Identifying Fraud from Enron Emails using Machine Learning

Enron was one of the prosperous companies in the United States in 2000. However, it had gone bankruptcy due to frauds committed by some of its CIOs by 2002. Therefore, a federal investigation was carried out to find the corruption in the company. As a result, a large amount of confidential details including email and financial details of top executives became public records. In this project, the machine learning algorithms along with Python data handling techniques used to identify persons of interests (POI) by analyzing financial and email data made public as a result of the Enron scandal.

Overview of the dataset

In the preprocessed dataset, the email and financial information are confined into followings 21 features for each person investigated.

'poi', 'salary', 'deferral\_payments', 'total\_payments', 'loan\_advances', 'bonus', estricted\_stock\_deferred', 'deferred\_income', 'total\_stock\_value', 'expenses', 'exercised\_stock\_options', 'other', 'long\_term\_incentive', 'restricted\_stock', 'director\_fees' to\_messages', 'email\_address', 'from\_poi\_to\_this\_person', 'from\_messages', 'from\_this\_person\_to\_poi', 'shared\_receipt\_with\_poi'

The dataset contains records of 146 persons (thus 146 records of for each feature (including missing values)).

Size of the dataset before,145 and, after, 136,running ‘featureFormat’ function

('nan count salary', 51)

('nan count bonus', 64)

('salary', 42)

('to\_messages', 50)

('total\_payments', 12)

('bonus', 55)

('expenses', 42)

('loan\_advances', 133)

('from\_messages', 50)

('from\_this\_person\_to\_poi', 70)

('poi', 118)

('deferred\_income', 88)

('shared\_receipt\_with\_poi', 50)

('from\_poi\_to\_this\_person', 62)

('number of poi:', 118)

('number of nonpoi:', 18)

Handling outliers

During the preliminary examination of the financial data in the dataset, some calculated values were identified. For example, the totals of the entire columns observed as records(that might have calculated when the dataset was in spreadsheet format). The totals of columns were shown in the main dictionary as ‘TOTAL’. The ‘TOTAL’ field deviate all the data because the total values are very high compare to other values. Therefore, the ‘TOTAL’ field removed form the dictionary using following Python code.

data\_dict.pop('TOTAL', None)

To identify whether there are outliers, boxplots were generated for each numerical feature in the dataset. The example boxplots are shown in Figure 1 and Figure 2. Based on the boxplots results, there are more potential outliers. However, these outliers maybe valid data points and can be helpful to identify a fraud if any. Therefore, these data points were kept in the dataset.

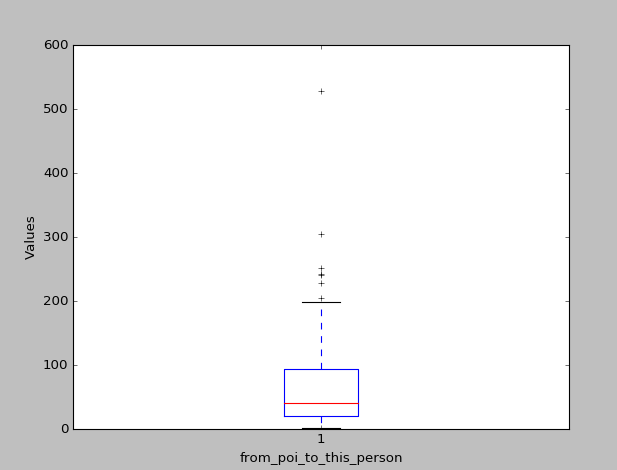


Figure 1: Boxplot for form\_poi\_to\_this\_person vs values

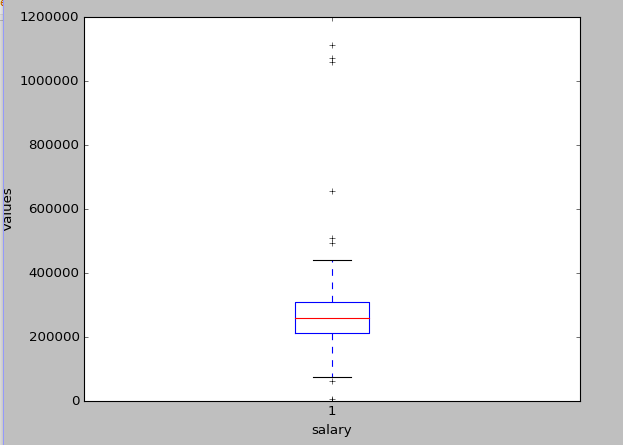


Figure 2: Boxplot for salary vs values

From above features following important feature list is selected in order to detect any fraud if any. All the selected features are numerical.

features\_list = ['poi','salary', 'deferral\_payments', 'total\_payments', 'loan\_advances', 'bonus', 'deferred\_income', 'total\_stock\_value','expenses', 'long\_term\_incentive', 'to\_messages', from\_poi\_to\_this\_person', 'from\_messages', 'from\_this\_person\_to\_poi', 'shared\_receipt\_with\_poi']

Fracture selection

Feature extraction

PCA: Unsupervised Dimentiality reduction. Find features most explaining labels

regularization