Machine Learning Engineer Nanodegree

Capstone Proposal

Rodrigo Miike da Silva
August 19th, 2017

Proposal

Domain Background

In this project, I'll try to solve a problem that affects most of the financial entities. Fraudulent transactions happen with a high frequency and the purpose is to obtain goods without paying for it. They are committed through identity/physical card theft or when personal data about accounts are leakaged[1]. Although the rate of fraudulent transactions are lower than 1%, the financial losses are huge due to the high amount value of the transactions[1].

Popular machine learning techniques such as supervised and unsupervised learning could be used to aid this problem. However, each approach have your own challenge and assumptions. Under the condition that fraudulent transactions incidence are highly unbalanced compared with legit transactions, anomaly detection is an amazing approach to find points that are not expected from the pattern.[3]

Chandola, Banerjee and Kumar[3] provide a research list with anomaly detection technique and its applications. Thus, providing a general idea about how to apply anomaly detection under a domain. Barkan and Averbuch[4] propose the application of a statistical model on an original representation of the data in order to detect anomalies.

Motivation

Nowadays I work in a place that handles with credit, debit and voucher cards. The company doesn't have any fraud detection system, so It depends a lot on outsourcing services. Since I am willing to jump on the machine learning world, I see this as an opportunity to develop what I learned on Nanodegrees program.

Problem Statement

The problem consists in find out which transaction is legit or fraudulent due to minimize financial losses. Since there is an unbalance between them, an approach through anomaly detection could be effective. Moreover, due to the unbalance, a recommended metric would be Area Under the Precision-Recall Curve (AUPRC), because the Confusion Matrix isn't significant over unbalanced classification[5].

Datasets and Inputs

The dataset is available in Kaggle[5] providing resource for study about fraud detection. The dataset contains credit card transactions made by european cardholders in September 2013. The data provides 497 frauds out of 284807 transactions in total (0.172%).

The variables were masked by PCA transformation in order to hide the original information due to confidentiality issues. The features named as V1 to V28 are the principal components of a PCA transformation, the amount and time variables were the only ones that have not been transformed by PCA. The last feature “Class” labels if the transaction is fraudulent or legit.

Solution Statement

The model should be able to give an insight about when a transaction could be or not a fraud through a probabilistic score. This score gives degrees of belief that a transaction is a fraud.[3] So, giving an opportunity for a future analysis.

Since the dataset is labeled, I could take an advantage of this feature to measure the degree of importance of each feature through decision trees.
Working on the most important features, we would be able to explore through visualizations in order to capture more insight about the data.

By choosing the most important feature, I will apply some statistical models such as Expectation Maximization with Normal Distribution and t Distribution in order to capture the sample parameters.

Benchmark Model

First of all, I’ll apply the same approach used on “Titanic Survival Exploration” to create an initial Benchmark Model by assuming all the labels are legit. Afterward, I’ll use some of popular machine learning models such as linear regression, SVM, decision trees as reference to measure the performance of the model.

Evaluation Metrics

Due to unbalanced nature of the dataset, the AOPRC would be a better approach to measure the performance since the precision is highly affected due to the false positives. The AOPRC is a scalar value that summary of the PR curve. The summary is calculated by integrating out the area under the PR curve.
On Scikit-learn library, the *average_precision_score* function calculates this area by receiving the labels and the predicted labels.

Project Design

First of all, I'll explore the dataset like plotting visualizations in order to take an overview about the data. Then, I'll try to do a feature selection. Since there are a lot of features, it would be efficient to use only the ones that most represent the data.
Then, I'll try to create a benchmark model by using linear regression, SVM and decision trees.
The most challenging part is how can I split the dataset into test and train set, since the ratio of fraudulent transactions is only 0.172%.
Since I am using as benchmark supervised learning techniques, I have no choice but splitting the dataset through the usual method. But, when I start to work with Expectation Maximization (unsupervised technique), I will work with two approaches:

- Generating a set with all the fraudulent transactions and then adding legit transactions about 4 times the fraudulent transactions.
- Generating a set with only legit transactions and then after estimating the parameters, I could add the fraudulent transactions to see how the model would behave.

After that, I will be applying two models of EM.

- One with a multivariate normal distribution (already implemented on scikit learn library).
- A multivariate t distribution implemented by me, but proposed by D. Peel and G. J. McLachlan[6].

Finally, I would be comparing and summarizing the results.

References

[1] Wikipedia - Credit Card Fraud
[2] Dunning, T., Friedman, E.: Practical Machine Learning - A new look at Anomaly Detection.
[3] Chandola, V., Banerjee, A., Kumar, V.: Anomaly Detection - A Survey.
[4] Barkan, O., Averbuch, Amir.: Robust Mixture Models for Anomaly Detection.
[5] Kaggle - Credit Card Fraud Dataset
[6] Peel D., McLachlan G. J.: Robust Mixture modelling using the t distribution.