14	counting, such as how many addresses are found to be associatedwith the payment card
D1-D15	timedelta, such as days between previous transaction, etc
M1-M9	match, such as names on card and address
Vxxx	Vesta engineered rich features, including ranking, counting, and other entity relations

TABLE I: Transaction Description

Feature	**Description**
Device Type	Type of machine customer uses
Device Info	Information of Machine
id_12 - id_38	Numerical features of identity

TABLE II: Identity Description

1) Data Description: The dataset is partitioned into two distinct categories: identity and transaction, linked by the TransactionID. The transaction table encompasses features outlined in Table 1, incorporating categorical variables such as ProductCD, card1-card6, addr, P_emaildomain, R_emaildomain, and M1-M9. The dimensions of the training transaction data are 590540 rows and 394 columns, and for the entire training set, the dimensions are 144233 rows and 434 columns.

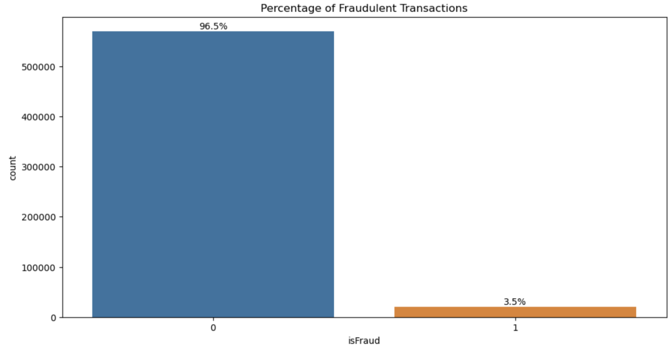


Figure 1: Imbalanced Dataset Visualization

The Identity table encompasses variables related to identity information, network connection details (IP, ISP, Proxy, etc.), and digital signatures (UA/browser/os/version, etc.) associated with transactions. Collected by Vesta's fraud protection system and digital security partners, the size of the training identity data is 144233 rows and 41 columns. A comprehensive summary of these variables is presented in Table 1 and Table 2 for clarity.

2) Exploratorty Data Analysis: The IEEE-CIS Fraud Detection dataset exhibits an imbalance, with 96.5% non-fraudulent transactions and 3.5% fraudulent transactions, as depicted in Figure 1. Imbalanced datasets pose challenges related to generalization and model bias.

In our data analysis, a noteworthy trend emerged, indicating a higher incidence of fraudulent transactions on Sundays compared to other days of the week. Furthermore, our investigation revealed a consistent decrease in fraudulent transactions during the afternoon hours across all days. It is also observed that non-fraudulent transactions exhibit a transaction amount cap below $10,000, whereas fraudulent reaches up to $5,000. It is also observed that the average of fraud transactions is higher than the average of non -fraudulent transactions.

The prevalence of fraudulent transactions exhibits a notable association with products categorized as W or C, particularly those settled through debit cards, credit cards (Visa or MasterCard). Interestingly, charge cards like American Express and Discover demonstrate minimal to no instances of fraudulent transactions, likely attributed to their less widespread usage. Moreover, a distinct pattern emerged concerning the origin of fraudulent activities, with protonmail.com and mail.com being identified as prominent email domains associated with such transactions.

Further insights reveal a concentration of fraudulent transactions on mobile platforms, even though desktops are more

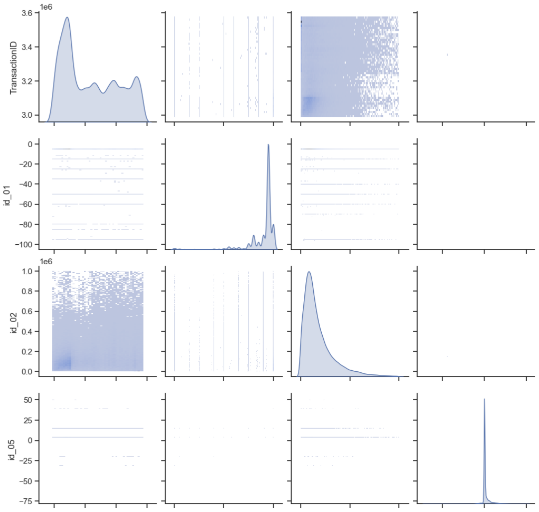


Figure 2: Pair plot of some identity features

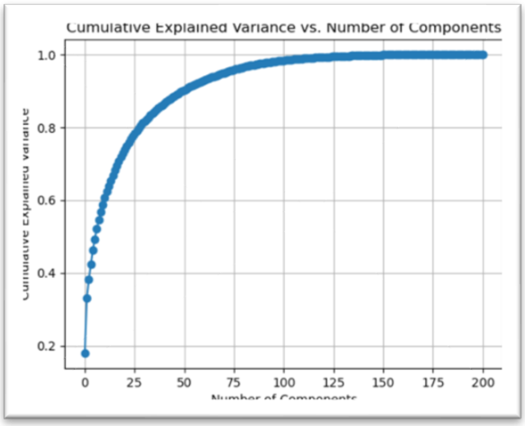


Figure 3: Explained Variance

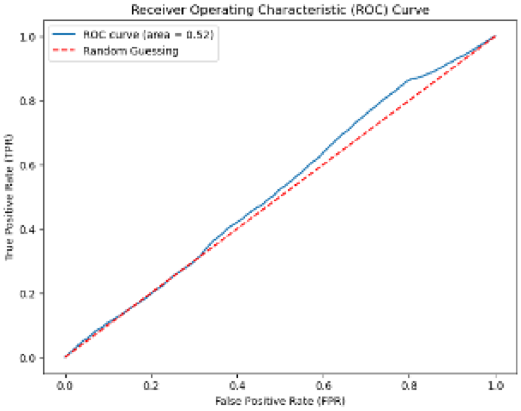


Figure 4: ROC Curve prior Sampling

commonly employed for legitimate transactions. This nuanced understanding underscores the need for tailored fraud detection measures in mobile transactions. Additionally, the examination of identification fields (id12 to id38) unveiled a substantial prevalence of null values, exceeding 70%. This suggests that these identifiers may not contribute significantly to the overall understanding of transactional patterns due to the limited available information.

The transaction and identity tables contain 394 and 41 columns, respectively. Figure 2 illustrates the distribution of data among TransactionID, DeviceInfo, DeviceType, id 36, id 37, and id 38.

IV. Implementation

This research focuses on optimizing memory usage and addressing class imbalance in a fraud detection dataset with a multitude of features and rows. Leveraging IEEE standards, we propose a comprehensive methodology that includes memory optimization, handling null values, treating outliers, and employing advanced machine learning techniques for improved fraud detection.

Due to the huge number of features and rows and their increased complexity, the dataset consumes a lot of memory storage to process the data. We optimized the memory usage of a dataset by casting appropriate data types. We followed the good practices of optimal memory usage like converting Boolean values to np.uint8 for better memory efficiency, converting object columns to ‘category’ type having less than 5% unique values, etc.

1) Data Preprocessing: To address null values, we utilized an iterative imputer, employing an iterative approach to estimate missing values by modeling each feature as a function of others. Exploratory Data Analysis (EDA) revealed non-uniform distributions and outliers in continuous data. We applied Power Transformation, specifically the Yeo-Johnson method, to mitigate outliers, improve data distribution, normalize the data, and mitigate skewness, enhancing the performance of machine learning algorithms on variables with a Normal Gaussian Probability distribution.

Since the original data has a large number of features, dimensionality reduction improves the computational efficiency of the system. Principal Component Analysis (PCA) [5] was thus employed to reduce feature dimensions to 72, explaining 95% of the data variance as depicted in Figure 3.

2) Imbalance Treatment: As shown in Figure 1, the IEEE-CIS dataset is highly imbalanced with 96.5% of non-frudulent transactions, and only 3.5 % of fraudulent transactions. If Imbalanced datasets are not treated, they may exhibit biased performance toward the majority class, and thus predicting the minority class becomes a challenge.

So, class imbalance in the dataset was addressed using the Synthetic Minority Oversampling Technique (SMOTE) [6], employing oversampling to balance target value counts. Initial attempts with Logistic Regression demonstrated underfitting due to duplicate records created during oversampling as shown in Figure 4. We introduced class weights during up sampling, resulting in improved performance with an AUC score of 0.84 as shown in Figure 5.

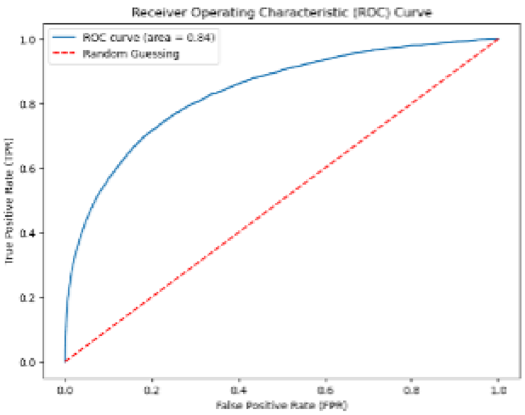


Figure 5: ROC Curve after Sampling

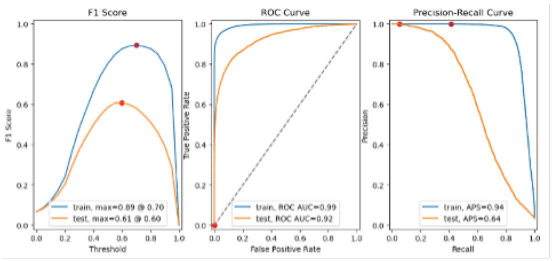


Figure 6: Random Forest Classifier

3) Ensemble Methods: Recognizing the need for improved shuffling, ensemble methods were employed. Ensemble methods are effective in improving model performance, reducing overfitting, and increasing generalization. Our approach commenced with the implementation of a hyper-optimized Random Forest Classifier, specifically configured with class_weight='balanced,' yielding outcomes detailed in Figure 6.

The challenge of class imbalance remains evident, a substantial disparity is observed between the test and training datasets. To address this issue, a hyperoptimized Balanced Random Forest Classifier, sourced from the imbalance library in Python, was employed. This classifier automatically rectifies imbalance concerns through the application of resampling and reshuffling techniques, ensuring the establishment of equitable target value counts and the normalization of scoring outcomes. Detailed results are presented in Figure 7, resulting in a remarkable AUC score of 0.93.

4) Advanced Models: GaussianNB demonstrated faster convergence but struggled with overfitting despite imbalance treatment. We used hyperoptimization of the CatBoost algorithm, specifying parameters such as 1000 iterations, a depth of 16, a learning rate of 0.73, and an evaluation metric of 'F1.' The results, depicted in the figure below, unequivocally illustrate that, despite maintaining a nearly identical F1 score as observed above, the model achieved a noteworthy improvement in the AUC score, reaching 0.95. This observation attests to a substantial enhancement in the overall efficiency of the model.

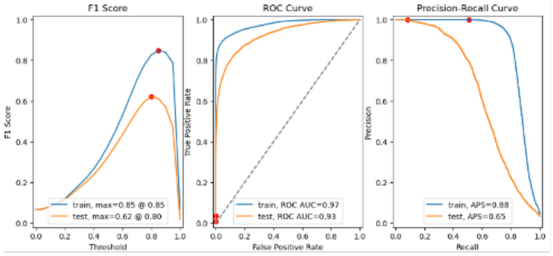
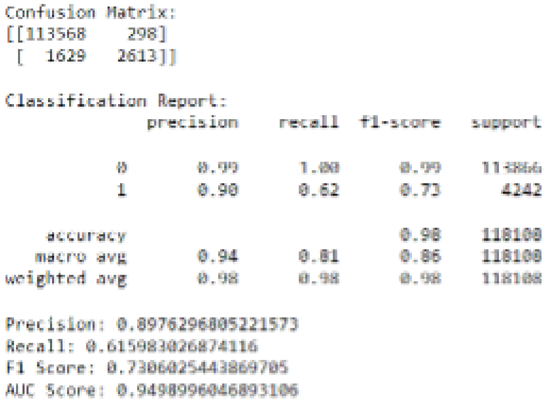


Figure 7: Balanced Random Forest Classifier



```
Confusion Matrix:
[[113568    298]
 [  1629   2613]]

Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      0.99    113866
           1       0.90      0.62      0.73      4242

    accuracy                           0.98    118108
   macro avg       0.94      0.81      0.86    118108
weighted avg       0.98      0.98      0.98    118108

Precision: 0.8976296805221573
Recall: 0.615983026874116
F1 Score: 0.7306025443869705
AUC Score: 0.9498996046893106
```

Figure 8: Results

We enhanced the capabilities of the Balanced Random Forest (Balanced RF) by incorporating the AdaBoost algorithm, employing a learning rate of 0.73, and setting the number of estimators to 50 to expedite convergence. The resulting model achieved a commendable public score of 91.39 during Kaggle testing against authentic data, establishing its proficiency as the optimal model for learning.

V. Conclusion

In this Research Project, we used several models to predict fraudulent transactions on the IEEE-CIS Dataset such as Logistic Regression, Random Forest Classifier, Balanced Random Forest Classifier, CatBoost Algorithm, and Gaussian NB. We addressed the imbalance in the dataset initially using SMOTE [6], and then employed a model that treats the imbalance. We observed that the best AUC Score was with a Balanced Random Forest Classifier incorporated with AdaBoost of 0.94. It reduced the FN (False Negatives) significantly which was our major concern.

VI. References

[1] W. Deng, Z. Huang, J. Zhang and J. Xu, "A Data Mining Based System For Transaction Fraud Detection," 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), Guangzhou, China, 2021, pp. 542-545, doi: 10.1109/ICCECE51280.2021.9342376.

[2] N. Nguyen et al., "A Proposed Model for Card Fraud Detection Based on CatBoost and Deep Neural Network,"

in IEEE Access, vol. 10, pp. 96852-96861, 2022, doi: 10.1109/ACCESS.2022.3205416.

[3] L. Chen, Q. Guan, N. Chen and Z. YiHang, ”A StackNet Based Model for Fraud Detection,” 2021 2nd International Conference on Education, Knowledge and Information Management (ICEKIM), Xiamen, China, 2021, pp. 328-331, doi: 10.1109/ICEKIM52309.2021.00079.

[4] H. Najadat, O. Altiti, A. A. Aqouleh and M. Younes, ”Credit Card Fraud Detection Based on Machine and Deep Learning,” 2020 11th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 2020, pp. 204-208, doi: 10.1109/ICICS49469.2020.239524.

[5] Ian T. Jolliffe and Jorge Cadima, ”Principal component analysis: a review and recent developments” 2016 doi: 10.1098/rsta.2015.0202d1e2845

[6] Dina Elreedy and Amir F. Atiya, ”A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance” 2019 doi: org/10.1016/j.ins.2019.07.070