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”]

   The goal of this project was to use machine learning to predict whether a person is a person of interest based on their financial and email data/features. Machine learning is useful since it can find patterns in the data/features and use them to classify whether they are a person of interest. The data included 133 data points, 18 of which were POIs and 115 were not. My algorithm used 7 features.

   There were features with missing data, denoted “NaN”, and were treated as 0 in the feature list. This actually proved useful to the algorithm since some features were missing mostly for one label, so if the algorithm saw that it was missing, it would be able to use that information for classification. There was one outlier with the data, which was the “TOTAL” data point. I removed that completely from the data. For my created feature (ratio between stock and payments), there were people who did not have a bonus or salary, so the ratio would be inflated. These were outliers for me so I set the ratio for those people to 0.

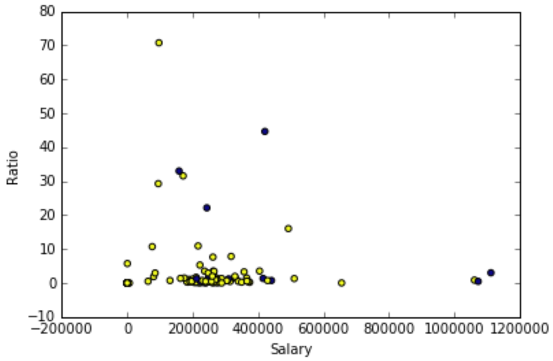
2. What features did you end up using in your POI identifier, and what selection process did you use to pick them? Did you have to do any scaling? Why or why not? As part of the assignment, you should attempt to engineer your own feature that does not come ready-made in the dataset -- explain what feature 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”]

   For my financial features, I wanted to use features that had a clear difference between POI and non-POI, and also features that were not missing many values. At first, I used salary and bonus and later added Exercised Stock Options since it was not missing many values, and Restricted Stock Deferred since it was a feature that mostly non-POI had. For email features, I used 'from_poi_to_this_person', 'from_this_person_to_poi', and 'shared_receipt_with_poi' because in the lesson, it was shown that these features were effective. I tried about 10 different combination of these before getting the performance I wanted. The combinations and performances are show below.

Features	F1	Recall	Precision
'poi','salary', 'bonus','from_poi_to_this_person', 'from_this_person_to_poi','shared_receipt_with_poi', 'ratio'	0.29094	0.29050	0.29137
'poi','salary', 'bonus','from_poi_to_this_person', 'from_this_person_to_poi','shared_receipt_with_poi'	0.26846	0.21000	0.37201
'poi','salary', 'bonus','ratio','from_poi_to_this_person', 'from_this_person_to_poi','shared_receipt_with_poi', 'exercised_stock_options','restricted_stock_deferred'	0.21551	0.18200	0.26415
'poi','salary', 'bonus','from_poi_to_this_person', 'from_this_person_to_poi','shared_receipt_with_poi', 'exercised_stock_options','restricted_stock_deferred'	0.30498	0.30650	0.30347

I did have to scale all these features using a minmax scaler to use SVM.

I also decided to create a feature which calculated the ratio between Total Stock Value and Total Payments. I noticed that for most POIs, this ratio was high and when combined with the salary, seemed to differentiate the two labels pretty well:



Unfortunately, it did not end up increasing the performance, so I left it out of the final feature set, but did append it to my_feature_list list.

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”]

   I tried SVM, Gaussian NB and Decision Trees. SVM performed the best, followed by Decision Trees. I was not able to tune anything on NB, which is probably why it performed the worst.

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”]

Tuning the parameters of an algorithm is essential to its performance. It dictates how much each data point affects the algorithm. If not done well, it will cause overfitting on one extreme, and an untrained/poorly trained algorithm on the other.

I tuned the PCA components, gamma and C parameters of the SVM, and “min_samples_split”, “max_depth”, “min_samples_leaf”, and “max_leaf_nodes” parameters of the decision tree. I used GridSearchCV to do this. I used a range of values from the SKLearn Documentation, and when the optimal parameters was on the edge of that range, I extended the range. I printed the best parameters and fed it into my clf pipeline.

The Grid returned the following optimal PCA components:

Component	Explained Variance Ratio
1	0.55336179
2	0.16567931
3	0.08130132
4	0.06943843
5	0.04855277
Total	0.918333626212

The PCA reduced the number of components from 7 to 5, but still kept about 92% of the variance.

~~Note: I did want to combine KFold and GridSearchCV so that GridSearchCV would optimize to an average score of the K runs, but did not know how to do that. I am guessing I would need to define a custom scoring function.~~ Previous reviewer explained how to do this. After doing further research, it looks like GridSearchCV uses StratifiedKFold by default anyway.

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 the process of assessing the performance of the algorithm on independent datasets. If done wrong, the data can be overfit and will not perform as well on real, separate data as it does on the training data. I validated my data by splitting it into independent training and testing sets, and also used KFold Cross Validation.

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”]

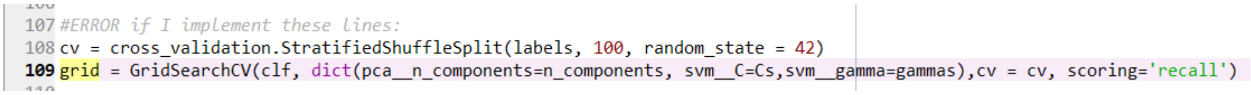
Since the majority of the data is non-POI, the accuracy score would be misleading for this case. I used Precision, Recall and F1 score to evaluate my performance. The precision was consistently high, with an average of about 0.4-0.5. I originally used F1 Score(around 0.4-0.5) as the objective of my algorithm, but found that it would give consistently high Performance score, but low Recall score (around 0.1-0.3), which meant that if classified as a POI, it is very likely a POI, but it would also fail to classify actual POIs correctly; there were many false negatives. This was most likely due to the lack of actual POIs in the original training set.

As a result, I used Recall as my objective and was able to achieve Recall and Precision above 0.3.

Summary

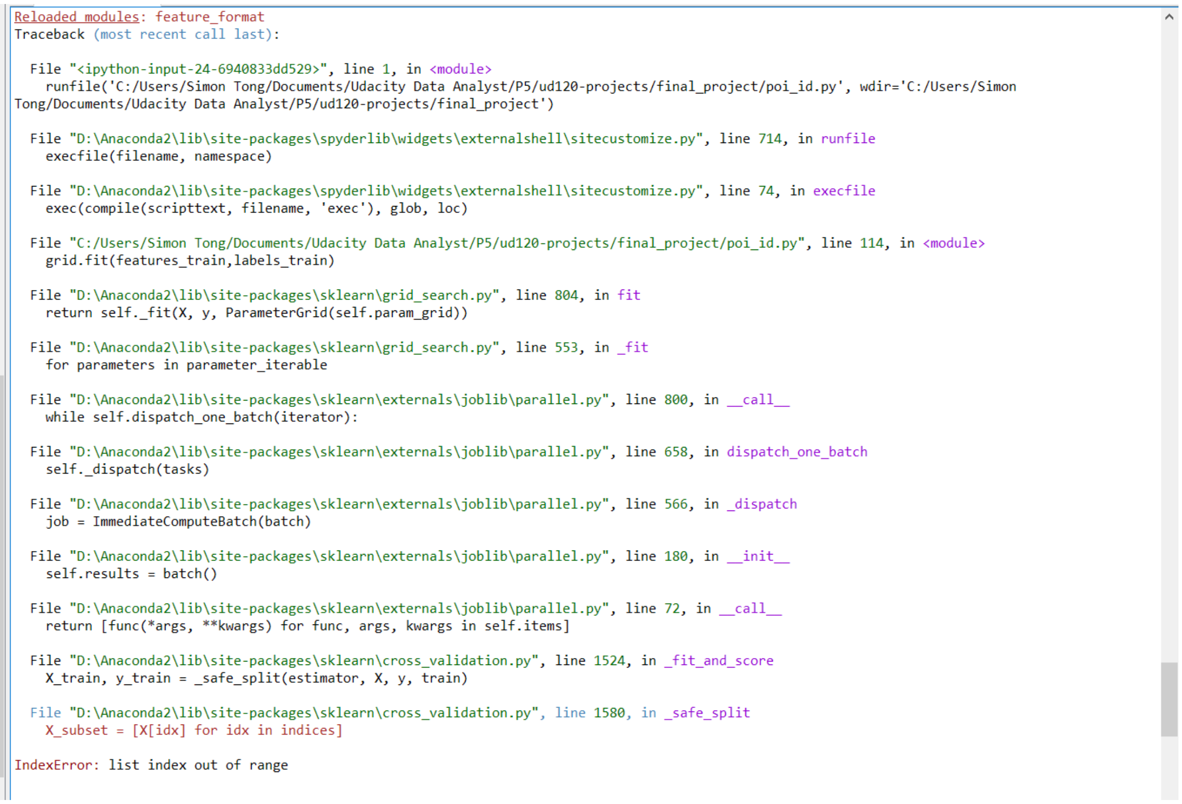
This project really demonstrated the difficulties of a Machine Learning project. While coding it up was straight forward, actually choosing features, algorithms and parameters to get a well performing algorithm was way more difficult than I thought.

Update: After the first review, I tried to implement the StratifiedSplitShuffle CV method in my GridSearchCV per the recommendation:



```
#ERROR if I implement these lines:
cv = cross_validation.StratifiedShuffleSplit(labels, 100, random_state = 42)
grid = GridSearchCV(clf, dict(pca__n_components=n_components, svm__C=Cs,svm__gamma=gammas),cv = cv, scoring='recall')
```

However, I got an error every time I tried to fit the grid to my training features and labels:



```
Reloaded modules: feature_format
Traceback (most recent call last):

  File "<ipython-input-24-6940833dd529>", line 1, in <module>
    runfile('C:/Users/Simon Tong/Documents/Udacity Data Analyst/P5/ud120-projects/final_project/poi_id.py', wdir='C:/Users/Simon
Tong/Documents/Udacity Data Analyst/P5/ud120-projects/final_project')

  File "D:\Anaconda2\lib\site-packages\spyderlib\widgets\externalshell\sitecustomize.py", line 714, in runfile
    execfile(filename, namespace)

  File "D:\Anaconda2\lib\site-packages\spyderlib\widgets\externalshell\sitecustomize.py", line 74, in execfile
    exec(compile(scripttext, filename, 'exec'), glob, loc)

  File "C:/Users/Simon Tong/Documents/Udacity Data Analyst/P5/ud120-projects/final_project/poi_id.py", line 114, in <module>
    grid.fit(features_train,labels_train)

  File "D:\Anaconda2\lib\site-packages\sklearn\grid_search.py", line 804, in fit
    return self._fit(X, y, ParameterGrid(self.param_grid))

  File "D:\Anaconda2\lib\site-packages\sklearn\grid_search.py", line 553, in _fit
    for parameters in parameter_iterable

  File "D:\Anaconda2\lib\site-packages\sklearn\externals\joblib\parallel.py", line 800, in __call__
    while self.dispatch_one_batch(iterator):

  File "D:\Anaconda2\lib\site-packages\sklearn\externals\joblib\parallel.py", line 658, in dispatch_one_batch
    self._dispatch(tasks)

  File "D:\Anaconda2\lib\site-packages\sklearn\externals\joblib\parallel.py", line 566, in _dispatch
    job = ImmediateComputeBatch(batch)

  File "D:\Anaconda2\lib\site-packages\sklearn\externals\joblib\parallel.py", line 180, in __init__
    self.results = batch()

  File "D:\Anaconda2\lib\site-packages\sklearn\externals\joblib\parallel.py", line 72, in __call__
    return [func(*args, **kwargs) for func, args, kwargs in self.items]

  File "D:\Anaconda2\lib\site-packages\sklearn\cross_validation.py", line 1524, in _fit_and_score
    X_train, y_train = _safe_split(estimator, X, y, train)

  File "D:\Anaconda2\lib\site-packages\sklearn\cross_validation.py", line 1580, in _safe_split
    X_subset = [X[idx] for idx in indices]

IndexError: list index out of range
```

In fact, when I try to set the “cv” attribute of the grid to my KFold, it throws the same error when fitting. Seems to be a similar problem to this: http://stackoverflow.com/questions/35998112/sklearn-grid-fitx-y-error-positional-indexers-are-out-of-bounds-for-x-tra, but I could not get it to work. I would like some advice on getting it to work, but it is not a big deal for this project since the GridSearchCV uses StratifiedKFold automatically anyway.