1. The goal of the project is to find the fraud Enron Employees based on email and financial dataset using machine learning algorithms. First I extracted the data from pickle file and saved in an excel sheet to observe what features has enough data to consider further. Then I used matplotlib to draw the scatter plot to see the patterns and find the outliers. Based on my observations, I finalized the features and removed the outliers from the data.
2. First I have eliminated features which has more NAN values for the POI examples. Then I used matplotlib scatter plot to see there is any pattern in the features. Then I looked at all emails under from,to sections and I noticed emails are categorizing under different folders like deleted items,discussion threads. So I applied above NAN filters on different sections and picked features which has more values for POI. Then I applied sklearn SelectKBest with k=12 to consider features which has good information. This helped to reduce my training time. Please find the scores on selector.scores\_

[ 1.85757033e+01 1.69882435e+00 8.86672154e+00 2.10600017e+01

9.34670079e+00 8.74648553e+00 2.44676540e+01 6.23420114e+00

4.20497086e+00 2.42650813e+00 5.34494152e+00 1.86593227e-02

1.27216315e-01 1.72121949e-01 6.56052834e-02 1.40925802e+00]

Then I noticed values for few features are small numbers compared to other. So I used sklearn MinMaxScaler for feature scaling.

1. I have tried the below algorithms with different parameters using GridSearchCV and noticed the accuracy. I first observed SVM accuracy is higher than other two. Then When I ran all three classifiers for precision and recall, I noticed SVM is not predicting correctly the examples when it is a POI. Even though the higher accuracy for SVM, based on the evaluation metrics, I have chosen Decision tree for this problem. Please find the accuracy for the above algorithms

|  |  |
| --- | --- |
| Algorithm | Accuracy |
| Support Vector Machine | 0.89 |
| Decision Tree | 0.86 |
| Gaussian Naive Bayes | 0.87 |

4) Picking the right parameter values for algorithm is one of the important criteria to get good accuracy. My plan is to tune C and gamma parameters for SVM and min\_samples\_split for Decision Tree. I want to apply PCA to reduce the dimensionality of features. So I used a pipeline with PCA and the classifier. I have tried the following parameters in the pipeline

|  |  |
| --- | --- |
| SVM | C = [2,5,10], kernel = ['linear','rbf'] |
| Decision Tree | min\_samples\_split=[2,5,10] |
| PCA | n\_components=[3,5,7] |

5) If we train a classifier with only training data, then there is a big chance of overfitting the data. That means the classifier can not give accurate results for data which it has not seen. So cross validation is very important for any machine learning algorithm. I split my data into train and test dataset using sklearn train\_test\_split with test\_size=0.2, random\_state=42. I trained my classifier with 80% of data and used 20% testing data for evaluating the accuracy of my classifier.

6) Once I trained and tested the classifier, I want to check the precision,recall on my classifier. I used test\_classifier in tester.py to get the metrics on my different algorithms. I noticed with SVM, true positives are 0 i.e my classifier is not predicting correctly when example is a POI. I observed naive bayes precision is 0.498027613412 i.e if classifier predicts correctly then there is almost 50% chance that predicted value is matching with actual . I see more balanced values using decision trees.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algorithm | precision | recall | f1 | f2 |
| Decision Tree | 0.381102362205 | 0.363 | 0.371830985915 | 0.366481574962 |
| Naive bayes | 0.498027613412 | 0.2525 | 0.335102853351 | 0.280119813623 |

Based on accuracy,precision and recall , I have picked decision tree over the other two