Fraud Detection Project.

This project aims to build a robust fraud detection system using machine learning techniques. The dataset is highly imbalanced, with fraudulent transactions being significantly rarer than legitimate ones. By leveraging various preprocessing steps, Isolation Forest, Local Outlier Factor, neural networks, and evaluation metrics, this project demonstrates a comprehensive approach to solving the fraud detection problem.

Data.

The dataset contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, … V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

Given the class imbalance ratio, we recommend measuring the accuracy using the Area Under the Precision-Recall Curve (AUPRC). Confusion matrix accuracy is not meaningful for unbalanced classification.

Here is a link to the data.

Methods.

1. **Exploratory Data Analysis.**

   Exploratory Data Analysis (EDA) is the process of analyzing and summarizing the main characteristics of a dataset, often using visual methods. It is a critical first step in any data analysis or machine learning project, as it helps to uncover patterns, spot anomalies, test hypotheses, and check assumptions through statistical summaries and visualizations

2. **Neural Network.**

   A neural network is a computational model inspired by the structure and functioning of the human brain. It is composed of layers of interconnected nodes (neurons) that process data by passing it through a series of transformations to learn patterns and make predictions.

3. **Isolation Forest**

   The Isolation Forest (iForest) algorithm is a machine learning technique used primarily for anomaly detection. It works on the principle of isolating anomalies rather than profiling normal data points, making it highly efficient for detecting outliers in high-dimensional datasets. The Isolation Forest (iForest) algorithm is a machine learning technique used primarily for **anomaly detection**. It works on the principle of isolating anomalies rather than profiling normal data points, making it highly efficient for detecting outliers in high-dimensional datasets

4. Local Outlier Factor (LOF)

The **Local Outlier Factor (LOF)** algorithm is a density-based anomaly detection method that identifies data points with significantly lower densities than their neighbors. It measures the local deviation of density for a given data point compared to its neighborhood.

5. Performance Comparison

Metric	Isolation Forest	Local Outlier Factor	Neural Network
Outliers Detected	675	935	0
Accuracy	0.9976	0.9967	0.7767
Precision (Class 1)	0.31	0.05	0.35
Recall (Class 1)	0.32	0.05	0.78
F1-Score (Class 1)	0.31	0.05	0.49

1. **Isolation Forest** performs significantly better than Local Outlier Factor in detecting outliers:
   - Higher precision, recall, and F1-score for Class 1.
   - Slightly better accuracy.
2. **Local Outlier Factor** struggles to identify outliers:
   - Very low precision and recall indicate it is not effectively distinguishing outliers from normal data in this context.
3. **Neural Network.**

Achieves the highest recall for Class 1 (0.78), indicating it is the most effective model for identifying actual outliers.F1-Score (0.49) is significantly higher than that of the Local Outlier Factor, showing a better balance between precision and recall.