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?**

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. The goal of this project is to create a prediction model to check whether an Enron employee is a "Person of Interest (POI)", i.e. whether an employee was involved in the fraud. By applying machine learning skills, I am trying to better a model which would serve for the goal for the project.

The data combines 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: 14 financial features, 6 email features and POI labels.

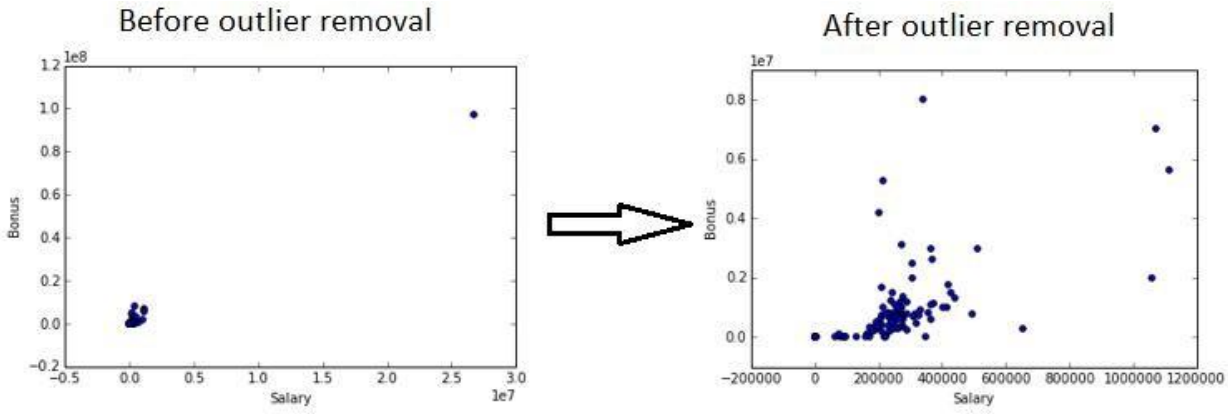
Financial_features: ['salary', 'deferral_payments', 'total_payments', 'loan_advances', 'bonus', 'restricted_stock_deferred', 'deferred_income', 'total_stock_value', 'expenses', 'exercised_stock_options', 'other', 'long_term_incentive', 'restricted_stock', 'director_fees'] (all units are in US dollars)

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)

There was a total of 146 observations (i.e. employees) with each observation having 21 features (14 financial + 6 email + POI label). Out of the 146 observations, 18 employees have the POI label. This is relatively a small sample size for a prediction model which could present a challenge in the analysis.

Making a scatterplot with the features, “salary” vs “bonus”, gave out a outlier, very far away from the rest of the data as seen in the below figure.



A review of the financial information found that this outlier corresponds to the “TOTAL” line item in the financial information, so it was removed by deleting it immediately after loading the dataset.

Taking in analysis, with a percentile of values, to remove outliers is not something to be done with the data and the question in hand, because the particular interest of this project is to find those high values so as to investigate the person in question with whether that employee was involved in the fraud or not.

Reviewing the file, line by line, also showed an observation for ““THE TRAVEL AGENCY IN THE PARK”. Since this does not correspond to an individual, it was also removed, leaving 144 individuals for the analysis.

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.**

Next step in my analysis, was to look out for missing values, i.e NaN values for the entries in the column values missing. The chart with each column’s percentage of values missing for each feature is shown below. If the percentage too big, then there’s no point in taking that feature into consideration for constructing the predictive model, I might as well drop it.

Then I created, three new features, two of these new features were calculated by finding the fraction of emails a person received (or sent) that involved a POI out of the total emails a person received (or sent). If a high percentage of an employee’s emails involved a POI, it may be more likely that they are also a POI. I named these two features as: “sent_to_poi_percent” and “rec_from_poi_percent.” The third feature I created determined the ratio of total payments to total stock value.

Feature	% NaN out of 144
salary	35%
deferral_payments	74%
total_payments	15%
loan_advances	98%
bonus	44%
restricted_stock_deferred	88%
deferred_income	67%
total_stock_value	13%
expenses	35%
exercised_stock_options	30%
other	37%
long_term_incentive	55%
restricted_stock	24%
director_fees	89%
to_messages	40%
from_poi_to_this_person	40%
from_messages	40%
from_this_person_to_poi	40%
shared_receipt_with_poi	40%
sent_to_poi_percent	40%
rec_from_poi_percent	40%
payments_to_stocks	26%