Capstone Project

Problem Statement: Credit Card Fraud Detection

Background

Credit card fraud is a significant issue faced by financial institutions and customers worldwide. Fraudulent transactions can lead to substantial financial losses and undermine consumer trust in digital payment systems. With the increasing volume and sophistication of fraudulent activities, there is a pressing need for robust and efficient detection mechanisms. Data science, particularly machine learning, offers powerful tools to identify suspicious transactions and reduce the incidence of fraud.

Objective

The primary objective of this project is to develop a machine learning model that accurately detects fraudulent credit card transactions using a historical dataset. The model should be capable of distinguishing between legitimate and fraudulent transactions in real time, minimizing false positives and false negatives to protect both customers and financial institutions.

Dataset Description

The dataset contains transactions made by European cardholders in September 2013. It includes transactions that occurred over two days, with 492 frauds out of 284,807 transactions, making the dataset highly unbalanced. The positive class (frauds) accounts for only 0.172% of all transactions.

The dataset is as follows:
https://drive.google.com/file/d/1u_9Zr5cEZYSCn-YhG4Ymrct6oHcNKTNC/view?usp=sharing

The dataset comprises only numerical input variables resulting from a PCA transformation. Due to confidentiality issues, the original features and more background information about the data are not provided. Features V1, V2, … V28 are the principal
components obtained with PCA. The only features not transformed with PCA are:

Time: The seconds elapsed between each transaction and the first transaction in the dataset.

Amount: The transaction amount, which can be used for example-dependent cost-sensitive learning.

Class: The response variable, taking a value of 1 in case of fraud and 0 otherwise.

Given the class imbalance ratio, it is recommended to measure the accuracy using the Area Under the Precision-Recall Curve (AUPRC), as confusion matrix accuracy is not meaningful for unbalanced classification.

Key Challenges

1. Imbalanced Data: Fraudulent transactions are rare compared to legitimate transactions, leading to a highly imbalanced dataset.

2. Feature Engineering: Working with PCA transformed features limits the ability to generate domain-specific features.

3. Model Performance: Achieving high accuracy while maintaining a low rate of false positives (legitimate transactions flagged as fraud) and false negatives (fraudulent transactions not detected).

4. Real-Time Detection: Ensuring the model can process transactions quickly and accurately in a real time environment.

5. Adaptability: The model should adapt to evolving fraud tactics over time.

Approach

Data Preprocessing:

Handle any potential missing values.

Normalize and standardize the 'Time' and 'Amount' features.

Ensure that the PCA-transformed features are appropriately scaled.

Exploratory Data Analysis (EDA):

Analyze transaction patterns.

Visualize the distribution of fraudulent vs. non fraudulent transactions.

Identify correlations between features.

Feature Engineering:

Utilize the 'Time' and 'Amount' features for additional insights.

Consider interactions between PCA components.

Model Development:

Evaluate various machine learning algorithms (e.g., Logistic Regression, Random Forest, Gradient Boosting, Neural Networks).

Address class imbalance using techniques such as SMOTE (Synthetic Minority Over-sampling Technique) or class weighting.

Perform hyperparameter tuning to optimize model performance.

Model Evaluation:

Use cross-validation to assess model robustness.

Evaluate performance using metrics such as Precision, Recall, F1-Score, and particularly the Area Under the Precision-Recall Curve (AUPRC).

Deployment and Monitoring:

Deploy the model in a real-time environment.

Implement monitoring to track model performance and update it as needed.

Success Metrics

AUPRC: The primary metric to measure the model's performance in handling imbalanced data. **Precision and Recall:** Measure the model's effectiveness in detecting fraud without generating excessive false alarms.

F1-Score: Balance between precision and recall.

ROC-AUC: Model’s ability to distinguish between classes.