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. 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 this project is to build an algorithm to identify Enron employees who may have committed fraud based on the public Enron financial and email dataset. The dataset provides financial and email data for a list of Enron employees, which will be the input data. In addition, the dataset provides a POI (Person Of Interest) label, which will be the output data. Machine learning will be very useful as it deals with supervised classification using the appropriate algorithms.
Several outliers have been identified and removed:

- 2 were obvious (and already identified in Udacity course: 'TOTAL', 'THE TRAVEL AGENCY IN THE PARK') since they are not employees
- 1 was Enron founder ('LAY KENNETH L') and had very high total payments and stock values. It could have substantially modified the POI identifier
- 1 was an employee who sent many emails ('KAMINSKI WINCENTY J') and was not POI.

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

First, all features with too many NaN values have been discarded (an arbitrary minimum number of 81 values has been set up). We could have kept them and assigned them a given value such as mean value but due to the low number of data points, we have chosen the option to remove them in order to not overfit the algorithm.
Feature scaling has been tested for KNN algorithm but feature scaling has not been done in the end as a decision tree type (AdaBoost) classifier has been selected and feature scaling is not useful for this type of algorithm. As observed in the dataset exploration, the ratio value instead of absolute value for emails sent or received seemed more useful for classification. Feature importances have been displayed for the decision tree classifier to show the relative importance of the features when making the classification (highest score was for feature ‘total_stock_value’).

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”]

The AdaBoost classifier has been selected as it showed a better performance than any of the other classifiers tested: Naive Bayes (NB), Decision Tree (DT), Random Forest (RF), k-Nearest Neighbors (KNN) and Support Vector Classification (SVC). Based on data observations (in particular for the features on ratios of emails sent to POI or received from POI), a preference for “decision tree” algorithm type was highlighted. AdaBoost classifier had the best overall performance. KNN and SVC algorithms have not been selected also due to the low number of data points and the overlap of classes (dataset probably too noisy). Finally, NB algorithm has not been selected despite the correct performance (and its speed efficiency) as the project required some parameter tuning and NB algorithm has no parameters. Please note that all metrics have been considered but a specific attention has been paid to the precision. Indeed, the idea is to get a good precision to capture the maximum of “false positive” so that we get all the employees who might have committed fraud. Further investigation will be then required.

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? 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”]

To tune the parameters of an algorithm is to find out the best parameters to get the best performance (in terms of given metrics) from an algorithm. If an algorithm, its performance will not be optimized and there will be less confidence in the results. 2 parameters were tuned on AdaBoost algorithm: ‘n_estimators’ and ‘learning_rate’ using cross validation. The best parameters have been determined by GridSearchCV technique focusing on the recall metric (as it was closer to 0.3 than the precision metric) and choosing the default 5-fold cross-validation.

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 items: "discuss validation", "validation strategy"]

The validation of the algorithm is necessary to ensure the algorithm will work correctly on its own. The idea is to train the algorithm on part of the dataset (named train data) and to validate it on the remaining part of the dataset (named test data). A classic mistake is to train and to test on the same data, which will probably give a good performance as the algorithm already saw the validation data during its training.
In this project, the cross validation from 'tester.py' has been used: StratifiedShuffleSplit with a very high number of folds (1000) in order to ensure the splitting procedure does not impact the overall performance. Indeed, the dataset has only a few points so the arbitrary splitting might impact the performance a lot.

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"]

The evaluation has focused on the 2 following metrics: accuracy and precision. The evaluation metrics give: the accuracy around 0.85 and the precision about 0.45. In this POI identifier, the accuracy quantifies the number of employees correctly identified (POI - true positive - or non POI - true negative) out of the total number of employees. The precision quantifies the number of employees correctly identified as POI out of the total number of employees identified as POI (whatever it is correct - true positive - or incorrect - false positive).