Project 5: Identify Fraud from Enron Emails

# 1. Introduction and data exploration

The goal of this project: given a dataset describing certain aspects of the employees (e.g. payment, stock, and email interaction data), by applying feature engineering techniques, we can build a model to predict who the POIs are among all the employees. Machine learning methods are important in this project due to its ability to learn the necessary features to distinguish POIs from non-POIs in an automated way.

## 1.1 Dataset description

The dataset used in this project contains 146 employees’ data and 21 columns. From the result of previous exercises we know that the dataset contains the following:

* + Label (1)
    - poi
  + Payment features (9+1)
    - salary, bonus, long\_term\_incentive, deferred\_income, deferral\_payments, loan\_advances, other, expenses, director\_fees, total\_payments
  + Stock features (3+1)
    - exercised\_stock\_options, restricted\_stock, restricted\_stock\_deferred, total\_stock\_value
  + Email/communication features (6)
    - to\_messages, from\_messages, from\_this\_person\_to\_poi, from\_poi\_to\_this\_person, shared\_receipt\_with\_poi, email\_address

20 out of 21 columns contain missing values with only poi column not containing any ‘nan’ values. 0 instead of ‘nan’ values are filled in to deal with missing values. Based on the insider pay document, this is reasonable since in the document ‘-‘ represents that the person does not have any value attached with the specified attribute, e.g. salary / bonus / loan advances etc.. In the script, this is done by specifying remove\_NaN to true when using function ‘featureFormat’.

Besides, one employee’s data are all 0. It does not make sense that this person does not have neither payment nor stock, we can infer this person’s information is missing. Thus his data are excluded from the dataset. Afterwards 145 employees’ data remain.

## 1.2 Dataset exploration

Below tables show the statistics of all features listed in the dataset. By applying describe function to the dataset, we got the mean, standard deviation, quartiles, min and max. It is easily seen that all features related to employee payments seem unusual when looking at the max value. For example the total payments – the 75% value is 1.98E+06, whereas the max value is 3.10E08, which is 157 times the 75% value. Also, this 75% value is smaller than the mean value. This indicates the mean value is actually influenced by one or few extremely large values. Therefore, the max of total payments can be considered as an outlier.

Similarly, we can find features such as salary, deferral payments, exercised stock options etc. also contain outliers. Here is a reference from Forbes[[1]](#footnote-1) regarding the Enron payment: “*Between 1996 and 2000, the average chief executive salary and bonus increased by 24% to $1.72 million, according to a Forbes study of proxy reports. Total CEO Compensation, including stock options and restricted stock grants, grew 166% to an average of $7.43 million. In the same period, corporate profits grew by 16%, and per capita income grew by 18%. Enron was at the cutting edge of this trend. The stated goal of its board of directors was to pay executives in the 75th percentile of its peer group. In fact, it paid them vastly more and on a scale completely out of whack with the company’s financial results–even if its reported financial results are accepted as accurate.*” In the article, the author analyzed Enron’s payment problem. The quote above mentioned that Enron paid executives much more than 75th percentile of the peer group. And the underlined features are the ones that can be confirmed by our dataset.

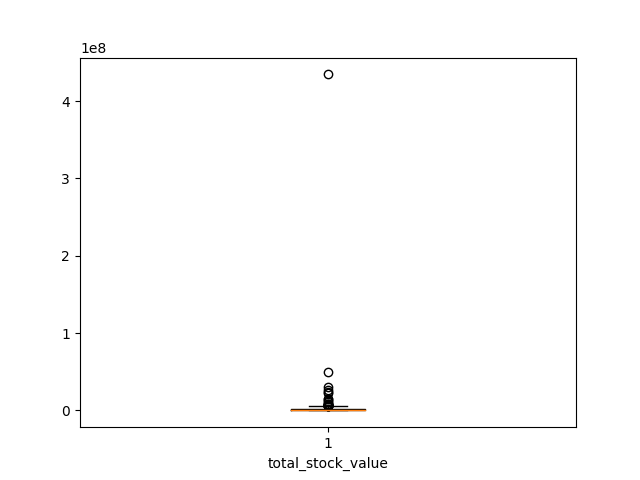
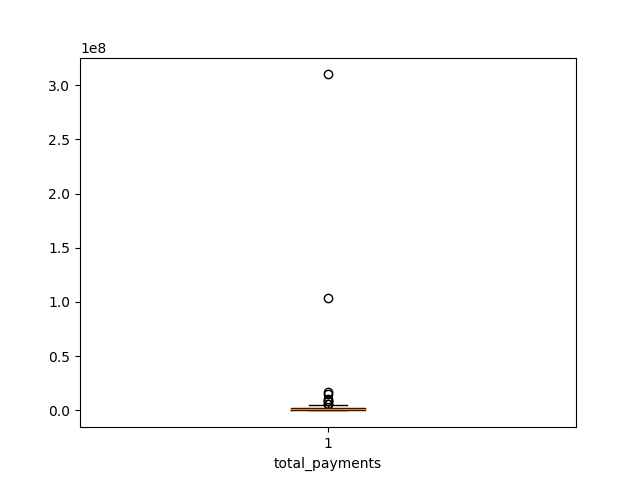
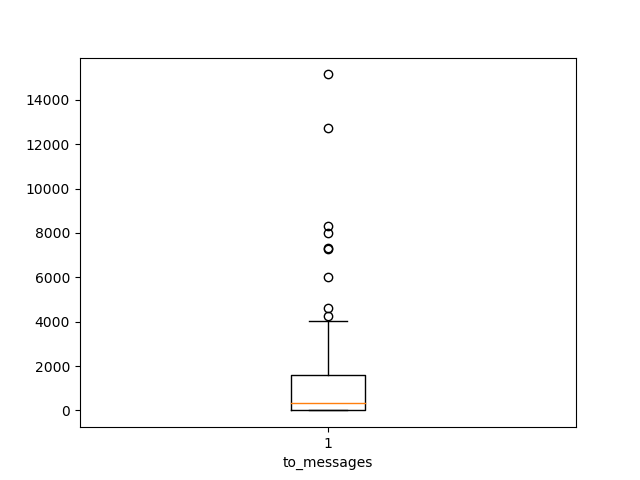
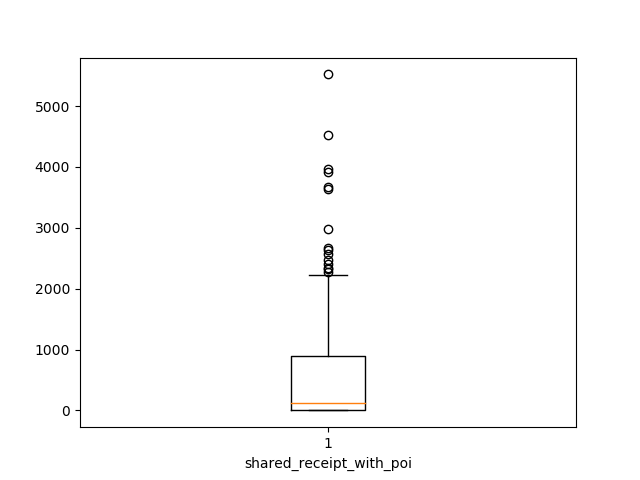
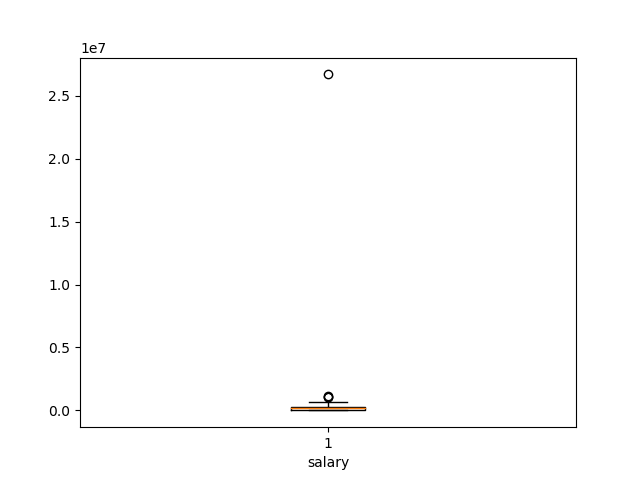
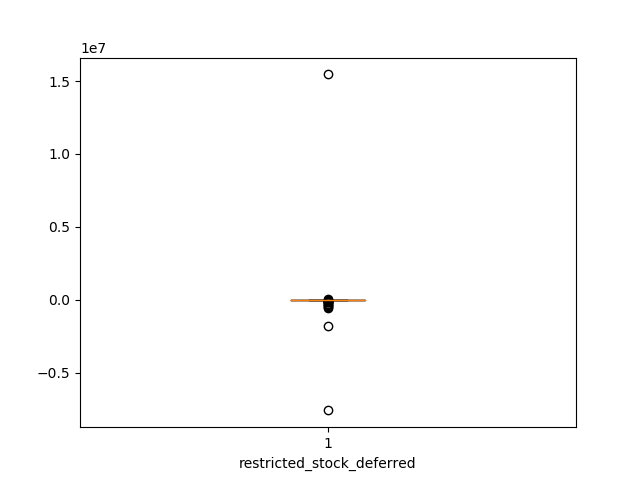
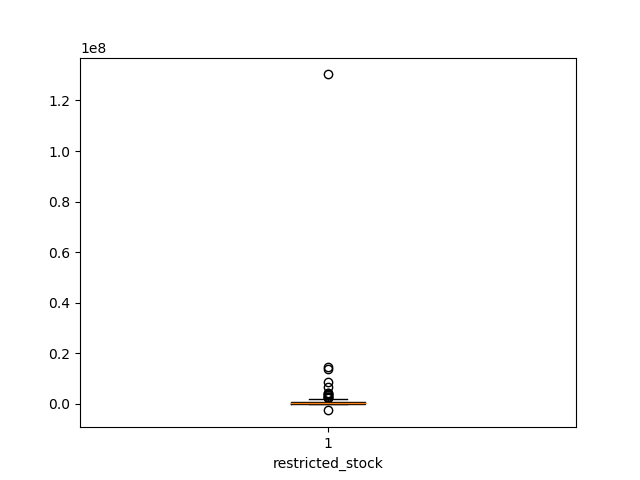
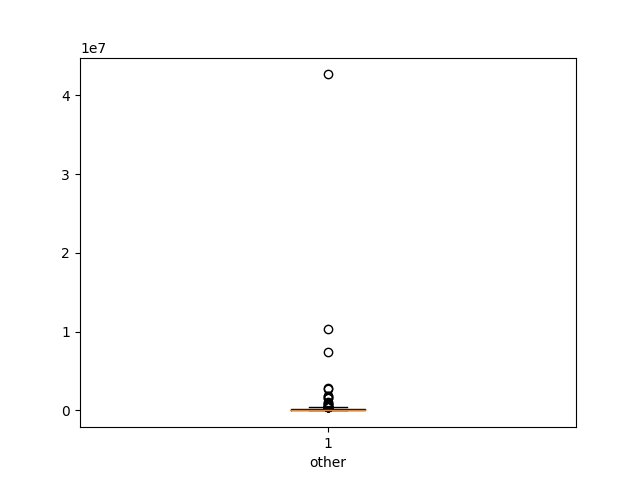
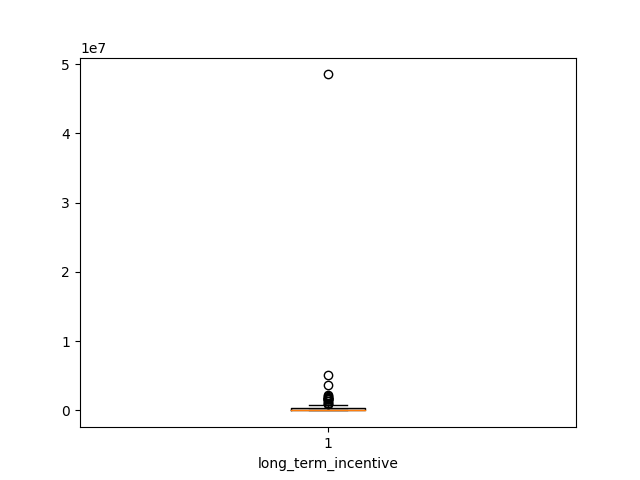
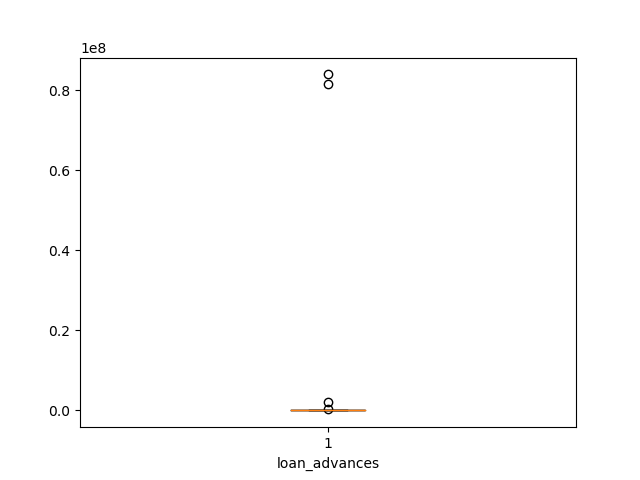
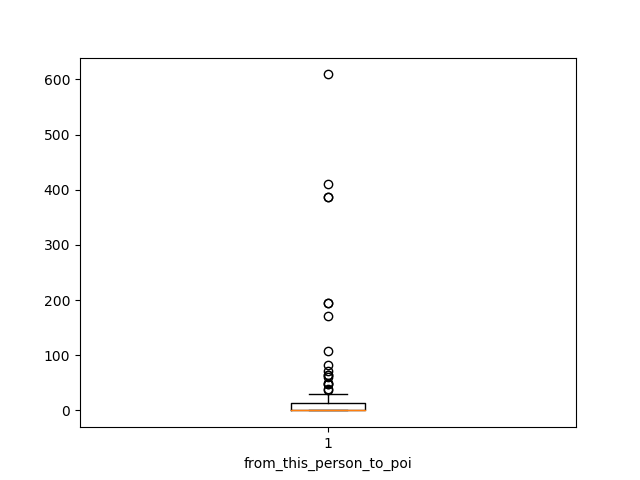
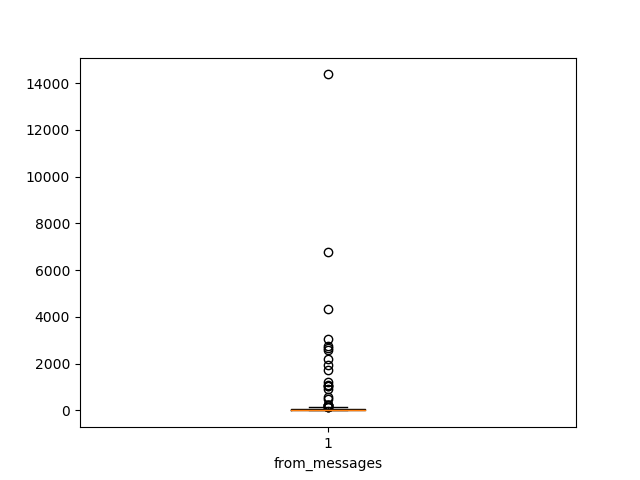
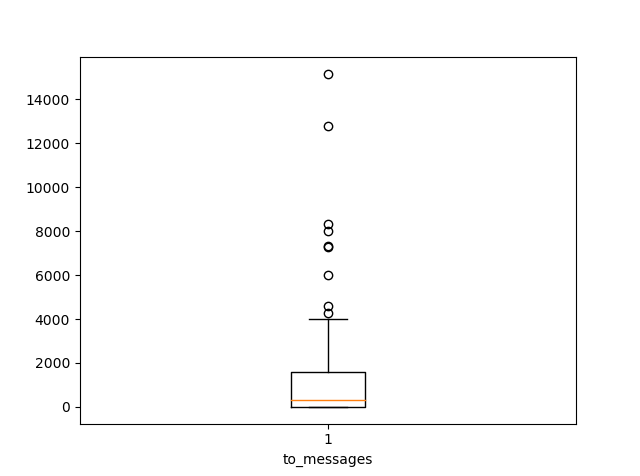
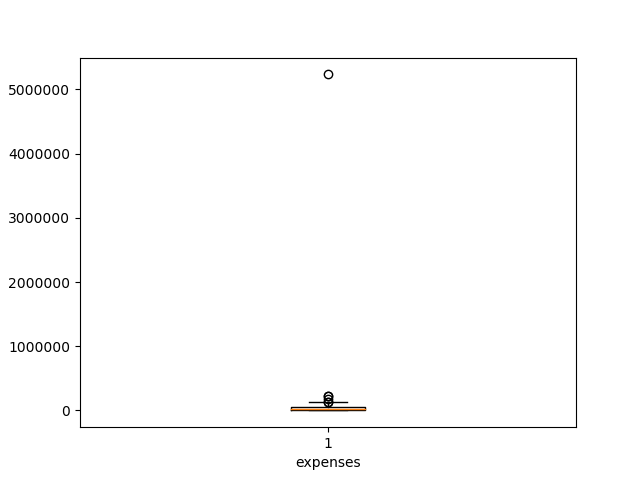
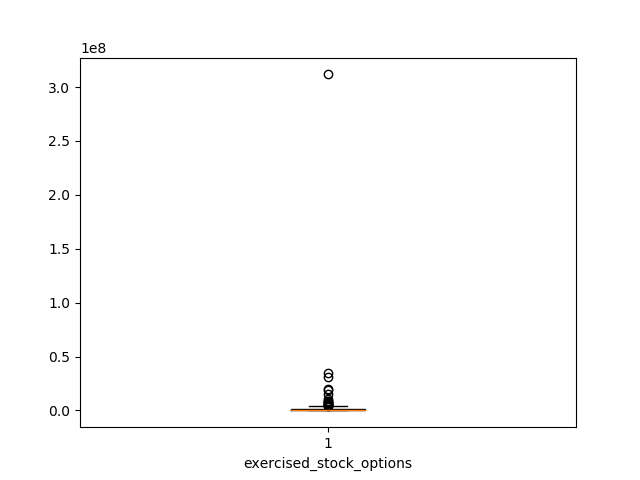
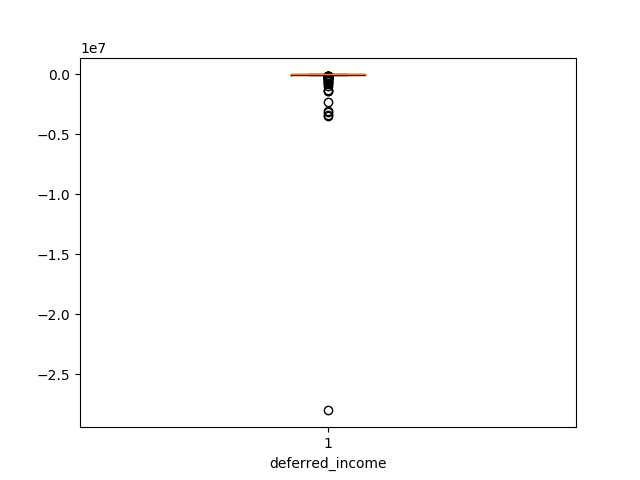
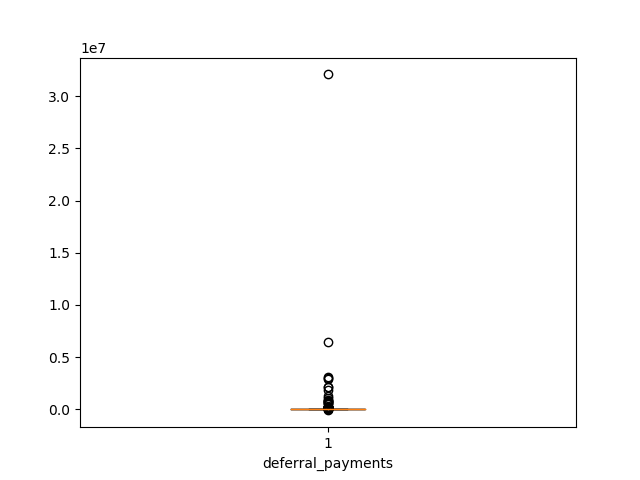
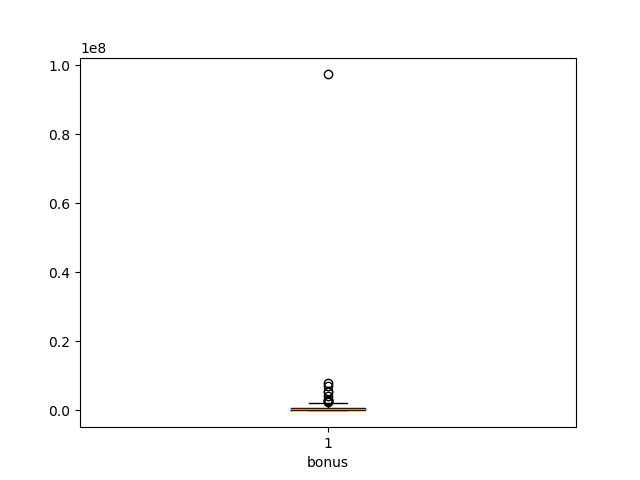
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **salary** | **to\_messages** | **deferral\_payments** | **total\_payments** | **exercised\_stock\_options** |
| **count** | 1.45E+02 | 1.45E+02 | 1.45E+02 | 1.45E+02 | 1.45E+02 |
| **mean** | 3.68E+05 | 1.23E+03 | 4.42E+05 | 4.38E+06 | 4.21E+06 |
| **std** | 2.21E+06 | 2.23E+03 | 2.75E+06 | 2.70E+07 | 2.62E+07 |
| **min** | 0.00E+00 | 0.00E+00 | -1.03E+05 | 0.00E+00 | 0.00E+00 |
| **25%** | 0.00E+00 | 0.00E+00 | 0.00E+00 | 1.03E+05 | 0.00E+00 |
| **50%** | 2.11E+05 | 3.12E+02 | 0.00E+00 | 9.67E+05 | 6.09E+05 |
| **75%** | 2.71E+05 | 1.61E+03 | 1.03E+04 | 1.98E+06 | 1.73E+06 |
| **max** | **2.67E+07** | 1.51E+04 | **3.21E+07** | **3.10E+08** | **3.12E+08** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **bonus** | **restricted\_stock** | **shared\_receipt\_**  **with\_poi** | **restricted\_stock\_**  **deferred** | **total\_stock\_value** |
| **count** | 1.45E+02 | 1.45E+02 | 145 | 1.45E+02 | 1.45E+02 |
| **mean** | 1.34E+06 | 1.76E+06 | 697.765517 | 2.07E+04 | 5.89E+06 |
| **std** | 8.12E+06 | 1.09E+07 | 1075.128126 | 1.44E+06 | 3.64E+07 |
| **min** | 0.00E+00 | -2.60E+06 | 0 | **-7.58E+06** | -4.41E+04 |
| **25%** | 0.00E+00 | 3.25E+04 | 0 | 0.00E+00 | 2.52E+05 |
| **50%** | 3.00E+05 | 3.61E+05 | 114 | 0.00E+00 | 9.76E+05 |
| **75%** | 8.00E+05 | 8.53E+05 | 900 | 0.00E+00 | 2.33E+06 |
| **max** | **9.73E+07** | **1.30E+08** | **5521** | **1.55E+07** | **4.35E+08** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **expenses** | **loan\_advances** | **from\_messages** | **other** | **from\_this\_person\_to\_poi** |
| **count** | 1.45E+02 | 1.45E+02 | 145 | 1.45E+02 | 145 |
| **mean** | 7.12E+04 | 1.16E+06 | 361.075862 | 5.89E+05 | 24.455172 |
| **std** | 4.34E+05 | 9.68E+06 | 1445.944684 | 3.69E+06 | 79.527073 |
| **min** | 0.00E+00 | 0.00E+00 | 0 | 0.00E+00 | 0 |
| **25%** | 0.00E+00 | 0.00E+00 | 0 | 0.00E+00 | 0 |
| **50%** | 2.15E+04 | 0.00E+00 | 17 | 9.72E+02 | 0 |
| **75%** | 5.39E+04 | 0.00E+00 | 52 | 1.51E+05 | 14 |
| **max** | **5.24E+06** | **8.39E+07** | **14368** | **4.27E+07** | **609** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **director\_fees** | **deferred\_income** | **long\_term\_incentive** | **from\_poi\_to\_this\_person** |
| **count** | 1.45E+02 | 1.45E+02 | 1.45E+02 | 145 |
| **mean** | 1.96E+04 | -3.85E+05 | 6.69E+05 | 38.489655 |
| **std** | 1.19E+05 | 2.39E+06 | 4.06E+06 | 74.088359 |
| **min** | 0.00E+00 | -2.80E+07 | 0.00E+00 | 0 |
| **25%** | 0.00E+00 | -3.83E+04 | 0.00E+00 | 0 |
| **50%** | 0.00E+00 | 0.00E+00 | 0.00E+00 | 4 |
| **75%** | 0.00E+00 | 0.00E+00 | 3.75E+05 | 41 |
| **max** | 1.40E+06 | 0.00E+00 | 4.85E+07 | 528 |

To get a better idea, below visualizes the statistics using boxplot. The boxplot considers values that is 1.5 times the mean value or more to be outliers. From the plots we can see based on this criteria, all of the features contain outliers. And in each of the plot, there is one or two outliers that are particularly much larger than other values.



Although in the above analysis many outliners are found, we still need to be cautious about how to process them. Whether to remove or not depends on the scenario. In Enron’s case, POIs tried to maximize their own benefits by transferring company assets to their own. Thus it makes sense to observe many outliers in salary, stock options, etc. Rather, having these outliers might be an indicator of the likelihood of being a POI. Hence the decision is not to remove these outliers.

However, keeping the outliers will result in another problem. The mean and interquartile values are high influenced by these outliers. For example in bonus, while 75% of the people have bonus no more than 80, 000, the maximum value is 973, 000, which is ~121 times the 75% value. As a result, the average value is also much larger than the 75% value.

To solve this problem, squared root is used for data transformation. Another option is to use logarithm but due to min values of 0 for many features, squared root is preferred over log.

# 2. Feature selection

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?

## 2.1 Adding new features

Apart from the existing features, following features are also considered.

1. from\_this\_person\_to\_poi\_fraction: calculates the portion of POI related emails among all the emails a person sent. My hypothesis is that if the percentage of POI communication overall the total communication is high, it means this person is closely connected with POIs. Then this person is also likely a POI.

2. from\_poi\_to\_this\_person\_fraction: calculates the portion of POI related emails among all the emails a person received. The reason is same as above.

3. bonus\_over\_payment\_ratio: calculates the portion of bonus over total payments. If a person is a POI, then it is likely that he has bonus as major income rather than salary or others.

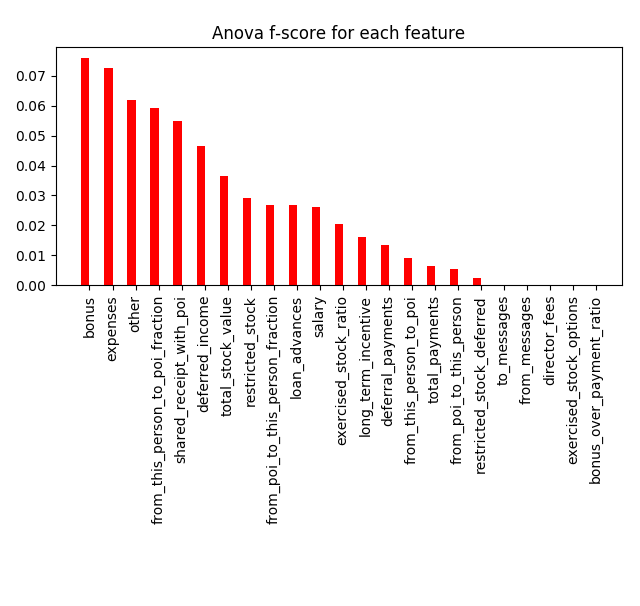
4. exercised\_stock\_ratio: calculates the portion of exercised stock over the total stock value. The hypothesis is a POI is likely to exercise the stock rather than leaving it still in the company’s hold.

## 2.2 Feature transformation

## 2.3 Selecting features

In your feature selection step, if you used an algorithm like a decision tree, please also give the feature importance 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”]

First I used univariate feature selection method. The idea of univariate feature selection is to analyze the relation between independent variable (in our case it is one of the features) and dependent variable (i.e. the labels). As a starting point, I chose the “select K best” method. In this method, there are a few score functions: chi-2, f\_classif and mutual\_info\_classif. Considering all the features are continuous variables, and we are having a classification problem, mutual\_info\_classif is used with number of selected features K=15. Chi-2 is not used because in order for it to apply, the features should be categorical. Below shows the feature score value in descending order. From this plot we can see the engineered feature ‘from\_this\_person\_to\_poi\_fraction’, ‘from\_poi\_to\_this\_person\_fraction’, ‘exercised\_stock\_ratio’ are selected features. Whereas the original features such as ‘to\_messages’, ‘from\_messages’ and ‘exercised\_stock\_otions’ are not selected as features. Overall, coming up with new features is indeed very helpful for us to understand the dataset.



# 3. Model building

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”]

## 3.1 Support Vector Machine

## 3.2 Naïve Bayes Classifier

# 4. Parameter tuning

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? What parameters did you tune? (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”]

# 5. Validation

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”]

# 6. Model evaluation

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 algorithms performance. [relevant rubric item: “usage of evaluation metrics”]

# Appendix

## Table 1. Feature scores

Feature scores using SelectKBest with scoring function ‘mutual\_info\_classif’.

|  |  |
| --- | --- |
| **feature** | **score** |
| bonus | 0.075807 |
| expenses | 0.072731 |
| other | 0.061883 |
| from\_this\_person\_to\_poi\_fraction | 0.059349 |
| shared\_receipt\_with\_poi | 0.054996 |
| deferred\_income | 0.046653 |
| total\_stock\_value | 0.036662 |
| restricted\_stock | 0.02905 |
| from\_poi\_to\_this\_person\_fraction | 0.02697 |
| loan\_advances | 0.026918 |
| salary | 0.026134 |
| exercised\_stock\_ratio | 0.020507 |
| long\_term\_incentive | 0.016159 |
| deferral\_payments | 0.013339 |
| from\_this\_person\_to\_poi | 0.008971 |
| total\_payments | 0.006572 |
| from\_poi\_to\_this\_person | 0.005326 |
| restricted\_stock\_deferred | 0.002403 |
| to\_messages | 0 |
| from\_messages | 0 |
| director\_fees | 0 |
| exercised\_stock\_options | 0 |
| bonus\_over\_payment\_ratio | 0 |

1. Pay Madness At Enron - <https://www.forbes.com/2002/03/22/0322enronpay.html> [↑](#footnote-ref-1)