Capstone Proposal: Anomaly Detection

Luis Chapa

May 29, 2018

Proposal

Domain Background

Banking and its transactions that they manage every day in the electronic medium, like loans, credit and debit purchases through internet and in the stores, cards that could be stolen from the clients pocket, payment orders to other accounts and other banks, etc. all of them are at risk of robbery or fraud.

So, with this project is intended to reduce that risk by detecting anomalies that could represent a risk to the client. An anomaly is defined as a discordant observation in the dataset, which is far away from the media or mean. This data point is generated by another process.

It is important to make a difference in between an outlier and anomaly. The outlier is also away from the mean or median, but it is still generated by the same process, it is just a valid data point in the data set, but away from the others. Anomaly is generated by a different process.

By detecting this anomaly points, the risk fo fraudulent transactions is reduced. But also detecting the outliers points it is import in order to let the clients to use their bank services for unusual purchases or transactions.

Furthermore, the anomaly detection algorithms could be applied to many other fields in the industry. One personal motivation I have for knowing them is to apply them in the enormous quantity of data generated by the manufacturing industry around the city I live in. So, detecting all those anomalies make improve the quality of the products they ensemble as well as get an optimal performance of those processes.

Problem Statement

Each user has a kind of constant behavior in their use of a bank services. That constant behavior might be modeled mathematically, so when a data point is outside of that model, that point might set on a kind of alarm to the bank managers by addressing that point as an anomaly. The bank managers might activate a protocol in order to validate the anomaly as an outlier, a totally authorized transaction by the user. Or, in the other hand set it as a fraudulent transaction and ban the account, not authorize the transaction or whatever it is necessary to avoid the robbery and catch the thieves if that is possible.

Fortunately for the bank has not suffered fraudulent operations yet. So, they recently are investing in this system as a preventing one, so no statistics have been collected yet in order to describe the problem.

Datasets and inputs

The bank has provided a sample of their database, which is composed as follow:

1. Clients, a list of some of the clients.
2. Those clients are bound to accounts.
3. Those accounts are bound to loans, transactions and payment orders.
4. Those clients are bound to a demography location.
5. Those clients are bound to a card type.

With the above anonymized database extract from the bank the benchmark model as well as the solution model will be trained. From that database which is normalized, a final one-table dataset will be retrieved and feed in the models.

The dataset was got from the following link:
https://data.world/lpetrocelli/czech-financial-dataset-real-anonymized-transactions
After having browsed a while around the internet, this database sample was the one most seemed to the one required to train the models.
As well, this database accomplishes the transactional featured I wanted to use for my project.

Solution statement

The solution model will be working as follows: a **deep auto-encoder** will be trained with the dataset, as if each observation was normal. In this context normal means non-anomaly. Then a threshold on the anomaly-score will be set in order to identify the anomalies among the dataset observations.

This solution is implemented from the following book citation:

> Zocca, Valentino, et.al. *Python Deep Learning*. Birmingham, UK. Pack Publishing LTD. April 2017.

Benchmark model

The benchmark model will be an semi-supervised clustering model. As the bank is not collecting information about what is a fraudulent transaction and what it is not, the assumption is made on this dataset that each one of the transactions are normal (non-anomaly) and then try to identify the ones that out stand from the median or mean fo the whole dataset.

Evaluation Metrics

It is proposed to evaluate the benchmark model versus the solution model the use of:

1. Confusion Matrix, to concentrate the results of the both models.
2. Accuracy in order to measure the false alarms.
3. Recall in order to measure the missed detections.
4. F-score for get a good balance from the both above metrics.

It was mentioned that the bank is not recording the fraudulent transactions until now. So, the benchmark versus solution models will be used to evaluate the performance of solution model.

Project design

This project will be containing the following sections:

1. **Intro**: The problem is described.

2. **Data collect**: The source of the dataset is described.

3. **Data preprocess**: The preprocessing needed in order to make the algorithms to work optimally is described and applied to the dataset in detail. Step by step the code and a result will by explained.

4. **Exploratory and visualization**: Some techniques will be applied to the multidimensional dataset in order to make it visualizable in 2D and plot some graphs that will help fully understand how the dataset is working. Thus, the exploratory analysis can be performed.

5. **Benchmark model**: Because the dataset haven't collected labeled data at all. It is proposed a benchmark model using Clustering algorithms. This will help to have a baseline in order to check up the performance of the solution model.

6. **Solution model**: It is proposed a deep auto-encoder model that will be monitoring transactions. The algorithm codes the observation through the deep auto-encoder by filtering a compressing. Then it retrieves the observation. Then, by comparing the distance, or difference between the original data observation and the result decoded one, and if it is over the predetermined threshold of the anomaly score an alarm might be set on for the managers of the bank to take action on the case.

7. **Metrics**: It is proposed for the evaluation of the solution model (versus de benchmark model) the use of a Concussion Matrix. There you can count the outcome of the solution model and compare against the outcome of the benchmark model in order to see how well the solution model it is performing.

Here there will be only two possible labels for each observation: fraud or normal transaction. So, by counting the findings of each label and comprising with the benchmark model you can calculate accuracy, recall and the F-score metrics.

8. **Conclusion**: Here, all conclusions will be made, in regards of the algorithms applied to the anomaly detection, and, as well, some conclusions about the general motivation of the entire nano degree.