FindDefault (Prediction of Credit Card fraud)

Problem Statement:

A credit card is one of the most used financial products to make online purchases and payments. Though the Credit cards can be a convenient way to manage your finances, they can also be risky. Credit card fraud is the unauthorized use of someone else's credit card or credit card information to make purchases or withdraw cash.

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.

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.

We have to build a classification model to predict whether a transaction is fraudulent or not.

Datasets

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-senstive learning. Feature 'Class' represents the class labelling, it takes value 1 in case of fraud and 0 otherwise.

Solution Statement

The objective is to create simple models like logistic regression, KNN, random forest and maybe others to compare how they perform regarding the metric chosen (AUC) for the task of predicting fraudulent credit card transactions. After that, I will create one or more ensemble model(s) using some of the simple models as a way of further enhancing the result.

Benchmark

The benchmark to be used is the best model among the simple models used. Usually, most people just try to find one unique model that has the best prediction and stick to this model.

Evaluation

The main metric to be evaluated is the Area Under the Receiver operating characteristic (ROC) curve after balancing the training set. The ROC curve is a plot of the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings, showing the trade-off between TPR and FPR. When using normalized units, the area under the curve (often referred to as simply the AUC) is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming 'positive' ranks higher than 'negative').

Project Design

Workflow is as follows:

- Download the dataset from Kaggle website
- Perform exploratory data analysis on the dataset to gain insights from the data structure (look for outliers, list the factors in order of relevance, etc).
- Balance the classes using one or more strategies (undersampling, oversampling or SMOTE).
- Create and train different simple models commonly used on supervised training tasks (like logistic regression, KNN, random forest, naïve Bayes, SVM).
- Use the best result of the previous models as a benchmark utilizing the area under the ROC curve as a metric.
- Use combinations of the simple models in an ensemble model to get a better result compared to the benchmark.