Identity Fraud From Enron Email

Awad Bin-Jawed

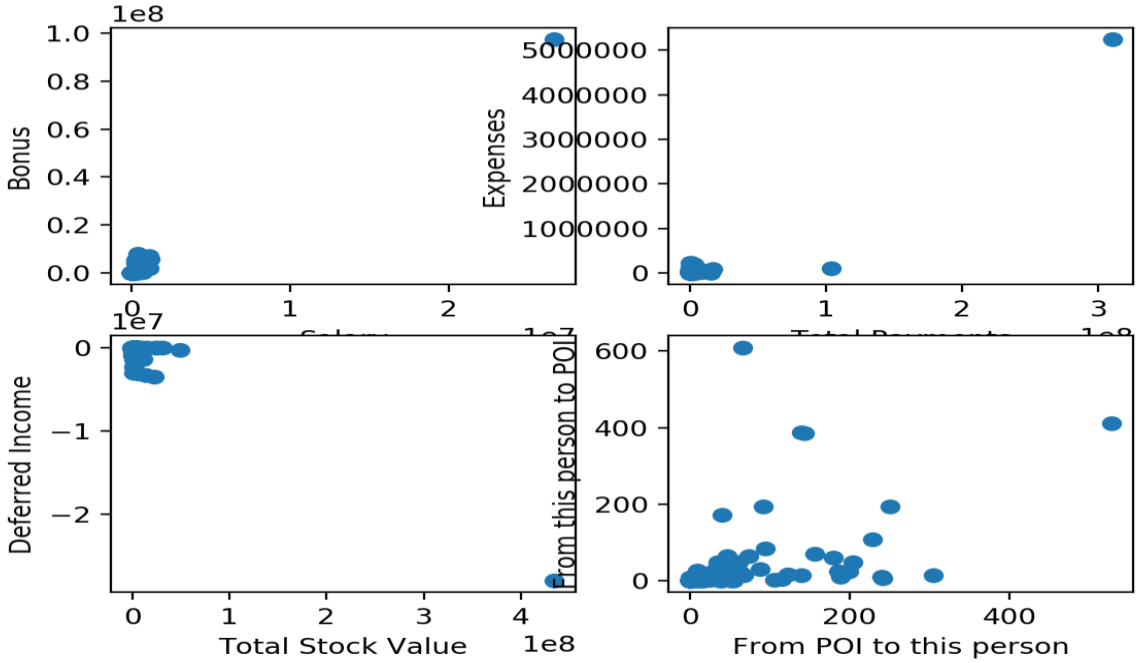
Introduction

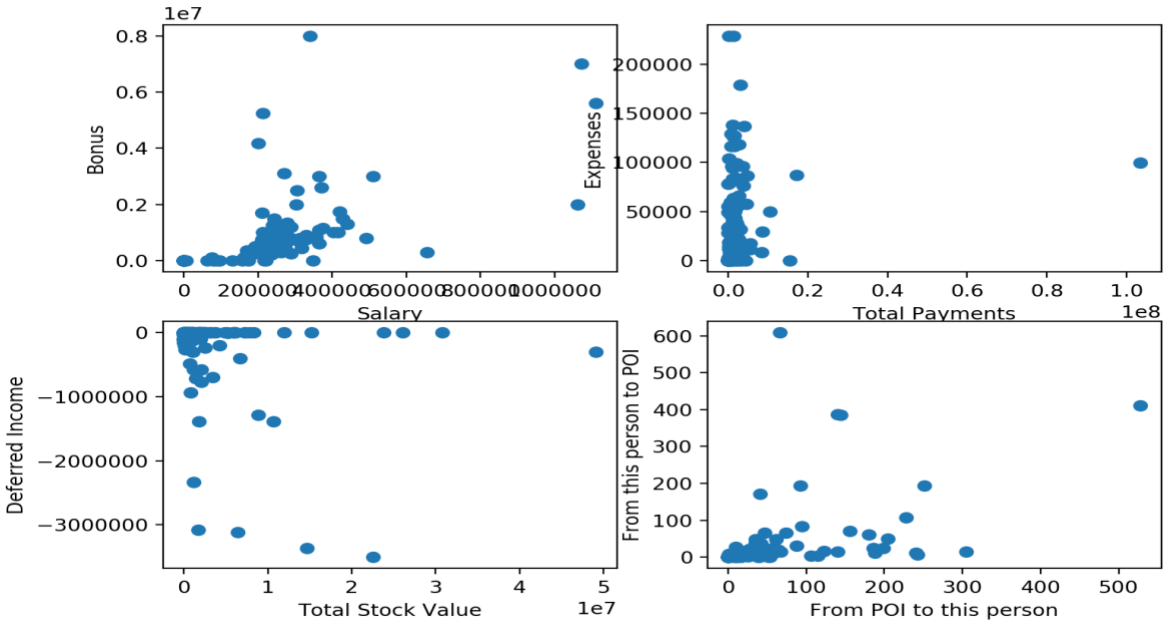
Enron Corporation was an American energy, commodities, and services company based in Houston, Texas. In 2001, it was revealed that Enron's reported financial condition was sustained by institutionalized, systematic, and creatively planned accounting fraud, known since as the Enron scandal. It was the largest case of corporate fraud in US history.

The purpose of the study is to identity Person of interest (POI) who were involved in the financial fraud based on the enron financial and email data set. POI can be one from these three categories indicted, settled without admitting guilt, or testified in exchange of immunity. In this project, capability of several machine learning algorithm are utilized in order to classify POI of each user. Data set contains 146 users each having 20 features.

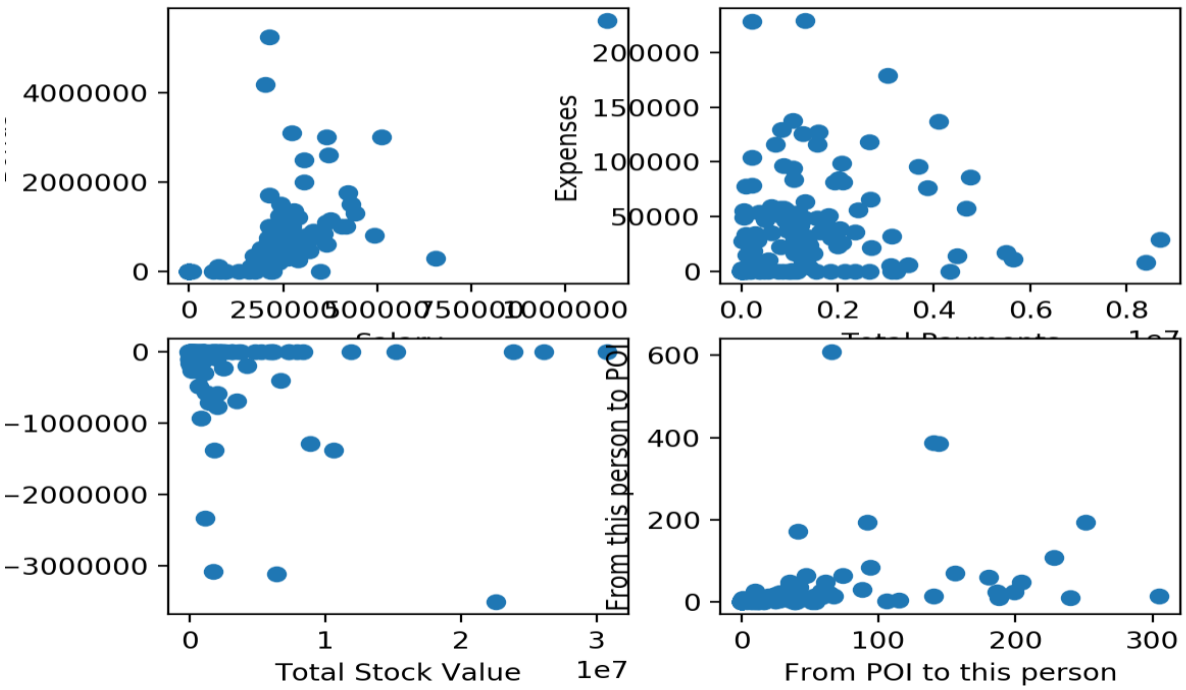
Outlier Removal

From the plots it is found that there is a unique value for bonus of a very high salary, this could be potential outlier. We have removed that value from the data.





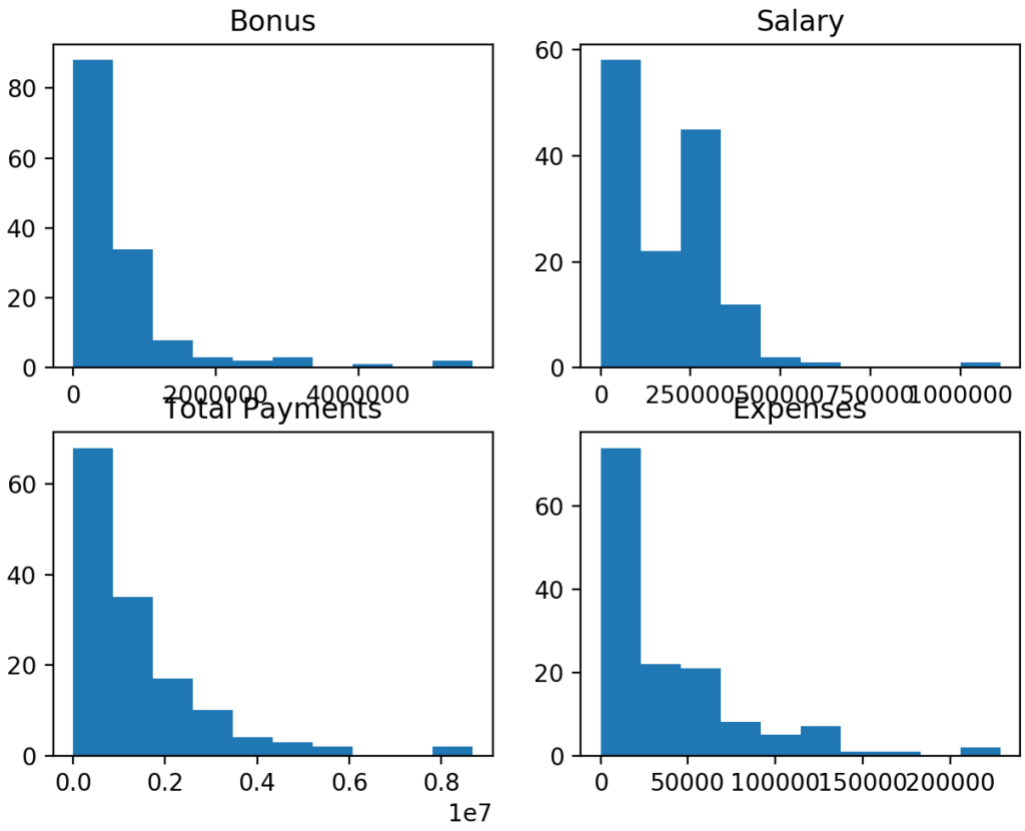
There is also a unique point denoting total payments for a very high expense. This could be the total payments made and total expenses in the company and hence, can be removed from the data set.



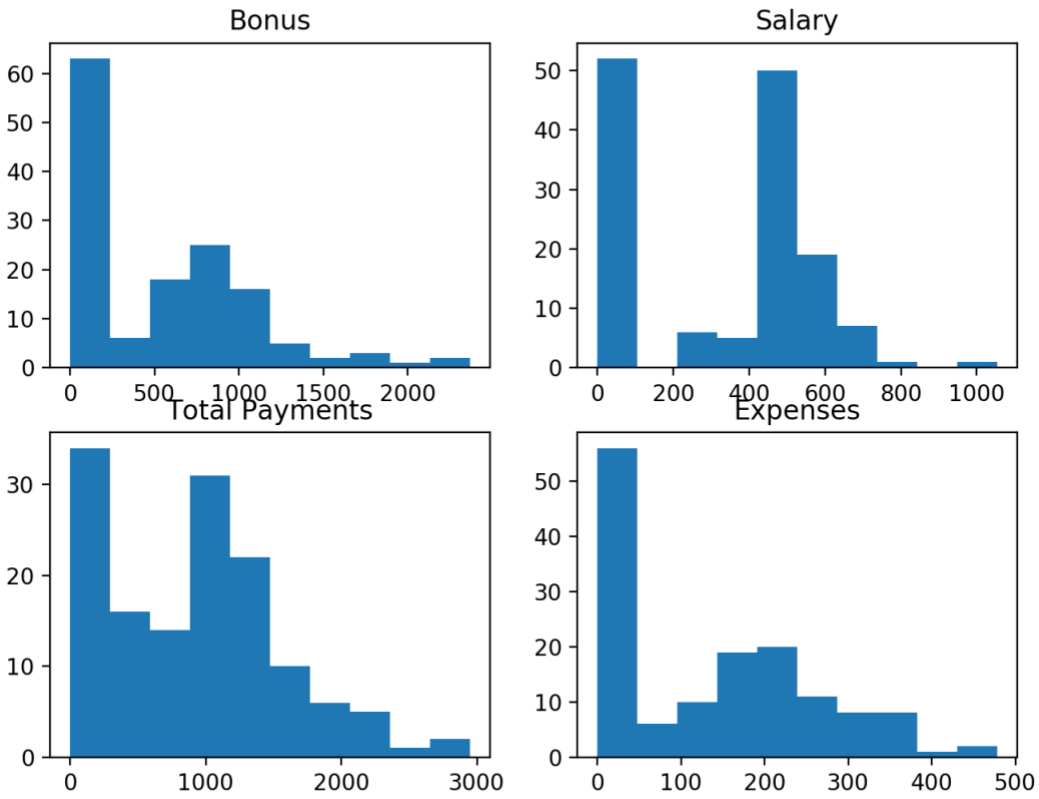
Final data set looks more cleaner and well spreaded after removal of potential outliers.

Feature Transformations

We have selected poi, salary, bonus, total payments, expenses, total stock value, deferred income, from poi to this person, from this person to poi, shared receipt with poi, restricted stock and long term incentive. From these data it is found that most of the features are right-skewed so we have transformed the bonus, salary, total payments, expenses and total stock value to square root.



Histogram showing right-skewness on bonus, salary, total payments and expenses feature.
It is not-normal, we have applied square root transformation. Log transformation can give infinity to some data points so we have not used the box-cox transformation which is generally more suitable for removing skewness from the data. The histogram below shows distributions are less skewed and looks more normally distributed after transformations.



Feature Engineering

We are interested to find out additional information from the given dataset. We would be interested to know how the salary, bonus and expense data are correlated, we feel that there would be a trend among POI and non-POI in these two relations with the assumptions that POI would be expending more, so we have constructed salary and bonus to expense ratio. We are also interested to know the ratio of messages sent among POI with non POI. Frequency ratio of such conversation would be a good feature to analyse. Also, total stock value of the person based on total income including salary and bonus, so we have added one more feature to find the ratio of total income to total stock value. As there would be a chance that POI might have invested more in the stock.

Model Fitting

Feature Scaling and Reduction

Some of the classifier are parametric and some are non-parametric. Parametric classifier assumes data is centered and normal. Considering this, we have applied centering and scaling

to each predictive variable. However, non-parametric classifier doesn't require data to be scaled or normally distributed.
Some feature may exhibit the similar information and can have high correlation. We have applied Principal Component Analysis (PCA) to reduce the feature space. We have selected first few Principal Component that explains maximum information about the feature space. Each PC is the linear combinations of features. PCA works well when data is centered and scaled.

Model Selection and Results

We have selected the following classification algorithm to fit the model:

1. Naive Bayes Algorithm
2. Logistic Regression
3. Support Vector Machine (SVM)
4. K-Nearest Neighbour Classification (kNN)
5. Linear Discriminant Analysis (LDA)
6. Random Forest Classifier (RF)

Naive Bayes is the parametric classifier with generally works well when data is normal. As we have transformed the data to mitigate the skewness from the data. Logistic Regression, LDA and Support Vector Machine are linear classifier, they build a linear boundary among two classes. k-KNN is non-linear classifier and random forest is also nonlinear classifier with ensembled learning approach. It builds multiple decision tree and predicts the class from the mode of the outputs from each tree.

We performed cross-validation on each model using 10-fold cross validation approach.
Results are summarized in the table below:

Model	**Precision**	**Recall**	**Accuracy**	**Tuning Parameters**
Linear Discriminant Analysis	0.31794	0.08950	0.84250	#PC=4, solver='lsqr', n_components= 7,shrinkage='aut o'
Random Forest Classifier	0.30168	0.09850	0.83864	#PC = 2
Logistic Regression	0.33036	0.05550	0.84900	penalty='l1' #PC = 2, max_iter=1000
Gaussian Naive Bayes	0.23681	0.10100	0.82507	#PC=3

The logistic regression achieved the maximum accuracy with highest precision, while naive bayes gives lowest accuracy. LDA is performed equally well with an accuracy of 84.25%, so both linear classifier working better than the non-linear models.

Conclusion

In this project, we have been investigating the enron data set in order to identify the person of interest (POI) who were involved in the famous scam happened in an american company. We have cleaned, removed outliers from the data set. We also transformed features and created additional features with an objective to build a strong predictive model. We applied several classification algorithms on the cleaned dataset and compared their performance. It is interesting to investigate that linear classifier performed well on the data set.

It is still open for further enhancement, class imbalances could also we investigated more closely because we assume there would be less number of people involved and marked as POI while majority would be non-POI. There could be an issue of class imbalances and hence stratified sampling could be used for data sampling.