Introduction

In late 2001 the Enron Corporation collapsed into bankruptcy due to corporate fraud in which Enron employees utilized accounting loopholes, special purpose entities, and poor financial reporting to hide billions of dollars of debt. During the scandal these corrupt employees received financial compensation, all of which became very apparent after a federal investigation was conducted.

Records pertaining to the scandal are publicly available and can be mined to discover patterns that distinguish a person of interest (POI) from someone who was not (a.k.a. non-POI). Machine Learning is well-suited for this task and provides a systematic framework for identifying these POIs. Furthermore, statistical metrics could be used to evaluate the machine learning classifier, thereby providing quantitative descriptions of its ability to distinguish POIs from non-POIs. The goal of the work presented here is to build a binary machine learning classifier that determines which Enron employees would have been considered "Persons of Interest" (POIs) for the infamous Enron Scandal.

Description of the Data

The data for this project consist of publicly available financial and email data for 146 Enron employees. The financial data were compiled by FindLaw and are summarized in an exhibit entitled "Payments to Insiders. Summary Schedule of all Debtors Combined". A snapshot of this data is presented in Figure 1. These financial data include various types of compensation including salary, bonus, exercised_stock_options, loan advances, etc.

PROVIDED BY FindLaw WWW.FINDLAW.COM

In re: Enron Corp.
Case No. 01-16034

Payments to Insiders
Summary Schedule of all Debtors Combined

EXHIBIT 3b.2

		Payments										Stock Value (13)			
Insider		Salary (1)	Bonus (2)	Long Term Incentive (3)	Deferred Income (4)	Deferral Payments (5)	Loan Advances (6)	Other (7)	Expenses (8)	Director Fees (9)	Total Payments	Exercised Stock Options (10)	Restricted Stock (11)	Restricted Stock Deferred (12)	Total Stock Value
ALLEN, PHILLIP K		$201,955	$4,175,000	$304,805	($3,081,055)	$2,869,717	-	$152	$13,868	-	$4,484,442	$1,729,541	$136,027	($126,027)	$1,729,541
BADUM, JAMES P		-	-	-	-	178,980	-	-	3,486	-	182,466	257,817	-	-	257,817
BANNANTINE, JAMES M		477	-	-	(5,104)	-	-	864,523	(b) 56,301	-	916,197	4,046,157	1,757,552	(560,222)	5,243,487
BAXTER, JOHN C	(a)	267,102	1,200,000	1,586,055	(1,386,055)	1,295,738	-	2,660,303	11,200	-	5,634,343	6,680,544	3,942,714	-	10,623,258
BAY, FRANKLIN R		239,671	400,000	-	(201,641)	260,455	-	69	129,142	-	827,696	-	145,796	(82,782)	63,014
BAZELIDES, PHILIP J		80,818	-	93,750	-	684,694	-	874	-	-	860,136	1,599,641	-	-	1,599,641
BECK, SALLY W		231,330	700,000	-	-	-	-	566	37,172	-	969,068	-	126,027	-	126,027
BELDEN, TIMOTHY N		213,999	5,249,999	-	(2,334,434)	2,144,013	-	210,698	17,355	-	5,501,630	953,136	157,569	-	1,110,705

Figure 1: A snapshot of the financial data provided by FindLaw.

The email data for each sample point include email addresses, the total number of messages sent, the total number of messages received, and the number of emails sent to or received from POIs. These heterogenous data can be found in the file named "final_project_dataset.pkl".

Data Exploration

A preliminary review of the data was conducted to determine the total number of sample points, the allocation of POIs and non-POIs, outliers, and missing values. The data were also visualized in histograms and scatter plots in order to obtain a quick overview of their distribution.

The raw data contain 146 records with 20 available starter features (14 financial features and 6 email features). There are 18 POIs and 128 non-POIs. There are several missing values in the data set. In particular, 50 records are missing 'salary' data and 63 records are missing values for 'bonus'.

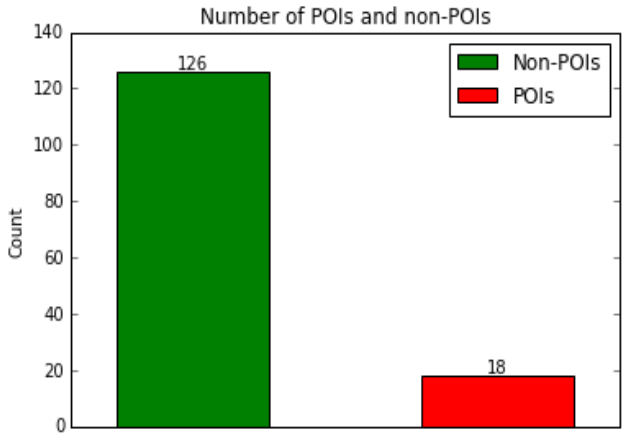

Figure 2: The raw data contain 18 POIs and 128 non-POIs.

Outliers

It was clear from the original histogram for salary (not shown here) that the raw data contain one obvious outlier, which was labeled 'TOTAL'. This record does not correspond to a person; in fact, it stores the total values of the 14 financial features for all of the other records and is easily identified in FindLaw's financial exhibit. This record was excluded from the data set.

The second outlier in the data was labeled 'THE TRAVEL AGENCY IN THE PARK'. It summarizes travel-related business expenses.

Because it does not correspond to a person and the travel expenses seem plausible, this record was excluded from analysis.

Fifty records were missing salary information, with only one of these records corresponding to POI status. So, of the initial 144 records roughly 34% (i.e. 49 of 144) were non-POI status who had missing salary. Thirty-four percent is a substantial portion of this small, imbalanced data set. The classifier could decide that missing salary is indicative of non-POI status, which is not necessarily true. Through experimentation, it was determined that removing these 50 records increased the predictive ability of this classifier.

There are records for which the bonus or salary is much higher than others. These records are outliers BUT should be included in the model, because they correspond to POIs. We need these POIs in order to train our POI identifier.

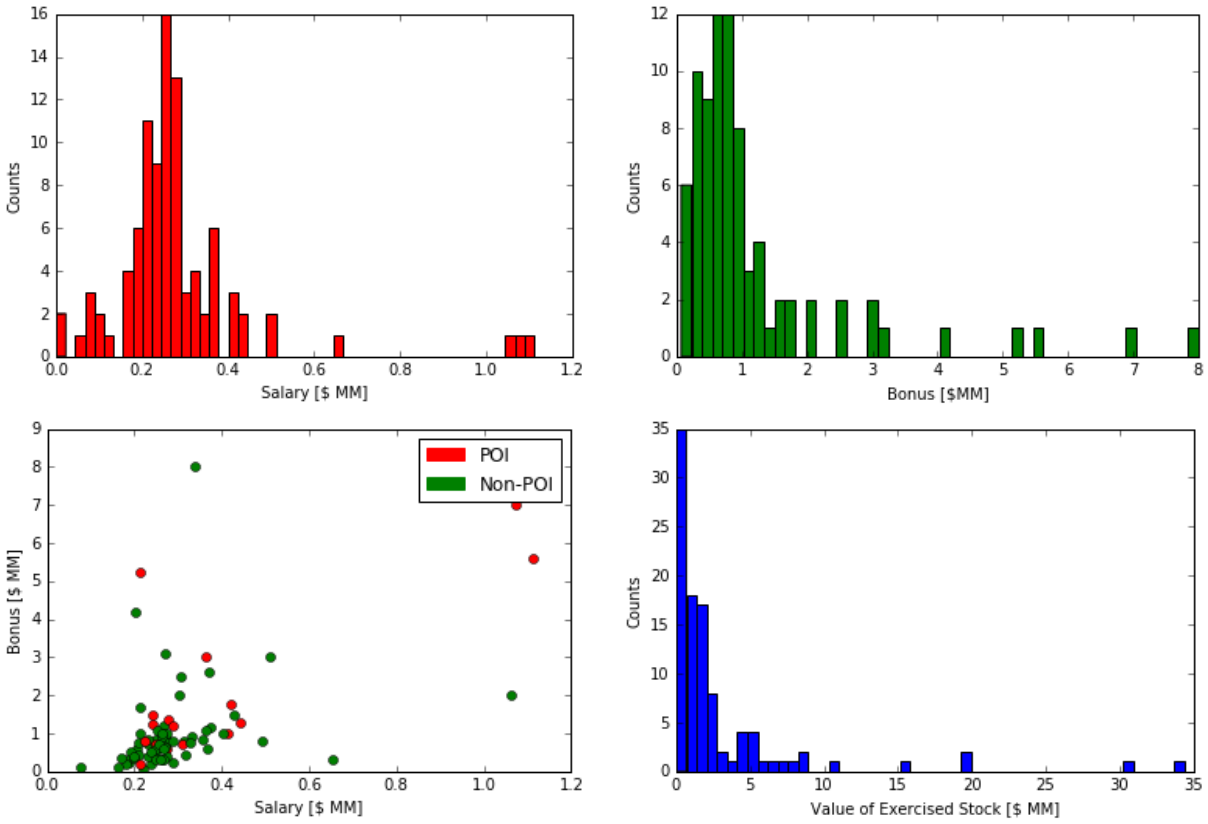

Figure 3: (Top Left) There are at least three high (outlier) salaries, corresponding to POIs.(Top Right) Histogram of bonuses reveals several outliers, also corresponding to POIs. (Bottom Left) Scatter plot of Bonus vs Salary does not reveal any obvious relationship between these two variables and shows that some outliers correspond to both POIs and non-POIs. (Bottom Right) Histogram of Exercised Stock reveals additional outliers.

After removing the two non-person outliers and the 50 records lacking salary information, the final data set contained 94 records: 17 POIs and 77 non-POIs.

Methods: Feature Selection, Classification Schemes, and Parameter Tuning

Data Transformation and Feature Selection

The natural logarithm was used to transform the data and resulted in uni-modal histograms for salary, bonus, and exercised stock options (Figure 4). This was particularly useful since many machine learning algorithms work very well when the data distributions are unimodal.

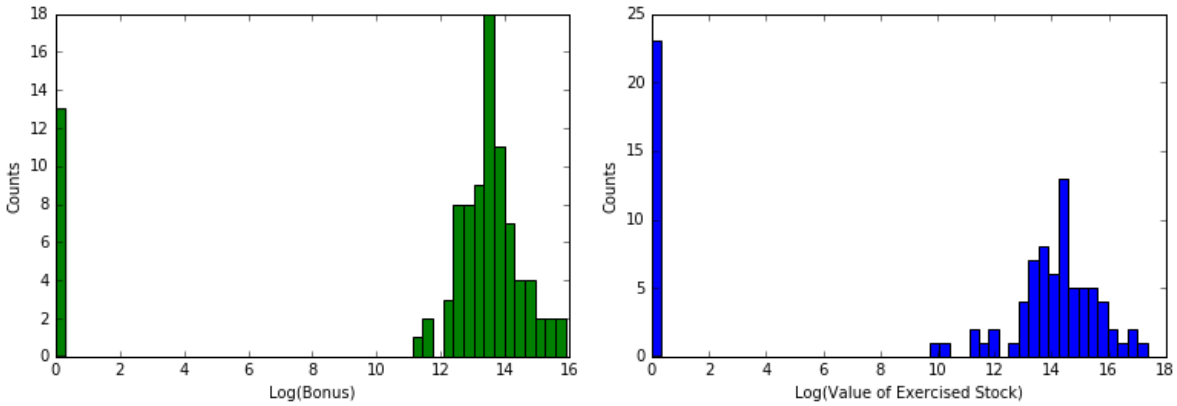

Figure 4: After logarithmic transformation the histograms of bonus and exercised stock became unimodal. See Figure 3 for comparison.

Principal Component Analysis for Selecting Financial Features

Principal component analysis (PCA) was used to reduce the 14 financial features to three principal components. PCA was performed only for the fourteen financial features since performing PCA on the combined financial and email data is not intuitive. The first, second, and third principal components comprise 32%, 23%, and 17% of the variation in the data, respectively. These components were used in the final model.

Univariate Feature Selection for Email Data

Three new email features were created from the original six, and the feature achieving the highest chi-squared statistic was

chosen for use in training. When combined with the three principal components, this yielded a total of four features for a data set consisting of 94 records.

Classification Models & Parameter Tuning

Three machine learning models were trained and tested as candidates for this POI identifier. These three models were Support Vector Classifier (SVC), Decision Tree, and Random Forest.

The hyper-parameters for each classifier were with Python's GridSearchCV method. This function tries all possible combinations of the specified parameters for each classifier and yields the optimal hyperparameter combination. The decision function for assessing the performance of each hyperparameter combination was the F1-score.

For the SVC algorithm, parameter tuning was performed for the penalty of the error term (i.e. C) and the kernel coefficient (i.e. gamma). The optimal values were determined to be C = 8 and gamma = 0.25. For the Decision Tree and Random Forest classifiers, tuning was performed for three hyperparameters: [1] the function that measures the quality of a split, [2] strategy for splitting, [3] number of features to consider for best splitting.

Model Assessment

Validation

After training the various models and choosing the optimal parameters, we allow them to make predictions using a set of data that it was not used for training. We then assess the performance of these models by calculating statistical metrics such as precision, recall, or F1 score. When performing validation, we expect a given model to perform slightly worse with the validation data than with the training data. This is because during training, the model may over-fit itself to patterns in the noise of the training data. These noise patterns are not real and will likely not exist in the validation data.

Stratified, shuffled cross-validation was utilized to facilitate validation. Cross-Validation is useful since it allows the model to be trained and tested with all of the available data, thereby providing means for the model to be exposed to the various intricacies in the data. Thirty folds were used for Cross

Validation, with each test validation set containing 10% of the data. This corresponds to a single test case per fold. The stratified nature of the folds preserves the ratio of the two target values and is especially useful when working with imbalanced data (only 17 of the 94 records used in this work were POIs). Overall, this stratified-shuffled cross validation approach scrutinizes the models well enough to allow them to be generalized.

Evaluation Metrics

Training / Cross Validation

Since the goal of this project is to build a person of interest identifier, it makes sense to utilize some measure of accuracy to compare the performances of the various models. An appropriate accuracy metric for imbalanced classes is the F1 score, and is a harmonic mean of precision and recall. The two models which achieved the highest cross-validated F1 score of 0.85 during training were the Support Vector Classifiers (Figure 5) that utilize one and two email features, respectively. The SVC model using one email feature was chosen for the final model since the two-email version yields no additional performance.

Model	No. Principal Components*	No. Email Features	F1 Score
SVC	3	0	0.74
SVC	3	1	0.85
SVC	3	2	0.85
DT	3	0	0.76
DT	3	1	0.79
DT	3	2	0.76
RF	3	0	0.77
RF	3	1	0.80
RF	3	2	0.80

Figure 5: The SVC model using one email feature was chosen as the final model. It was the simpler of two models that yielded the highest F1 score during training.

Final Classifier: Parameters and Evaluation

The final model was a Support Vector Classifier with a radial basis function as the kernel, class weights that are inversely proportional to their class frequency, a penalty value of 8 (i.e. C=8), and a kernel coefficient of 0.25 (i.e. gamma=0.25).

This final SVC model was tested with a script which creates 1,000 random folds from the final data set. The script uses the

classifier to make 10,000 predictions and calculates precision, recall, F1, and F2 scores. Out of 10,000 tests there were 960 true positives, 488 false positives, 1040 false negatives, and 7512 true negatives (Table 1).

		PREDICTED		
		Non-POI	POI	
TRUTH	Non-POI	7512	488	8000
	POI	1040	960	2000
		8552	1448	

Table 1: Confusion matrix for the final SVC model in this work.

The precision of the final SVC method was 0.66, which means that if the classifier identified an individual as being a POI there is a 66% chance that (s)he actually is a POI. The SVC method achieved a recall score of 0.48. That is, 48% of POIs were actually identified as being POIs. The F1 and F2 scores were 0.56 and 0.51, respectively.

Summary

A Support Vector Classifier was developed to identify persons of interest in the Enron Scandal. The identifier presented here yielded satisfactory scores for precision, recall, and F1 score, especially since there were few available data, many samples with missing values, and imbalanced classes.

Appendix

Python code for this project can be found in the following GitHub repository.