UDACITY

PROJECT

Identify Fraud from Enron Email

A part of the Data Analyst Nanodegree Program

PROJECT REVIEW

CODE REVIEW 4

NOTES

SHARE YOUR ACCOMPLISHMENT!

Requires Changes

4 SPECIFICATIONS REQUIRE CHANGES

Dear Student,

Congratulations on your first submission!. This is a challenging project and you address most of it, well done!. You'll find some sections are marked as not meeting specifications, I hope you find my argumentation reasonable and clear and helps you to continue your work on such sections. However, if you still have questions/comments, please don't hesitate to reach us, we'll be glad to help you.

Keep up your good work!

Quality of Code

Code reflects the description in the answers to questions in the writeup. i.e. code performs the functions documented in the writeup and the writeup clearly specifies the final analysis strategy.

Your written response perfectly describes your strategy, includes the required level of detail when required and your code includes all the processes mentioned. Machine learning is not just about building good models, it is also about communicating results to your audience in a clear and direct way. You achieve both goals. Well done!.

As a suggestion, this project is a great opportunity for you to create a new repository in Github that becomes part of your online portfolio and allow potential employers to review your work. This report defines your credentials, so it is important that you put special attention not just to the technical side of the project but also the communications side since this is a critical characteristic for any data scientist. For your reference, check this Kaggle post for further reference, as you can see this is really a hot topic in the data science world! 😉

poi_id.py can be run to export the dataset, list of features and algorithm, so that the final algorithm can be checked easily using tester.py.

All required `.pkl` files are included and `poi_id.py` worked without problems.

Understanding the Dataset and Question

Student response addresses the most important characteristics of the dataset and uses these characteristics to inform their analysis. Important characteristics include:

- **total number of data points**
- **allocation across classes (POI/non-POI)**
- **number of features used**
- **are there features with many missing values? etc.**

Your report includes a description of the main characteristics of the dataset.

Note these numbers are particularly important since they describe the main dataset characteristics:

1. The small data set is why the tester.py file uses StratifiedShuffleSplit instead of a simpler cross-validation method such as TrainTestSplit. StratifiedShuffleSplit will make randomly chosen training and test sets multiple times and average the results over all the tests.

2. The data is unbalanced with many more non-POIs than POIs. StratifidShuffleSplit also makes sure that the ratio of non-POI:POI is the same in the training and test sets as it was in the larger data set.
3. The unbalanced data is also why we use precision and recall instead of accuracy as our evaluation metric.

For your reference, some techniques to handle unbalanced datasets and this repo with plenty of tools.

Student response identifies outlier(s) in the financial data, and explains how they are removed or otherwise handled.

Nice job finding the TOTAL row. As a suggestion, the dataset comes in a json format that makes it difficult to handle and explore. My suggestion is to use Pandas:

```
data_dict = pickle.load(open("final_project_dataset.pkl", "r") )
###creating dataFrame from dictionary - pandas
df = pandas.DataFrame.from_dict(data_dict, orient='index', dtype=np.float)
print df.describe().loc[:,['salary','bonus']]
```

Check the docs for more information on how pandas can read data from different sources.

Optimize Feature Selection/Engineering

At least one new feature is implemented. Justification for that feature is provided in the written response, and the effect of that feature on the final algorithm performance is tested. The student is not required to include their new feature in their final feature set.

Good work engineering your features, including your reasons and testing their impact over your classifier, however correlation is not an appropriate test to determine their importance since POI is a categorical variable, you can find more info here on this topic . An appropriate analysis to estimate the relevance of these features would be to calculate the proportions of each value of the independent variable (`POI`) vs all the values of the dependent variable and then use a test that takes into account these proportions. A Contingency Table is an excellent tool for this purpose and Chi Squared Test is the right test to choose since it will test the observed values of each cell against the expected value of each cell in the contingency table and return a test result with a p-value. For your reference, have a look at this link where it is explained the Chi-Squared test and Contingency tables.

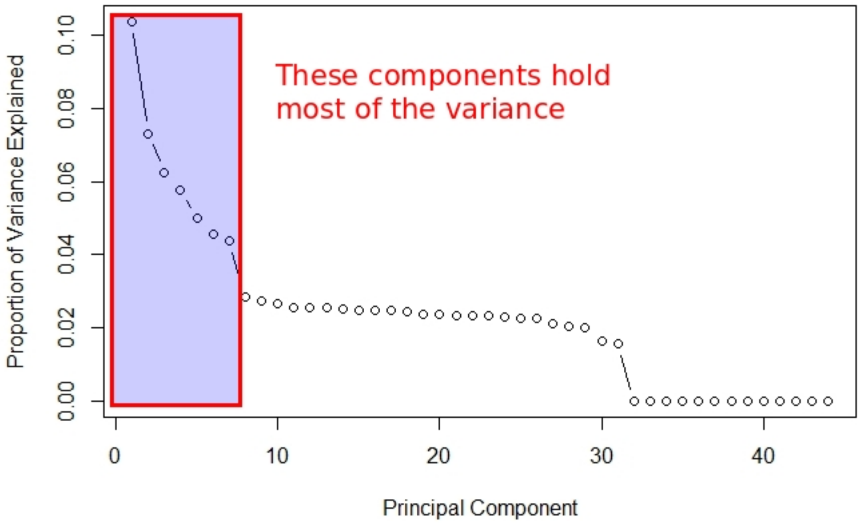
Another option is to simply test your classifiers with/without these features.

Univariate or recursive feature selection is deployed, or features are selected by hand (different combinations of features are attempted, and the performance is documented for each one). Features that are selected are reported and the number of features selected is justified. For an algorithm that supports getting the feature importances (e.g. decision tree) or feature scores (e.g. SelectKBest), those are documented as well.

Good work in this section. Note feature selection process is key in Machine Learning problems, the idea behind it is that you want to have the minimum number of features than capture trends and patterns in your data. Your machine learning algorithm is just going to be as good as the features you put into it. For that reason, this is definitely a critical step into any ML problem and the methodology deployed must be scientific and exhaustive without room for intuition.

Having said so, your methodology is scientific and exhaustive as required, PCA is evaluated in the context of a Pipeline, well done!. Now, since you repeated this process for different classifiers, you identified which is the best performing classifier for this dataset, that goes beyond the project expectations, but definitely, it is worth to do in the real world. Well done!

However, the reason for marking this section off is because PCA components variance is not included and required in the rubric. Try to include a table with the variance explained by each component. The variance associated with each component is critical to understand the predictive power of each component. Here you can find more info about this topic.



STEPS TO PASS THIS SECTION:

1. Include the variance associated to each component.

If algorithm calls for scaled features, feature scaling is deployed.

It is great you scaled your features since some of the classifiers attempted calls for it. However, it would be great if you could include few words in your written response describing which of these algorithms call for scaled features.

Also, great work using a log transform for your features and testing its impact! 👍

Pick and Tune an Algorithm

At least 2 different algorithms are attempted and their performance is compared, with the more performant one used in the final analysis.

Outstanding work testing several algorithms and comparing their performance in terms of f1 using different validation techniques.

As a side comment, note F1 is a subclass of the F-Scores as here described. For example, F0.5 puts more importance over Precision than Recall, it is up to your problems needs to decide any F-Score between 0 (only considers Precision) and infinite (only consider Recall). Here you can find more information.

Response addresses what it means to perform parameter tuning and why it is important.

I have not been able to find the section regarding the importance of parameter tuning, please add some comments about it in your next submission. This is a minor issue but still it is part of the rubric. Lecture videos are a good resource, but also this link or this link.

At least one important parameter tuned with at least 3 settings investigated systematically, or any of the following are true:

- **GridSearchCV used for parameter tuning**
- **Several parameters tuned**
- **Parameter tuning incorporated into algorithm selection (i.e. parameters tuned for more than one algorithm, and best algorithm-tune combination selected for final analysis).**

Good work using SearchGridCV, and as a suggestion to enhance its performance, use the `cv` parameter you can pass a cross-validation object to validate your search results that best adapt to your dataset characteristics.

For example:

```
# Set up cross validator (will be used for tuning all classifiers)
    cv = cross_validation.StratifiedShuffleSplit(labels, 100, random_state = 42)
    a_grid_search = GridSearchCV(clf, param_grid = clf_params,cv = cv, scoring = 'recall')
```

This GridCV object validates algorithm performance using a cross-validation object that best adapts to the dataset characteristics and searches for those parameters that maximize recall.

For your reference, note you can also visualize your grid results.

Validate and Evaluate

At least two appropriate metrics are used to evaluate algorithm performance (e.g. precision and recall), and the student articulates what those metrics measure in context of the project task.

Good work using precision&recall to evaluate your classifier. Also, good definition of both in terms of POI detection.

As a side comment: Note accuracy is not a good score in this case since the dataset in this project is very small and the ratio of negatives to positives is highly skewed (18 POIs out of 146 total cases), a classifier that predicted only non-POIs as output, would obtain an accuracy score of 87.4%. In this regard, accuracy is not a strong validation metric for our classifier. For your reference, this interesting post.

Response addresses what validation is and why it is important.

What we are looking for here are some general comments about what is validation and why is it important, for example, what is the main goal of validation?, why dataset is split into train/test sets?, what is their purpose?, Which is the typical error if we don't perform validation properly?, etc. Lecture videos are good resources, but also this link or this link. For your reference, have a look at this interesting entry on validation

Performance of the final algorithm selected is assessed by splitting the data into training and testing sets or through the use of cross validation, noting the specific type of validation performed.

For this particular problem, a stratified shuffle split is preferred for a number of reasons. Since the dataset is small, a shuffle split will randomly assign entries to test and training sets in a fold. However, the stratification preserves the percentage in the target class as in the complete set. This is particularly important Validation for our investigation as we have such a small percentage of PoIs - if we did not stratify we could easily have a test or train set which contained no entries that were PoIs. If we had no entries that were PoIs in either a test or train set then the model would perform very badly.

When tester.py is used to evaluate performance, precision and recall are both at least 0.3.

Well done!. Your classifier is over 0.3 for precision&recall scores.

RESUBMIT

DOWNLOAD PROJECT

4 CODE REVIEW COMMENTS ›

Best practices for your project resubmission

Ben shares 5 helpful tips to get you through revising and resubmitting your project.

Watch Video (3:01)

Have a question about your review? Email us at review-support@udacity.com and include the link to this review.

RETURN TO PATH

Rate this review