Identify Fraud from Enron Email Dataset

1. Goal of this project is to Identify the fraud people from Enron Company's email dataset by using Supervised Machine learning method. Enron company email dataset is huge and Identifying person of interest(fraud) was difficult. There are few outliers, I have identified them using a scatter plot, I have plotted one of the features of the dataset, Salary against Person if Interest and I found that the value 'Total' is the outlier in the dataset and I have immediately removed it

2. I have chosen the following features: poi, salary, total stock value, from poi to this person, from this person to poi, bonus and I have used PCA to pick the most important features, I did scaling for salary, bonus, total stock value, from poi to this person, from this person to poi because salary and from poi to this person are two different values. I have scaled all of them in the range 0 to 1.

   I was interested in learning about percent of emails sent by a person to poi and percent of emails received from poi, so I have created two new features: percent to email from poi and percent from email to poi

3. I have used Naïve Bayes Algorithm for this project. I have also tried Support Vector Machine, logistic regression classifier, Decision Tree, Random Forest and Ada boost Algorithms. Logistic Regression Classifier has got the highest accuracy of 0.86, but its Recall is less than 0.3, other classifiers also have accuracy in the range 0.82 to 0.86, Recall values are not above 0.3, Naïve Bayes got accuracy of 0.85, precision 0.5 and Recall 0.4, so I have chosen Naïve Bayes Algorithm

4. Tuning of an Algorithm means getting the best performance out of a model by using appropriate parameters of that model for that dataset, if we don't do this well, it can result in bad results, I did not tune parameters in this project as I have used Naïve Bayes algorithm which did not have parameters to tune. But initially I have tested Support Vector Machine. I have tuned its parameters: C, gamma and Kernel using GridSearchCV, I have taken list of random values for C , gamma and Kernel and passed them to the GridSearchCV along with Classifier and then I have validated this model by calculating Accuracy , precision and Recall , I tried different values and observed for the improvement in validation results

5. Validation means checking the performance of the model, if we don't validate well, we may end up overfitting the data. I have validated my data by splitting the whole data into 70% training data and 30% testing data, so that after fitting the training data into the model, I have predicted values using testing data and compared the actual values with predicted values, which helped my model from overfitting

6. The evaluation metrics chosen are Precision and Recall. The average Precision for Naïve Bayes algorithm on this data is 0.41, which means predicted data is 41% true out of all the predicted true values. and Average Recall is 0.37, which means predicted data is 37% true out of all the true values in the dataset