Credit Card Fraud Detection

Abstract:

It is important that credit card companies are able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase.

Design:

This project originates from Kaggle's "Credit Card Fraud Detection" dataset. The data is provided by Machine Learning Group – ULB.

Data:

The dataset contains 284,807 records with 31 features. It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, they 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-dependent cost-sensitive learning. Feature 'Class' is the response variable, and it takes value 1 in case of fraud and 0 otherwise.

Algorithms:

Feature Engineering

1. Converted raw data into categories to answer multiple questions.
2. Balanced the dataset.

Models

Logistic Regression, K-Means were used before settling on Logistic Regression as that model with the highest score.

Model Evaluation and Selection

Logistic Regression Model:

Accuracy on Training Data: 0.940279542566709

Accuracy on Test Data: 0.9187817258883249

Tools:

- Numby and Pandas for data manipulation.
- Matplotlib for plotting.
- Scikit-learn for modeling

Communication:

Slides will be presented next to this document.