UDACITY

Back to Machine Learning Engineer Nanodegree

Predicting Boston Housing Prices

REVIEW

HISTORY

Requires Changes

4 SPECIFICATIONS REQUIRE CHANGES

Your answers show the great amount of effort you have put into learning the concepts. Just a few concepts to pick up/additions to be made in your answers and you will be good to go ahead on your journey to becoming a Machine Learning Expert. All the best. Happy learning. 👍 😄

Data Exploration

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

Great work getting the summary statistics. 😄 However, the purpose of this section is to introduce you to the efficient numpy package. Please use the functions available in numpy package and rewrite the respective code lines. Currently numpy is used only for standard price.

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

Wow...You have captured the concept of correlation accurately here!
There are many types of correlation in statistics actually. Do you know what kind of correlation you have described in your answer? Check out the following link to find out:
http://www.statisticssolutions.com/correlation-pearson-kendall-spearman/

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.

Correct answer 😄
But for the 'Why/Why not?' part, we need to demonstrate our understanding of R^2 value here.
Try focusing on defining what R^2 means, how is it calculated, what is the range of values it can assume and then compare the output with the range to justify your answer.

R-squared value is the percentage of the response variable variation that is explained by a linear model:
R-squared = Explained variation / Total variation
In the target variable, there is some variation contained. The model is trained to learn pattern from the independent variables and try to explain the variation in target variable.
Total sum of squares (SST)= sum of squares explained by regression model(SSR) + sum of squared errors(SSE)
1 = (SSR/SST) + (SSE/SST)
(SSR/SST)=1-(SSE/SST)
R^2 = (SSR/SST) =1-(SSE/SST)
Please modify your answer to demonstrate this understanding.

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.

Great explanation! You captured the core behind this accurately 😄

When it comes to measuring generalization, we do it by observing the model performance on a data set which has not been seen by the ML model. When we are talking about the model evaluation on the unseen data, we would need the true values for that unseen data in order to compare the model predictions. If we train the model on the whole available data set, would we be able to check its performance on unseen data set? This is the only reason we split the data ie to be able to check for generalization of trained model.

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.

Excellent observation !! Try to search for answer to the following question for deeper understanding:
In general, what would happen to the model performance, if you increase the number of columns (not training points) in the training data?

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.

Excellent observation again 😄
Can you list down what should be done to improve the model evaluation metric if you encounter a

1. High bias model for a data set?
2. High variance model for a data set?

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

Great answer!

Evaluating Model Performance

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

Nice attempt, but kindly provide more detailed answer using the reference of Hint section which states that "When explaining the Grid Search technique, be sure to touch upon why it is used, what the 'grid' entails and what the end goal of this method is. To solidify your answer, you can also give an example of a parameter in a model that can be optimized using this approach."

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 explanation. To reiterate points of this concept:
K-fold CV is an algorithm validation technique: whether a given algorithm will train properly or not. When you get different models from different folds, what you do is average out the evaluation metric of all the models to get what? Well, to get an 'unbiased estimate of model generaliztion on unseen data'. That is the main purpose of k-fold cross validation.

Student correctly implements the `fit_model` function in code.

Awesome implementation!

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

Perfect!

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

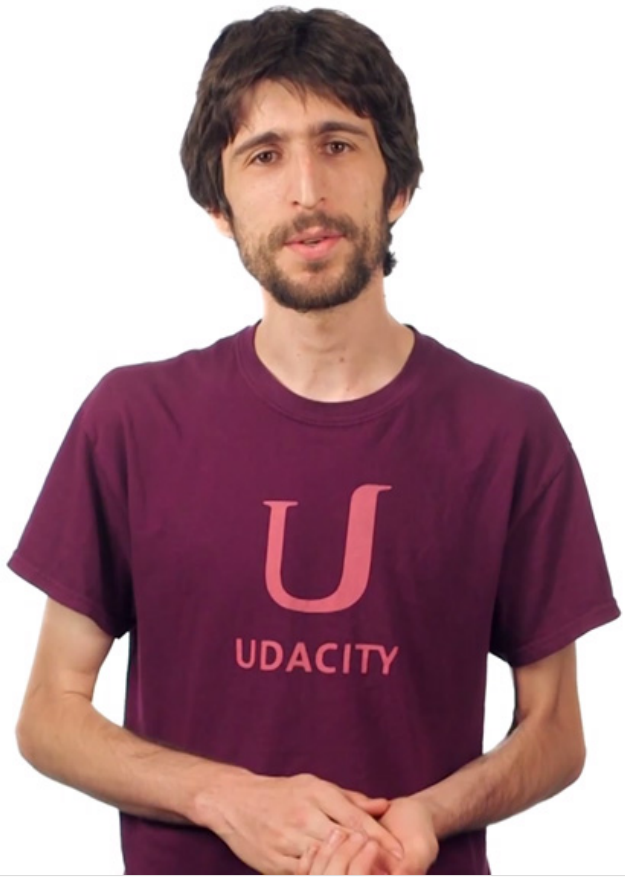
Awesome observation and great detailed answer !!
To deepen you understanding try to find answers to the following questions:
What kind of relationship do the independent variables have with the dependent variable? Decision Tree, and in general tree based models, handle a kind of relationship between the IVs and DV better than non-tree based models. Can you find which kind of relationship it is and how tree based models handle it?

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

It seems you have provided answer to only few question. Please try to think about answers to questions given in the hint section. Although they are provided as a guideline for you to frame your answer, finding their answers would help you gain a deeper theoretical understanding.

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)

RETURN TO PATH

Rate this review