**1. Summarize for us the goal of this project and how machine learning is useful in trying to accomplish it. As part of your answer, give some background on the dataset and how it can be used to answer the project question. Were there any outliers in the data when you got it, and how did you handle those? [relevant rubric items: “data exploration”, “outlier investigation”]**

The aim of this project is to identify the person who has involved in fraud happened in the Enron company. It has been observed that most frauds that happen in organization can be detected by peep down into the insight of financial data and communication data.

Using Machine learning, it is possible to investigate and find the pattern in data which could be helpful to identify the fraudulent.

The dataset used for analysis and classification is used about email data and financial data like stock exchange information, salary, number of emails exchanged to each other etc. Here comes the importance of using machine learning which used this information to build a model to find the relationship between this information and the person of interest(guilty).

The dataset consists of 146 observations and 21 features for each observation for two types of classes poi and non\_poi. The number of poi in the dataset is 18 and non\_poi is 128. There are certain outliers exhibited during analysis by each feature. There was one observation 'TOTAL' which distorted the distribution of data turned out to be outlier. After removing this observation, the distribution was significant enough to carry the further analysis. There was certain inconsistency in data entries for two observations 'BELFER ROBERT' and 'BHATNAGAR SANJAY' which was resolved using original pdf data of financial information.

To find out the more possible outliers, a multivariable linear regression model was created. By removing data points with top 10% variance between predicted and real values, an outlier-cleaned dataset was built for the analysis. However, after examining there were lots person of interest appearing in the outliers, so, I have used only the original to train the model.

**2. What features did you end up using in your POI identifier, and what selection process did you use to pick them? Did you have to do any scaling? Why or why not? As part of the assignment, you should attempt to engineer your own feature that does not come ready-made in the dataset -- explain what feature you tried to make, and the rationale behind it. (You do not necessarily have to use it in the final analysis, only engineer and test it.) In your feature selection step, if you used an algorithm like a decision tree, please also give the feature importance of the features that you use, and if you used an automated feature selection function like SelectKBest, please report the feature scores and reasons for your choice of parameter values. [relevant rubric items: “create new features”, “properly scale features”, “intelligently select feature”]**

The feature selection method was used **SelectKBest** which gives the following result

['poi', 'salary', 'from\_this\_person\_to\_poi\_ratio', 'bonus', 'total\_stock\_value', 'exercised\_stock\_options']

The iterative approach has bees used to obtained the final result out of all features in the data set using

params\_k\_best = {"k\_best\_\_k": [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]}.

Feature scaling is important to standardize the dataset. Standardization involves rescaling the features such that they have the properties of a standard normal distribution with a mean of zero and a standard deviation of one so that all feature values fall within the range. Feature scaling is also important to achieve the highest score of the model. There is also good source of information by Andrew Ng that feature scaling helps us to achieve the gradient decent more quickly, if the data is more spread out, that means if it has a higher standard deviation, it will relatively take more time to calculate the gradient decent compared to the situation when we scale our data via feature scaling.

Two new features, from\_this\_person\_to\_poi\_ration, from\_poi\_to\_this\_person\_ratio are created when analyzing the data. It is assumed that the number of mails exchanged between person and POI and vice versa are more likely to high as the ratio are.

**3. What algorithm did you end up using? What another one(s) did you try? How did model performance differ between algorithms? [relevant rubric item: “pick an algorithm”]**

The final algorithm used is adaBoost. There were three algorithms being picked and evaluated, considering the accuracy along with computation time, adaBoost gave the fastest performance on the same scores. There were other algorithms performing significant good, like LinearSVC which I have used to run poi\_id.py in order to get the result. The ada\_boost algorithm was taking too long to run.

**4. What does it mean to tune the parameters of an algorithm, and what can happen if you don’t do this well? How did you tune the parameters of your particular algorithm? (Some algorithms do not have parameters that you need to tune -- if this is the case for the one you picked, identify and briefly explain how you would have done it for the model that was not your final choice or a different model that does utilize parameter tuning, e.g. a decision tree classifier). [relevant rubric item: “tune the algorithm”]**

Tuning the parameters of machine learning algorithm means to find a optimized solution specific for the specific problem. If an algorithm is not tuned well, it could lower down the accuracy score, precision and recall scores, while increasing the runtime.

To tune the algorithm, a list of available parameters are put into a dictionary as keys, with a list of possible values as their values. Then GridSearchCV was conducted to search for the best solution among the given parameters. So the tuning result is limited by the range of values given to parameters, it is important to provide reasonable amount of values to be tuned while considering the time cost.

**5. What is validation, and what’s a classic mistake you can make if you do it wrong? How did you validate your analysis? [relevant rubric item: “validation strategy”]**

Accuracy is not the reliable metric to estimate the performance of machine learning model.

The drawbacks of accuracy metric are:

\* Classification accuracy is the easiest classification metric to understand

\* But, it does not tell the underlying distribution of the response values

\* it does not tell you the types of error classifier is making.

\* It is not reliable metric in the case of imbalance data.

Thus, the validation is required to ensure that classification model generalizes well. If it is not performed or wrongly performed, the overfitting problem could be encountered during the test of algorithm. Overfitting means that algorithm performs well on training data but does not perform well on testing or unseen data.

To prevent overfitting performance of algorithm, a cross validation is needed. I have used 10- folds cross validation method. Since the data set is imbalance the splitting method is stratified shuffling.

Steps for k-fold cross validation

1. Split the dataset into k equal partitions (or "folds")

2. Use fold 1 as a testing set and the union of others folds as the training set.

3. Calculate testing accuracy.

4. Repeat steps 2 and 3 k times, using different folds as the testing set each time.

5. Use the average testing accuracy as the estimate of Out-of-sample accuracy.

**6. Give at least 2 evaluation metrics and your average performance for each of them. Explain an interpretation of your metrics that says something human-understandable about your algorithm’s performance. [relevant rubric item: “usage of evaluation metrics”]**

The evaluation metrics in this model building are accuracy score, f1 score, precision score and recall score. The final model has an accuracy score of 0.94 on cleaned dataset, which means 94% of the predictions are found to be true. With a precision score of 1, it tells us that if this model predicts a person as a poi there is 100% chance that person is poi. on the other hand, with a recall score of 0.5 only 50% all the poi could be found out by this model. Finally, there's always a tradeoff between precision and recall, f1 comes and measure how well the tradeoff is. With a f1 score of 0.5. This model is fairly robust, while more optimization method could be applied to improve the model.