Fraudulent Transaction Detection

Project Description:

The goal of this project is to build a machine learning model that can detect fraudulent transactions in a dataset consisting of 1.75 million transactions made by simulated users through various terminals from January 2023 to June 2023. The data is highly imbalanced, with only a small percentage (0.1345%) of transactions classified as fraudulent. Due to the uneven distribution of classes in the dataset, the performance of the model is evaluated using AUPRC (Area Under the Precision-Recall Curve).

The project involves the following steps:

1. Data Collection: The dataset is obtained from Kaggle, which consists of 10 columns, including transaction ID, date and time of the transaction, customer ID, terminal ID, transaction amount, duration of the transaction in seconds and days, whether the transaction is fraudulent or not, and the type of fraudulent scenario if any.

2. Data Pre-processing: The data is inspected for missing values, outliers, and inconsistencies. The columns with no relevance to the prediction task are removed. The categorical features are one-hot encoded, and the numerical features are standardized.

3. Data Visualization: Exploratory data analysis is performed to understand the distribution of each feature, correlation among the features, and class imbalance. Visualization tools such as histograms, scatter plots, and heatmaps are used for this purpose.

4. Model Training: Several machine learning models are trained on the pre-processed data, including Logistic Regression, Decision Tree, Random Forest,

XGBoost, and LightGBM. The models are trained using a 5-fold cross-validation procedure to optimize the hyperparameters and prevent overfitting.

5. Model Evaluation: The trained models are evaluated on the test set using AUPRC, ROC-AUC, and F1-score. The performance of each model is compared, and the best-performing model is selected for deployment.

Tools and Technologies Used:

- Python 3.8
- Jupyter Notebook
- Pandas
- NumPy
- Scikit-learn
- Matplotlib
- Seaborn
- Plotly

Conclusion:

In this project, we built a machine learning model that can accurately detect fraudulent transactions in a highly imbalanced dataset. We evaluated the performance of several models using AUPRC, ROC-AUC, and F1-score, and selected the best-performing model for deployment. The deployed model can be used to detect fraudulent transactions in real-time, which can help prevent financial losses and improve the security of the payment system.