Enron Submission Free-Response Questions

A critical part of machine learning is making sense of your analysis process and communicating it to others. The questions below will help us understand your decision-making process and allow us to give feedback on your project. Please answer each question; your answers should be about 1-2 paragraphs per question. If you find yourself writing much more than that, take a step back and see if you can simplify your response!

When your evaluator looks at your responses, he or she will use a specific list of rubric items to assess your answers. Here is the link to that rubric: Link to the rubric Each question has one or more specific rubric items associated with it, so before you submit an answer, take a look at that part of the rubric. If your response does not meet expectations for all rubric points, you will be asked to revise and resubmit your project. Make sure that your responses are detailed enough that the evaluator will be able to understand the steps you took and your thought processes as you went through the data analysis.

Once you've submitted your responses, your coach will take a look and may ask a few more focused follow-up questions on one or more of your answers.
We can't wait to see what you've put together for this project!

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 goal of the project is to use features within the data set to identify person of interests ("POI"). Within the dataset, there are 146 data points and 21 features. The dataset is divided into 18 POIs and 128 non-POIs. With such an unbalanced dataset, cross validation techniques will be important in our upcoming analysis or processing. Moreover, we have to use additional evaluation metrics such as recall and precision instead of only accuracy score, as most of our data are in one class and our model may be tuned to be always predicting the same class. Since our data set here is relatively small in size, we can use the technique stratified shuffle split with gridsearchCV to deal with the problem.

   There are a lot of missing data points as exemplified by the below table.

Features	Number of missing data point
total_stock_value	20
total_payments	21
email_address	35
restricted_stock	36
exercised_stock_options	44
salary	51

expenses	51
other	53
to_messages	60
shared_receipt_with_poi	60
from_messages	60
bonus	64
from_poi_to_this_person	72
long_term_incentive	80
from_this_person_to_poi	80
deferred_income	97
deferral_payments	107
restricted_stock_deferred	128
poi	128
director_fees	129
loan_advances	142

Machine learning is used to train up classifier models and identify if a person is POI per his features. The dataset includes remuneration and email communication related information. Assuming that POIs have higher remuneration and are tightly connected with other POIs as indicated by email communication, we can use the dataset to accomplish our objective.

There are two outliers, namely “TOTAL” and “THE TRAVEL AGENCY IN THE PARK”. I have deleted these data points from the dataset.

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 selected features with SelectKBest and ended up with 5 features, namely 'salary', 'bonus', 'deferred_income', 'total_stock_value' and 'exercised_stock_options'.

I did not conduct scaling for most of the algorithms except for the supported vector machine algorithm. The scaling is important to prevent large values from dominating the result. I have used a min-max scaler to conduct the process

I created three features, namely email to poi percentage, email from poi percentage and

total remuneration. I hypothesized that POIs should have a closer connection with each other and they should have a high proportion of emails either from/to POIs. On the other hand, I suspect that POIs will have abnormally high total remuneration, as these people will be more actively betting on the stock market, and they may internally manipulate the remuneration system to increase their own salaries. By testing out the Naive Bayes algorithm with original and new data set, we can see that the precision score has been improved.

Naive Bayes Classification Report

	precision	recall
Not POI	0.92	0.95
POI	0.5	0.4

Naive Bayes Classification Report with new features

	precision	recall
Not POI	0.93	0.97
POI	0.67	0.4

Feature Score Table

Features	Feature Score
exercised_stock_options	20.79225205
salary	18.28968404
deferred_income	8.77277773
total_stock_value	7.184055658
bonus	0.224611275

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 the Naive Bayes Classifier. I tried Naive Bayes, SVM, Decision Tree, Random Forest and Adaboost. After comparing the models' recall and precision score, it is found that Naive Bayes > Adaboost = Decision Tree > Random Forest = SVM. Thus I have chosen Naive Bayes for further fine tuning and optimization.

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 means to find the combination of parameters which could generate the best precision and recall score fitting the purpose of the investigation. There aren't parameters to be tuned within Naive Bayes. However, I have employed the technique within selectKbest to choose the best number of features to be included in the model. Here I have plugged in a range of number of features (i.e. SKB__k : range(5:10)). The function will plug in the list of possibilities and return the best solution.

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 the process of assessing how well a statistical analysis generalize to an independent data set. To be more precise under the current context, it is assessing the performance of the predictive model which we have generated. Data is split into training and testing set to simulate the scenario of having an unseen independent data set to be tested by a model. The training data set is used to train the predictive model which will then be applied to the "independent" test set to understand how well the model generalize.

   A common mistake is that the original model developed with the training data set may predict the training data perfectly but ultimately fail to predict testing data set. My strategy is to use the stratified shuffle split cross validator ("SSS"). Data are usually split into the training set and testing set. However, due to the very small sample size in this project, overfitting may easily occur and the model cannot generalize well to independent sample. SSS can overcome this challenge, the function divides the data randomly into multiple sets while rearranging the data to ensure that each fold is representative of the whole dataset. Multiple fitting of different data set to the model is carried out. Overfitting effect will be minimized after the average outcome of each fitting result is calculated.

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"]

   Recall and precision are used to measure the performance of algorithms. In general, "recall" measures how well the model can detect a positive sample, given the sample being measured is positive. "Precision" measures how well a model predicts a correct positive sample, given the predicted result is positive.

   For a more concrete example, we can assume a model for classifying POI has 60% recall and 40% precision. Given that the person is a POI, there is a 60% chance that the model can correct detect it. And when the model predicts a person as a POI, there is 40% chance that it is correct.