Credit Card Fraud Detection Project Report

1. Introduction

Context and Importance:

Credit cards are extensively used for online purchases and payments due to their convenience. However, they come with significant risks, particularly in the form of credit card fraud. This fraud occurs when someone uses another person's credit card or card information without authorization to make purchases or withdraw cash. Therefore, it is essential for credit card companies to swiftly identify and address fraudulent transactions to prevent unauthorized charges on customers' accounts.

Goal:

The main goal of this project is to create a machine learning model designed to identify fraudulent credit card transactions. By implementing this model, credit card companies can reduce risks and safeguard their customers.

Dataset:

The dataset comprises transactions made by European cardholders in September 2013, spanning a two-day period. It includes 492 fraudulent transactions out of a total of 284,807 transactions. This results in a highly imbalanced dataset, with the positive class (frauds) representing just 0.172% of all transactions.

---

2. Data Exploration and Preprocessing

Data Loading and Initial Exploration:

- The dataset is loaded using pandas, and initial exploration involves checking the first few rows, summary statistics, and data types of the columns.

Summary Statistics:

- Summary statistics provide insights into the distribution and scale of the data.

Checking for Missing Values:

- The dataset is checked for missing values, and it was found that there are no missing values in any of the columns. This ensures that the dataset is complete and ready for analysis without the need for imputation.

Distribution of Classes:

- The distribution of the classes (fraudulent vs. non-fraudulent transactions) is examined to understand the class imbalance, which is crucial for model training and evaluation.

Handling Data Imbalance:

Given the heavily imbalanced nature of the data, several approaches are employed to address this imbalance:

- Check and Mitigate Skewness: Explaining the techniques used to handle skewness in the data.
- Handling Data Imbalance: Describing the methods used to balance the dataset (e.g., SMOTE, ADASYN). Handling data imbalance as we see only 0.17% records are the fraud transactions

---

3. Model Building

Algorithms Used:

- The model will be trained using a variety of algorithms, including Logistic Regression, Decision Tree, Random Forest, and XGBoost. Each algorithm has unique strengths.
- by comparing their performance, we can determine the most effective one for detecting fraudulent transactions.

Hyperparameter Tuning:

- The process of hyperparameter tuning will involve using Grid Search Cross Validation to identify the optimal values for the model's hyperparameters.
- This technique systematically searches through a predefined set of hyperparameters and evaluates their performance using cross-validation, ensuring that the model achieves the best possible performance on unseen data.

---

4. Model Evaluation and Results

Evaluation Metrics:

- Accuracy is not a suitable metric for this imbalanced dataset, as it can be misleading. Instead, we will focus on Precision, Recall, and ROC-AUC. Precision measures the proportion of correctly identified frauds among all transactions flagged as fraudulent, while Recall measures the proportion of actual frauds correctly identified by the model.
- Balancing Precision and Recall is crucial to minimize false positives and false negatives.
- Additionally, the ROC-AUC metric will be used to assess the model's ability to distinguish between fraudulent and non-fraudulent transactions.
- A good ROC score is indicated by a high True Positive Rate (TPR) and a low False Positive Rate (FPR), which reduces misclassifications and improves overall model performance.

Results:

- The evaluation results for each model will be presented, highlighting the key metrics of Precision, Recall, and ROC-AUC. These metrics will provide a comprehensive view of the model's performance, emphasizing its effectiveness in identifying fraudulent transactions while maintaining a balance between false positives and false negatives.
- The results will help determine which model and hyperparameter settings achieve the best performance in detecting fraud within this imbalanced dataset.

---

5. Conclusion and Summary

Summary of Findings:

- Our study has underscored the significant impact of balancing techniques on model effectiveness, as well as the varying performances observed across different model architectures.
- Key findings highlight the critical role of balanced datasets in enhancing predictive accuracy and reducing bias in model outcomes.
- Additionally, our analysis has provided insights into the comparative strengths of various machine learning algorithms under different balancing strategies.

Conclusion:

- Moving forward, further advancements can be pursued in the realm of balancing techniques, with exploration into more sophisticated methods such as ensemble learning and adaptive sampling.
- Additionally, the integration of deep learning models presents an exciting avenue for future research, promising enhanced capabilities in handling complex and high-dimensional data.
- Continual refinement and validation of these approaches will be crucial in advancing the field and achieving robust, reliable predictive models.