Project Overview

In 2000, Enron was one of the largest companies in the United States. By 2002, it had collapsed into bankruptcy due to widespread corporate fraud. In the resulting Federal investigation, there was a significant amount of typically confidential information entered into public record, including tens of thousands of emails and detailed financial data for top executives. The goal of this project is to use machine learning to analyze this data to predict who the people of interest are why they are suspicious.

The purpose of the author distributing this dataset is to serve “as a resource for researchers who are interested in improving current email tools, or understanding how email is currently used. This data is valuable; to my knowledge it is the only substantial collection of "real" email that is public. The reason other datasets are not public is because of privacy concerns. In using this dataset, please be sensitive to the privacy of the people involved (and remember that many of these people were certainly not involved in any of the actions which precipitated the investigation.)” It is available for anyone to download here: https://www.cs.cmu.edu/~./enron/

Purpose

The goal of this project is to use machine learning to play detective and identify Enron Employees who may have committed fraud based on the dataset.

Dataset Details-Data Exploration

There are 146 individuals in the dataset each containing 21 features. Here are also some more details from datasets_questions\explore_enron_data.py:
POI True: 18
PRENTICE JAMES total stock value: 1,095,040
Email messages from Wesley Colwell to persons of interest: 11
Jeffrey Skilling (ceo at the time) total stock value: 19,250,000
Enron chairman of the board of directors: Kenneth Lay
CFO Andrew Fastow
SKILLING CEO: 8,682,716
LAY BOD: 103,559,793
FASTOW CFO: 2,424,083
Quantified Salary not blank 95
Email Addresses not blank 111

Outliers Investigation:

Through examining the dataset using pprint I found three outliers which stood out in the dataset:
THE TRAVEL AGENCY IN THE PARK-not a person
TOTAL-not a person
LOCKHARD EUGENE E-showed all NaN features

Features

Here are the full list of features available to choose from and highlighted are the ones I found most interesting to use in my analysis:

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'] (all units are in US dollars)

email features: ['to_messages', 'email_address', 'from_poi_to_this_person', 'from_messages', 'from_this_person_to_poi', 'poi', 'shared_receipt_with_poi'] (units are generally number of emails messages; notable exception is 'email_address', which is a text string)

POI label: ['poi'] (boolean, represented as integer)

For the main features I selected and thier purpose are list below:
poi-main feature which was supplied
salary-my intuition told me higher salaries means more fraud
total payments-this is total compensation which should be related to the biggest poi
total stock value-assumed poi would be well vested in the company
frompoitothisperson-if a person directly recieves a lot of emails from a poi they must be in the know
fromthispersontopoint-if they are communicating back and forth this would show it
sharedreceiptwithpoi.-is on many distribution lists that the poi would be included on

I did not use scaling as it was not needed.

I attempted to create my own three features:
xsalary_to_stockvalue-this devices salary/total stock value to make a ratio of how much someone is getting paid vs how much stock they get. Would suggest higher salary/stock= higher poi rate
xfrom_to_poi-from poi to this person/from this person to poi should give the ratio of how much conversation this person has with a poi, high the number, the more likely the should be involved in suspicious activities
xto_receipt_poi- from this person to poi/shared receipt with poi, this person might not be as involved but still a good ratio to run tests on
xpayout-just for fun to see which people are getting the most payments+stock values

Algorithm and Tuning Parameters

For my algorithm I experimented with Naive Bayes and Decision Trees.

Parameter tuning is essential to configure the algorithms in such a way to not misinterpret the data.

Validation and Analysis

Validation is verifying the correct accuracy, precision, recall and F1/F2. If these are measured wrong it could lead to an invalid analysis.

Conclusion

Further updates need to be done to improve on the accuracy scores and the analysis of the data. Digging into each individual's message would be interesting to looking into. Word frequency to and from POI might be an interesting research topic.

References

https://docs.python.org/2/tutorial/datastructures.html