Enron Person Of Interest Identifier

Introduction

This project consists of finding an algorithm to flag individuals as being a potential Person Of Interest (POI) in the Enron fraud of the late 1990s. The project is based around a dataset drawn from the Enron financial scandal investigation, including a corpus of emails, pay and incentive information, and a flag for each individual indicating if they were a POI in the Enron investigation.

The goal of this project is to demonstrate use of machine learning by deriving a POI prediction algorithm that has accuracy and precision both greater than 0.3.

Data Exploration and outlier investigation

Some basic summary information of the dataset is shown in Table 1. It is immediately clear from this summary data that we will need to deal with the large number of missing data points and deal with over fitting due to the small number of POI individuals.

Table 1: Raw Data properties

Property	Value
Total points	3066
Missing Values	1358
Features	21
POI individuals	18
Non-POI individuals	128

The top three features with missing values were "loan_advances"(142), "director_fees"(129) and "restricted_stock_deferred"(128).

Outliers for each feature were examined via histograms. A label titled "TOTAL" was removed due to being an extreme outlier. "'THE TRAVEL AGENCY IN THE PARK'" was also removed because it did not appear to be a person.

Feature Engineering and Selection

One problem raised during the initial data exploration is the large number of missing values for some features. These missing values could be handled by imputing values of 0, but imputing is undesirable because it will throw some important information away. One way to retain the information is to engineer new features for each column that is true/false based on the presence of an imputed value. These features were added to the dataset before feature selection.

Univariate feature selection was performed using the F-statistic. It was found that using more than three features resulted in over-fitting, so the top three features were retained: 'exercised_stock_options', 'total_stock_value', and 'bonus'.

Multivariate feature selection was performed using a decision tree. In multivariate feature selection, features are recursively removed to determine relative importance. It was again found that more than three features resulted in overfitting, so just the top three were retained: 'exercised_stock_options', 'restricted_stock', and 'total_payments'.

Finally, feature importances from a fit with a random forest classifier were also used.

Technique	Features
RF importance	'expenses','exercised_stock_options','bonus'
Univariate	'exercised_stock_options','total_stock_value','bonus'
Multivariate	'exercised_stock_options','restricted_stock','total_payments'

Algorithm Selection And Tuning

Two of the classifiers used in the UD120 course lessons were used for comparison: naïve Bayes and decision tree. For the decision tree, the minimum samples to split were optimized in the grid [2,5,10,20,40,80] based on accuracy.

All optimized combinatons were evaluated using the methodology described in the "Algorithm Validation" section to produce the table below:

Feature Set	Classifier	Precision	Recall
RF importance	Naïve Bayes	0.482	0.289
RF importance	Decision Tree	0.392	0.385
Univariate	Naïve Bayes	0.486	0.351
Univariate	Decision Tree	0.356	0.338
Multivariate	Naïve Bayes	0.387	0.179
Multivariate	Decision Tree	0.752	0.177

In a fraud dectection application, precision(flagged POIs are actually POIs) is probably more important than recall, although recall should also be as high as possible. Based on that standard, the champion algorithm is the Univariate feature set with a Naïve Bayes classifier.

Algorithm Validation

The standards for a good algorithm are high precision(flagged POIs are actually POIs) and high recall (large percentage of POIs are flagged). These metrics were evaluated using the course-specified methodology of averaging over 1000 splits of 50% training data and 50% testing data.

The final result is summarized below:

Classifier	Naïve Bayes
Features	'expenses','exercised_stock_options','bonus'
Average Precision	0.486
Average Recall	0.351

Conclusion

Application of machine learning to a sample problem has been demonstrated, including data exploration, feature engineering, algorithm selection and validation. At the end, we have a person of interest identifier that has about a 40% chance of flagging a POI and about a 50% change of every flagged person actually being a POI. This result exceeds the original requirement of 0.3 accuracy and precision.