

FindDefault: Prediction of Credit Card Fraud

Problem Statement

Credit cards are widely used for online purchases and payments, providing convenience but also posing risks such as fraud. Credit card fraud involves the unauthorized use of a credit card or its details to make purchases or withdraw cash. Detecting fraudulent transactions is crucial for credit card companies to prevent financial losses and protect customers.

In this project, we aim to build a classification model that can predict fraudulent transactions. The dataset consists of transactions made by European cardholders in September 2013, covering a two-day period. Out of 284,807 transactions, 492 are fraudulent, making the dataset highly imbalanced (fraud cases account for only 0.172% of all transactions). Addressing this imbalance is key to building an effective model.

Project Objectives

The project follows a structured approach to solving the problem:

1. Exploratory Data Analysis (EDA): Analyze and visualize the dataset to identify trends, relationships, and potential anomalies.
2. Data Cleaning: Handle missing values, standardize formats, and remove outliers.
3. Dealing with Imbalanced Data: Apply appropriate methods such as oversampling or undersampling to balance the dataset.
4. Feature Engineering: Create and transform features to enhance model performance.
5. Model Selection: Choose the most suitable machine learning model.
6. Model Training: Split the dataset into training and testing sets and optimize model parameters.
7. Model Validation: Evaluate performance using appropriate metrics to ensure generalization.
8. Model Deployment: Make the trained model available in a production environment.

Deliverables

Submit the following items in a zip file:

- Report (PDF)
 - Description of design choices and performance evaluation
 - Discussion on future improvements
- Source Code
 - Code implementing the pipeline

Tasks and Activities

The project includes the following key activities:

- Load and process the time-series data from the CSV file.
- Conduct Exploratory Data Analysis (EDA) to assess data quality and handle missing values/outliers.
- Convert date features to the correct data type.
- Balance the dataset to address class imbalance.
- Perform Feature Engineering and feature selection.
- Split data into train/test sets using an optimal sampling distribution.
- Select appropriate evaluation metrics.
- Implement model selection, training, prediction, and assessment.
- Tune hyperparameters for improved performance.
- Plan for model deployment.

Success Metrics

The project will be considered successful if:

- Model accuracy on the test dataset is greater than 75%.
- Hyperparameter tuning is performed.
- Proper model validation is conducted.

Bonus Points

- Providing a well-documented zip file with a README explaining installation and execution.
- Documenting how the solution benefits the company.

Data Source

The dataset for this project can be accessed via `creditcard.csv`.