Xiaolong Yang
Feb 28,2017

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”]

Enron is an American energy company, bankrupt after fraud scandal. The dataset of the email metadata and financial data created by 144 employees of Enron. The goal is to find who was fraud, or was one of the Person of Interest in Enron Scandal. Since a person in the dataset is either one of the POI or not and we had the POI labels in the dataset, the problem became a Binary Classification question. Using features such as ‘salary’ and ‘exercised stock options’, provided by the file “enron61702insiderpay.pdf”, we can use classification algorithm to find who was the Person of Interest. Because the project requires highest accuracy, recall and precision, I need to optimize the result of the model, using different algorithms and select specific features.

The dataset contains 13 features and the label POI and 146 data points. There are 18 POI person in the dataset, 12.32% of total dataset. We can find it is an unbalanced dataset for training, we need to resample the dataset and try evaluate matrix like recall and precision over merely accuracy. For the 13 features of the dataset, we can find all of them contains at least 44 missing value. The top 3 features are “loan advances”, “director fees” and “restricted stock deferred” have 142,129 and 128 missing values. salary', 'to messages', 'deferral payments', 'total payments', 'exercised stock options', 'bonus', 'restricted stock', 'shared receipt with poi', 'restricted stock deferred', 'total stock value', 'expenses', 'loan advances', 'from messages', 'other', 'from this person to poi', 'poi',

'director fees', 'deferred income', 'long term incentive', 'email address', 'from poi to this person' contains 0.349, 0.733, 0.548, 0.973, 0.438, 0.877, 0.349, 0.301, 0.363, 0.664, 0.247, 0.884 missing value rate.

There are two principles for the outliers, find those not an actual people and those irrelevant from the selected features. I plot the data and find the “TOTAL” is the combination of all other data, it did not indicate an actual person. We would also exclude those has almost all features 'NaN' point, include 'WODRASKA JOHN', 'WHALEY DAVID A', 'LEWIS RICHARD', 'PIRO JIM', 'WROBEL BRUCE', 'LOCKHART EUGENE E', 'THE TRAVEL AGENCY IN THE PARK', 'BROWN MICHAEL', 'HAYSLETT RODERICK J', 'SCRIMSHAW MATTHEW', 'GRAMM WENDY L’. The further analysis based on selected features, if the one of the data has all these features ‘NA’, I would exclude that as an outlier. The data popped out were 'HAUG DAVID L', 'CLINE KENNETH W', 'WAKEHAM JOHN'.

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”]

I tried to collect the TF-IDF of key words by POI and Non-POI in mail. I find the top 3 influential key words are “Vinc”, “Vkamin” and “Vincent”, which all indicated the name “Vincent Kaminski”, or “KAMINSKI, WINCENTY J” showed in the dataset. Vince worked as Managing Director for Research at Enron and raised strong objections to the financial practices of Enron’s CFO, who fraud concealed company’s burgeoning debt. The TF-IDF model would have all the accuracy, precision and recall close to 1.

Although the evaluation looks perfect and the new features are intuitively relevant, the TF-IDF model conclude only 14 POI and 2 Non-POI (since we did not know others actual email address based on their name in the dataset), the prediction won’t help the whole dataset. In this case, I won't use this data in further analysis.

I ran the decision tree model to find influential features, using “feature importance” to sort the feature list. However, the features selected each time are different, in this way I test the model 10,000 times to show all possible result, and union all result together. The initial feature list contains 'from messages', 'salary', 'other', 'deferred income', 'deferral payments', 'expenses', 'exercised stock options', 'bonus'. their significant points are 0.21, 0.14, 0.27, 0.25, 0.21, 0.18, 0.32, 0.14. The closer point is to 1, the more important the feature is, all features selected have higher than 0.1 importance.

Since we did not know how many feature used would optimize the performance, I have tried all combination of the features from all selected to 2 highest. We can find use all selected features would give us the highest accuracy 0.823 with precision and recall

higher than 0.3. In other words, I would use 'from messages', 'salary', 'other', 'deferred income', 'deferral payments', 'expenses', 'exercised stock options', 'bonus' for further analysis.

Since the features selected might correlated (each time decision tree choose a different features list for the classification), I use PCA to get the most important features in the final analysis. The PCA algorithm also do the data normalization and new features creation job.

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”]

The final decision based on K Nearest Neighbors, other algorithms include GaussianNB, Adaboost and Random Forest. KNN has the highest accuracy as 0.86900, highest precision as 0.75296 and highest recall as 0.31850. Adaboost also meet the rubric, but its accuracy and precision is lower than KNN. Random Forest has a recall lower than 0.3.

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”]

Tuning the parameters would help you optimize the model performance. If you did not tune well, even if you used the correct model, it won't give the best result. The result should not only consider accuracy, but also recall and precision.

For KNN, you need to tune the numbers of neighbors. For Random Forest, you need to tune the number of estimators, number of trees in the forest and number of features to consider for each split.

I used the "optunity" packages to auto tune the parameters. It showed that using KNN with a 2 neighbors would reach the best result, compared to all the models and all the parameters I set in the "select".

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"]

Validation is used to split the data into train and test part in the first place. It helps the model not over fit in training data but still get enough data for train.

I started the validation using cross validation for the feature selection part, but then I used the StratifiedShuffleSplit function in test_classifier since it was the evaluation function the project request. The fold was 1,000, the dataset was random selected 1,000 times and the model ran 1,000 times for each model and parameters I tuned. It took quiet a long time for the auto tuning part.

Xiaolong Yang
Feb 28,2017

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"]

Even though the result output gave Precision, Recall, F1, F2, I used the Precision, Recall for the evaluation.

T = True, F = False, P = Positive, N = Negative

Precision = TP / TP+FP = 637/ (209+637) = 0.75296

Precision means the ratio of correct positive predictions compared to total positive prediction (both true positive predictions and false positive predictions). For this project, it means the correct predicted POI out of the total predicted POI. It showed that from 865 predicted POI, 651 correct predicted are relevant.

Recall means the ratio of correct positive prediction compared to total actual positive (both true positive predictions and false negative predictions). For this project, it means the correct predicted POI out of the total actual POI. It showed that from 2000 actual POI, 651 relevant POI are selected.

T = True, F = False, P = Positive, N = Negative

Recall = TP / TP+FN = 637/ (637+1363) = 0.31850