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

Dear student

Great job on this project! You're project now meets all specifications. You've shown that you can diagnose problems with a model using the training/testing performance and that you have a strong understanding of the relationship between model complexity and behavior. These are principles that you can use in any area of machine learning so pat yourself on the back! Congratulations on passing the project and good luck with the next section of the course!

Cheers!

Data Exploration

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

Great job using NumPy to document the dataset statistics!

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

Great analysis. Your reasoning matches what we see when each feature is plotted vs. the housing prices.

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.

> Answer: It gets a 0.923 R2_score, that is close to 1, and based on it we could say it has a good performance making the predictions.

It's true that this is certainly a positive result, but keep in mind that we only have 5 data points so far and we don't know anything about the model.

https://en.wikipedia.org/wiki/Coefficient_of_determination#Caveats

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.

> Answer: The perks of splitting the data in training and test subsets is to validate the model by testing the predictions against no previously saw data. It will show us the capacity of adaptation and to predict unseen data. If it is not done this way we could not anticipate the behavior of the model in unseen cases and it could get bad predictions for bias or variance errors.

Exactly ! We need *independent* data in order to tell if our model is generalizing well to patterns within the data or simply memorizing the training data (overfitting).

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.

> No, it would not be beneficial to have more training points, seeing the graph we could see that from 150 points and on the training score decrease and the same happen to the testing score. So at first glance, it doesn't seems to be helpful to have more training data.

Correct! We can't always fix a model by throwing more data at it. Sometimes, we need to manipulate the data or change the model complexity.

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.

Perfect answer!

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.

Great job! We need to use CV to help prevent the model from overfitting on a single split of the dataset.

Student correctly implements the `fit_model` function in code.

Nice job tuning your model!

Suggestion:
I'd recommend setting random states for your algorithms whenever possible to improve reproducibility:

```
regressor = DecisionTreeRegressor(random_state=42)
```

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

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.

Great job using the dataset statistics to bolster your answer! You've also noted that the features of the house are correlating with the predicted prices. This type of analysis can act as a crucial 'sanity check' on our models when they become large and complex.

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

DOWNLOAD 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