**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**

Careful introspection regarding the original data features revealed that although they are many features, the finance-related features center around remunerations, whereas email-related features center around intensity of communications. Therefore I engineered new features by constructing the following ratios:

New features: ['poi', 'salary/director\_fees ratio', 'salary/total\_stock\_value ratio', 'salary/other ratio', 'from\_this\_person\_to\_poi/from\_messages ratio', 'from\_poi\_to\_this\_person/from\_messages ratio', 'salary/long\_term\_incentive ratio', 'to\_messages/from\_messages ratio', 'salary/total\_payments ratio', 'salary/loan\_advances ratio', 'salary/bonus ratio', 'shared\_receipt\_with\_poi/from\_messages ratio', 'salary/deferred\_income ratio', 'salary/exercised\_stock\_options ratio', 'salary/restricted\_stock ratio', 'salary/expenses ratio', 'salary/deferral\_payments ratio', 'salary/restricted\_stock\_deferred ratio']

PCA was then deployed to reduce the dimensionality of the engineered features.

**Models/Algorithms Used**

Two models were fitted: one with a decision tree classifier, the other with support vector machines. Both returned good results however I have chosen to SVM 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 list of values for and (for SVM) and math depth of (for decision tree), and adopted cross validation to pick the best values that maximizes the 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 a good result (see Table below). With a recall of 1.0, and precision of 0.96, it means that there is a POI in the data set, my algorithm will pick it out. It also means that there is a strong probability that a person is truly a POI given that he or she has been identified as a POI by my algorithm.