Identifying Fraud from Enron Email

Does Not Meet Specifications

Quality of Code

SPECIFICATION

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.

MEETS SPECIFICATION

Reviewer Comments

Your submission is really good, it is clear and well structured. The machine learning pipeline is properly organized.
Ideally every step/choice of parameters and algorithm selection, tuning and validation should be followed by a table including the adequate figures and reporting the impact of that specific step/choice on the chosen performance metrics and a consequent rationale for the choices made in the light of those results.
Doing so would help the reader properly understanding the machine learning pipeline implemented and the consequent decision process. It is very important, in every scientific work, to explain and justify every step/choice as thoroughly as possible, no matter how obvious to you the various steps might look.

This seldom happens in your report, in particular some improvements are required in:

1. Documenting the impact of newly created features on the performance metrics and explaining the rationale for the new features.
2. Documenting the feature selection process and its impact on the performance metrics.
3. Documenting the algorithm selection process and its impact on the performance metrics.
4. Defining the performance metrics and their meaning in the specific context.
5. Documenting why this specific problems requires to perform validation with special care and using specific techniques.

The project rubric mandatorily requires some definitions to be given and the importance of some factors to be discussed.

I will, section by section, point out the areas where more information is required, please invest some more time on crafting the best possible project for you to showcase in your portfolio or to your prospect employer.

Your actual results in terms of metrics are definitively good, well done, keep up your excellent work!

SPECIFICATION

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.

MEETS SPECIFICATION

Understanding the Dataset and Question

SPECIFICATION

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.

MEETS SPECIFICATION

SPECIFICATION

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

MEETS SPECIFICATION

Reviewer Comments

Good job spotting the otulier TOTAL, please note that there is another one labeled TRAVEL AGENCY IN THE PARK that should possibly be removed as well. It is quite difficult to find it though without manually checking the list of names in the dataset.

Optimize Feature Selection/Engineering

SPECIFICATION

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.

DOES NOT MEET SPECIFICATION

Reviewer Comments

Please include your classifier's results for precision and recall (and/or for any other relevant metric) before and after the newly implemented features are added.

SPECIFICATION

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.

DOES NOT MEET SPECIFICATION

Reviewer Comments

Feature selection is performed and it is done in a smart and effective way but it is not documented well enough. In order meet requirements please include the figures that justify the feature selection. The process needs to be transparent and the numerical rationale for the selection needs to be provided. Regarding this the rubric requires: *"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."*

SPECIFICATION

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

MEETS SPECIFICATION

Pick and Tune an Algorithm

SPECIFICATION

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

DOES NOT MEET SPECIFICATION

Reviewer Comments

Well done testing multiple algorithms, please provide the performance results as well, as from the rubric: "At least 2 different algorithms are attempted and their performance is compared, with the more performant one used in the final analysis." The comparison needs to be shown in the document.

SPECIFICATION

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

MEETS SPECIFICATION

SPECIFICATION

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).

MEETS SPECIFICATION

Reviewer Comments

Well done!

Validate and Evaluate

SPECIFICATION

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.

DOES NOT MEET SPECIFICATION

Reviewer Comments

To pass this section please provide a definition of each used metric and describe the meaning for the problem at hand: For instance Precision can be interpreted as the likelihood that a person identified as a POI is actually a true POI as you correctly state in the last answer: The definition of recall is missing as well as the reasoning over the meaning of the two metrics in the context of specific problem. (Which one would you favor and why?)

SPECIFICATION

Response addresses what validation is and why it is important.

MEETS SPECIFICATION

Reviewer Comments

A generic definition of validation is given, it would be better to fit it to the specific problem at hand.

SPECIFICATION

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

DOES NOT MEET SPECIFICATION

Reviewer Comments

To meet requirements it is necessary to explicitly explain why SSSCV is used and why it is proper to do so in the context of the specific dataset.

http://scikit-learn.org/stable/modules/generated/sklearn.cross_validation.StratifiedShuffleSplit.html

SPECIFICATION

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

MEETS SPECIFICATION

Reviewer Comments

Well done!

Download project

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.