Credit Card Fraud Detection

Sai Sindhu Thondapu (002785016)

College of Engineering

Northeastern University

Problem Statement:

Every day, billions of credit card transactions are processed worldwide. Given the extensive usage of cellphones and the Internet on the planet, an increasing number of individuals are utilizing their credit cards to make online purchases, payments via applications, and so on. In a situation like this, it is critical that credit card organizations can quickly distinguish whether a transaction is the result of a fraud or a genuine purchase, preventing customers from being charged for products they did not purchase. In this project, I'll utilize the scikit-learn package to create a prediction model that can learn and recognize whether a transaction is fraudulent or real. I plan to utilize a different classification models.

Rational of Target variable selection:

Credit Card Fraud Detection covers credit card transactions conducted by European clients during two days in September 2013. The dataset includes the feature time, which displays the number of seconds elapsed between each transaction and the first transaction in the dataset. The feature amount, which includes the transaction amount and the feature class, indicates whether the transaction is legitimate or fraudulent, with 1 indicating fraud and 0 indicating genuine. Features V1, V2,... V28 are numerical input variables resulting from a PCA transformation, the contents of which cannot be published due to their confidential nature. According to the given dataset, the target variable is the Class column, which fits with the project's objective. The presented dataset is an unbalanced dataset.

Associated Tasks	Number of Instances	Variables	Target Variable
Binary Classification	284,807	V1-V28, Time & Amount(Total 30)	Class (0-fraud, 1-genuine)

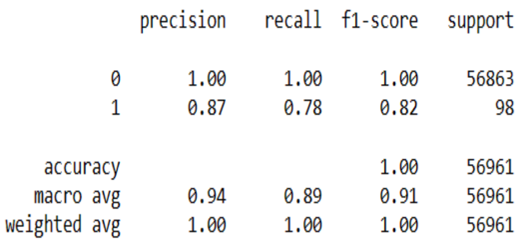
Table 1: Basic feature of the dataset

Suitable Machine learning algorithm:

The selected model employs the SMOTE (Synthetic Minority Over-sampling Technique) oversampling technique in combination with StratifiedKFold cross-validation. This approach is used to enhance the performance of the predictive model in credit card fraud detection. The model utilizes the XGBoost algorithm and has demonstrated remarkable accuracy, achieving an accuracy score of 99.94%. The ROC-AUC score, which indicates the model's ability to distinguish between fraud and non-fraud cases, is at an impressive level of 98.16%. This high ROC-AUC value suggests that the model can effectively discriminate between positive and negative cases, showcasing its robustness in detecting fraudulent transactions. The chosen threshold for classification is set at 0.027%, which helps strike a balance between precision and recall. This threshold determines the point at which a transaction is classified as fraudulent or legitimate based on the predicted probabilities.

Conclusion:

Incorporating SMOTE oversampling to address class imbalance, along with Stratified K-Fold cross-validation, strengthens the model's ability to generalize well to new data while effectively handling the challenges posed by imbalanced datasets. This selected approach, coupled with the XG Boost algorithm's strong predictive capabilities, yields a high-performing model for credit card fraud detection with impressive accuracy and robustness. Below are the visuals of selected model XG Boost



```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     56863
           1       0.87      0.78      0.82        98

    accuracy                           1.00     56961
   macro avg       0.94      0.89      0.91     56961
weighted avg       1.00      1.00      1.00     56961
```

Figure 1: Classification Report

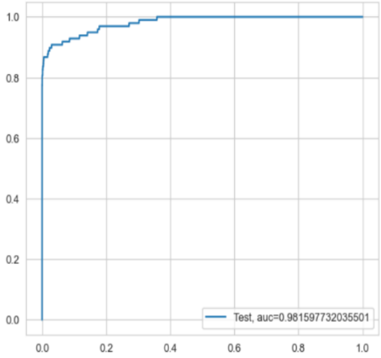


Figure 2: ROC

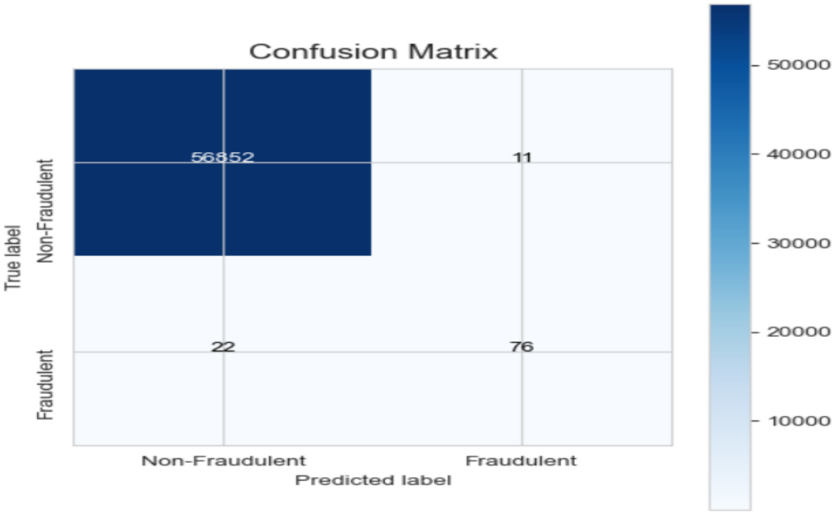


Figure 3: Confusion Matrix

Scope for future work:

While we did not achieve our aim of 100% accuracy in fraud detection, we did create a system that, given enough time and data, can come extremely close. As with every such undertaking, there is opportunity for improvement here. The dataset contains more opportunities for development. The accuracy of machine learning algorithms such as XG Boost, Random Forest, and others should be subjectively verified on other data sets for credit card fraud detection. Because the total dataset consists of only two days of transaction records, it is only a portion of the data that can be made available if this research is used commercially. Based on machine learning methods, the program's efficiency will only rise over time as more data is fed into it. Furthermore, the performance of modern machine learning algorithms for credit card fraud detection can be evaluated.

Additional work:

I used two cross validation methods, Repeated K-Fold and Stratified K-Fold on the dataset to see which model works best with the imbalance dataset. I also trained an additional model for both the balanced and imbalanced datasets. Furthermore, I used two sampling approaches, Random Sampler and SMOTE, to forecast the model that performs best with the balancing dataset.