Subject:

# DATA SCIENCE PROJECT

Title:

**Detecting Credit Card Fraud Using**  **Machine**

**Learning**

Presented by: KAVIBHARATHI SOMASUNDARAM

(ks23aad@herts.ac.uk)

Student ID: 22027052

Date: 17.10.2024

**Research Question**:

How can machine learning techniques be used to accurately detect credit card fraud, and what methods are effective in addressing the imbalance between fraudulent and legitimate transactions?

**Introduction & Project Overview**

**What is the project about?**

The primary goal of this project is to develop a **machine learning model** capable of detecting fraudulent credit card transactions.

**Why is this important?**

* Fraud detection is crucial for financial security.
* The growing volume of online transactions increases the risk of fraud.

The dataset contains **284,807 transactions** where only **492 transactions (0.17%) are fraud**, making it a **highly imbalanced dataset**.

* The model will need to accurately identify fraudulent transactions despite the overwhelming number of legitimate ones.

* **Challenges**: o Imbalanced data makes it difficult for models to learn fraud patterns effectively.

o Fraud detection systems must be highly precise to avoid false positives (flagging legitimate transactions as fraud) and false negatives (missing actual fraud cases).

A green circle with numbers and a red line

Description automatically generated

**Anonymized Features**:

* **Time**: Time elapsed between a transaction and the first transaction in the dataset, used to track transaction intervals.
* **Amount**: The monetary amount of each transaction, which is scaled for analysis to help in identifying fraud patterns.
* **Class**: The target variable indicating whether a transaction is **fraud (1)** or **non-fraud**

**(0)**.

## Data Preprocessing & Class Imbalance

**Data Preprocessing Steps:**

* **Handle Missing Values**: o Ensure any missing data is appropriately filled or removed to maintain data integrity.
* **Scale Time and Amount Features**:
  + **Time** and **Amount** are not on the same scale as other features, so they need to be normalized or scaled to ensure they contribute effectively to the model.
* **Split the Data**: o Divide the dataset into **training** and **testing** sets to train the model on a portion of the data and evaluate its performance on unseen data.

**Addressing Class Imbalance:**

* **SMOTE (Synthetic Minority Over-sampling Technique)**:
  + This technique generates synthetic samples of the minority class (fraudulent transactions) to balance the dataset, ensuring the model has enough fraud examples to learn from.
* **Undersampling**: o Involves reducing the number of legitimate transactions in the dataset to balance the classes. This focuses the model's attention on learning fraud patterns but risks losing valuable data.

## Model Development & Evaluation

**Baseline Models:**

* **Logistic Regression**: o A simple yet effective classification algorithm that estimates the probability of a transaction being fraudulent based on the input features. It’s easy to interpret and often serves as a good starting point for binary classification problems like fraud detection.
* **Decision Trees**:
  + A tree-like model that splits data into branches based on feature values, working well with smaller datasets but prone to overfitting. It offers clear interpretability but can lack the complexity needed to detect subtle fraud patterns.

**Advanced Models:**

* **Random Forest**:
  + An ensemble learning method that builds multiple decision trees and combines their outputs for more accurate and stable predictions. Random Forests help reduce overfitting and improve performance, especially in complex datasets like this one.
* **XGBoost**: o A highly efficient implementation of gradient boosting that builds trees sequentially, optimizing performance by correcting the errors made by previous trees. XGBoost is known for its superior speed and accuracy, especially in handling imbalanced datasets.
* **Neural Networks**: o A deep learning model inspired by the human brain, consisting of multiple layers of neurons (nodes) that learn intricate patterns in the data. Neural Networks are particularly useful for large, complex datasets, but they require more computational power and are less interpretable compared to tree-based models.

**Model Evaluation Metrics:**

"Model Evaluation Metrics" refers to the quantitative measures used to assess the performance of machine learning models, particularly in the context of classification problems like fraud detection. These metrics help you understand how well your model is predicting the target class (e.g., fraudulent vs. non-fraudulent transactions). Here are the key components typically included in model evaluation metrics:

* **Precision**:
  + Measures how many of the predicted fraudulent transactions were fraud. High precision ensures fewer false positives (i.e., legitimate transactions misclassified as fraud).
* **Recall**: o Measures how many actual fraudulent transactions were correctly identified by the model. High recall is crucial in fraud detection, as missing fraud cases can lead to significant financial losses.
* **F1-Score**: o A balanced metric that considers both precision and recall, useful when the dataset is highly imbalanced, as it captures how well the model balances detecting fraud and minimizing false positives.
* **AUC-ROC**:
  + Stands for Area Under the Receiver Operating Characteristic Curve. It measures the model’s ability to distinguish between fraudulent and nonfraudulent transactions. A higher AUC value indicates better discrimination between the two classes.

**Feature Analysis:**

* **SHAP (Shapley Additive explanations)**:
  + SHAP is used to explain the contribution of each feature (like transaction amount and time) to the model's predictions. This helps in understanding which features are most important in identifying fraudulent transactions, providing transparency to the model’s decision-making process.

## Ethical Considerations & Data Management

* **GDPR (**General Data Protection Regulation**) Compliance**: o Dataset contains no personal identifiers, as it is anonymized.

O Complies with GDPR regulations.

* **Data Management Plan**:

**Data Source**: Kaggle, publicly available for research.

**Storage**: Data and code will be stored on GitHub and backed up on OneDrive.

**Version Control**: Weekly commits to GitHub, following consistent file naming conventions.

## Conclusion

In this project, we developed a machine learning model to detect fraudulent credit card transactions, addressing a critical need in financial security. We handled class imbalance using techniques like SMOTE and evaluated multiple models, including advanced algorithms, with metrics such as precision, recall, and AUC-ROC.

Additionally, we employed feature analysis using SHAP to improve model interpretability and provide insights into important predictors. This work contributes to effective fraud detection and enhances our understanding of transaction patterns.