UDACITY

PROJECT

Predicting Boston Housing Prices

A part of the Machine Learning Engineer Nanodegree Program

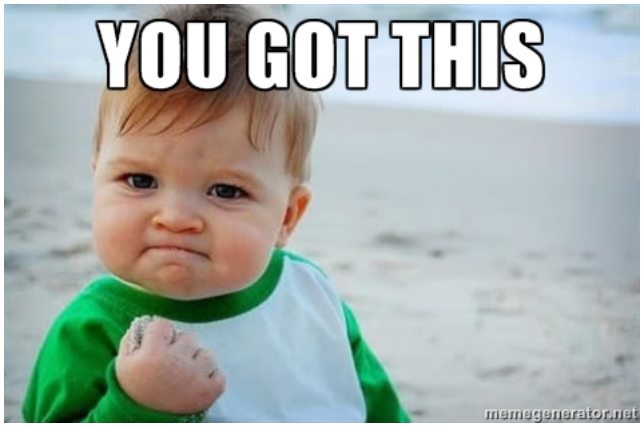
PROJECT REVIEW

CODE REVIEW

NOTES

SHARE YOUR ACCOMPLISHMENT!

Meets Specifications



Well done addressing the remaining issues! Congratulations on passing your exam! 👏👏

Data Exploration

✓

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

Awesome

Exceeds expectations with the additional calculations!

✓

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

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.

✓

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.

That's correct! We are especially interested in checking if the model is overfitting to the training set. The only way to check that is by having an independent dataset (test set) that was not used during the training process.

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.

Comment

> However, prior to this point of optimal convergence -- i.e. before the model has trained on a sufficient amount of data (left side of the graphs) -- the model tends to be very biased. In other words, it has a high error rate and low accuracy in predicting targets based on the test data. As the model trains on more data, the bias decreases and testing accuracy improves. Additionally, the model at this point also suffers from high variance as you see there is a large difference bewteen the training and testing scores.

I would argue that for very few points, the model is suffering from overfitting (high variance). It sounds confusing, but a *very biased model* in this context means high variance, not high bias. High bias (in a model) occurs when both training and testing scores are very low. That's what happens for the model with max_depth = 1 for more than 50 training points.

> In fact, adding too much data could decrease model peformance due to overfitting.

I am not sure if I agree. Although the graph seems to start diverging, usually, more points are helpful to decrease the variance of a model with high variance (overfitting). Check this class from Andrew Ng. Class about Learning Curves. In the context of this project, though, it is not worth to add more points in any of the models, since all of them seem to have converged and more points wouldn't bring a significant change.

✓

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.

✓

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

Evaluating Model Performance

✓

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

✓

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.

That's correct! Well done addressing the issue here!

✓

Student correctly implements the `fit_model` function in code.

Awesome

Great job using correctly the `range` function!

✓

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

Awesome

Very impressive arguments here:

> So, considering the graph above, it appears the 'best_estimator' favors the slightly higher R^2 score (lower bias) that max_depth=4 yields despite the slight trade-off in higher variance and model complexity. This is reasonable.

That's indeed correct! GridsearchCV tends to favor lower bias over lower variance in the tradeoff. It's algorithm seeks to optimize the validation score (without considering the distances between validation and training scores).

✓

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.

✓

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

DOWNLOAD PROJECT

RETURN TO PATH

Rate this review

Student FAQ