Bryan Smith - bryantravissmith@gmail.com
Project 4 Analysis

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 is to develop and tune a supervised classification algorithm to identify persons of interest (POI) in the Enron scandal based on a combination of publically available Enron financial data and email records. The compiled data set contains information for 144 people employed at Enron, a ‘Travel Agency in the Park’, and the total compensations for all of these sources. Additionally, the Udacity.com course designer created email features that give the total number of e-mails sent and received for each user, and the total number of emails sent to and receieved from a POI. Of the 144 people, 18 of them are labeled as POIs.

In regards to the financial information, I defined outliers as having values that are more than 3 standard deviations from the mean value for the group. This is not the traditional definition or criteria for an outlier which is 1.5 times the interquartile range below the first quartile or Above the third quartile. I used my definition after I have replaced missing values with zero. Using this definition of outliers there are 25 people in the data set that are financial outliers:

['ALLEN PHILLIP K','BELDEN TIMOTHY N', 'BHATNAGAR SANJAY', 'BLAKE JR. NORMAN P','FREVERT MARK A','GRAMM WENDY L','HANNON KEVIN P','HIRKO JOSEPH', 'HORTON STANLEY C', 'HUMPHREY GENE E', 'JAEDICKE ROBERT', 'LAVORATO JOHN J','LAY KENNETH L','LEMAISTRE CHARLES', 'MARTIN AMANDA K', 'MCCLELLAN GEORGE', 'PAI LOU L', 'RICE KENNETH D','SAVAGE FRANK','SHANKMAN JEFFREY A', 'SKILLING JEFFREY K', 'URQUHART JOHN A', 'WAKEHAM JOHN','WHITE JR THOMAS E', 'WINOKUR JR. HERBERT S']

This accounts for 17% of the data being considered an outlier and also contains 33% of the POI. Ultimately I decided that financial outliers were relevant information, and decided not to remove them from the data.

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 doesn’t 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.) If you used an algorithm like a decision tree, please also give the feature importances of the features that you use. [relevant rubric items: “create new features”, “properly scale features”, “intelligently select feature”]

3. What algorithm did you end up using? What other one(s) did you try? [relevant rubric item: “pick an algorithm”]

I started off by creating two features I thought were be informative to the data: the ratio of emails received from a POI to the total number of emails received and the ratio of emails sent to a POI to the total number of emails sent. After I created these features I started with a systematic search of the results of pairs of features from the data set.

I trained two classifiers, a Decision Tree and a Gaussian Naive Bayes, on all pairs of features using 10 fold cross validation, and examined the recall and precision scores on the test set. At this point I did not tune any parameters, nor normalize the data. In fact both Decision Trees and Naive Bayes are not affected by normalization. If i chose an SVM, I would have normalized.

The best results of the scan is as follows:

Type	Variable 1`	Variable 2	Recall	Precision
Tree	bonus	expenses	41.1%	38.8%
Tree	bonus	total_stock_value	45.8%	38.8%
Tree	total_stock_value	expenses	36.8%	38.8%
NB	total_stock_value	deferred_income	53.8%	38.8%
Tree	expenses	others	50%	50%
Tree	expenses	from_this_person_to_poi	43.8%	38.8%
Tree	expenses	from_poi_to_this_person	43.8%	38.8%

The decision tree consistently performed better than Naive Bayes, so I decided to to use Decision Trees for this analysis.

I next looked at my created features for the ratio of mails from or two POI. I performed the same scan including one or two of my created features. The best results follow:

In conjunction with from_poi_to_this_person_ratio

Variable 1`	Variable 2	Recall	Precision
restricted_stock	other	47.6%	55.6%

In conjunction with from_this_person_to_poi_ratio

Variable 1`	Variable 2	Recall	Precision
salary	exercised_stock_options	41.0%	50.0%
salary	restricted_stock_deferred	47.1%	44.4%
exercised_stock_options	other	52.6%	55.6%
exercised_stock_options	deferred_income	50.0%	44.4%
total_stock_value	other	42.9%	50.0%

In conjunction with from_poi_to_this_person_ratio and from_this_person_to_poi_ratio

Variable 1`	Variable 2	Recall	Precision
deferral_payments	exercised_stock_options	44.4%	44.4%
exercised_stock_options	loan_advances	44.4%	44.4%
exercised_stock_options	from_this_person_to_poi	44.4%	44.4%
exercised_stock_options	long_term_incentive	41.0%	50.0%
restricted_stock	other	41.0%	50.0%

In both cases using a single one of my created features in conjunction with two financial features performed better than using both when used with a decision tree. They did perform better than a decision tree with two variables, I am curious how these values compare to using the three best features. That scan produces the following.

Variable 1`	Variable 2	Variable 3	Recall	Precision
expenses	other	from_this_person_to_poi	68.7%	61.1%

This performed much better than using my features, so I opted to not use my features for the rest of the analysis.

In looking for the best features to use I scanned for 4 variables, importance are in parentheses:

Variable 1`	Variable 2	Variable 3	Variable 4	Recall	Precision
expenses (0.356)	other (0.329)	from_this_person_to_poi (0.316)	restricted_stocks_deferred (0.00)	70.6%	66.7%
expenses (0.356)	other (0.329)	from_this_person_to_poi (0.316)	loan advance(0.00)	68.7%	61.1%
expenses (0.251)	other (0.331)	from_this_person_to_poi (0.311)	from_messages (0.107)	70.6%	66.7%
expenses (0.305)	other (0.373)	from_this_person_to_poi (0.321)	director_fees (0.00)	68.7%	61.6%

The results are not significantly improved by an additional feature, and with the exception of 'from_messages', the relative importance is 0. The best features to use for a decision tree seem to be 'expense', 'other' and 'from_this_person_to_poi'.

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 don't 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 if you used, say, a decision tree classifier). [relevant rubric item: "tune the algorithm"]

Tuning an algorithm involving changing the parameters of a particular algorithm in the attempt to increase its performance, hopefully in a context broader than the training. If this is not done well, the algorithm can overfit the training data and perform significantly more poorly on test or real data. In the case of Decisions trees, the classification algorithms I used, this is a real danger because Decision trees tend to over fit.

After I decided on the features in the previous section, I used sklearn Grid Search to scan through the parameters to give the best performance on the training data using the following parameters over the listed values:

Split Criteria: Gini Index, Mutual Information/Entropy
Max Depths: None, 1, 2, 4, 8, 16, 32
Min Sample Split: 1, 2, 4, 8, 16, 32
Min Sample Leaves: 1, 2, 4, 16, 32

The search was performed over half of the total data set to avoid any possible overfitting. The best parameters found were ('gini',None,32,1). This classifier was passed to the classifier tester for final analysis.

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 an attempt to confirm that a model will give reasonable or consistent results on new, untrained data. A classic mistake is to test the results of a model on the data used to train the model. This is no doubt give the best possible score, but can over fit the data leading to less than desired results on new data. Validation protects against this mistake by training the model on one set of data and testing on yet another.

I used 10-Fold Cross Validation for investing and comparing algorithms in this analysis. This is where there the algorithm is trained on the on the data 10 times using 90% of the data as a training set and 10% of the data as a testing set. Each time this is done, the training and testing set are shuffled to create a new 90/10 split on the data. I then used the average performance as an estimate of its performance on new data. It is through this method I settled on the Decision Tree algorithm over Naive Base, and selected the three features I choose for this analysis.

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"]

In making decisions in this analysis, I choose the metric Recall and Precision. Recall is the ratio of the number of POI's correctly identified out of the total number of POI. Precision is the number of POI's correctly identified out of all the values predicted to be POI's.

For my analysis I found that my average recall was 36.5%, and my average precision was 47.8%.

In terms of the data, the recall means that, on average, it correctly identifies 6.5 out of the 18 POI in the data set. The precision means about half of all the people it predicts are POI, turn out to actually be POI.