ENRON PERSON OF INTEREST IDENTIFIER

By G Adityan

Introduction

Enron, an energy, commodities and services company was one of the largest companies in the United States in 2000. By 2002, it had collapsed into bankruptcy due to widespread corporate fraud. The financial and email data of the employees was made public during investigation and this data has been used to find person of interest (POI) who may have been a part of the fraudulent behaviour.

Machine learning is a powerful tool that can be used to draw predictions based on the dataset available. Here, financial data of the employees were used as features and different machine learning algorithms were tested on them to find a correlation between the financial data and the POI.

There were 146 persons in the financial dataset out of which entries named 'TOTAL', 'THE TRAVEL AGENCY IN THE PARK' and 'SAVAGE FRANK' were popped from the data dictionary after considering them as outliers. There were 21 features per person and 18 of them were labelled as POI. Some financial features had 'NaN' values and these values were replaced by 0 for calculations. The task was to train a classifier that had precision and recall both > 0.3.

Feature Selection

6 features used for the analysis were salary, exercised stock options, total payments, ratio of deferred income to total payments, percentage of emails sent from a person to POI and the percentage of emails sent from POI to a person. Deferred income feature on its own had a very low feature importance and hence was removed. Unusually high values of salary and exercised stock options compared to other entries in the dataset may act as an indicator that the person under analysis is a POI. Feature scaling was not employed and apart from salary and exercised stock options, the other 3 features were newly created. High ratio for the email related features mentioned above mean that there is a lot of communication between the person and POI. This could act as an indicator that the person is also a POI.

Features	salary	total payments	exercised stock options	% from this person to poi	% from poi to this person	Ratio of deferred income to total payments
Feature Importance	0.26	0.16	0.11	0.27	0.12	0.08

Algorithms used

The various machine learning algorithms used to classify a person as POI or non-POI were AdaBoost Classifier, Decision Tree Classifier, Random Forest Classifier and KNeighbors Classifier. The performances of all the algorithms were compared based on different evaluation metrics and AdaBoost Classifier gave the best results for classifying the data.

Algorithm	Accuracy	Precision	Recall	F1 Score	F2 Score
AdaBoost Classifier	0.86300	0.52859	0.37900	0.44147	0.40174
RandomForest Classifier	0.88113	0.64333	0.24350	0.35328	0.27086
Decision Tree Classifier	0.83500	0.41539	0.38050	0.39718	0.38700
KNeighbors Classifier	0.86180	0.42940	0.11100	0.17640	0.13033

Tuning the Classifier

Each algorithm mentioned above uses a specific set of parameters to classify the given data and these parameters can take up a range of values. The set of values that give the best performance for an algorithm is desired and the task of finding this set for implementation is what is called tuning the classifier. If the classifier is not tuned properly, the performance won't be the best and the reliability of classification reduces.

GridSearchCV from the model_selection module of sklearn was used to tune the parameters of each of the above mentioned algorithms. The parameters that were tuned for AdaBoost Classifier were n_estimators and random_state with values 100 and 1 respectively

Validation

The use of particular data to train the classifier and testing it on a separate set of values in the data is called validation. A classic mistake is to train and test the classifier on the same set of values which leads to overfitting (high variance). A proper balance between bias (too much generalization) and variance is what gives a good classifier. The project divides data into training set and testing set. The classifiers are trained on the training set and the result is tested on the testing set. Apart from this, another python file called tester.py uses stratified shuffle split cross validation to test the classifier and returns the average values of all the evaluation metrics.

Evaluation Metrics

- Accuracy measures the % of data points that were correctly classified among all the data that were classified. In other words, it tells how accurate our classifier is in classifying a given data point correctly
- Precision measures the % of data points that were classified as POI were actually POI. That is, among all the persons flagged as POI, how many of them were actually POI. The value was 0.52859 for AdaBoost Classifier.
- Recall measures the % of data points that were classified as POI given that they are POI. That is, out of all the POIs, how many were correctly identified as POI by the classifier. The value was 0.37900 for AdaBoost Classifier.
- F1 score is a weighted average of precision and recall. It reaches its best value at 1 and worst value at 0.
- F2 score is calculated by giving twice the weight to recall as opposed to precision. In other words, it places more emphasis on those persons who were wrongly classified as non-POIs.

Conclusion

The Enron POI identifier created using machine learning techniques performed the best when a tuned AdaBoost Classifier was used. The accuracy of classification was high (0.863). Even though the precision and recall values were not so high (0.52859 and 0.37900 respectively), they were much better than what was required (both > 0.3) to meet the specifications of the project.