How do I Complete this Project?

This project is connected to the Intro to Machine Learning course, but depending on your background knowledge of machine learning, you may not need to take the whole thing to complete this project.

A note before you begin: the mini-projects in the Intro to Machine Learning class were mostly designed to have lots of data points, give intuitive results, and otherwise behave nicely. This project is significantly tougher in that we're now using the real data, which can be messy and does not have as many data points as we usually hope for when doing machine learning. Don't get discouraged--imperfect data is something you need to be used to as a data analyst! If you encounter something you haven't seen before, take a step back and think about a smart way around. You can do it!

Project Overview

In 2000, Enron was one of the largest companies in the United States. By 2002, it had collapsed into bankruptcy due to widespread corporate fraud. In the resulting Federal investigation, a significant amount of typically confidential information entered into the public record, including tens of thousands of emails and detailed financial data for top executives. In this project, you will play detective, and put your new skills to use by building a person of interest identifier based on financial and email data made public as a result of the Enron scandal. To assist you in your detective work, we've combined this data with a hand-generated list of persons of interest in the fraud case, which means individuals who were indicted, reached a settlement or plea deal with the government, or testified in exchange for prosecution immunity.

Resources Needed

You should have Python 2.7 and sklearn running on your computer, as well as the starter code (both Python scripts and the Enron dataset) that you downloaded as part of the first mini-project in the Intro to Machine Learning course. You can get the starter code, which uses Python 2.7, on git:

```
git clone https://github.com/udacity/ud120-projects.git
```

The starter code can be found in the final_project directory of the codebase that you downloaded for use with the mini-projects. Some relevant files:

`poi_id.py` : Starter code for the POI identifier, you will write your analysis here. You will also submit a version of this file for your evaluator to verify your algorithm and results.

`final_project_dataset.pkl` : The dataset for the project, more details below.

`tester.py` : When you turn in your analysis for evaluation by Udacity, you will submit the algorithm, dataset and list of features that you use (these are created automatically in `poi_id.py`). The evaluator will then use this code to test your result, to make sure we see performance that’s similar to what you report. You don’t need to do anything with this code, but we provide it for transparency and for your reference.

emails_by_address : this directory contains many text files, each of which contains all the messages to or from a particular email address. It is for your reference, if you want to create more advanced features based on the details of the emails dataset. You do not need to process the e-mail corpus in order to complete the project.

Steps to Success

We will provide you with starter code that reads in the data, takes your features of choice, then puts them into a NumPy array, which is the input form that most sklearn functions assume. Your job is to engineer the features, pick and tune an algorithm, and to test and evaluate your identifier. Several of the mini-projects were designed with this final project in mind, so be on the lookout for ways to use the work you’ve already done.

As preprocessing to this project, we've combined the Enron email and financial data into a dictionary, where each key-value pair in the dictionary corresponds to one person. The dictionary key is the person's name, and the value is another dictionary, which contains the names of all the features and their values for that person. The features in the data fall into three major types, namely financial features, email features and POI labels.

financial features: ['salary', 'deferral_payments', 'total_payments', 'loan_advances', 'bonus', 'restricted_stock_deferred', 'deferred_income', 'total_stock_value', 'expenses', 'exercised_stock_options', 'other', 'long_term_incentive', 'restricted_stock', 'director_fees'] (all units are in US dollars)

email features: ['to_messages', 'email_address', 'from_poi_to_this_person', 'from_messages', 'from_this_person_to_poi', 'shared_receipt_with_poi'] (units are generally number of emails messages; notable exception is ‘email_address’, which is a text string)

POI label: [‘poi’] (boolean, represented as integer)

You are encouraged to make, transform or rescale new features from the starter features. If you do this, you should store the new feature to my_dataset, and if you use the new feature in the final algorithm, you should also add the feature name to my_feature_list, so your evaluator can access it during testing. For a concrete example of a new feature that you could add to the dataset, refer to the lesson on Feature Selection.

In addition, we advise that you keep notes as you work through the project. As part of your project submission, you will compose answers to a series of questions (also given on the next page) to understand your approach towards different aspects of the analysis. Your thought process is, in many ways, more important than your final project and we will by trying to probe your thought process in these questions.