Synthetic Financial Datasets for Fraud Detection

(Synthetic datsets generated by the PaySim mobile money simulator)

Purpose of Project:

When it comes to fraud detections, we required faster results. Machine learning is like having several teams of analysts running hundreds of thousands of queries and comparing the outcomes to find the best result - this is all done in real-time and only takes milliseconds. As well as making real-time decisions, machine learning is assessing individual customer behaviour as it happens. It is constantly analysing 'normal' customer activity, so when it spots an anomaly it can automatically block or flag a payment for analyst review.

Machine learning systems improve itself with larger datasets because this gives the system more examples of fraud and normal transactions or genuine and fraudulent customers. This means the model can pick out the differences and similarities between behaviours more quickly and use this to predict fraud in future transactions.

Machine learning is like having several teams running analysis on hundreds of thousands of payments per second. The human cost of this would be unaffordable on the contrary the cost of machine learning is just the cost of the servers running. Machine learning does all the repetitive work of data analysis in a fraction of the time and gives the result. Unlike humans, machines can perform repetitive, tedious tasks 24/7 and only need to escalate decisions to a human when specific insight is needed.

Problem Statement:

Use machine learning to detect fraudulent transactions from online payment methods.

There are a few publicly available datasets on financial services and specially in the emerging mobile money transactions domain. Financial datasets are important to many researchers and to us performing research in the domain of fraud detection. Part of the problem is the intrinsically private nature of financial transactions, that leads to no publicly available datasets.

The idea of the project and required dataset is taken from Kaggle.com website.

Source: https://www.kaggle.com/ntnu-testimon/paysim1

Kaggle contains a synthetic dataset generated using the simulator called PaySim as an approach to fraud detection type of problem. PaySim uses aggregated data from the private dataset to generate a synthetic dataset that resembles the normal operation of transactions and injects malicious behaviour to later evaluate the performance of fraud detection methods.

PaySim produce a model of mobile money transactions based on a sample of real transactions extracted from one month of financial logs from a mobile money service implemented in an African country. The original logs were provided by a multinational company, who is the provider of the mobile financial service which is currently running in more than 14 countries all around the world. This synthetic dataset is scaled down 1/4 of the original dataset and it is created just for Kaggle.

Scope of work:

I am going to use anaconda with jupyter notebook for perfuming machine learning and data analysis activities. Main language for this project will be python. Initially this project will be running on my local computer and as an outcome I will be predicting that whether the transaction is fraud or genuine.

The available dataset has only one CSV file with 470 MB size. This .csv file contains 63,62,620 rows and 11 columns in it. We will focus on generating "isFraud" label after applying machine learning algorithm. This dataset is huge in size and required lots of data cleaning and data preparation process. That will be done using pandas and numpy library functions. After that I must show some visualizations to get the insights of dataset and relation between features and labels. In this part I will use mostly matplotlib and seaborn.

After that I will split the dataset for training and testing. For this I will use scikit learn library. Then, using cross validation among various classification algorithms I will find out the most suitable algorithm for given dataset to build better model. Here for evaluation process I will use confusion matrix to check model's accuracy.

References

This work is part of the research project "Scalable resource-efficient systems for big data analytics" funded by the Knowledge Foundation (grant: 20140032) in Sweden.

PaySim first paper of the simulator:

E. A. Lopez-Rojas , A. Elmir, and S. Axelsson. "PaySim: A financial mobile money simulator for fraud detection". In: The 28th European Modeling and Simulation Symposium-EMSS, Larnaca, Cyprus. 2016