1) The goal of this project is to utilize machine learning techniques in order to analyze a data set and build a predictive model. Specifically, we analyze the Enron data set which contains financial and email data from a number of employees of that former corporation which suffered bankruptcy due to employee fraud. Our model attempts to predict whether an individual in the data set is a person of interest (i.e. whether they may have committed fraud). Every individual in the data set possesses a boolean value (or class) indicating whether or not they are a POI and as such this is a supervised learning task. The problem of identifying POIs requires that we split our data into training and test sets in order to avoid overfitting and accurately verify our results. Because of the numerous features in the data set, machine learning will help to uncover the underlying basis for those people who are persons of interest. As we build our model, accuracy, precision, recall and F1 scores guide our decisions (primarily accuracy and F1). The benchmark model used to arrive at these scores is a Gaussian Naive Bayes Classifier.

An exploration of the data set revealed the number of instances to be relatively scant. There are 146 individuals of which only 18 are persons of interest. There are 21 features (including the target), with 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']

email features:

['to_messages', 'from_poi_to_this_person', 'email_address', 'from_messages', 'from_this_person_to_poi', 'shared_receipt_with_poi']

and target: ['poi']

The features have various levels of missing/non-existent values, and several of the email features contain information about interactions with POIs constituting data leakage. Because of this I chose to exclude all of the email features for the remainder of the project except for 'email_address' which is later used to engineer features from the email corpus. The email corpus was also deficient of data, for example Andrew Fastow, who was a principal architect of the scandal, did not have any email data and

his communication with accountant Arthur Andersen would have proved to be valuable.
Before removing any outliers or individual features programmatically, I used the entire data set with all features (15 financial features) to get a benchmark score:

GaussianNB()
Accuracy: 0.33473 Precision: 0.15420 Recall: 0.88950 F1: 0.26284 F2: 0.45529
Total predictions: 15000 True positives: 1779 False positives: 9758 False negatives: 221
True negatives: 3242

To impute missing numerical and text data, I used a zero value (instead of an average) and empty string respectively. There were several outliers in the data set that were removed because they were either not a person but an organization, represented a cumulative total, or did not have enough relevant data. The first outlier removed was based upon the insiderpay.pdf file provided to us. From this, the organization 'THE TRAVEL AGENCY IN THE PARK' was removed because it is not a person (also not POI). By plotting 'total_payments' by 'total_stock_value' I was able to identify the 'TOTAL' outlier in the data set. This was also removed because it is not a person. Finally, I chose to remove all individuals without any financial data. There was only one person, 'LOCKHART EUGENE E', who was not a POI. The scores slightly decreased after removing the outliers but the amount was negligible:

GaussianNB()
Accuracy: 0.32793 Precision: 0.15279 Recall: 0.88900 F1: 0.26076 F2: 0.45272
Total predictions: 15000 True positives: 1778 False positives: 9859 False negatives: 222
True negatives: 3141

After removing all outliers I split the data into training and testing sets to ensure that the distribution of POIs was equal. The training set contained 100 people with 13 POIs and the test set contained 43 people with 5 POIs. Both had roughly the same distribution of POIs to non-POIs (13% and 12%).

2) Feature selection and engineering was the most involved part of the project. I pruned financial features based upon available data, engineered text features from email data, and selected the best features using a combination of methods.
Reducing the number of features was done on the entire data set and naturally followed from the

previous data exploration. The first step in pruning was to compare the amount of available data between features. I tallied the total number of non-nulls in each feature giving extra weight to POIs by their inverse proportion. I selected 8 features with the highest scores to keep:

['total_stock_value', 'total_payments', 'restricted_stock', 'expenses', 'other', 'salary', 'bonus', 'exercised_stock_options']

and excluded the rest with limited data:

['deferral_payments', 'loan_advances', 'restricted_stock_deferred', 'deferred_income', 'long_term_incentive', 'director_fees']

After reducing the number of features, accuracy and precision more than doubled. Recall decreased significantly although the F1 score remained nearly the same:

GaussianNB()
Accuracy: 0.84773 Precision: 0.36450 Recall: 0.19100 F1: 0.25066 F2: 0.21110
Total predictions: 15000 True positives: 382 False positives: 666 False negatives: 1618
True negatives: 12334

Text features were engineered based upon training data (as the process involved a model to predict written email words indicative of a POI). The features are a numerical count of the frequency of a particular word used by every individual. The features were created by aggregating text from a specified number of emails, stemming each word (linguistic root), computing term frequency-inverse document frequency (TF-IDF) of all words for all people in the data set, fitting a decision tree model using the TF-IDF to predict POIs, extracting the most important words from the model, and using those important words to count how many times each was written by a person in the original text aggregate. The important words vary based upon the number of emails processed and the particular self-identifying and nonsense words that are excluded. I chose to exclude any words including numbers and those that I believed to be self identifying or nonsense. After processing 100 emails per person, the words that I consistently encountered were:

[u'blown', u'boardroom']

I did not go through the emails to see if any of these were in fact self-identifying, so it remains possible that they are. 'Boardroom' may have been a place were meetings took place to discuss things that were not meant to be documented, and 'blown' may have been a way to describe the situation of how finances were managed as things came to an end. Adding these text features increased the total feature count to 10. Evaluation showed that accuracy, precision, and recall all improved:

GaussianNB()
Accuracy: 0.88667 Precision: 0.61244 Recall: 0.40850 F1: 0.49010 F2: 0.43765
Total predictions: 15000 True positives: 817 False positives: 517 False negatives: 1183
True negatives: 12483

The feature selection process involved a combination of Pearson correlation analysis, recursive feature selection, and tree-based classifiers. Each step added a score to rank the importance of each feature and a cumulative total was used to select the best ones. The cumulative feature ranks were almost consistently (although scores varied):

[('total_stock_value', 10), (u'boardroom', 10), ('bonus', 9), ('exercised_stock_options', 9), ('salary', 5), (u'blown', 5), ('restricted_stock', 5), ('total_payments', 4), ('expenses', 4), ('other', 2)]

I reduced the number of features to get the best 6:

['total_stock_value', u'boardroom', 'bonus', u'blown', 'exercised_stock_options', 'restricted_stock']

After reducing the number of features to these 6, precision and recall both improved:

GaussianNB()
Accuracy: 0.87785 Precision: 0.65125 Recall: 0.44350 F1: 0.52766 F2: 0.47372
Total predictions: 13000 True positives: 887 False positives: 475 False negatives: 1113
True negatives: 10525

3) Among several classifiers that were tested, k-nearest-neighbors (KNN) and adaBoost were the most promising. K-nearest-neighbors requires range scaling in order to preprocess data and so this was used in a pipeline along with a min-max scaler. Using this pipeline but without any parameter specifications (5 nearest neighbors by default), the k-nearest-neighbor classifier scored:

Pipeline(steps=[('scaler', MinMaxScaler(copy=True, feature_range=(0, 1))), ('knn', KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski', metric_params=None, n_neighbors=5, p=2, weights='uniform'))])
Accuracy: 0.85438 Precision: 0.60075 Recall: 0.15950 F1: 0.25207 F2: 0.18697
Total predictions: 13000 True positives: 319 False positives: 212 False negatives: 1681
True negatives: 10788

This was a decline over the benchmark classifier but I decided to keep it, as tuning parameters in the next step might allow this to outperform the benchmark. The adaBoost classifier was initially used outside of a pipeline and scored:

AdaBoostClassifier(algorithm='SAMME.R', base_estimator=None, learning_rate=1.0, n_estimators=50, random_state=None)
Accuracy: 0.84931 Precision: 0.51386 Recall: 0.38000 F1: 0.43691 F2: 0.40089
Total predictions: 13000 True positives: 760 False positives: 719 False negatives: 1240
True negatives: 10281

I decided to use a pipeline and combine adaBoost with a standard scaler and principle component analysis. It performed better with these than alone (although still a decline compared to the benchmark) but previous successes with this classifier and the promise of parameter tuning kept it in my repertoire:

Pipeline(steps=[('scaler', StandardScaler(copy=True, with_mean=True, with_std=True)), ('reducer', PCA(copy=True, n_components=None, whiten=False)), ('ada', AdaBoostClassifier(algorithm='SAMME.R', base_estimator=None, learning_rate=1.0, n_estimators=50, random_state=None))])
Accuracy: 0.86331 Precision: 0.57189 Recall: 0.44350 F1: 0.49958 F2: 0.46435
Total predictions: 13000 True positives: 887 False positives: 664 False negatives: 1113

True negatives: 10336

4) Parameter tuning is an important part of the model building process as it allows us to optimize our model's machinery. In some cases the parameters are very sensitive to the nature of the data and only certain values will allow the model to perform well at all. In other cases the model may perform fairly well but we want to optimize it for a particular evaluation metric. Parameters that are not optimized within the model itself are referred to as hyperparameters. They can be found by fitting the model on a parameter space and choosing those with the best fit. In order to find these optimal parameters, I used a grid search and cross-validated over a stratified shuffle split on the training portion of the data only (again all model generation was done with the same training set). I also optimized the search to find the best 'recall' score.

For KNN, the parameter space was searched to optimize the number of nearest neighbors as well as the weighting of distance to neighbors. The number of nearest neighbors used can drastically affect the predicted class of an event as can the weighting of a neighbors distance. The best number of neighbors proved to be 1 and the best weight function 'uniform':

Pipeline(steps=[('scaler', MinMaxScaler(copy=True, feature_range=(0, 1))), ('knn', KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski', metric_params=None, n_neighbors=1, p=2, weights='uniform'))])
Accuracy: 0.87631 Precision: 0.62596 Recall: 0.48700 F1: 0.54781 F2: 0.50963
Total predictions: 13000 True positives: 974 False positives: 582 False negatives: 1026
True negatives: 10418

I varied the parameter space of the base estimator and learning rate for the adaBoost classifer. AdaBoost is an ensemble method which uses many estimators, and the learning rate determines how much each individual estimator contributes to the whole ensemble. The base estimator is the model type for each individual estimator, and changing the nature of the individual model can obviously affect the outcome of the whole. The best base estimator was a default decision tree and the learning rate was best at 20:

Pipeline(steps=[('scaler', StandardScaler(copy=True, with_mean=True, with_std=True)),

('reducer', PCA(copy=True, n_components=None, whiten=False)), ('ada',
AdaBoostClassifier(algorithm='SAMME.R',
base_estimator=DecisionTreeClassifier(compute_importances=None, criterion='gini',
max_depth...om_state=None, splitter='best'),
learning_rate=20.0, n_estimators=50, random_state=None))])
Accuracy: 0.84869 Precision: 0.50835 Recall: 0.50250 F1: 0.50541 F2: 0.50366
Total predictions: 13000 True positives: 1005 False positives: 972 False negatives: 995
True negatives: 10028

Both of the algorithms that I chose accomplished their best accuracy and F1 scores after parameter tuning and final feature pruning. Changing the default hyperparameters for KNN drastically improved the F1 score while adaBoost improved marginally.

5) Validation is the process by which we evaluate our model's performance. It allows us to verify whether or not to remove particular outliers, include some features, or use one model over another. To do this we often train the model on a portion of the entire data set and test on another (just as we have developed all predictive capabilities of our model on a training set only). By testing our model on unseen data we can see how well it performs in a 'real-world' situation. Because of the limited size of our data set, and in order to accurately test our model's performance, we use a stratified shuffle split. This allows us to randomly select train/test portions of the entire data set, fitting the model on the training set and testing on the test set to generate evaluation scores. With stratified shuffle split, this random selection is done numerous times, and each time the proportion of POIs is preserved across both the train and test sets.

6) The benchmark gaussian naïve bayes classifier was the most accurate and precise model:

Accuracy: 0.87785 Precision: 0.65125 Recall: 0.44350 F1: 0.52766 F2: 0.47372
Total predictions: 13000 True positives: 887 False positives: 475 False negatives: 1113
True negatives: 10525

The k-nearest-neighbor classifier was very close in accuracy and precision, but had the highest F1 score

of all the models:

Accuracy: 0.87631 Precision: 0.62596 Recall: 0.48700 F1: 0.54781 F2: 0.50963
Total predictions: 13000 True positives: 974 False positives: 582 False negatives: 1026
True negatives: 10418

The adaBoost classifier had the highest recall score although it's accuracy and F1 scores were lower than the other two:

Accuracy: 0.84869 Precision: 0.50835 Recall: 0.50250 F1: 0.50541 F2: 0.50366
Total predictions: 13000 True positives: 1005 False positives: 972 False negatives: 995
True negatives: 10028

Accuracy is the proportion of time that we're able to correctly predict the class of all observations. In terms of our data set, it means that if you give us a person, our model can correctly predict whether or not they are a POI some proportion of the time. Precision represents the proportion of time we are correct given our model has chosen a positive classification. Statistically speaking, it is the number of true positives divided by the total number of true positives and false positives. In other words, when we evaluate somebody and our model says that they are indeed a person of interest, our chances of being correct are equal to the precision. Recall represents the proportion that the model can correctly identify a positive class, given it actually has one under scrutiny. It is the number of true positives divided by the total number of false negatives and true positives. So, if we are in fact evaluating a POI and the model has not yet tried to make a classification, our model's chance of correctly identifying them when it does make a classification is equal to the recall.

Ultimately, each of the 3 classifiers had their strengths and weaknesses. Which one is best depends upon the evaluation metrics that are considered most important for real-world applications. Given the context of this problem, I would argue that recall is the most important metric following accuracy. For a federal trial I imagine that it would be logistically impractical to subpoena all employees of Enron. Those that are brought in for questioning and are in fact guilty, should be found and charged with as little as error as possible. Maximizing recall for our model accomplishes this and the adaBoost classifier had the highest recall score of all the models that also achieved higher than 80% accuracy.