Capstone Proposal

Abhishek Ghosh
October 21st, 2024

Proposal

Domain Background

Credit card fraud is a type of crime that is on the rise (link) and this is a risk to everyone who uses this payment method. I surely know how annoying is to find that your credit card was frauded and used by criminals to buy anything they want with your hard-earned money. It already happened to me 2 times even though I am not an easy target for scams. I have always taken all the recommended measures for a regular citizen to avoid frauds and yet it wasn't enough to protect me against this crime.

Despite several papers on this theme were written in the last years, banks still struggle to correctly classify normal and fraudulent transactions because swindlers are always coming up with new strategies to disguise fake transactions. Delamaire, Abdou & Pointon (2009) list ways of detecting fraudulent transactions like pair-wise matching, decision trees, clustering techniques, neural networks, and genetic algorithms.

Problem Statement

The problem chosen for this project is to predict fraudulent credit card transactions by using machine learning models. The models are going to be trained using supervised learning. A dataset containing thousands of individual transactions and their respective labels was obtained from Kaggle website (link).

Datasets and Inputs

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 Model

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. I hope I will get a better performance using an ensemble model.

Evaluation Metrics

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'). (Fawcett, Tom (2006))

Project Design

(approx. 1 page)

The summarized workflow of this project is as follows:

- Download the dataset from Kaggle website (link).

- 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 (under sampling, 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.
