Machine Learning Engineer Nanodegree

Capstone Proposal

Mustafa Adel
April 29th, 2019

Domain Background

Context

Credit card plays a very important rule in today's economy. It becomes an unavoidable part of household, business and global activities. Although using credit cards provides enormous benefits when used carefully and responsibly, significant credit and financial damages may be caused by fraudulent activities. Many techniques have been proposed to confront the growth in credit card fraud. However, all of these techniques have the same goal of avoiding the credit card fraud; each one has its own drawbacks, advantages and characteristics. In this project, I'm going to work on a real dataset has been collected in September 2013 to build *classification model (using supervised learning) to predict the fraudulent transaction.*

Acknowledgements

The dataset has been collected and analyzed during a research collaboration of Worldline and the Machine Learning Group (http://mlg.ulb.ac.be) of ULB (Université Libre de Bruxelles) on big data mining and fraud detection. More details on current and past projects on related topics are available on https://www.researchgate.net/project/Fraud-detection-5 and the page of the DefeatFraud project.

- *Check more about the dataset here.*
- *Related research paper here.*

Problem Statement

In this project. I'll deal with anonymized and imbalanced dataset. My goal in some steps

1. *Understanding the dataset.*
2. *Dealing with imbalanced dataset.*
3. *Preprocessing.*
4. *Evaluating model.*
5. *Using Evaluation Metrics.*

Datasets and Inputs

The datasets 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-dependent cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

Solution Statement

In this project, I'll use Correlation matrix to understand the correlation coefficients between features (especially V1 to V28) and target label. And a resampling method to deal with imbalanced dataset to create a 50-50 ratio (fraud-real) transactions. Using SVM classifier to fit the resampled data and predict the labels.

Benchmark Model

Fitting SVM classifier model on imbalanced data produces a model with high accuracy matrix but poor recall score and high precision score "Recall score is which we care about in this project". After resampling the dataset, the recall score became high score and very low precision score which we needed. And these steps will be shown in project.

Evaluation Metrics

Precision-Recall is a useful measure of success of prediction when the classes are very imbalanced. In information retrieval, precision is a measure of result relevancy, while recall is a measure of how many truly relevant results are returned

- *Precision Score: True Positives/(True Positives + False Positives).*
- *Recall Score: True Positives/(True Positives + False Negatives).*
- *Precision as the name says, says how precise (how sure) is our model in detecting fraud transactions while recall is the amount of fraud cases our model is able to detect.*

Project Design

1) Understanding our data
   a) Data Exploration
   b) Data Visualization
2) Preprocessing
   a) Data Cleaning
   b) Data normalization
   c) Shuffle and split the data
   d) Evaluating model on imbalanced data
   e) Resampling data
3) Evaluating model
4) Test our model using evaluation metrics