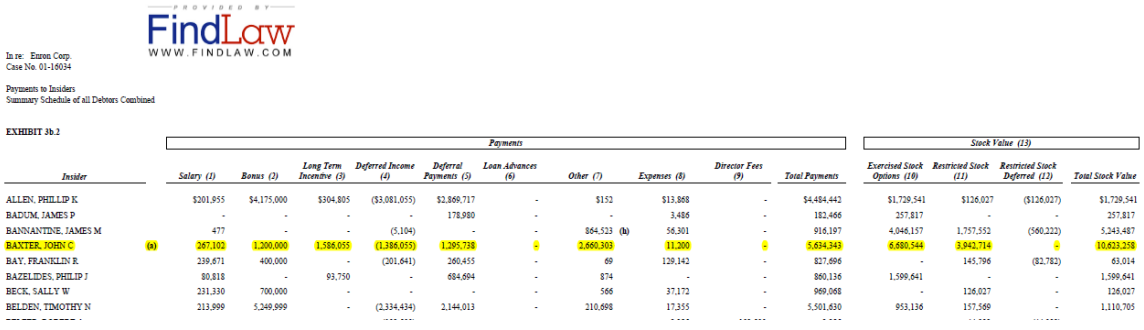
Introduction

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 in 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 involved in the scandal. 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" for the infamous Enron Scandal.

Description of the Data

The data for this work 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". The financial data provided in this exhibit include various types of compensation, including salary, bonus, exercised_stock_options, loan advances, etc. The email content includes each employee's email address, 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".



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	$126,027	($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.

Data Exploration

The data were examined for total number of records, 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 21 available starter features. There are 18 POIs; the remaining 128 records are for 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'.

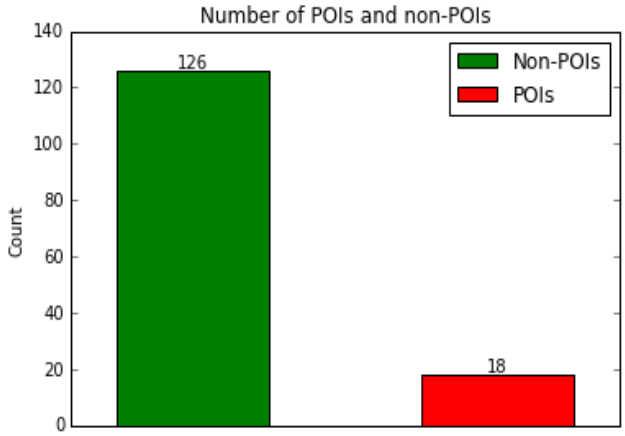


Figure 1: The raw data contain 18 POIs and 128 non-POIs.

Outliers

It was clear from the original visualizations (not shown here) that the raw data contain one obvious outlier, which was labeled 'TOTAL'. This record corresponds to the grand total, or sum, of all the other records and should obviously be 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.

The 50 records without salary information were excluded from analysis since only one of these cases was POI. This action is important for a small data set since it prevents the classifier from determining that missing salary indicates non-POI status.

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.

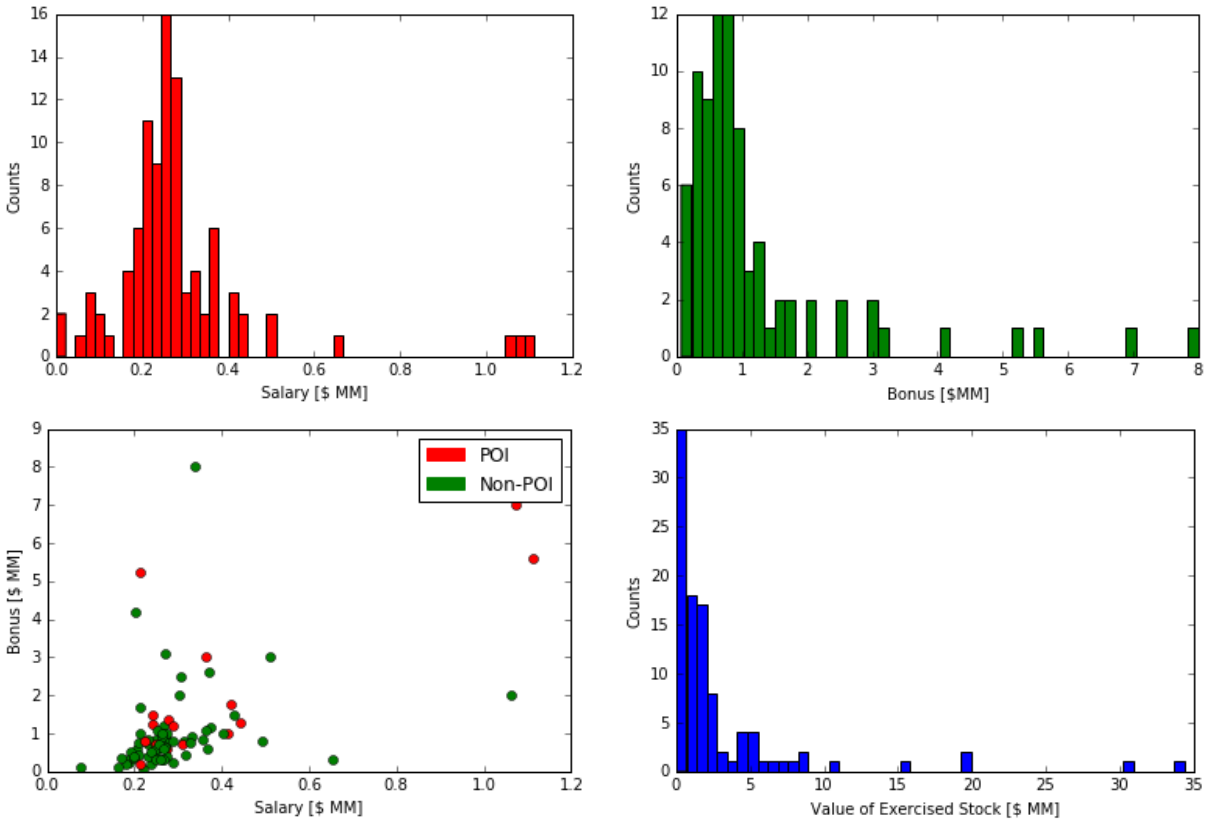


Figure 2: (Top Left) There are at least three high (outlier) salaries, corresponding to POIs.(Top Right) Histogram of bonuses reveals several outliers, also corresponding to POIs. (Bottom Left) Scatter plot of Bonus vs Salary does not reveal any obvious relationship between these two variables and shows that some outliers correspond to both POIs and non-POIs. (Bottom Right) Histogram of Exercised Stock reveals additional outliers.

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. This was particularly useful since many machine learning algorithms work very well when the data distribution are unimodal.

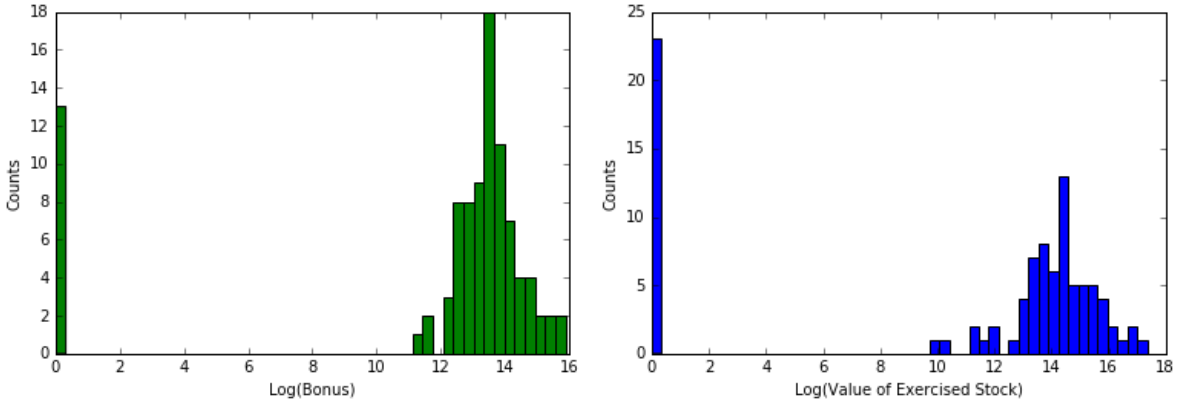


Figure 3: After logarithmic transformation the histograms of bonus and exercised stock appear unimodal. See Figure 2 for comparison.

Principal component analysis (PCA) was used to reduce the fourteen 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.

Three new email features were created from the original six and the best feature was selected when training the model. However, no email features were included in the final model since they did not improve its performance.

Classification Algorithms Tested

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

Parameter Tuning

The hyper-parameters for each of the tested algorithms were tuned in order to determine the optimal model since poor choice of parameters could lead to poor predictive performance. In this work, tuning was achieved with the Python utility GridSearchCV, which tries all possible combinations of suggested parameters and uses cross-validation (along with a decision function) to determine the optimal combination.

For the SVC algorithm--which became the final model--parameter tuning was performed for the penalty parameter (i.e. C) of the error term and the kernel coefficient (i.e. gamma). The optimal values were determined to be C = 8 and gamma = 0.25.

Model Assessment

Validation

After training a model and choosing the optimal parameters, we allow it to make predictions using a set of (testing) data that it has not been trained with. We then assess the model by calculating statistical metrics such as precision, recall, and F1 score. When performing validation, we expect the model to perform slightly worse with the validation data than it did 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 patterns will generally not exist in the new testing data.

Stratified, shuffled cross-validation was utilized to validate models in this work. 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 97 records used in this work were POIs). 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. One thousand folds were used in the Cross Validation, with each test validation set containing 10% of the data. This corresponds to a single test case per fold. This approach scrutinizes the final model well enough to allow it to be generalized.

Evaluation Metrics

Since the goal of this classifier is to identify persons of interest, it made the most sense to utilize precision and recall to evaluate the machine learning classifiers. The model which achieved the highest precision and recall was the SVC. The precision of the final SVC method was 0.66, which means that if an individual is identified as 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 actual POIs were actually identified as being POIs.

Each method was also compared using the F1 score, an accuracy metric that works well for imbalanced classes. The SVC classifier achieved an F1 score of 0.56 during testing.

For completeness a confusion matrix for the SVC model is presented here:

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

Confusion matrix for the final SVC model in this work.

Out of 10,000 tests there were 960 true positives, 488 false positives, 1040 false negatives, 7512 true negatives.

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, especially since there were few data points available and the data were imbalanced, with missing many values.

Appendix

Code

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