**Cristian Biondi**

**Introduction:**

In this project, we aimed to develop and evaluate a machine learning model for fraud detection using a dataset containing various transaction features. The primary objective was to create a robust model capable of accurately identifying fraudulent transactions while minimizing false positives and false negatives.

**Data Preprocessing:**

We began by loading the dataset and performing essential preprocessing steps. This included handling missing values, encoding categorical variables, and generating new features such as the transaction amount (amt), transaction hour (trans\_hour), and geographical distances (distance). We also ensured that the dataset was split into training and testing sets with an appropriate class balance between fraudulent and non-fraudulent transactions.

**Model Training:**

A Random Forest classifier was selected for its robustness and ability to handle imbalanced datasets using class weights. The model was trained using the preprocessed training data, with an emphasis on ensuring that the features used for training were consistent across both the training and testing datasets. During training, we paid particular attention to potential issues such as overfitting, where the model might perform well on the training data but poorly on new, unseen data.

**Model Evaluation:**

The model's performance was evaluated on the test set using key metrics such as accuracy, precision, recall, and F1-score. The results showed an accuracy of 97%, with high precision (93%) and recall (82%) for identifying fraudulent transactions. Despite the strong overall performance, we explored whether the model exhibited signs of overfitting by comparing its performance on the training data and considering cross-validation techniques.

**Simulated Data Testing:**

To further challenge the model, we generated simulated data based on the distributions observed in the original dataset. The simulated data was used to test the model’s ability to generalize beyond the original dataset. However, we encountered issues where the model struggled to correctly identify fraudulent transactions in the simulated data, likely due to discrepancies in feature distributions and potential overfitting to the training data.

**Feature Importance:**

An analysis of feature importance revealed that the transaction amount (amt) and transaction hour (trans\_hour) were the most influential features in predicting fraud. This insight can be used to refine the model further or guide future data collection efforts.

**Conclusion:**

The Random Forest model demonstrated strong performance in detecting fraudulent transactions, with high accuracy and precision. However, the challenges faced with simulated data highlighted the importance of ensuring that the model is well-generalized and robust to variations in data. Future steps could include further tuning of the model, exploring more sophisticated data simulation techniques, or applying ensemble methods to enhance model performance.

This project provided valuable insights into the complexities of fraud detection and the importance of careful data preprocessing, model training, and evaluation to achieve reliable results.