Enron Submission Free-Response Questions

ML project5
By: Fareeda saleh

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”]

The goal of this project is to build predictive model for Enron company which is a famous fraud case in US .We identify the person of interested POI by using financial and email information of the Enron dataset. Using Machine learning in this case is useful as Ml technics are faster and powerful more than human to identify the relationships and pattern from financial and email information for those who may sheared in the fraud to predict the POI. There is 146 employees in this data set 18 of them is POI and 128 is non POI . There are 21 features divide to 3 parts:

-13 financial features (all units are in US dollars):

['salary',
'deferral_payments',
'total_payments',
'loan_advances',
'bonus',
'restricted_stock_deferred',
'deferred_income',
'total_stock_value',
'expenses',
'exercised_stock_options',
'other',
'long_term_incentive',
'restricted_stock', 'director_fees'
'director_fees']

-6 features for email information units number of emails messages exception is 'email_address', which is a text string):

['to_messages',
'email_address',
'from_poi_to_this_person',
'from_messages',
'from_this_person_to_poi',
'shared_receipt_with_poi']

-One boolean feature which is label represented as integer 0 or 1:
[poi]

Missing features:

All the features has missing values which is represent as NaN except the poi features

Feature Name	Number of Missing Values
salary	51
to_messages	60
deferral_payments	107
total_payments	21
long_term_incentive	80
loan_advances	142
bonus	64
restricted_stock	36
restricted_stock_deferred	128
total_stock_value	20
shared_receipt_with_poi	60
from_poi_to_this_person	60
exercised_stock_options	44
from_messages	60
other	53
from_this_person_to_poi	60
deferred_income	97
expenses	51
email_address	35
director_fees	129

Outliers :
Before I add any features **I** identify the outliers by drawing scatterplot and I found a clear outlier point and after review the enron61702insiderpay pdf I found it related to the TOTAL record which is not employees so I choose to remove it and also I remove THE TRAVEL AGENCY IN THE PARK for the same reason in addition to

'LOCKHART EUGENE E' because it has no non NaN values, then I draw another scatterplot and also I found a clear outliers but it related to (SKILLING JEFFREY K , LAY KENNETH L) so I kept them as they valid data point.

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 importance 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 create 4 new features:

- fraction_from_poi:
  create this feature to check the fraction of emails send by POI to all other emails.
- fraction_to_poi: create this features to check the fraction of emails send by all to the POI.
- fraction_shared_receipt: This feature is drive from shared_receipt_with_poi feature that identify the relationship between POI and other POI by assuming that POIs contacted to each other more than Non POI.
- Salary_Bounace: I create this feature to identify the fraction of bounce to salary so if there is a big difference between bonus and salary so there is a high probability to be this person POI.

I use scaling feature with all algorithms that I try it in this model (SVC, AdaBoost, KNeighbors, naive bayes) except decision tree because it don't usually require scaling however, applying feature scaling will allowed allowed for greater flexibility and performance in using algorithms especially SVC.

For chosen the features I used SelectKBest , in a pipeline with grid ,the following is the 6 features which selected with their scores :

Features	Scores	P-Values
loan_advances	30.73	0.000
fraction_shared_receipt	21.12	0.000
poi	15.86	0.000
fraction_from_poi	15.84	0.000
from_this_person_to_poi	10.72	0.001
deferred_income	10.63	0.002

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”]

I try 5 algorithms and I used classification report and tester.py to evaluate the algorithms performance :

	naive bayes	decision tree	SVC	AdaBoost	KNeighbors
precision	0.40767	0.23707	0.82667	0.26654	0.17700
Recall	0.31350	0.56400	0.03100	0.20550	0.17700

We can see the SVC is poor in recall but in other hand it is the best in precision ,KNeighbors is worst in both precision and recall and decision tree is the best in recall, so I choose the naive bayes algorithm in this model as it has precision and recall >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? What parameters did you tune? (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”]

Parameters tuning is the last step of the process of applied machine learning before presenting results ,it means the change of algorithm parameters input to effect the performance of this algorithm if we don't do this step well then we would see the bad impact in our algorithms (recall, accuracy..). I didn't tune the naive bayes as it doesn’t

have parameter. For other algorithms I used the GridSearchCV to tune the algorithms parameter with using stratified shuffle split cross-validation to guard against bias This is the parameters that I used with SVC algorithm:

- kbest__k=[2, 3, 5, 6,12]--6
- SVC__C=[0.1,1, 10, 100, 1000,10000]--10000
- SVC__kernel=['rbf']--rbf
- SVC__gamma=[0.001, 0.0001] –0.001

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"]

The validation is use to ensure the ML algorithms generalizes well by keeping a validation data set out from a training data set ,if we use the test set to evaluate the ML algorithms then it might lead to overfitting the training set and this will effect on the ability predictive of our model.

In the poi_id.py the data splitting into 70% as training set and the rest is testing set and as our dataset is small I used the StratifiedSuffleSplit to random split and create multiple dataset and this will help us to have more accurate result than the single split .

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"]

As our data set imbalance 18 POI vs 128 Non POI we will ignore the accuracy score and focuses in Recall and precision scores .Recall mean simply how many POI are actually selected while precision is how many selected records are POI . in our model we found the precision is 0.40767 (40% are actually POI) and the recall is 0.31350 (can capture 31% of POI) So I think I should emphasize my model to rise the recall because we want to unmissed any POI.

References :

Udacity: https://udacity.com

Sklearn: http://scikit-learn.org

What is the best way to understand the terms "precision" and "recall"? from Quora :

https://www.quora.com/What-is-the-best-way-to-understand-the-terms-precision-and-recall