Udacity

Intro to Machine Learning

Project

IDENTIFY FRAUD FROM ENRON EMAIL

---

by Omoju Miller

July 9, 2016

IDENTIFY FRAUD FROM ENRON EMAIL

PROJECT OVERVIEW

In early 2000s, the Enron corporation of Houston, Texas was considered one of the most profitable energy companies in the United States. Fortune magazine named Enron one of the most innovative companies in the US for six years in a row, 1996 - 2001. However, by late 2001, the Enron corporation filed for bankruptcy after it was revealed that the company overstated its profits and defrauded it share holders of a considerable amount of wealth. Several person were later indicted and convicted of fraud in the Enron case.

The goal of this project is to use applied machine learning, based on released Enron data—Enron emails and financial accounts—to identify persons of interest in the Enron debacle.

PROBLEM STATEMENT

The problem that we are interested in solving is building an algorithm that can help us identify persons who might be of interest with regards to fraud at Enron. These persons are predominantly Enron employees and consultants who worked for the corporation.

The dataset that we have available for this task was collected over two weeks in 2002 by the Federal Energy Regulatory Commission (FERC) during its investigation into the case. The dataset contains over 600,000 emails of 158 high-level Enron employees as well as the financial records of stock payments, salary and so forth of those employees.

To identify persons of interest that might have been party to fraud at Enron we suggest the following strategy:

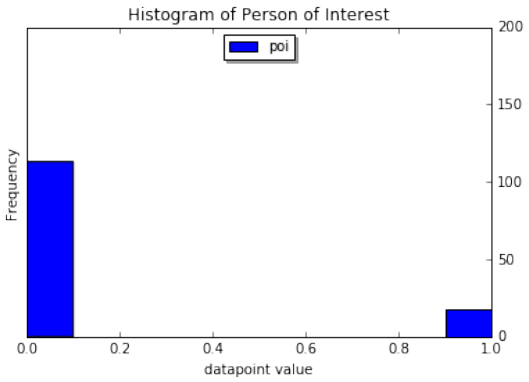
(a) Explore the dataset to ensure its integrity and understand the context.

(b) Identify features that may be used. If possible, engineer features that might provide greater discrimination.

(c) With the understanding that this a "classification" task, explore a couple of classifiers that might be well suited for the problem at hand.

(d) Once classifiers have been identified, tune them for optimality.

The Enron dataset while it contains a robust amount of emails, contains data for only about 150 people. An interesting aspect of the Enron data is that the class labels for our classification is heavily unbalanced at a ratio of around 6:1 in favor of negative examples as can be seen in figure 0.1a.



(a)

Figure 0.1: **Person of Interest plot.** *By visualizing this feature which acts as the label for our classifiers, we can see that the dataset is extremely skewed in favor of non-persons of interest.*

For this reason, using "Accuracy" as a performance metric leads to misleading information. A more apt measure of the performance of a learner should take into account the results of a confusion matrix and calculate "precision" and "recall," noting that precision is more a measure of a classifiers exactness.

ANALYSIS

DATA EXPLORATION

The Enron dataset used in the project was created by the Udacity team. It was done by combining the Enron email and financial data. The data is stored in a python dictionary where each key-value pair in the dictionary corresponds to one person. For example shows the corresponding key-value data point of Enron executive Pai Lou, the only executive to reap and keep substantial wealth from Enron. Table 0.1 shows a list of persons of interest that were generated by the Udacity team.

```
data_dict['PAI LOU L']
{'bonus': 1000000,
 'deferral_payments': 'NaN',
 'deferred_income': 'NaN',
 'director_fees': 'NaN',
 'email_address': 'lou.pai@Enron.com',
 'exercised_stock_options': 15364167,
 'expenses': 32047,
 'fraction_from_poi': 0,
 'fraction_to_poi': 0,
 'from_messages': 'NaN',
 'from_poi_to_this_person': 'NaN',
 'from_this_person_to_poi': 'NaN',
 'loan_advances': 'NaN',
 'long_term_incentive': 'NaN',
 'other': 1829457,
 'poi': False,
 'restricted_stock': 8453763,
 'restricted_stock_deferred': 'NaN',
 'salary': 261879,
 'shared_receipt_with_poi': 'NaN',
 'to_messages': 'NaN',
 'total_payments': 3123383,
 'total_stock_value': 23817930}
```

- 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, email_address, from_poi_to_this_person, from_messages, from_this_person_to_poi, shared_receipt_with_poi`
- POI label:
  `poi`

PERSONS OF INTEREST
BELDEN TIMOTHY N
BOWEN JR RAYMOND M
CALGER CHRISTOPHER F
CAUSEY RICHARD A
COLWELL WESLEY
DELAINEY DAVID W
FASTOW ANDREW S
GLISAN JR BEN F
HANNON KEVIN P
HIRKO JOSEPH
KOENIG MARK E
KOPPER MICHAEL J
LAY KENNETH L
RICE KENNETH D
RIEKER PAULA H
SHELBY REX
SKILLING JEFFREY K
YEAGER F SCOTT

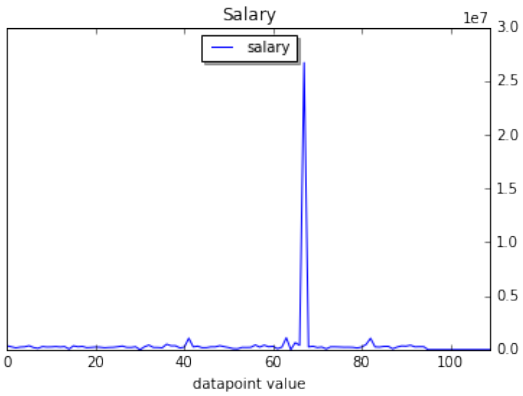
Table 0.1: Persons of interest in the dataset

While the original Enron dataset contained a robust amount of emails, this pared down one contains data for only about 150 people. Furthermore, the dataset is extremely skewed as can be seen from figure 0.1a.

Outlier detection

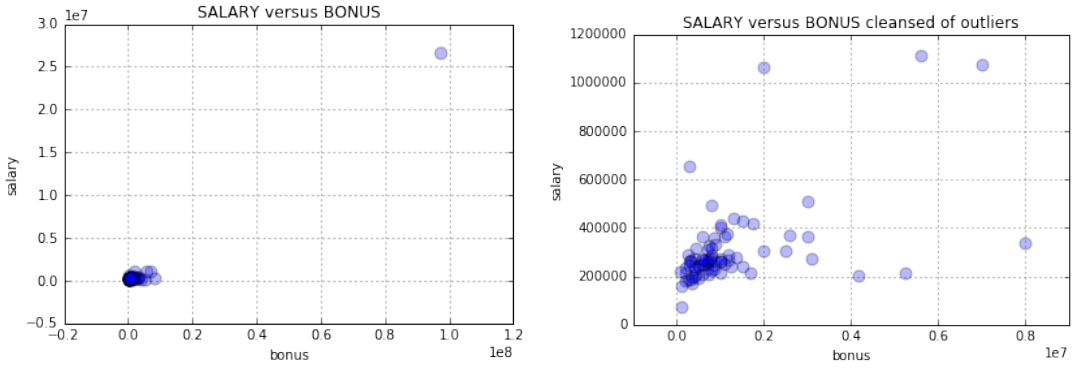
As part of the data exploration process, we were careful to analyze the data for potential outliers. One of the first tasked we did was visualize the salary of Enron executives 0.2a. From that visualization, it was clear that there was an outlier in the dataset. We found a datapoint that was completely outside of a reasonable range as can be seen from figures 0.2a and 0.3a.

A closer look revealed that the datapoint was an input error which included the TOTAL of all salaries as its own row. The datapoint was removed leaving the dataset in a more realistic state as can be seen from figure 0.3b.



(a)

Figure 0.2: **Plot of Enron employee sataries.** *From this visualization, we can see there is a huge spike to the right of datapoint 60, this spike corresponds with an outlier of value $26,704,229.00.*



(a) (b)

Figure 0.3: **TOTAL insertion.** *Figure (a) Bonus versus salary with outlier present. Figure (b) Same dataset but without the outlier. By visualizing these two features in a scatter plot we were able to clearly detect the existence of an outlier.*

One of the areas we concentrated on was that of compensation. We were very interested in the compensation packages as represented in the financial data of the executives. We moved ahead with a working hypothesis, that if fraud was indeed occurring at Enron, then more than likely, the money was probably going to be funneled out through paid bonuses and stocks. A more rigorous hypothesis would then correlate stock options granted and exercised by the Enron executives with the sales of the shares on the open market. However, such an investigation, is outside the scope of this project.

NAME	SALARY	BONUS
ALLEN PHILLIP K	$201,955.00	4,175,000
BELDEN TIMOTHY N	$213,999.00	5,249,999
SKILLING JEFFREY K	$1,111,258.00	5,600,000
LAY KENNETH L	$1,072,321.00	7,000,000
LAVORATO JOHN J	$339,288.00	8,000,000
NAME	SALARY	EXERCISED STOCK OPTIONS
FREVERT MARK A	$1,060,932.00	10,433,518
PAI LOU L	$261,879.00	15,364,167
SKILLING JEFFREY K	$1,111,258.00	19,250,000
RICE KENNETH D	$420,636.00	19,794,175
LAY KENNETH L	$1,072,321.00	34,348,384
NAME	SALARY	RESTRICTED STOCK
YEAGER F SCOTT	$158,403.00	3,576,206
IZZO LAWRENCE L	$85,274.00	3,654,808
BAXTER JOHN C	$267,102.00	3,942,714
KEAN STEVEN J	$404,338.00	4,131,594
FREVERT MARK A	$1,060,932.00	4,188,667
SKILLING JEFFREY K	$1,111,258.00	6,843,672
PAI LOU L	$261,879.00	8,453,763
WHITE JR THOMAS E	$317,543.00	13,847,074
LAY KENNETH L	$1,072,321.00	14,761,694

Table 0.2: Financial data on some of Enron's top paid employees.

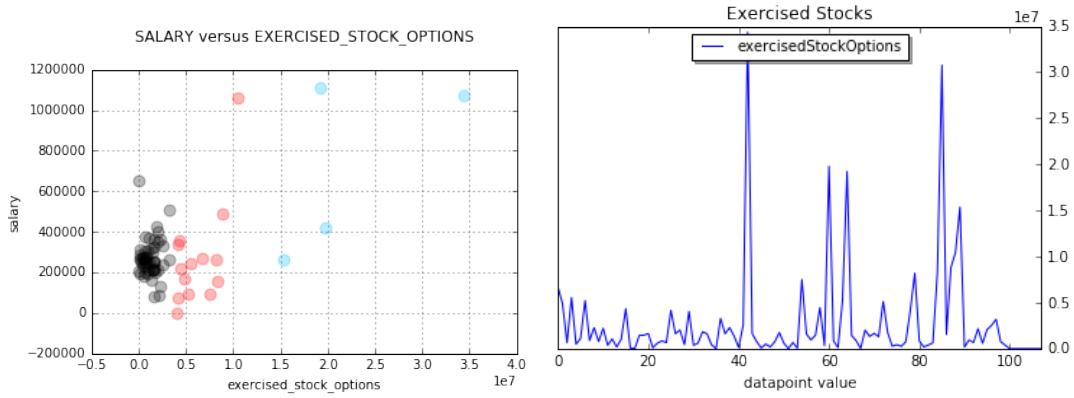
To zoom in on the compensation, we focused on three features 'bonus,', 'exercised_stock_options' and 'restricted_stock'. Table 0.2 shows

some financial data for some of the highest compensated employees. It comes as no surprise that we see that the sitting CEO at the time of the collapsce, Ken Lay, had the second highest bonus paid, that wasn't surprising. What was surprising was the name "John Lavorato." He was the employee that got the highest bonus. So who was John Lavorato? He was the former head of Enron's trading operations.

Another interesting character that showed up in the financial data was "Pai Lou." From table 0.2, one can see that he had some of the largest exercised stock options. Apart from the past CEOs and chairmen, he was *the* employee that sold the most Enron stock. He also got some of the largest shares of restricted stock. Its interesting that these men aren't on the POI list in table 0.1. I firmly believe they are probably on the co-conspirators unindicted list.

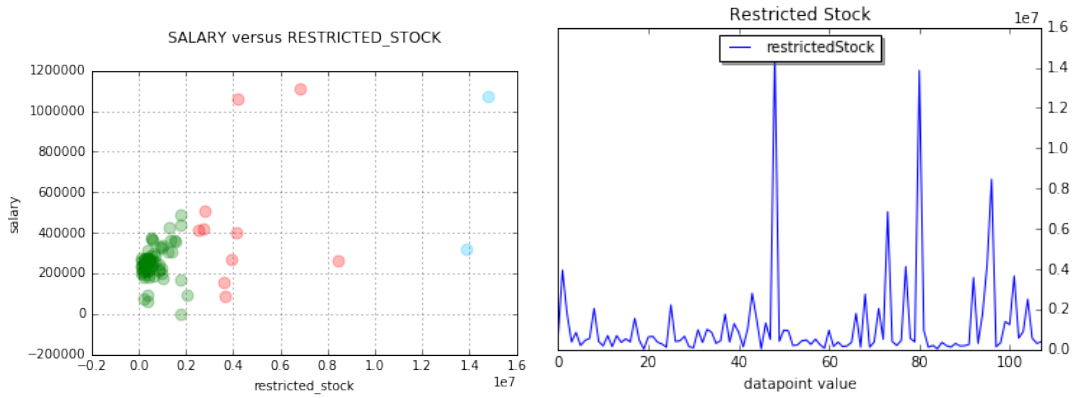
To gain insights into the relationship between salary and exercised stock options, we took that subset of data and ran a KMeans clustering algorithm on it to see how the data clustered; we focused on three clustered as can be seen from figure 0.4a. The clusters generated matched our intuition about the executives, mostly all the five executives with the highest exercised stock options were clustered as one, as can be seen in the visualization in figure 0.4a by the markers colored blue. It is important to note we were not able to infer a relationship between salary and exercised stock options as we had thought.

We proceeded on, with a hypothesis that more than likely, the executives with the highest salaries, probably also had the highest restricted stock. Just as with exercised stock options, it wasn't the case. From figure 0.5a one can see that there is a datapoint with a salary close to $200,000, but with over 8 million shares, when most of the other folks around that salary bracket had less than a million shares. Its important to note that for both our visualizations, the KMeans clustering algorithm, clustered the datapoints along both the exercised stock options and restricted stocks.



(a) (b)

Figure 0.4: **A closer look at salary and exercised stock options.** *Figure (a) A scatter plot of salary versus exercised stock options. Figure (b) A plot of exercised stock options. From these two plots one can see that there is a small subset of senior executives who exercised a significant amount of Enron shares.*



(a) (b)

Figure 0.5: **A closer look at salary and restricted stock.** *A scatter plot of salary versus restricted stock. Figure (b) A plot of restricted stock.*

For the problem of identifying fraud from the Enron financial and email data, we experimented with three different classifiers:

(a) A RandomForestClassifier
We selected this learner because it is considered one of the best off-the-shelf learning algorithm, and requires almost no tuning.

(b) An ExtraTreesClassifier
Its another ensemble learner like Random Forest, with one caveat. When creating a branch in a tree, Random Forest chooses the most discriminant value, whereas ExtrRandTrees split point is arbitrarily. This helps to increase the bias slightly and lower the variance even more.

(c) A LogisticRegression (Robust, with Lp = L1)
Its computational simpler than the ensemble methods. Furthermore, it has been used in real life to predict adverse risk events that have relatively small chances of occurring like credit card fraud and so on. The dataset composition of those events are similar to the Enron dataset, where the labels are heavily skewed towards one class. We selected the robust LogisticRegression based on the fact that it handles outliers better. While the data has been cleaned of "outliers," there are a few executives whose datapoints fall outside of the mean by several standard deviations.

We implemented these three learning algorithms. For each of the learners we implemented the baseline algorithm using 10 fold cross validation to get a the accuracy score. These scores proved to be misleading. We went ahead a used a stratified shuffle split cross validation and calculated the precision, recall and F_1 scores respectively.

From our trials we generated the following scores as can be seen in table 0.3. From these trials, it was evident that the robust Logistic regression beat out both the ensemble learners on all three metrics.

Table 0.3: Result of training with baseline **Learners**

	Metrics		
	Precision	Recall	F1
RandomForestClassifier	0.52197	0.19600	0.28499
ExtraTreesClassifier	0.50438	0.20150	0.28796
LogisticRegression	0.63027	0.20200	0.30594

BENCHMARK

In this section, you will need to provide a clearly defined benchmark result or threshold for comparing across performances obtained by your solution. The reasoning behind the benchmark (in the case where it is not an established result) should be discussed. Questions to ask yourself when writing this section:

- Has some result or value been provided that acts as a benchmark for measuring performance?
- Is it clear how this result or value was obtained (whether by data or by hypothesis)?

METHODOLOGY

DATA PREPROCESSING

In this section, all of your preprocessing steps will need to be clearly documented, if any were necessary. From the previous section, any of the abnormalities or characteristics that you identified about the dataset will be addressed and corrected here.

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"]

Questions to ask yourself when writing this section:

- If the algorithms chosen require preprocessing steps like feature selection or feature transformations, have they been properly documented?
- Based on the Data Exploration section, if there were abnormalities or characteristics that needed to be addressed, have they been properly corrected?
- If no preprocessing is needed, has it been made clear why?

IMPLEMENTATION

In this section, the process for which metrics, algorithms, and techniques that you implemented for the given data will need to be clearly documented. It should be abundantly clear how the implementation was carried out, and discussion should be made regarding any complications that occurred during this process. Questions to ask yourself when writing this section:

- Is it made clear how the algorithms and techniques were implemented with the given datasets or input data?

- Were there any complications with the original metrics or techniques that required changing prior to acquiring a solution?
- Was there any part of the coding process (e.g., writing complicated functions) that should be documented?

REFINEMENT

In this section, you will need to discuss the process of improvement you made upon the algorithms and techniques you used in your implementation. For example, adjusting parameters for certain models to acquire improved solutions would fall under the refinement category. Your initial and final solutions should be reported, as well as any significant intermediate results as necessary.

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"]

Questions to ask yourself when writing this section:

- Has an initial solution been found and clearly reported?
- Is the process of improvement clearly documented, such as what techniques were used?
- Are intermediate and final solutions clearly reported as the process is improved?

RESULTS

MODEL EVALUATION AND VALIDATION

In this section, the final model and any supporting qualities should be evaluated in detail. It should be clear how the final model was derived and why this model was chosen. In addition, some type of analysis should be used to validate the robustness of this model and its solution, such as manipulating the input data or environment to see how the model,Äôs solution is affected (this is called sensitivity analysis).

Questions to ask yourself when writing this section:

- Is the final model reasonable and aligning with solution expectations? Are the final parameters of the model appropriate?
- Has the final model been tested with various inputs to evaluate whether the model generalizes well to unseen data?
- Is the model robust enough for the problem? Do small perturbations (changes) in training data or the input space greatly affect the results?
- Can results found from the model be trusted?

JUSTIFICATION

In this section, your model,Äôs final solution and its results should be compared to the benchmark you established earlier in the project using some type of statistical analysis. You should also justify whether these results and the solution are significant enough to have solved the problem posed in the project. Questions to ask yourself when writing this section:

- Are the final results found stronger than the benchmark result reported earlier?
- Have you thoroughly analyzed and discussed the final solution?
- Is the final solution significant enough to have solved the problem?

CONCLUSION

FREE-FORM VISUALIZATION

In this section, you will need to provide some form of visualization that emphasizes an important quality about the project. It is much more free-form, but should reasonably support a significant result or characteristic about the problem that you want to discuss. Questions to ask yourself when writing this section:

- Have you visualized a relevant or important quality about the problem, dataset, input data, or results?
- Is the visualization thoroughly analyzed and discussed?
- If a plot is provided, are the axes, title, and datum clearly defined?

REFLECTION

In this section, you will summarize the entire end-to-end problem solution and discuss one or two particular aspects of the project you found interesting or difficult. You are expected to reflect on the project as a whole to show that you have a firm understanding of the entire process employed in your work. Questions to ask yourself when writing this section:

- Have you thoroughly summarized the entire process you used for this project?
- Were there any interesting aspects of the project?
- Were there any difficult aspects of the project?
- Does the final model and solution fit your expectations for the problem, and should it be used in a general setting to solve these types of problems?

IMPROVEMENT

In this section, you will need to provide discussion as to how one aspect of the implementation you designed could be improved. As an example, consider ways your implementation can be made more general, and what would need to be modified. You do not need to make this improvement, but the potential solutions resulting from these changes are considered and compared/contrasted to your current solution. Questions to ask yourself when writing this section:

- Are there further improvements that could be made on the algorithms or techniques you used in this project?
- Were there algorithms or techniques you researched that you did not know how to implement, but would consider using if you knew how?
- If you used your final solution as the new benchmark, do you think an even better solution exists?