Udacity

Intro to Machine Learning

Project

IDENTIFY FRAUD FROM ENRON EMAIL

---

by Omoju Miller

July 7, 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. As such,

In this section, you will need to clearly define the metrics or calculations you will use to measure performance of a model or result in your project. These calculations and metrics should be justified based on the characteristics of the problem and problem domain. Questions to ask yourself when writing this section:

- Are the metrics you've chosen to measure the performance of your models clearly discussed and defined?
- Have you provided reasonable justification for the metrics chosen based on the problem and solution?

accuracy, precision, recall, f1, f2 true positives, false positives, false negatives, true negatives)

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. The features in the data fall into three major types, namely financial features, email features and POI labels.

- 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`

```
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}
```

Figure 0.1: **Data point for Enron Employee Pai Lou.**

while it contains a robust amount of emails, contains data for only about 150 people. Furthermore, the dataset is extremely skewed.

In this section, you will be expected to analyze the data you are using for the problem. This data can either be in the form of a dataset (or datasets), input data (or input files), or even an environment. The type of data should be thoroughly described and, if possible, have basic statistics and information presented (such as discussion of input features or defining characteristics about the input or environment). Any abnormalities or interesting qualities about the data that may need to be addressed have been identified (such as features that need to be transformed or the possibility of outliers).

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"].

Questions to ask yourself when writing this section:

- If a dataset is present for this problem, have you thoroughly discussed certain features about the dataset? Has a data sample been provided to the reader?
- If a dataset is present for this problem, are statistics about the dataset calculated and reported? Have any relevant results from this calculation been discussed?
- If a dataset is not present for this problem, has discussion been made about the input space or input data for your problem?
- Are there any abnormalities or characteristics about the input space or dataset that need to be addressed? (categorical variables, missing values, outliers, etc.)

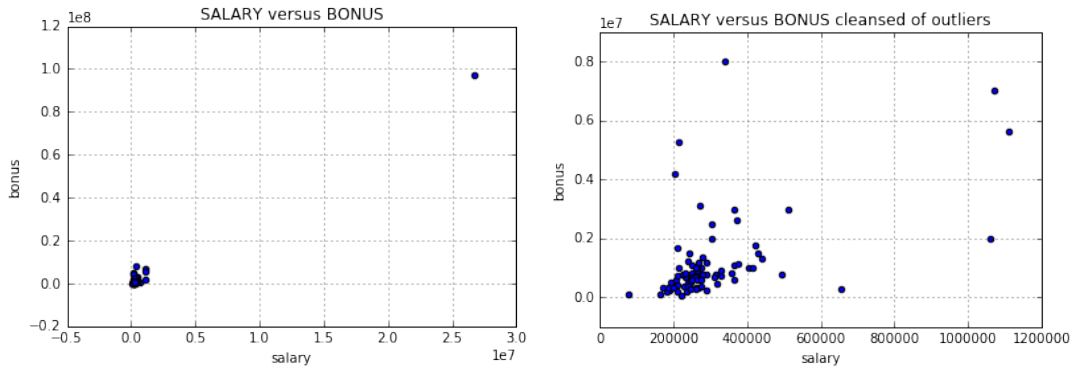
Outlier detection

So whose bonus is 97343619?
`TOTAL salary is $26,704,229.00 bonus $97,343,619.00`
Found the source of the outlier. It was the TOTAL row that was mistakenly read into the data dict

EXPLORATORY VISUALIZATION

In this section, you will need to provide some form of visualization that summarizes or extracts a relevant characteristic or feature about the data. The visualization should adequately support the data being used. Discuss why this visualization was chosen and how it is relevant. Questions to ask yourself when writing this section:

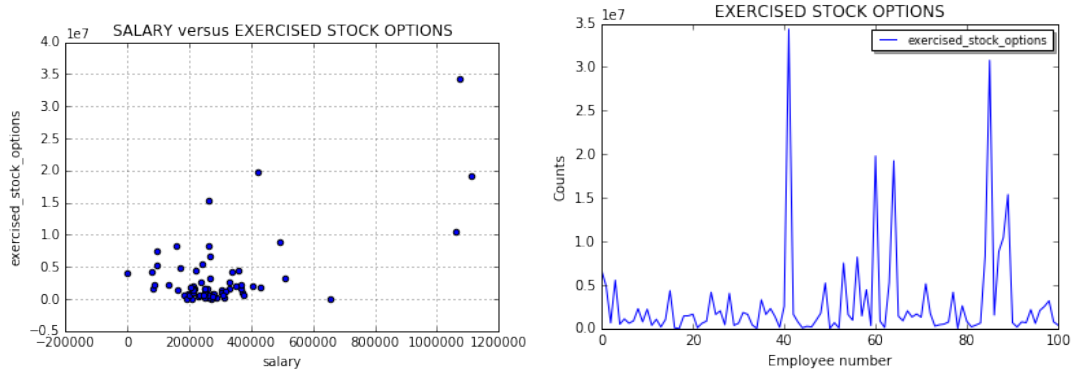
- Have you visualized a relevant characteristic or feature about the dataset or input data?



(a) (b)

Figure 0.2: **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.*

- Is the visualization thoroughly analyzed and discussed?
- If a plot is provided, are the axes, title, and datum clearly defined?

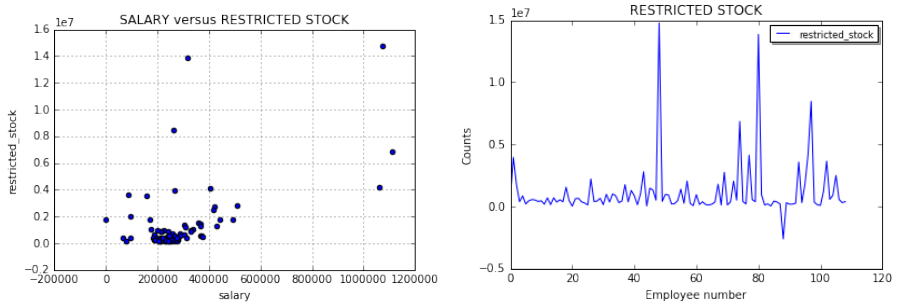


(a) (b)

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.*

ALGORITHMS AND TECHNIQUES

In this section, you will need to discuss the algorithms and techniques you intend to use for solving the problem. You should justify the use of each one based on the characteristics of the problem and the problem domain.



(a) (b)

Figure 0.4: **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.*

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"]

Questions to ask yourself when writing this section:

- Are the algorithms you will use, including any default variables/parameters in the project clearly defined?
- Are the techniques to be used thoroughly discussed and justified?
- Is it made clear how the input data or datasets will be handled by the algorithms and techniques chosen?

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?