Final Project

1. Nowadays, fraud detection is one key aspect of many businesses and Machine Learning algorithmns are helping Data Scientist and Engineers tackle this problem. Banks, credit card providers, online shops and many other types of services have to face daily problems with attempts of fraud throughout the world. One of the most famous cases of corporate fraud in our recent history is the Enron scandal that happened in 2001. Enron Corporation was an American energy company with a market value of billions of dollars. The fraud occurred specially on the financial division of corporation, with key directors of the company falsifying numbers to investors and general public. Some directors and high position employees were involved in the fraud and were late charged with crimes. The goal of the current project is to access data generated from employees that were part of the fraud inside Enron (labeled Person of Interest, a.k.a POI) and regular employees that were not prosecuted criminally. The dataset contains different types of numeric data like salary, number of email exchanged with POI, amount of stocks, etc and also some text data contained on the emails exchanged by respective employees. We will focus on the numeric data to do our exploratory data and model prediction, with the goal to predict the tested subjects as being POI or not.

As our database had many features, spotting outliers is not straight foward. One way to do it is by using PCA to reduce the number of dimensions and try visualizing it.

PCA plot

Also we can measure the z-score to find values that show 3 standard deviations more or less in relation to the mean. By doing this we spotted a row named ‘TOTAL’, which is probably an aggregation of all the other rows of the Dataframe, so it was removed.

2)