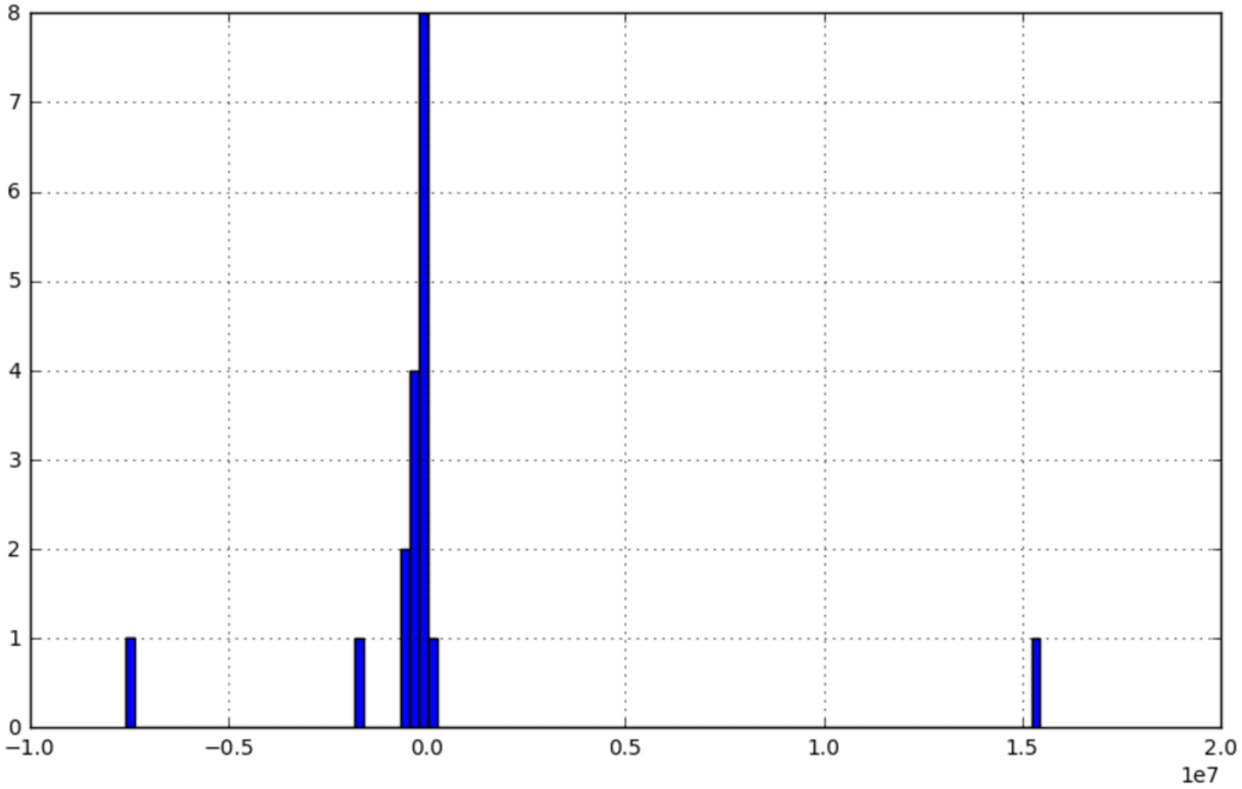
1. **Summarize for us the goal of this project and how machine learning is useful in trying to accomplish it. As part of your answer, give some background on the dataset and how it can be used to answer the project question. Were there any outliers in the data when you got it, and how did you handle those? [relevant rubric items: “data exploration”, “outlier investigation”]**

Back in the early 2000’s, Enron was one of the largest companies in the United States to go bankrupt due to corporate fraud. Since their collapse, the Federal Energy Regulatory Commission made this data available during Enron’s investigation. For this project, we are using the data to identify persons of interest (POI) based on a person’s financial and email data.

The dataset is comprised of 146 individuals and 21 features. Most of these individuals were senior level management. Of these 146, 18 are identified as POIs. Skimming through the list of individuals, I noticed that "THE TRAVEL AGENCY IN THE PARK" and “TOTAL” was part of the dataset. Since these records are not people, I removed both from the dataset. Also through inspection, I notice that “LOCKHART EUGENE E” had no data and was also removed.

Using exploratory data analysis, there was a noticeable outlier in “restricted_stock_deferred”. Upon further investigation, it looks like there was a data entry error in the stock features for "BELFER ROBERT" and "BHATNAGAR SANJAY".



After removing the aforementioned individuals, the dataset is now at 141.

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. [relevant rubric items: “create new features”, “properly scale features”, “intelligently select feature”]**

E-mail Address, Director Fees, and Loan Advances were manually excluded. Director Fees and Loan Advances were removed because 90.1% and 97.9% of the observations were ‘NaN’ respectively. Also, POI’s were not prevalent in those features. E-mail Address was removed because it was a non-numeric feature. Using “SelectKBest” for automated feature selection, the following features and scores were selected

Selected Features	Scores
exercised_stock_options	24.43
total_stock_value	23.61
bonus	20.26
salary	17.72
deferred_income	11.18

When exploring the data, salary and bonus are correlated. This makes sense because the more you make, the higher bonus you get. With this insight, I created a new feature called “salary_bonus_ratio”. By adding this feature, my classifier improves.

Feature List	Precision	Recall
w/o salary_bonus_ratio	0.41945	0.33850
w/ salary_bonus_ratio	0.42572	0.36250

3. **What algorithm did you end up using? What other one(s) did you try? How did model performance differ between algorithms? [relevant rubric item: “pick an algorithm”]**

The algorithm that performed the best was the Decision Tree Classifier in comparison to Naïve Bayes. Below is the performance of the classifiers:

Algorithm	Precision	Recall	F1 Score
Decision Tree	0.42572	0.36250	0.39157
Naïve Bayes	0.60654	0.25050	0.35456

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 do not 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 for the model that was not your final choice or a different model that does utilize parameter tuning, e.g. a decision tree classifier). [relevant rubric item: "tune the algorithm"]**

The purpose of tuning the parameters is to optimize the performance of the algorithm for the given dataset. If not done well, over fitting may occur. To tune my algorithm, I used "GridSearchCV" to work through several combinations of parameters and determine which combination performed the best. The parameters I was interested in tuning for my decision tree was "min_samples_split" and "criterion"

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 used to separate your data into separate training and test data sets. By separating your data into separate data sets, you can check how your algorithm performs against data it never seen before (the test set). Depending on how well it performs, this gives you feedback on how you should tune or adjust the algorithm. If you do not create an independent data set and test the algorithm against itself, over fitting may occur. A classic mistake one can make with validation is not randomizing the test and training sets.

StratifiedShuffleSplit was used for validation. Using this method creates multiple and different iterations of test/train splits. Also, StratifiedShuffleSplit keeps the proportion of POI's / non-POI's when splitting the data.

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"]**

The precision of my algorithm was 0.42572. This means out of all individuals I identified as a POI, 42.6% of those individuals were actually a POI.

The recall of my algorithm was 0.36250. This means out all actual POIs, my algorithm is able to identify 36.3% of them. 63.7% were identified as non-POI.

Citation:

Cohen, W. W., MLD, CMU. (2015, May 8). Enron Email Dataset. Retrieved January 10, 2017, from https://www.cs.cmu.edu/~./enron/

Galkin, A. (2011, Nov). What is the difference between test set and validation set? [Msg 2]. Message posted to http://stats.stackexchange.com/questions/19048/what-is-the-difference-between-test-set-and-validation-set?noredirect=1&lq=1

Gibney, A., & McLean, B. (n.d.). Enron: The Smartest Guys in the Room. Retrieved December 12, 2016, from https://www.netflix.com/watch/70024087?trackId=13752289&tctx=0%2C0%2Cfbfdf2e2e887b674a784fc4f17b1b249cfc9423b%3A4268967489d55d843cd28e64ec772f7cd4a0bd15

Capell, H. (2016, Mar). How to Get KBest features kept [Msg 3]. Message posted to https://discussions.udacity.com/t/how-to-get-kbest-features-kept/160463/3

Sklearn.model_selection.GridSearchCV. (n.d.). Retrieved January 10, 2017, from http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html