**INTRODUCTION TO MACHINE LEARNING, FINAL PROJECT**

**(POI IDENTIFICATION IN THE ENRON SCANDAL USING THE ENRON DATASET)**

**INTRODUCTION**

This document is with reference to the final project for Udacity’s Intro to Machine Learning course. The objective of this project is to develop a classifier to identify the persons of interest who might be involved in the Enron Scandal. The Enron dataset provided the information of various employees at the company.

**UNDERSTANDING THE DATASET**

A walkthrough of the Enron dataset provided the following insights-

* Number of data points = 146
* Allocation across classes:
* POI labels = 18
* Non-Poi labels = 128
* Total Features = 21
* Features with missing values: {[‘salary’: 51], [‘deferral\_payments’: 107], [‘total\_payments’: 21], [‘loan\_advances’: 142], [‘bonus’: 64], [‘restricted\_stock\_deferred’: 128], [‘deferred\_income’: 97], [‘total\_stock\_value’: 20], [‘’expenses’: 51], [‘exercised\_stock\_options’: 44], [‘other’: 53], [‘long\_term\_incentive’: 80], [‘restricted\_stock’: 36], [‘director\_fees’: 129], [‘to\_messages’: 60], [‘from\_messages’: 60], [‘from\_poi\_to\_this\_person’: 60], [‘from\_this\_person\_to\_poi’: 60], [‘email\_address’: 35], [‘shared\_receipt\_with\_poi’: 60]}
* Further, scatter plots between salary and bonus, salary and total\_payments indicated the presence of an outlier ‘TOTAL’ which was then removed from the dataset.

**FEATURE SELECTION & FEATURE SCALING**

Features with large number of missing values were omitted. These features include-

* deferral\_payments
* loan\_advances
* restricted\_stock\_deferred
* deferred\_income
* long\_term\_incentive
* director\_fees

Therefore, the features which seem to have a direct impact on the identification of POI were chosen-

* poi
* salary
* bonus
* total\_payments
* exercised\_stock\_options
* from\_poi\_to\_this\_person
* from\_this\_person\_to\_poi
* shared\_receipt\_with\_poi

Also, two new features were created and added in the dataset, ‘frac\_from\_poi’ and ‘frac\_to\_poi’ indicating the fraction of messages sent from poi to a person and fraction of messages sent to a poi from a person respectively. This is created on the fact that the rate at which a poi will send/receive emails to/from another poi will be higher as compared to the other employees.

Feature scaling is deployed because there is a wide difference between the range of the financial features and the email features.

The feature importance of the features used are as follows-

Feature Ranking:

* 1 feature salary (0.301368701787)
* 2 feature bonus (0.242076034983)
* 3 feature total\_payments (0.203101003849)
* 4 feature exercised\_stock\_options (0.118748465881)
* 5 feature from\_poi\_to\_this\_person (0.0679553264605)
* 6 feature from\_this\_person\_to\_poi (0.0413605648445)
* 7 feature shared\_receipt\_with\_poi (0.025389902194)
* 8 feature frac\_from\_poi (0.0)
* 9 feature frac\_to\_poi (0.0)

Since, the added features are of least importance, they aren’t used in the analysis.

**Performance with the added features-**

Accuracy = 0.89655

Precision = 0.33

Recall = 0.5

F\_score = 0.4

Best parameter = {'min\_samples\_split': 10, 'splitter': 'best', 'criterion': 'gini'}

**Performance without the added features-**

Accuracy = 0.93103

Precision = 0.5

Recall = 1

F\_score = 0.67

Best parameters = {'min\_samples\_split': 15, 'splitter': 'random', 'criterion': 'entropy'}

**PICK AND TUNE AN ALGORITHM**

Several algorithms were tested like the Naive Bayes, Support Vector Machines, Decision Tree and Adaboost. The performances given by these algorithms with the default parameters were-

1. Naive Bayes

Accuracy = 0.896551724137931

Precision = 0.4

Recall = 1.0

1. SVM

Accuracy = 0.9310344827586207

Precision = 0.0

Recall = 0.0

1. Decision Tree

Accuracy = 0.8275862068965517

Precision = 0.2

Recall = 0.5

1. Adaboost

Accuracy = 0.8620689655172413

Precision = 0.3333333333333333

Recall = 1.0

For the final analysis, Decision Tree Classifier was used as it gave better results.

The algorithms do not always give the best results with the default parameters, as a result of which parameter tuning is required to optimise the results.

GridSearchCV() was used to tune the algorithm in which a set of parameters are passed and the combination of these values is chosen to run the algorithm, the process being repeated until the optimised result is obtained.

The parameters and their values used were:

'criterion': ('gini', 'entropy'),

'splitter': ('best', 'random'),

'min\_samples\_split': [5, 10, 15]

The optimal parameters were found to be:

'min\_samples\_split': 15,

'splitter': 'random',

'criterion': 'entropy'

**VALIDATION AND EVALUATION**

Validation is the process in which we first train out model using a subset of the dataset and then test/evaluate our model using the complementary dataset.

‘train\_test\_split’ of the ‘cross\_validation’ package was used to split the dataset into 80% training data and 20% test data.

Accuracy, precision and recall were calculated to evaluate the results.

Accuracy = 0.93103

Precision = 0.5

Recall = 1

F\_score = 0.67

Precision denotes the likelihood that a person that is identified as a POI, is actually a POI and Recall measures how likely is it that a POI will be identified correctly by the algorithm.