1. The goal of this project was to explore machine learning techniques and apply them to the Enron Email dataset. The dataset consists of 146 executives and outlines people identified as POIs (Persons of Interest). A POI was someone who was indicted for fraudulent activity, reached a settlement, or testified for immunity. The data was mostly clean but had some gaps that needed to be handled in order to make processing and analyzing easier. This will be explained below.
2. Features that I was interested in using were messages to and from a POI. I also created the feature "Bonus to Salary Ratio" that explored the relationship between the bonus an employee was given and their respective salary. Even with this additionally feature that was added, precision and recall requirements could be met with out without this additional feature.
3. I ended up favoring the Ada Boost algorithm. It is an algorithm that is simple to work with and easy to understand. For a beginner in this field, I wanted to choose something that allowed me to easily understand the "behind the scenes" on the data manipulation that was occurring. This will allow me to scale up complexity in the future without feeling overwhelmed. Additionally, processing time for this algorithm was small which allowed me to quickly make changes in my code and see the results.
4. Tuning the parameters of an algorithm means to tweak specific settings to get an output that fits the relationship that you are trying to explore. Tuning can also mean time optimization (i.e. processing time) or modifying parameters to handle outliers or non-standard values.
5. Validation is the process of feeding test data into your algorithm to test how your algorithm responds to real-world data. Validation is important to ensure that your algorithm is behaving as you expect. This plays a larger factor when your data size increases drastically or when you want to add more variables/features to analyze. One mistake is not using a representative sample of data to validate your algorithm. Using a non-representative sample can lead to false positives or false negatives in your analysis of your data.
6. Feature scaling was not used in this project. Feature scaling allows the comparison of two or more variables to be standardized. For example, one variable could have a range of 1 to 10 and another would have a range of 1 to 1000. Feature scaling would "map" both this variables to a range of 0 to 1 to allow for comparisons within the same domain. However, feature scaling is not appropriate for all algorithms. The algorithms I used would not be positively impacted by features scaling.
7. Precision and recall are two metrics that can be used to identify the quality of the algorithm chosen. Precision is equitable to a low false positive rate whereas recall is equitable to a low false negative rate. An optimal precision and recall value is one in which false values are kept to a minimum. It should be noted that false positives and false negatives are only two of many metrics that can be used to measure algorithm performance.
8. "Splitting" your data refers to how much data you want to use to train your algorithm and how much data you want to use to test your algorithm's results. It is commonly expressed as a decimal where the split values are values that span between 0 and 1 that when added together sum up to 1. For example, 0.8 to train and 0.2 to test. The relationship of these values and the data that you choose can impact your algorithm. Non-uniform, non-representative data samples used to train can erroneously impact your algorithm. A common approach is to randomly

choose which data will be used to train and to test. This is the approach I used here. Although there may be imbalances due to the small number of POIs compared to the total number of persons, I still felt this was a justifiable approach due to the recall and precision values I was able to obtain. Further optimizations can be made by considering the impact of class imbalances in this dataset.