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 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 coach evaluates 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 the rubric. If your response does not meet expectations, you will be asked to resubmit.

Once you've submitted your responses, your coach will take a look and 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 identify the employees from the Enron incorporation who committed fraud. We label this kind of employees as POIs. The dataset we have contains information about the employees. There are two kinds of information available for analysis. The first one is the financial information, including the salary information, bonus information, stock information, etc. The second one is the email information, including all the email texts, the number of total sending and receiving email, the number of email sent to the POIs, the number of email received from the POIs.

   Machine learning algorithms leverage these information to identify possible POIs. There were two outliers in the dataset. I identify them through inspecting the length of the name of the employees. Since one outlier is the summary information of all employees and the other is difficult to understand, I removed these two outliers.

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 doesn't 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.) If you used an algorithm like a decision tree, please also give the feature importances of the features that you use. [relevant rubric items: “create new features”, “properly scale features”, “intelligently select feature”]

This dataset has totally 146 data points. There are 18 of them are labelled as POI, while the other 128 are labelled as non-POI. The number of features in the original dataset is 21. Feature ‘loan_advances’ has 142 missing values. Feature ‘from_messages’ and ‘to_messages’ has 60 missing values. Feature ‘salary’ has 51 missing values.

I ended up using the following eight features:

- Financial features: ‘exercised_stock_options', 'total_stock_value', 'bonus', 'salary', 'deferred_income', 'long_term_incentive', and 'restricted_stock'.
- Email features: 'toPoiRatio'.

I picked them through the process of selecting k best features. I used the StandardScaler to scale the dataset. Because inverse of regularization parameter C was used. If I did not scale the dataset, some parameters may have huge impacts on parameter C, which was not desirable.

I created 4 new features: ‘hasEmail’, ‘fromPoiRatio’, ‘toPoiRatio’, and ‘total_money’. ‘hasEmail’ indicates if an employee’s email address is available or not. Since I suspect that POI should be higher positions in the company, an email for them is necessary to do business, I add this feature. ‘fromPoiRatio’ and ‘toPoiRatio’ show the percentage of the email an employee receives from or sends to the POI out of all the emails he receives or sends, respectively. I add these two features since I think if an employee has high proportion of emails from or to POI, he is probably a POI. ‘total_money’ is the total money an employee gets from salary, bonus, all kinds of stocks, etc. I think this features could be useful since the only reason POIs commit fraud is money.

Without new features, 8 best features selected are : ['exercised_stock_options', 'total_stock_value', 'bonus', 'salary', 'deferred_income', 'long_term_incentive', 'restricted_stock', 'total_payments']. And the scores of these 8 best features are: [24.815, 24.183, 20.792, 18.29, 11.458, 9.922, 9.213, 8.773].

With new features, 8 best features selected are: ['exercised_stock_options', 'total_stock_value', 'bonus', 'salary', 'toPoiRatio', 'deferred_income', 'long_term_incentive', 'restricted_stock']. And the scores of these 8 best features: [25.098, 24.468, 21.06, 18.576, 16.642, 11.596, 10.072, 9.347]. We can see that the new feature ‘toPoiRatio’ has as score of 16.642. That means the new feature is effective.

Table I Results of Different Classifiers

	W/O new features		W/ new features	
	Precision	Recall	Precision	Recall
Logistic Regression	0.47549	0.45736	0.42480	0.44717
Naïve Bayes	0.38316	0.32053	0.44586	0.38521
SVC	0.21969	0.20340	0.33680	0.36344
K-Nearest Neighbors	0.05956	0.01229	0.28596	0.14121

We can see from the above table that except for Logistic Regression, with the new features, all other classifiers' performance has been improved. That justifies the new features I created.

3. What algorithm did you end up using? What other one(s) did you try? [relevant rubric item: "pick an algorithm"]

   I ended up using Logistic Regression. Others I have tried include: Naïve Bayes, Random Forest Classifier, AdaBoost Classifier, K Nearest Neighbor Classifier, Support Vector Classifier, Gradient Boosting Classifier.

   The reason I chose logistic regression is two-folds. Firstly, according to Table I, it provided the highest precision and recall. Secondly, it is a simple algorithm. The training time of logistic regression is much shorter than these ensemble methods such as random forest and adaboosting.

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 don't 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 if you used, say, a decision tree classifier). [relevant rubric item: "tune the algorithm"]

   Tuning the parameters of an algorithm means to adjust the values of the parameter so that the algorithm could achieve better performance. If I don't do this well, the algorithm may turn out to be useless and I may lose the chance of doing better predictions.

   For each algorithm, I used the GridSearchCV to test out the best parameter set for the current value of k (number of best features to use). Then I manually tried different values of k using the best parameter set to determine the final parameters to use.

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 using the test dataset to test the performance of the trained classifier. A classic mistake is overfitting is not shown up. I validate my analysis using 1000 fold cross validation.

Since the dataset is small, I used StratifiedShuffleSplit. That is shuffle the data, split the data into several folds, train the data on the training data and then test the model on the testing data. Then the testing result is recorded. After many times of testing, I take the average of the testing result and report it.

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"]

Two evaluation metrics I used are Precision and Recall. The average performance is Precision: 0.425, Recall: 0.447. Precision being 0.425 means that of all the POIs I claim, 42.5% of them are real POIs. Recall being 0.447 means that of all the POIs in the dataset, I have identified 44.7% of them.