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. Training data set is 79 and testing data set is 34, which are respectively including 12 poi and 5 poi. My data set has four features. Shared_receipt_with_poi has missing values.

   The goal of this project is to predict poi person as much as possible. In order to evaluate my prediction model, I used accuracy 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”]

   I used four features to build a model, such as salary, fraction_from_poi, fraction_to_poi, and shared_receipt_with_poi. From financial features, I chose salary feature. It enables me to compare everyone because salary does not have NaN. From email features, I focused on from_poi_to_this_person, from_this_person_to_poi, and shared_receipt_with_poi since I suspect these features somehow connected poi. I did scaling especially for from_poi_to_this_person and from_this_person_to_poi, devided by to_messages and from_messages respectively because I wanted to distinguish person who used email a lot from person who really related to poi. Feature importance are 0.30689971, 0.13736006, 0.31864691, and 0.23709332 each for salary, fraction_from_poi, fraction_to_poi, and shared_receipt_with_poi. According to feature importance fraction_to_poi is most effective feature for predicting poi.

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 best precision. I tried SVM, decision tree and random forest. Although SVM showed best accuracy score, I could not improve precision from 0. Random forest show better scores than decision tree, so I picked up random forest.

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”]

   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.

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”]

I used accuracy score and precision score. My training data set has only five poi opposite to 29 non-poi, so it is really important to get higher precision because this model easily gets high accuracy score by assuming that all predicted value as non-poi, which is a classic mistake. I considered both accuracy score and precision score as equally important.

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"]

   Accuracy score and precision score. Accuracy score is an important measurement because my goal is to predict poi. At the same time, precision score is also important since data set rarely has poi values, I needed to make my model to predict them correctly.