Bank Account Fraud (BAF) Detection Using Machine Learning

Group Members:

Narendra Reddy Reddymalla - 700743852
Akhil Patlori - 700744211
Vikram Boppana - 700742246

Introduction

Fraudsters try to either pretend to be someone else through stealing identities or create fake individuals to gain access to banking services. Once they get access to a new bank account, they exploit the available credit or use the account for receiving illegal payments. The bank incurs all the costs related to the fraud since it's difficult to trace the true identity of the fraudster.

In this important area, machine learning plays a vital role. If a fraudulent application is predicted, the customer's bank account application is rejected. On the other hand, a negative prediction allows the customer to open a new bank account and receive a credit card. Since having a bank account is considered a fundamental right in the European Union, detecting fraud is significant socially. Due to concerns about biased decision-making using machine learning systems, banks and merchants are adopting fair ML methods.

Each instance in the dataset represents an individual application made on an online platform where applicants agreed to store and process their data. The label "is_fraud" indicates whether an application is fraudulent (positive instance) or legitimate (negative instance). The dataset covers a period of eight months from February to September.

The occurrence of fraud in the dataset ranges from 0.85% to 1.5% across different months. It is observed that these values tend to be higher in the later months. Additionally, the distribution of applications fluctuates from 9.5% to 15% of the total applications each month. These variations are important for determining the approximate number of legitimate and fraudulent instances to sample for each month in the dataset variants.

There has been a surge in publicly available unstructured data resources for computer vision and natural language processing tasks, the domain of tabular data, which is prevalent in many critical domains, has not progressed at the same pace. To bridge this gap, Bank Account Fraud (BAF) was introduced, an innovative and privacy-preserving suite of realistic tabular datasets accessible to the public.

BAF is built using cutting-edge techniques for generating tabular data applied to an anonymized, real-world dataset specifically focused on detecting fraud in bank account openings. This approach addresses the challenges commonly faced in real-world applications, including the influence of time and significant imbalances between fraudulent and legitimate instances. Each variant of the BAF dataset incorporates specific types of data bias, enabling practitioners to thoroughly evaluate the performance and fairness of their ML methods.

GITHUB LINK: https://github.com/AkhilPatlori/BankAccountFraudDetectionUsingMachineLearning

Primary objective of BAF is to provide the research community with a more authentic, comprehensive, and robust platform for evaluating both novel and existing methods. This resource fills a crucial void by offering a diverse range of tabular datasets that accurately represent the intricacies and nuances present in high-stakes domains. By utilizing BAF, researchers and professionals can significantly enhance the development and assessment of ML techniques in the analysis of tabular data.

This dataset provides valuable insights for developing fraud detection models in the banking industry, where ensuring fairness in machine learning predictions is crucial.

Objective :

With the BAF dataset, various types of analysis can be performed, including descriptive analysis, prescriptive analysis, and predictive analysis. Here's how each analysis can be applied to the dataset:

Descriptive Analysis: Descriptive analysis involves summarizing and exploring the data to gain insights into its characteristics. With the BAF dataset, we can perform descriptive analysis to understand the overall distribution of fraudulent and legitimate applications. This analysis can include calculating summary statistics, creating visualizations (such as histograms or bar charts) to examine the prevalence of fraud across different months, and exploring any temporal trends or patterns in the data.

Prescriptive Analysis: Prescriptive analysis focuses on providing recommendations or actions to optimize outcomes. In the context of the BAF dataset, prescriptive analysis can be used to identify key factors or features that contribute to fraudulent applications. By analyzing the dataset and applying techniques such as feature importance, we can determine which variables have the most significant impact on fraud detection. This information can then be used to develop strategies and guidelines for banks to prevent and detect fraud more effectively.

Predictive Analysis: Predictive analysis involves using historical data to build models that can make predictions about future events or outcomes. In the case of the BAF dataset, predictive analysis can be employed to develop machine learning models that accurately classify new bank account applications as fraudulent or legitimate. By training these models on the historical data, they can learn patterns and relationships between variables to make predictions on unseen data. Predictive analysis can help banks automate the process of fraud detection, enabling them to identify potential fraudulent applications in real-time and take appropriate actions to mitigate risks.

Dataset Source & Description

The BAF suite of datasets is a valuable resource for evaluating ML methods in bank account fraud detection. The datasets offer realistic scenarios, controlled biases, imbalanced data, and dynamic temporal dynamics, providing a comprehensive and privacy-preserving platform for researchers and practitioners in the field. For our analysis, we will be using the base dataset of the BAF suite, which is a synthetic fraud dataset containing 1 million instances.

The base dataset also includes three protected attributes, namely age group, employment status, and % income, which can be used to evaluate fairness in machine learning. The dataset provides a solid foundation for analyzing fraud and developing fair ML methods. We have used the base dataset to build our model and evaluate its performance by splitting the dataset into train and test sets.

Data Preparation

I. To gain an understanding of the overall characteristics of the data, a review and summary were conducted.

II. The dataset was examined for the presence of null values, and variables with more than 50% null values were removed from further analysis.

III. For the remaining variables with null values, imputation using the median value was performed, as it can effectively handle outliers.

IV. Variables that showed constant variance and did not vary across the dataset were removed from consideration since they do not provide useful information for analysis.

V. Initially, the analysis and model building were conducted using the base dataset, which consisted of six data files. The data was then split into a training set and a testing set, with a 80% and 20% split, respectively. The training set was allocated 80% of the data, while 20% was reserved for testing the models' performance.

VI. Since the problem involved imbalanced data, where one class was more prevalent than the others, various sampling techniques were explored to address the imbalance and improve the model's performance by providing balanced representations of the different classes.

Exploratory Data Analysis

Univariate Analysis

In the univariate analysis, we looked at each continuous variable in the dataset. First, we checked the distribution of these variables to understand their shape and behavior. For categorical variables, we counted the frequency of each category to see how often they appeared.

Next, we wanted to find any outliers, which are extreme values that don't fit the overall pattern of the data. We used a technique called the 3 standard deviation rule to identify outliers. Basically, if a value was more than 3 times the standard deviation away from the average, we considered it an outlier.

When we found outliers, we decided to impute them in the dataset by median.It's important to note that removing outliers should be done thoughtfully and with caution, as it can affect the overall understanding of the data.

By going through these steps for each continuous variable and examining the distribution and frequencies of categorical variables, we gained a better understanding of each variable on its own. This helped us identify any data issues and prepare the data for further analysis or modeling.

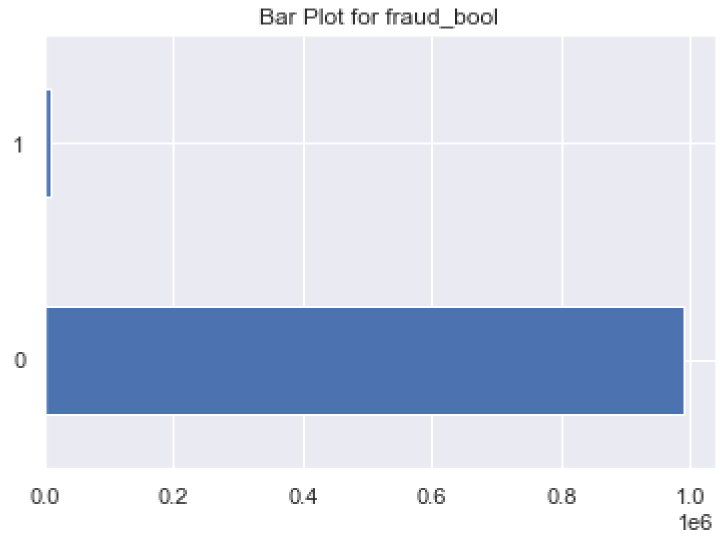


Fig 1. Bar Plot for variable fraud_bool

Based on Fig 1, the fraud_bool variable exhibits an imbalanced distribution. The percentage of instances labeled as 0 (non-fraudulent) is 98.8971%, while the percentage of instances labeled as 1 (fraudulent) is 1.1029%.

Imbalanced problems like this can pose challenges during analysis and modeling, as the minority class (fraudulent) may be underrepresented, making it harder for a model to learn and predict accurately.

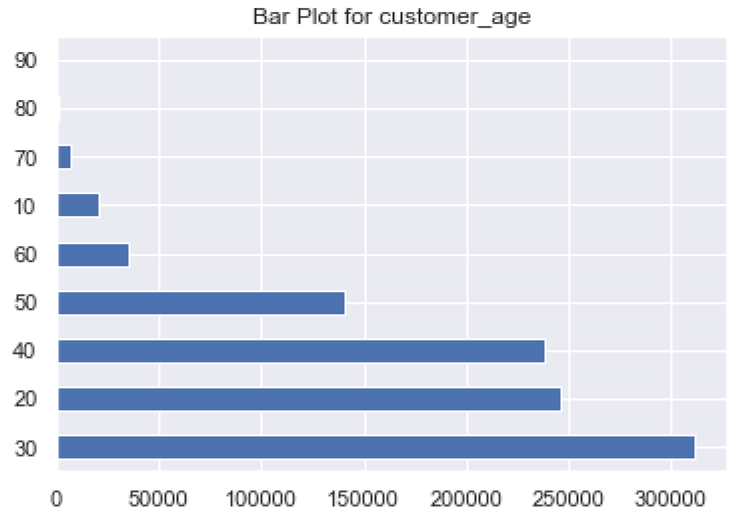


Fig 2 Bar Plot of customer age

As Fig 2 describes the age distribution of applicants,more than 93% of the applicants fall within the age range of 20-50. Less than 1% of the applicants are aged 60 or above. Approximately 2% of the applicants are aged 10.

These statistics highlight the age composition of the applicant pool, emphasizing that a significant majority (over 93%) are between the ages of 20 and 50. Meanwhile, a very small proportion (less than 1%) consists of individuals aged 60 or older. Additionally, around 2% of the applicants are found to be 10 years old.

Outlier Analysis:

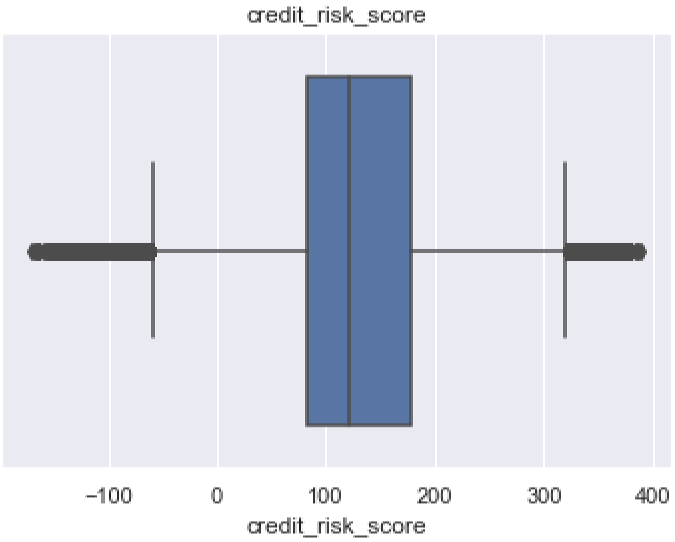


Fig 3 Box Plot of credit_risk_score with Outliers

To identify outliers, we applied the 3 standard deviation rule. If a value in a particular column was more than 3 times the standard deviation away from the average, we classified it as an outlier. Upon analysis, we observed outliers in certain columns, such as "credit_risk_score."

To address these outliers, we performed outlier treatment on all the features that exhibited extreme values. This involved removing the outliers from the dataset. As a result, we obtained a modified dataset where extreme values were no longer present.

As depicted in Figure 3, the column "credit_risk_score" initially displayed outliers. However, after outlier treatment, as shown in Figure 4, the outliers were successfully eliminated. Similar outlier treatment techniques were applied to other features in the dataset that also contained extreme values.

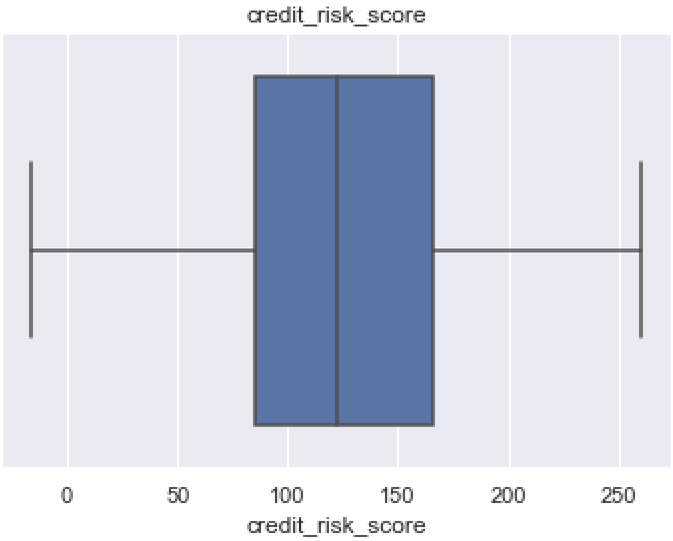


Fig 4 Box Plot of credit_risk_score without Outliers

Bivariate Analysis

Bivariate analysis involves examining the relationship between two variables in a dataset. It aims to understand how changes in one variable relate to changes in another variable. By analyzing the interaction between these two variables, we can gain insights into any patterns, trends, or associations that may exist between them.

Correlation Graph(Numerical vs Numerical)

Between the month and velocity_4w, we have the maximum correlation, which can be observed from Figure 5.

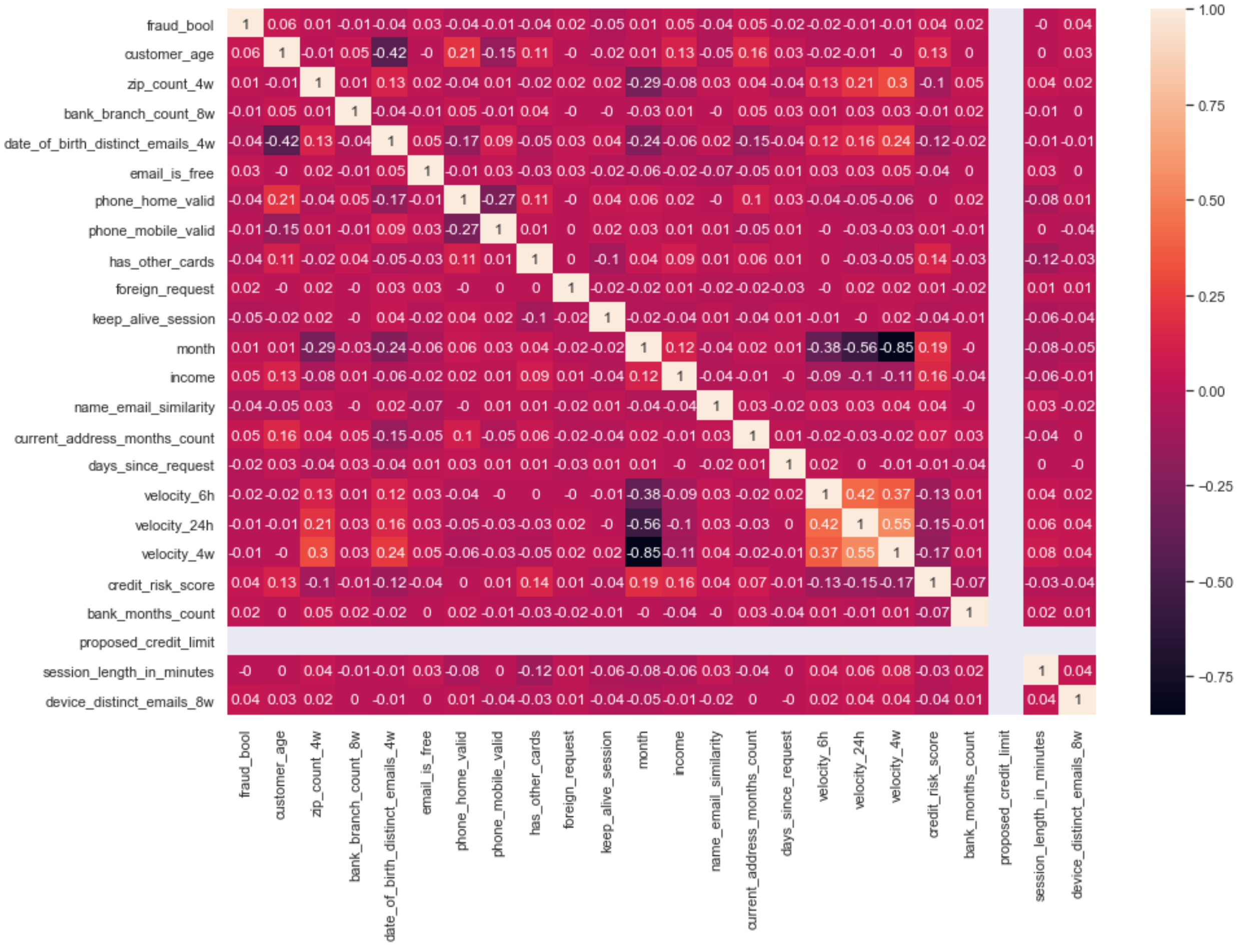


Fig 5. Correlation Graph(heat map)

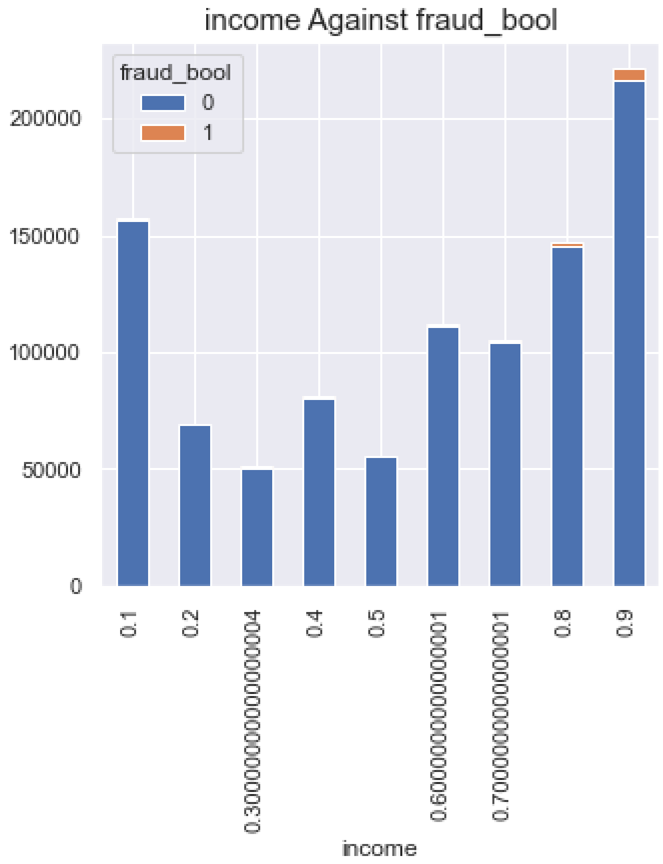


Fig 6 Bar Graph Income Vs Fraud_bool

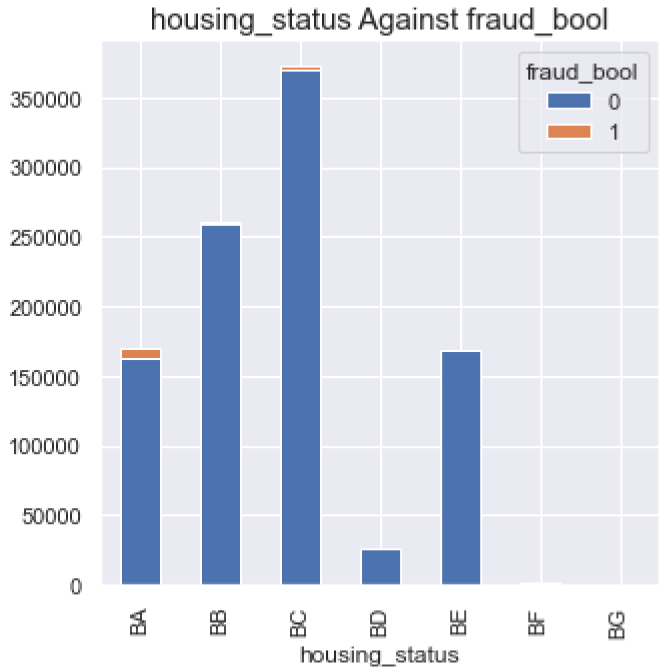


Fig 7 Bar Graph housing status vs Fraud_bool

During the analysis, it was observed that when the income falls within the range of 0.9 (decimal value), the chances of fraud increase. This indicates that individuals with incomes close to 1 are more likely to be associated with fraudulent activities.

Additionally, it was found that individuals with a housing status labeled as "BA" have a higher likelihood of being involved in fraud. This categorical variable, indicating a specific housing status, shows a strong association with the target variable (fraud).

Similar analyses were conducted for all the categorical variables in relation to the target variable. By examining the relationship between each categorical variable and fraud, patterns and correlations were identified. These findings help in understanding the factors that contribute to fraud and provide insights for further investigation or predictive modeling.

Modeling:

Logistic Regression:

Logistic Regression is a popular classification algorithm used to model the relationship between a dependent variable and one or more independent variables. It estimates the probability of an event occurring by fitting a logistic function to the input variables. Logistic Regression is widely used for binary classification problems and can be extended to handle multi-class classification as well.

Decision Tree Classifier:

A Decision Tree Classifier is a machine learning algorithm that creates a tree-like model of decisions and their possible consequences. It splits the data based on different features to create branches and nodes that represent decision rules. Each leaf node represents a class label. Decision Tree classifiers are interpretable and can handle both categorical and numerical data. They can capture complex interactions between variables but are prone to overfitting.

Random Forest Classifier:

Random Forest Classifier is an ensemble learning method that combines multiple Decision Trees to make predictions. It constructs a forest of trees by randomly selecting subsets of features and data samples. Each tree in the forest independently makes a prediction, and the final classification is determined by voting or averaging the predictions. Random Forests are robust, handle high-dimensional data well, and can handle both classification and regression tasks.

Light GBM:

Light GBM (Gradient Boosting Machine) is a gradient boosting framework that uses a tree-based learning algorithm. It is designed to be efficient and scalable while providing high accuracy. Light GBM builds trees in a leaf-wise manner, reducing the number of splits and optimizing memory usage. It uses gradient-based learning with a focus on improving training speed and handling large datasets. Light GBM is commonly used for classification and regression tasks.

Balanced Bagging:

Balanced Bagging is an ensemble learning technique specifically designed for imbalanced datasets, where one class is significantly underrepresented. It combines the concepts of bagging and undersampling to create a balanced training set. It trains multiple classifiers on different bootstrap samples of the balanced data and combines their predictions through voting or averaging. Balanced Bagging helps address class imbalance issues and can improve the performance of classifiers on minority class prediction.

Easy Ensemble:

Easy Ensemble is an ensemble learning method that tackles imbalanced classification problems by generating multiple balanced training sets through random undersampling of the majority class. It trains multiple classifiers on these balanced sets and combines their predictions through voting or averaging. Easy Ensemble aims to overcome the challenges posed by imbalanced datasets and give more attention to the minority class, thereby improving classification performance on the minority class while maintaining a reasonable accuracy on the majority class.

The selection of the most suitable algorithm depends on the specific characteristics of the problem and the desired balance between interpretability, accuracy, and scalability.

Evaluation:

When evaluating a fraud vs non-fraud classification model, several metrics can be used to assess its performance. The choice of the best metric depends on the specific goals and priorities of the application. However, there are a few key metrics that are commonly used in fraud detection scenarios:

Precision: This metric measures the proportion of correctly predicted positive instances (fraudulent cases) out of all instances predicted as positive. A high precision indicates a low rate of false positives, which means fewer non-fraudulent cases being classified as fraud.

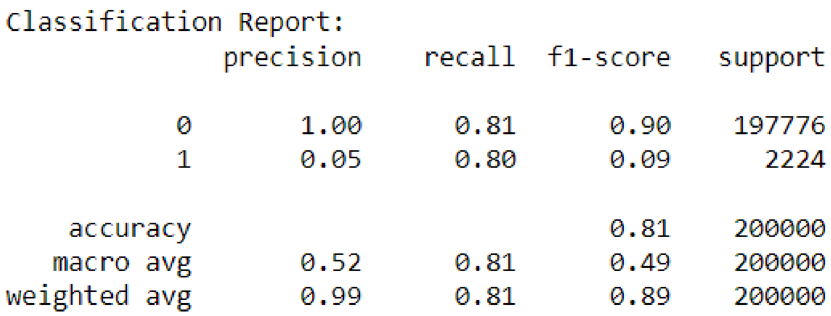
Recall: This metric measures the proportion of correctly predicted positive instances (fraudulent cases) out of all actual positive instances. High recall indicates a low rate of false negatives, meaning fewer cases of actual fraud being missed or not detected.

F1 Score: This metric is the harmonic mean of precision and recall. It provides a balanced measure of a model's performance by considering both precision and recall. It is particularly useful when there is an imbalance between the number of fraud and non-fraud cases in the dataset.

Area Under the ROC Curve (AUC-ROC): The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) at various classification thresholds. AUC-ROC measures the overall performance of the model by calculating the area under the ROC curve. A higher AUC-ROC indicates a better discrimination capability of the model in distinguishing between fraud and non-fraud cases.

It is important to choose the best metric based on the specific requirements of the fraud detection system. For example, a high precision may be more important to minimize false positives and avoid inconveniencing legitimate customers in some cases. In other cases, a high recall might be prioritized to capture as many instances of actual fraud as possible. It is crucial to consider the trade-offs between precision and recall and select the metric that aligns with the specific needs of the fraud detection system.

After evaluating various models, the Easy Ensemble model emerged as the top performer, demonstrating the following performance metrics



```
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.81      0.90    197776
           1       0.05      0.80      0.09      2224

    accuracy                           0.81    200000
   macro avg       0.52      0.81      0.49    200000
weighted avg       0.99      0.81      0.89    200000
```

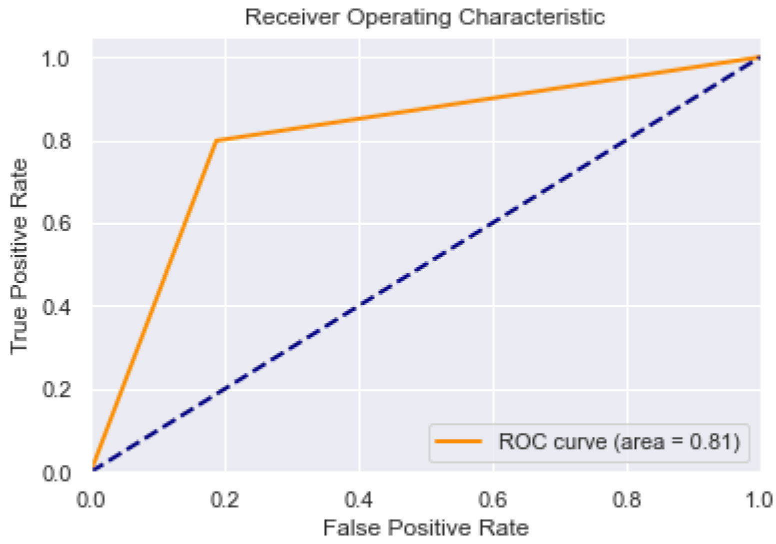


Fig 8 AUC-ROC Curve

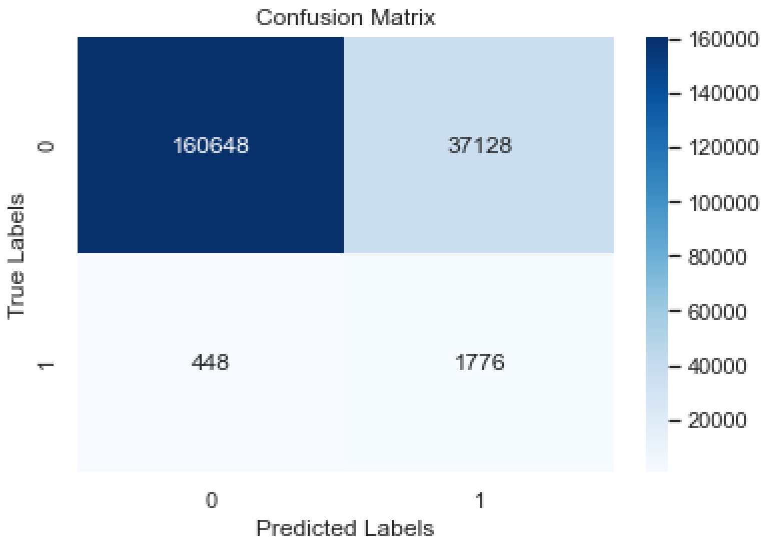


Fig 9 Confusion Matrix

CONCLUSION AND FUTURE WORK

In conclusion, our study focused on conducting thorough research using the BAF base data to develop an effective model for predicting fraud labels. Our initial testing with the baseline model revealed that the extensive range of features led to significant variance and poor performance on the validation set compared to the training set. This lack of accuracy is concerning, particularly considering the imbalanced nature of the data.

To address these challenges, future research on this topic can explore alternative feature selection strategies and apply various sampling techniques. Additionally, incorporating machine learning algorithms such as Support Vector Machines (SVM), Naive Bayes, and K-Nearest Neighbors (KNN) could provide valuable insights. Furthermore, experimentation with different neural network designs could also be beneficial.

By investigating these avenues, we hope to find solutions to the imbalanced dataset issue and develop models that can be effectively trained in the future. This research is crucial for improving fraud detection and mitigating the impact of imbalanced data on the performance of predictive models.

Reference

[1] https://cs229.stanford.edu/proj2018/report/261.pdf

[2] https://pubmed.ncbi.nlm.nih.gov/11376540/

[3] https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9999220

[4] https://journalofbigdata.springeropen.com/articles/10.1186/s40537-022-00573-8

[5] https://www.irjmets.com/uploadedfiles/paper//issue_1_january_2022/18561/final/fin_irjmets1643111967.pdf

[6]https://www.researchgate.net/publication/227441142_Logistic_regression_in_data_analysis_An_overview

[7] Hanley J, McNeil B. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148:839–43.

[8] Harrell F, Lee K, Mark D. Multivariable prognostic models: issues in developing models, evaluation assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361– 87.

[9]Phua, C., Alahakoon, Damminda, Lee, Vincent: Minority report in fraud detection: classification of skewed data. ACM SIGKDD Explor. Newsl. 6, 50–59 (2004)

[10] https://arxiv.org/abs/2101.08030

[11] Allen D. The relationship between variable selection and data augmentation and a method of prediction. Technometric 1977;16:125–7

[12]Pérez, J.M., Muguerza, J., Arbelaitz, O., Gurrutxaga, I., Martín, J.I.: Consolidated tree classifier learning in a car insurance fraud detection domain with class imbalance. In: Pattern Recognition and Data Mining (ed), pp. 381–389. Springer (2005)

[13]Phua, C., Alahakoon, D., Lee, V.: Minority report in fraud detection: classification of skewed data. ACM SIGKDD Explor. Newsl. 6, 50–59 (2004)

[14]https://www.sciencedirect.com/science/article/abs/pii/S0952197619302714

[15] http://www.cs.put.poznan.pl/jstefanowski/pub/ABBag2017.pdf

[16] Wojciechowski, S., Wilk, Sz.: The generator of synthetic multi-dimensional data. Poznan Univ. of Technology Report RB-16/14 (2014).

[17] He, H., Yungian, Ma (eds): Imbalanced Learning. Foundations, Algorithms and Applications. IEEE - Wiley, (2013).

[18] https://www.researchgate.net/publication/220198530_Roughly_balanced_Bagging_for_Imbalanced_data

[19]Vapnik V. The nature of statistical learning theory. 2nd ed. New York: Springer; 2000

[20] Branco, P., Torgo, L., Ribeiro, R.: A Survey of Predictive Modeling on Imbalanced Domains. ACM Computing Surveys (CSUR), 49(2), 31 (2016).