Enron Submission Free-Response Questions

Introduction

The goal of the project is to see if we can identify the POI in Enron fraud via machine learning technique using financial and email data. Person of interest here is defined as anyone who was indicted, settled or testified in exchange for immunity in the Enron fraud case. Because there are many variables to interpret, machine learning techniques will be used to identify pattern in the data to identify POI.

Dataset was review for outlier prior to classification. There was a 'total' entry because of spreadsheet conversion. There are also a entries where number appear very small or negative relative to rest of dataset. This type of extreme outliers were removed prior to the classification.

Feature Selection

A few different feature selection approaches explored. These approach includes KBest, DecisionTree, and RFECV. RFECV was chosen because the systematically feature elimination processes that identify optimal number of features while KBest and DecisionTree only provide importance of the features.

Scaling of the data was performed with MinMaxScaler. This was necessary because mix of features scale between email, financial and the new ratio features.

New financial ratio features were created. In order to identify financial irregularity, key variables was normalized to allow for better comparison. For example, large amount of exercised stock options is not a clear signal because every insiders has different amount being granted depending on their position. Exercised stock options / total stock value would be more accurate base of comparison across the insider. Two email ratio (from poi and to poi) from class were also recreated. These follow the logic that person who has more contact from POI or to POI is also more likely to be a POI.

Algorithm Selection

Decision tree and adaptive boost was the first two algorithm choose initially because they don't require scaling and they also provides insight into feature importance. After the code was flushed out SVM was ran for comparison. SVM got much better result so it was used that in the final classier along with scaling.

Parameter Tuning

After choosing the classifier, parameter tuning is needed to tell the classifier how closely it should try to match the training data. Without this step it's possible for the classifier to cause overfitting or underfitting. Documentation was reviewed to see what parameters the function support and then GridSearchCV was used to test a range of parameter for the best F1 score. Parameter tuning was performed on the whole dataset prior to breaking the data into test/train.

Validation
Validation step is needed to verify classifier will perform well on non-training data with reasonable precision and recall. If validation step is skipped it's possible to get a classifier that performs very well on training data but not when new data is introduced.

Validation was tricky in this case because POI was sparse. Stratified Shufflesplit was applied for validation to ensure adequate amount of POI exist in every fold.

Final Result
The optimization step with GridSearchCV and RFECV use F1 score for evaluation metric. F1 scored was chosen because its a weighted average of the precision and recall so it simplify the optimization step. Optimization was performed to increase the weighted F1 rather than precision or recall.

The final result has a precision of .7, recall of .315 and F1 score of .434. Precision score of .7 means there is 70% chance the POI identified are correct. Recall of .315 means that there is 68.5% chance that the non-POI identified could actually be a POI.