Identify-Fraud-from-Enron-Email

1: Data exploration

This data set has 146 data points. There has 21 features, which include 'salary', 'to_messages', 'deferral_payments', 'total_payments', *etc*. The feature “poi” is the label for this study to identify that a person is a poi or not. The job for this project is using 20 features to make a classifier, which could be used to decide whether a person is a poi or not. By check the data, there has 18 people which is poi from total 146 data points. Thus this data set is unbalanced. This means that when we choose the training and testing data set, it is important to make sure that the ratio of POI and non-POI is same. Besides that, the accuracy is not a very good way to evaluate the classifier performance for this data.

By check the data, the key --- ‘TOTAL’ is removed from the dataset. So current number of data points is 145.

2: Intelligently select features

I create two new features from “from_poi_to_this_person” and “from this person to poi”. The reason to create such two new features is that if a people communicate with poi frequently, the probability to be a poi is also very high. But I do a normalization by “to_messages” and “from_messages” to generate those such features include 'fraction_from_poi_email' and 'fraction_to_poi_email'.

Later I combine pipeline, SelectKBest, and gridsearchCV method to decide how many features should be selected for this model. Based on the evaluation result, I choose the 7 features, which are 'salary', 'exercised_stock_options', 'bonus', 'total_stock_value', 'deferred_income', 'long_term_incentive', and 'fraction_to_poi_email'.

Feature scaling is a method used to standardize the range of independent variables or features of data. It is very important to get a better machine learning model for some algorithms including the SVM, K-mean cluster. Since we choose the decision tree algorithm in this study, it is not necessary to do any feature scaling work.

3: Data cleaning

By check the data, the key --- ‘TOTAL’ was removed from the dataset. So current number of data points is 145. Later I also check the missing data points for each features. There are 62 void data points selected from 146 data points. All the missing data points are removed from the data set.

4: Pick an algorithm

Then I tried the naïve bayes algorithm at first and compare with decision tree algorithm. Through the comparison, it is concluded that the decision tree method is much better.

5: Tune the algorithm

My goal in this project is to develop a classifier which has the high performance to identify poi. To obtain better performance, I need to optimize the algorithm parameter in order to enable the algorithm to have better performance. Since the algorithm I choose is the decision tree, an important parameter is “min_samples_split”, which specifies the minimum number of samples required to split an internal node. A decision tree is classically an algorithm that can be easy to overfitting. Thus the tuning of such parameter helps us to obtain a classifier with good performance without overfitting. I also choose “GridSearch” method to find the best parameter. After tuning, the performance is much better than before.

6: Validation strategy

Validation is very important in machine learning since it evaluate the working ability of a trained model. In this way, we can obtain a final model with best performance. For classifier model, there have several important parameter for evaluation which include accuracy, precision, and recall score, which are applied in this model. Since the data set for this study is not balanced, accuracy is not a very good way to evaluate the classifier performance. So precision and recall score are used in this study. The precision is the fraction of true POI data points among the data points which are identified as POI. The recall is the fraction of true POI data points which are identified among all the POI data points.

If we learn the parameters of a prediction function and test it on the same data, the model would have a perfect score and fail to predict anything useful. This situation is called overfitting. So normally, we apply cross-validation strategy to split data, training model and perform test separately.

In this study, I try two different methods to perform the cross-validation, which are “K-fold” and “Stratified-KFold” methods. In this dataset the majority of data points are non-poi. Thus the StratifiedKFold method is much better, since the folds are made by preserving the percentage of samples for each class based on this method. Thus the StratifiedKFold method is chose as the final method.

7: Usage of evaluation metrics

After I tune the parameter, I use tester file to evaluate the result. The final results I got is 0.71400 for accuracy, 0.49941 for precision and 0.42150 for recall.