Credit Card Fraud Detection

Table of Contents

1. Problem Statement

Credit card fraud is a significant concern in the financial sector, with increasing cases of unauthorized credit card transactions leading to substantial financial losses. Detecting fraudulent transactions effectively and efficiently is essential to maintain trust in the credit card system and to protect consumers from financial harm.

2. Project Objective

The objective of this project is to develop a machine-learning model that can accurately identify fraudulent transactions in a credit card dataset. The goal is to minimize false positives and false negatives, ensuring genuine transactions are processed smoothly while flagging or blocking fraudulent ones.

3. Data Description

The dataset contains 11,665 transactions, each described by 31 features. These features include Time, Amount, and 28 anonymized variables (V1 to V28). The 'Class' variable is the response variable, indicating whether a transaction is fraudulent (**1**) or genuine (**0**). The dataset is highly imbalanced, with a vast majority of transactions being genuine.

4. Data Pre-processing Steps and Inspiration

- **Null Value Analysis:** Checked for any null values in the dataset.
- **Feature Normalization:** Used **StandardScaler** to normalize the 'Amount' feature for more efficient model training.
- **Feature Selection:** Dropped the 'Time' and original 'Amount' columns to focus on more relevant features.
- **Data Splitting:** The data was split into features (X) and the target variable (Y). It was further divided into training and test sets.

5. Choosing the Algorithm for the Project

Two algorithms were chosen for this project:

- **Decision Tree Classifier:** A simple, interpretable model.
- **Random Forest Classifier:** An ensemble method that improves upon the decision tree by reducing the risk of overfitting.

6. Motivation and Reasons For Choosing the Algorithm

- Decision Trees are easy to interpret and good for initial benchmarking.
- Random Forests are robust, less prone to overfitting, and generally provide high accuracy.

7. Assumptions

- Features are sufficiently representative of the behaviour of fraudulent and genuine transactions.
- The dataset, despite being imbalanced, provides enough instances of fraud for the model to learn effectively.

8. Model Evaluation and Techniques

Evaluation Metrics:

- **Accuracy:** Measures the overall correctness of the model.
- **Precision:** Indicates the proportion of identifications that were correct.
- **Recall:** Measures the proportion of actual positives that were identified correctly.
- **F1-score:** Balances precision and recall.

Confusion Matrix:

- Used to visualize the performance of the Decision Tree model.

9. Inferences from the Same

- The performance of each model on these metrics will provide insight into their effectiveness in fraud detection.
- A balance between recall and precision is crucial, as both false positives and false negatives carry significant costs in this context.

10. Future Possibilities of the Project

- Implementing more advanced algorithms like Neural Networks or Gradient Boosting.
- Using techniques to handle imbalanced data more effectively.
- Incorporating more features, like user behaviour data, for richer insights.

11. Conclusion

This project demonstrates the feasibility of using machine learning algorithms to detect credit card fraud. The right balance of precision and recall, along with an understanding of the data, is essential for developing an effective fraud detection system.