**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**](https://review.udacity.com/#!/projects/3174288624/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 identify POI persons using the database available. Dataset includes financial data and e-mail logs of different persons in the company. Data include some outlier points such as TOTAL and TRAVEL AGENCY, which are actually not persons and therefore are outliers as expected. In addition to these; there are also many people with very little information about them. These are also treated as outliers because there is no useful information included.

1. 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”, “intelligently select features”, “properly scale features”]

By usind SelectKBest feature selection method, I end up using the following highest scoring 7 features: 'salary', 'total\_payments', 'bonus', 'total\_stock\_value', 'exercised\_stock\_options', 'long\_term\_incentive', 'restricted\_stock'. To reduce and combine characteristics; I tried to get the ratio of exercised stocks to total stocks ratio but it was eliminated during feature selection process.

1. 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 have tried GaussianNB, Support Vector Machine (linear and normal), Decision Tree, Gaussian Process Classifier, Random Forest and AdaBoost. Decision Tree is eliminated due to it tends to overfit and lower score on test features (0.75). Linear SVM also scores lower than others. GaussianNB, SVM and Gaussian Process models score 0.88. Random Forest and AdaBoost score 0.86. These values are obtained for default setting without tuning any of the models. If a simple model is required, GaussianNB is the obvious choice. SVM has close score to GaussianNB. Random Forest has slightly better score than the others (both score of the classifier and f1 score). So the conclusion is if even 2-3% accuracy is better than one should choose Random Forest otherwise Gaussian NB is much easier to handle and no tuning is required.

One can also check and see the differences between different classification models from the link below, which I found very useful:

<https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html>

1. 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? What parameters did you tune? (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 items: “discuss parameter tuning”, “tune the algorithm”]
2. 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 items: “discuss validation”, “validation strategy”]
3. 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”]