UDACITY

PROJECT

Predicting Boston Housing Prices

A part of the Machine Learning Engineer Nanodegree Program

PROJECT REVIEW

CODE REVIEW

NOTES

SHARE YOUR ACCOMPLISHMENT!

Requires Changes

4 SPECIFICATIONS REQUIRE CHANGES

Hello Student,
You almost get it done, keep it UP!
There are only few minor mistakes you need to amend, I believe you already understand most of the concept in this project.

Data Exploration

All requested statistics for the Boston Housing dataset are accurately calculated. Student correctly leverages NumPy functionality to obtain these results.

Great job briefly exploring the Housing dataset and calculating it's key statistics! You also made good use of NumPy functionalities to calculate your results.

Pro Tips

- Checking your dataset statistics is an very useful routine in applying a predictive model. This is because:
  - It helps us to check if the key assumptions of our algorithms hold (thereby helping us choose which model to apply).
  - These statistics tend to be very handy when you obtain a prediction, to check whether the predictions are reasonable, and not off-chart, compared to central values of the dataset.

Student correctly justifies how each feature correlates with an increase or decrease in the target variable.

You clearly understand the relationship between each parameter with the house price. Great job!

Developing a Model

Student correctly identifies whether the hypothetical model successfully captures the variation of the target variable based on the model's R^2 score. The performance metric is correctly implemented in code.

Nice attempt here, however, we are expecting you to explain why the R^2 scores is valid.

Required:

- As required by the question, please write about whether you would consider this model to have successfully captured the variation of the target variable, and why.

Student provides a valid reason for why a dataset is split into training and testing subsets for a model. Training and testing split is correctly implemented in code.

You gave great reasons for splitting a dataset, and nice implementation using sklearn's `train_test_split`!

Suggestions and Comments:

Here's some more background or information regarding this section:

- You may have a look at this Wikipedia page about test sets, which gives excellent background about test sets and why they are used.
- Test sets in Machine Learning are created in order to test the *trained model's* ability to *generalise* beyond to predict other points which it didn't encounter in the training set.
- The Wikipedia page also mentions that test sets are used to assess the strength and utility of a predictive model.
    - Kindly note that splitting of data also serves as a check on overfitting.
    - This page from Amazon gives really ample information about why a dataset is to be split into a training and evaluation set.

Analyzing Model Performance

Student correctly identifies the trend of both the training and testing curves from the graph as more training points are added. Discussion is made as to whether additional training points would benefit the model.

Nice job in indicating adding more training points would not benefit the model. However, we are expecting a more thorough description on the trend of training / testing curve.

Required:

In this session, you only both trend became flattened as more training points increase, however, we would like to know their trend respectively. We would like to know whether their scores go upward or downward since there are more data points.

- Please describe the trend of both training and testing curves when the training sizes increase.
- Hints:
    - You might look for the curves converge point and what happened before and afterward.

Student correctly identifies whether the model at a max depth of 1 and a max depth of 10 suffer from either high bias or high variance, with justification using the complexity curves graph.

Great job identifying that the model suffers from high bias when `max_dept = 1` and that the model suffers from overfitting (high variance) when the `max_dept = 10`.

Suggestions and Comments

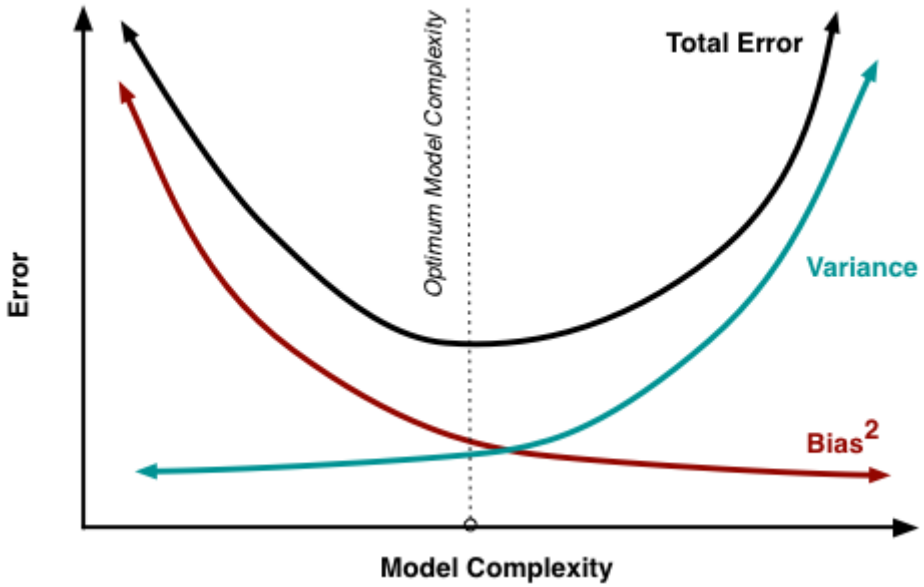
- Please check out this link if you want to understand more about model bias-variance tradeoff.

Student picks a best-guess optimal model with reasonable justification using the model complexity graph.

Great! it is a good guessing. You made this guess because of the explanation and understanding about overfitting and underfitting. It sounds reason. Let's see whether it matches the result.

Pro Tip:

Check out the following diagram, which amazingly summarises the concept of bias-variance tradeoff in Machine Learning:



Evaluating Model Performance

Student correctly describes the grid search technique and how it can be applied to a learning algorithm.

Awesome explanation of the grid search algorithm.

Pro Tips:

- Another very powerful parameter tuning algorithm is RandomizedSearchCV. In contrast with GridSearchCV, not all parameters are tried out, but rather a fixed number of parameter settings is sampled from the specified distributions.
- One particular advantage of RandomizedSearchCV is that it is much faster than GridSearchCV, and it is theoretically proven to find models that are as good; or even better than grid search.

Student correctly describes the k-fold cross-validation technique and discusses the benefits of its application when used with grid search when

optimizing a model.

Great attempt in describing k-fold cross-validation. Here are my comments regarding this section:

Required:

Your description on the benefit of `k-fold cv` is great, however we are expecting a more thorough explanation on it.

- Please explain how does it achieve the best result.
- Hints:
  - After we ran the test for `k` times and got `k` results, how do we extract the best result from it?
  - Here is a Link that explain what is k-fold cv.

Student correctly implements the `fit_model` function in code.

Awesome!! a good implement here.

Student reports the optimal model and compares this model to the one they chose earlier.

Great job here! Your max_depth matched exactly the best-guess optimal model max_depth you gave earlier.

Pro Tips

In order to have a more robust estimate of the best `max_depth` parameter, you might want to run the grid search algorithm multiple times.
Below I provide the code to help you do so:

```
max_depths = []
for i in range(500):
    reg = fit_model(X_train, y_train)
    max_depths.append(reg.get_params()['max_depth'])
best_max_depth = np.mean(max_depths)
```

```
print "The Best model, on average, has a max depth of:", best_max_depth
```

In general, if you had good intuition in picking your max_depth parameter from the complexity curves, your result from running the code above should be very close to your best-guess estimate of max_depth.

Student reports the predicted selling price for the three clients listed in the provided table. Discussion is made as to whether these prices are reasonable given the data and the earlier calculated descriptive statistics.

Well done predicting, these are valid prices for your clients' houses! However, you might want to elaborate more on your discussion, using the guidance below:

Required:

- Please compare this price you obtained to the dataset statistics, just as requested by the question
- What categories of houses do you think these features represent? For example, do you think these are rich, poor or average home owners? Do you think they represent the same categories or different categories? Please use the features of each different client to justify whether you think the prices are reasonable or not, and the category they belong to.
- Please also make a discussion for each client.

Student thoroughly discusses whether the model should or should not be used in a real-world setting.

Nice discussion here :)

Suggestions and Comments:

I would like to share my point of views of the dataset for your interest, let's see if these questions would inspire you:

- Could the Boston-House pricing model apply on other city?
- Do you think the economic condition did change a lot in these few decades, and what impact did it bring on the price level?(e.g. Boston is suffering from economic trough like Detroit and nobody would like to live here, etc.)
- Do you think the population structure did change a lot in these few decades, and what impact did it bring on the price level? (e.g. Boston population

dropped 20%, or Boston people do not like to have children, etc.)

Hope it could inspire you to have more idea on choosing the right variables, as machine learning could not perform well without choosing the right data. Choosing the right dataset is the fundamental of building a good model. And I believe you already had a great logical mind to analyze the dataset. Keep it UP :)

RESUBMIT

DOWNLOAD PROJECT

Learn the best practices for revising and resubmitting your project.

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

RETURN TO PATH

Student FAQ