**Enron Submission Free-­Response Questions and Answers**

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 choose a combination of features of former Enron*

*employees and choose an appropriate machine learning algorithm to predict whether*

*that person is considered a person of interest (POI) or not. This is considered a*

*Supervised Classification problem as we are trying to predict the discrete outcome,*

*which is binary (2 options only), and we are actually given a data set that contains*

*the right “answers” (whether the person is actually a POI or not). The goal will be to*

*get the most accurate predictions when we apply our ML algorithm to test this model.*

*For example, if we correctly predicted all the POIs in the test data set we would get*

*an accuracy of 1.0.*

*This dataset is from Enron, which is an infamous American company known for its*

*extensive fraud. The actual dataset itself consists of email metadata and financial*

*data for created by about 150 employees of Enron (mostly senior management).*

*I initially thought about using a strategy of removing the top 10% of values with the*

*largest residual errors (as highlighted in the instructor videos). However, that*

*strategy seemed appropriate for a regression problem where we are trying to arrive*

*at a predicted quantitative value vs this problem which is a classification problem*

*with only two results: Non-­POI or POI.*

*Using a more manual strategy of inspection, I found that there were actually just two*

*outliers. ‘Total’ was definitely not a valid person which was found in the dataset after*

*exploring all the employee names. The second invalid record was for a travel agency*

*called ‘THE TRAVEL AGENCY IN THE PARK’. Most of the values for each field*

*were invalid (‘NaN’) and thus this record was also removed. The final count of*

*records in this relatively small dataset was 144.*

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”]

*I used ‘domain knowledge’ to select the features for the model. After watching the*

*film “Enron: The Smartest Guys in the Room”, I became somewhat knowledgeable*

*about the features found in a few of the people of interest such as Lou Pai, Jeff*

*Skilling, Ken Lay, and Andy Fastow. I noticed that Lou had a really large expense*

*account;; he was known to spend a lot of the company’s money on corporate jet trips*

*and strip clubs for example. Another thing that was frequently mentioned in the*

*movie was that a lot of the POIs exercised a lot of options as they knew that the*

*company’s fortunes were bound to eventually turn around due to the fraudulent*

*activities that were commonplace. So for the initial feature selection process, I chose*

*the features related to POIs and engineered three new features fraction\_from\_poi, fraction\_to\_poi and from\_specific\_email(this option I thought would benefit after watching the documentary as said before.).*

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 the Decision Tree Classifier algorithm in a supervised classification context. I considered other algorithms such as Naïve Bayes and SVM and when checked all of these three algorithms, I got better precision and recall values from Decision Tree Classifer.*

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 is nothing but using the algorithm to its fullest so that we could get optimized result when executed our model. i.e., we will get more optimized precision and recall values when compared with the first run before changing the parameters. The depth and max\_no\_of\_splits in the algorithm played a significant role in deciding the optimal precision of the developed model.*

*In this context, I have tuned parameters manually by playing around with some combinations of max\_depth and min\_samples\_split as I have used Decision Tree Algorithm. Obviously, selecting a wrong number of these parameters can have a profound effect on the performance of the model. So I started with max\_depth and max\_split as 2 for both and after various combinations I found that the optimal number to get accuracy, precision and recall is 12 for min\_samples\_split and 6 for max\_depth.*

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 process of determining how well your model performs or fits, using a specific set of criteria. Cross-validation is used to ensure that the model generalizes well, avoiding over or under-fitting. Essentially you break the data set into a test set (often 20-30% of the data) and a training set (the remainder of the data). You fit on the training set and predict on the test set. Metrics from the test set determine your performance. When using cross-validation on small data set, it is helpful to perform this process multiple times, randomly splitting each time.I validated my dataset with the help of the test\_classifier () function which is provided in tester.py by Udacity that used the sklearn StratifiedShuffleSplit () function to do the splitting of the data into a training set and test set.*

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: Accuracy refers to the ratio of correct predictions out of the total*

*predictions made. In this context, it means how many POIs the model was able to*

*correctly predict. The average performance for the model I tuned was 0.* *85986.*

*This means nearly 86% of predictions the model made were correct*

*Precision: Precision refers to the ratio of correct positive predictions made out of the*

*total positive predictions made (Positive predictions means predicting that the*

*employee is a POI). The precision of the model was just around 52%.*

*Recall: Recall is a more difficult thing to conceptualize for many (I have attached an*

*image from Wikipedia below that should help). Recall refers to the ratio of correct*

*positive predictions made out of the actual total that were indeed positive (correct*

*positive predictions + incorrect false negative predictions). We are basically looking*

*only at the actual POIs in the test data set and seeing what proportion of them were*

*correctly identified. The model was able to achieve a recall of about 41%.*

*You can see that there is a tradeoff between precision and recall which needs to be*

*balanced in all models depending on what is more important for the model. However,*

*in cases such as this which contains a lot more of one class over another class (way*

*more non-­POI than POI), recall and precision are both better measures than*

*accuracy. Just predicting that everyone is not a POI would get a pretty high accuracy*

*due to the fact there are just a lot more non-­POIs in the dataset. This is a very*

*important takeaway I learned from this entire process!! Just because a model has a*

*very high accuracy doesn’t necessarily mean it is a great model. However considering the imbalance of the features and variables in the given dataset and the precision is higher than .3, I believe my model is a good fit for identifying the POIs.*