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”]

   Total data points are 145. The number of poi is 18.

   Two features are included in my data set.

   Here is the number of ‘NaN’ values for each variable.
   {poi': 0,
   ''bonus': 64,
   'exercised_stock_options': 44}

   The goal of this project is to predict poi person as much as possible. In order to evaluate my prediction model, I used recall score and precision score. I removed one key named TOTAL since it just sums up the data and it interrupts precise modeling.

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 importances 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 number of ‘NaN’ value for each new variable
   {'from_poi_to_this_person': 0

'from_this_person_to_poi': 0}
I made new features such as from_poi_to_this_person and from_this_person_to_poi. I made these new features based on the hypothesis that person who often contacts with poi person might be poi person.

The precision and accuracy score with engineered features:
Precision: 0.4473
Recall: 0.26550

The precision and recall score without engineered features:
Precision: 0.4747
Recall: 0.40850

As we can see, new features did not have positive effect to the final classifier.

I choose 2 features by using below process.

1. calculate feature importances
2. remove features with low score
3. test your classifier in terms of precision&recall
4. repeat the process for the remaining features

I especially focus on improving recall score since it reflects on how accurately my model predict poi among the data with a few poi.

{The number of features: precision and recall score}
[{2: Precision: 0.4747 Recall: 0.40850},
{3: Precision: 0.5558 Recall: 0.38600},
{5, Precision: 0.6017 Recall: 0.34000},
{21, Precision: 0.4809 Recall: 0.10100}]

The importance for my features.

Importance {'bonus': 0.46998094273401092,
'exercised_stock_options': 0.53001905726598908}

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

   I ended up using random forest because it shows good precision and recall score. I tried SVM, and random forest. Although SVM showed best recall, I could not improve precision. Random forest showed better scores than SVM, so I picked up random forest.

   Random forest, Precision: 0.4747 Recall: 0.40850
   SVM, Precision: 0.20987 Recall: 0.92900

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”]

   When conducting machine learning analysis, tuning is important because it helps us to build better machine learning model. In random forest, I tuned two parameters, like the number of trees of splitgroup and the number of features to pick up within the split group which are respectively n_estimators, and max_features. Without tuning, the result scores got worse. I tried three value for each parameters, which are n_estimators (10, 50, and 100) and max_features(2, 3, and 5). And I found best scores 50 and 2 for n_estimators and max_features.

In addition, I used grid research and min max scaling when building SVM model.

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"]

Validation is a technique for assessing the results of predicting model. Doing validation is important because it helps us to avoid over-fitting or helps us to realize how poor my prediction model is.

The data only has 18 poi out of 145 data points, so it is really important to get high precision and recall score because this model easily gets high accuracy score by assuming that all predicted value as non-poi, which is a classic mistake.

I used StratifiedShuffleSplit function to evaluate performance of the algorithms. StratifiedShuffleSplit is a merge of StratifiedKFold and ShuffleSplit. StratifiedKfold might be suitable for this model since the data only has 18 poi, which means I need 18 poi to be distributed equally to each folds, but the data only has 145 points. Therefore, if I split my data to k fold, the size might not be enough to evaluate the performance. I need to shuffle my data and get more than 145 samples. By using StratifiedShuffleSplit, the data is assigned to folds with equal percentage of composition and the amount of data is enough to evaluate performance.

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"]

   I used precision score and recall score to evaluate my results. I avoided to use accuracy score since it does not work as a good evaluation due to the small number of poi. Precision score is represented as the number of true poi predicted as poi divided by the number of poi predicted as poi. Recall score is defined in this data as the number of true poi predicted as poi divided by the number of poi and not poi predicted correctly.