1.Summarize for us the goal of this project and how machine learning is useful in trying to accomplish it. As part of your answer, give some background on the dataset and how it can be used to answer the project question. Were there any outliers in the data when you got it, and how did you handle those? [relevant rubric items: "data exploration", "outlier investigation"]

Answer

The goal of this project is to identify the point of interest in the fraud case, which means individuals who were indicted, reached a settlement or plea deal with the government, or testified in exchange for prosecution immunity. The dataset contains 145 employees information and 20 features including financial information and email features such as the number of from/to emails. 18 out of 145 are poi and the others are not poi. If the features were only small number, we could have checked the relationship between the featuers and then, we could identify the poi. However, since there are many features, it would be difficult to identify the poi by plotting the features into the graph.(Three dimensions are the maximum for us to understand.) On the other hand, by utilizing machine learning, we can understand which features are important, then by using some of machine learning algorithm, we can identify poi.

The dataset includes some outliers both in poi and non-poi. The number of poi is 18 and non-poi is 127. Since the number of poi is much smaller than non-poi, it should not be eliminated from the dataset. I removed "TOTAL" which is clearly not person. I removed "FREVERT MARK A" since he isn't "poi" and is far much higher salary than any other "not-poi".

2.What features did you end up using in your POI identifier, and what selection process did you use to pick them? Did you have to do any scaling? Why or why not? As part of the assignment, you should attempt to engineer your own feature that does not come ready-made in the dataset -- explain what feature you tried to make, and the rationale behind it. (You do not necessarily have to use it in the final analysis, only engineer and test it.) In your feature selection step, if you used an algorithm like a decision tree, please also give the feature importances of the features that you use, and if you used an automated feature selection function like SelectKBest, please report the feature scores and reasons for your choice of parameter values. [relevant rubric items: “create new features”, “properly scale features”, “intelligently select feature”]

Answer

The dataset contains many "NaN" in some features. Those features are not suitable to predict whether the person is poi or not. Therefore, I removed those features which include more than 50% of "NaN". (Specifically,loan_advances,restricted_stock_deferred,deferred_income,long_term_incentive and director_fees) At first, I selected "poi","salary", "deferral_payments","total_payments","bonus","total_stock_value","expenses","exercised_stock_options","other","restricted_stock","to_messages","from_messages" ,"from_this_person_to_poi" and "shared_receipt_with_poi" as features(14 features). In my poi_id_project5, I used SelectKBest to check which variables are important. To do this, I put SelectKBest into pipeline and set the "K" from 1 to 14(all possible features). From the result of all possible numbers(1 to 14), K= 2 is the highest score. Therefore I chose 2 .
After setting those features, I changed the scale. I set the max value to be 1 and minimum value to be 0 since the figure of financial features(some of them are more than 1000000) is much higher than that of email features(most of them are less than 1000).

I added two new features which are "fraction_from_poi" and "fraction_to_poi". Those values are caluculated from dividing "from_poi_to_this_person" with "from_messages" and "from_this_person_to_poi" with "to_messages",respectively.

3.What algorithm did you end up using? What other one(s) did you try? How did model performance differ between algorithms? [relevant rubric item: “pick an algorithm”]

Answer

I used DecisionTreeClassifier. I tried Support Vector Machine, Adaboost, KNeighborsClassifier and RandomForest. When I used Support Vector Machine, I couldn't get high value of precision and recall. However DecisionTreeClassifier returns the best evaluation metrics.

4.What does it mean to tune the parameters of an algorithm, and what can happen if you don’t do this well? How did you tune the parameters of your particular algorithm? (Some algorithms do not have parameters that you need to tune -- if this is the case for the one you picked, identify and briefly explain how you would have done it for the model that was not your final choice or a different model that does utilize parameter tuning, e.g. a decision tree classifier). [relevant rubric item: “tune the algorithm”]

Answer

Most of the algorithms take several parameters and taking the appropriate value is quite important to acquire good result. If we set the paremeter poorly, it can be highly possible to get poor result or take so much time to calculate.

I chose Decision Tree Classifier and set the criterion "entropy" and other parameters are default.

5.What is validation, and what’s a classic mistake you can make if you do it wrong? How did you validate your analysis? [relevant rubric item: “validation strategy”]

Answer

Validation is the way whether we can predict the outcome. The classic mistake is to train all the data and test all the data. The result will be high. However if we add several new data and try to test the algorithm which are already trained by all of the data, the result will be very low. This is because of overfitting.

In the tester.py, stratified shuffle split cross validation is used. This cross validation is better than StratifiedKFold in this project. Since the number of "poi" is much smaller than "not-poi", it is highly likely to split data without containing any "poi" in training or test set. Stratified shuffle split cross validation preserves the percentage of samples for each class. Therefore, there is no chance to have no "poi" in training data set. However, StratifiedKFold is just to split the data into K (without shuffling), so it may include no "poi" in training or test data set.

6.Give at least 2 evaluation metrics and your average performance for each of them. Explain an interpretation of your metrics that says something human-understandable about your algorithm’s performance. [relevant rubric item: “usage of evaluation metrics”]

Answer

Through the algorithm I made, I got the following result.

Accuracy:0.83887, Precision:0.37927, Recall:0.3275

	Predicted		
True		P	N
	P	True Positive (655)	False Negative (1345)
	N	False Positive (1072)	True Negative (11928)

In the explanation below, I use True Positive as TP, False Negative as FN and so on.

Accuracy is defined as (TP+TN)/(TP+FN+FP+TN). This value shows that the classifier can predict the fact correctly.

Precision is defined as TP/(TP+FP). This value show that the classifier can predict the
If the rate is high, that means the classifier can predict the truth correctly when the classifier predict Positive. On the other hand, if the rate is low, that means the classifier can't predict the truth correctly(negative examples are misclassified as positive). In this case, classifier predicted the data as "poi" even though the truth is "not-poi".
Recall is defined as TP/(TP+TN). This rate shows how much the positive examples are classified correctly. In this case, high rate of recall means that high number of "poi" is correctly classified as "poi" and small number of "poi" is misclassified as "not-poi". Low rate of recall means that small number of "poi" is correctly classified as "poi" and high number of "poi" is misclassified as "not-poi".