Why Fraud Detection

Fraud detection whether in commerce, finance or any other domain has become critical. Beymani provides a set of algorithms for fraud detection. The solutions and algorithms in *beymani* are equally applicable for cyber security breach detection.

Introduction

*Beymani* offers a suite of fraud detection algorithms on Hadoop and Storm. As far as I know, *beymani* as an open source solution has the largest offerings of fraud detection algorithms.

Any data point detected as fraudulent is only a prediction. Only a manual process can verify whether it’s a fraudulent and make the final call.

Bit of theory

Fraud detection is a use case for outlier detection in machine learning. In layman’s term, outliers are data points that are not like most other data points in a data set. There are various algorithms for detecting outliers.

They come under two broad categories, depending on whether the detection algorithm works on individual data points or sequence of data points.

For individual data point based algorithms, beymani supports the following

1. Distribution model
2. Proximity analysis
3. Relative density analysis

For data sequence based algorithms, beymani supports

1. Markov model

More will be added soon

Real time fraud detection

Many of the algorithms listed above are not applicable for real time prediction. The algorithms either build model or work with the data directly to detect outliers.

Real time detection is only possible for those algorithms that build a model in Hadoop batch mode, so that the model can be used to detect using the model and the incoming real time data stream.

For example, proximity based techniques use all the data to detect outliers. It’s not feasible to process all the in real time. So these techniqies can not be used for real time detection.

On the other hand, using sequence based algorithms, we can build a Markov model offline using Hadoop and historical data. This model can be used in real time for incoming data stream to detect fraud.

Configuration

*Beymani* believes in giving power in the hands of the end users. There are many configuration parameters to control the behavior of the different algorithms

Data Format

The input and output data format is CSV. The meta data i.e., the meaning of different columns in the input data is specified through a JSON file.