UDACITY

PROJECT REVIEW

CODE REVIEW

NOTES

SHARE YOUR ACCOMPLISHMENT!

# Requires Changes

2 SPECIFICATIONS REQUIRE CHANGES

Dear student,

well done significantly improving your submission, there is still an issue with the pros and cons of the algorithms and the possible misunderstanding regarding the choice of the algorithm to be optimized: your selection should be based on the un-optimized results, after you have chosen the best performing algorithm you should optimize that with grid search.

That being said keep up your good work, you're almost there!

## Classification vs Regression

**Student is able to correctly identify which type of prediction problem is required and provided reasonable justification.**

## Exploring the Data

**Student response addresses the most important characteristics of the dataset and uses these characteristics to inform their decision making. Important characteristics must include:**

- Number of data points
- Number of features
- Number of graduates
- Number of non-graduates
- Graduation rate

Pro Tip:
When dealing with the new data set it is good practice to assess its specific characteristics and implement the cross validation technique tailored on those very characteristics, in our case there are two main elements:

1. Our dataset is **small**.
2. Our dataset is slightly **unbalanced**. (There are more passing students than on passing students)

We could take advantage of K-fold cross validation to exploit small data sets. Even though in this case it might not be necessary, should we have to deal with heavily unbalance datasets, we could address the unbalanced nature of our data set using Stratified K-Fold and Stratified Shuffle Split Cross validation, as stratification is preserving the preserving the percentage of samples for each class.
http://scikit-learn.org/stable/modules/generated/sklearn.cross_validation.StratifiedShuffleSplit.html
http://scikit-learn.org/stable/modules/generated/sklearn.cross_validation.StratifiedKFold.html

## Preparing the Data

**Code has been executed in the iPython notebook, with proper output and no errors.**

**Training and test sets have been generated by randomly sampling the overall dataset.**

## Training and Evaluating Models

**The pros and cons of application for each model is provided with reasonable justification why each model was chosen to explore.**

There might be possible misunderstanding regarding the meaning of this question: For at least 3 algorithms you should clearly indicate at least one pro and one con for each of them, and you should indicate as well why you have chosen to try that algorithm in our context. Your section is discussing some advantages and disadvantages of some algorithms though it is not thoroughly specifically discussing at least one advantage and one disadvantage for each of the algorithms mentioned.

For instance:

1. I've not been able to find a con for the nearest neighbors.
2. When discussing support vector machine it is stated: "Since the number of features is small relative to the data size in this problem SVC is also a viable option" please note that it is one of the advantages of support vector machines to be able to handle a relatively high number of features compared to the size of the data set.
3. I have not been able to find a con for naïve Bayes.

**All the required time and F1 scores for each model and training set sizes are provided within the chart given. The performance metrics are reasonable relative to other models measured.**

## Choosing the Best Model

**Justification is provided for which model seems to be the best by comparing the computational cost and accuracy of each model.**

There is a misunderstanding regarding what is required in this section: here you should compare and contrast the results in terms of computational costs and F1 scores **before** tuning the algorithm with grid search. Once you've made your comparison and chosen your algorithm then you should apply to search on it.

**Student is able to clearly and concisely describe how the optimal model works in laymen terms to someone what is not familiar with machine learning nor has a technical background.**

**The final model chosen is correctly tuned using gridsearch with at least one parameter using at least three settings. If the model does not need any parameter tuning it is explicitly stated with reasonable justification.**

**The F1 score is provided from the tuned model and performs approximately as well or better than the default model chosen.**

### Quality of Code

Code reflects the description in the documentation.

☑ RESUBMIT

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.

RETURN TO PATH

Rate this review