Identifying Person of Interest in Enron Fruad Case

Jenny Hung

Project Goal

We analyzed the Enron data set, which composes of both the finance-related data (such as salary, bonus and stocks) as well as email-related data (such as communication to/from a Person of Interest (POI). Our goal is to help identify all POIs from the available Enron data. This goal is achieved by several iterations of data explorations, detecting and removing outliers, and engineering meaningful features before constructing and experimenting different machine learning algorithms.

Features Used

Given that finance-related features center around remunerations, whereas email-related features center around intensity of communications, I engineered new features by first constructing the finance- and email-ratios by dividing each of them by 'salary' and 'from_messages', then sum them up. The new features (and final features list) are therefore: 'finAll', 'emailAll', 'from_poi_ratio'.

Models/Algorithms Used

Two models were fitted: one with a decision tree classifier, the other with support vector machines. Both returned reasonable results however I have chosen to Decision Tree Classifier as my final model. Please see Evaluation Metrics for model summary.

Hyper-Parameter Tuning

Many of the algorithms I have tried need to be tuned by parameters. If an algorithm that requires tuning was used without proper tuning, we could get a result that is far worse than not using this algorithm at all. To tune and determine the final parameter values, I have used a max depth of 2 (for decision tree) and a list of values for C and γ (for SVM), and adopted cross validation to pick the best values that maximizes the $f1$ score.

Validation

Validation refers to separating the data into training and test sets, where the test sets are used for testing of model performances only. A simple-minded mistake could happen where one trains the model on all the available data (especially when the quantity of data is limited). A good approach, apart from the separation of training and test sets, is cross-validation, which we use in this project.

Evaluation Metrics

My model (using SVM) produces reasonable results (see Table below). With a recall of 0.433, and precision of 0.497. It means that there is a POI in the data set, my algorithm will pick it out half the time.

It also means that a person is truly a POI approximately half the time given that he or she has been identified as a POI by my algorithm.

Table 1: Comparison of Evaluation Metrics

	Precision	Recall	F1
Decision Tree	0.49799	0.43300	0.46323
Support Vector Machine	0. 73376	0.23150	0.35196