REVIEW

Meets Specifications

Udacity student,

your project shows how committed you are to the course. You probably spent hours or days on this project and really should be proud of your work here and I'm proud of being part of your journey!

Keep the great work on the next sections of this amazing Udacity course!!!

I share with you some extra links from Medium:

A guide to start your path in Data Science and Machine Learning:

Fundamental Python Data Science Libraries

This last one is not for only a job interview, but it contains a lot of useful information about some great topics for a machine learning professional:

Data Science and Machine Learning Interview Questions

😄

Data Exploration

✓

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

Correct! 😄

You have answered all statistics questions using the `Numpy` library. Nice job! 👍

It is important here to know why Udacity ask students to use the `NumPy` library:

NumPy is the fundamental package for scientific computing with Python. It contains among other things:

- a powerful N-dimensional array object
- sophisticated (broadcasting) functions
- tools for integrating C/C++ and Fortran code
- useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined. This allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

Numpy Documentation

It's always important to be aware of tools you use. For example, the Pandas' `Series.std()` will by default give you different result than `numpy.std()`. It's because **Numpy** takes in count the whole population while **pandas** assumes that you are evaluating the standard deviation for a sample of your dataset.

This article has a very good explanation about it.

✓

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

Correct! 😄

You have correctly justified how each feature correlates with an increase or decrease in the target variable.

Awesome job plotting you data the get some insights!

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.

Great job! 😄

Since R2 Score is close to one we may say it is a good model.

The R-Squared is very common score used in Data Science / Machine Learning. But we can have another type of R-squared called Adjusted R-Squared.

Also, in this article from Duke University you can find a very nice opinion about how good R2 Score can be in a ML mode.

✓

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.

Correct! 😄

You have implemented a `random_state` for the `train_test_split` function properly.

In this article about "what’s the purpose of splitting data up into test sets and training sets?" you may go deeper on your studies.

Splitting the dataset into a reasonable ration, we can generalize the model and avoid to perform worse when an unseen data is inputted.

Great job!

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.

Perfect! 😄

As more data we add the training score stays constant (small decrease) and the testing score increases. Nice job about the overfitting in this learning curve!

Moreover, adding even more points will not benefit the model and only "make the computer work harder" in terms of processing the data. I recommend this reading about this topic: "How much data is enough?".

In a short video you may go deeper in your studies about learning curves.

✓

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.

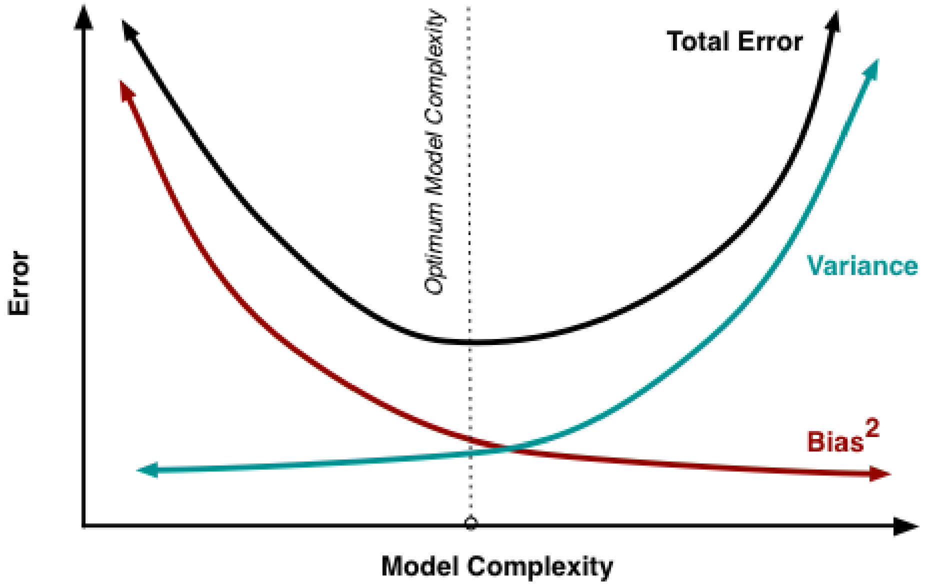
Correct! 😄

Good job identifying high bias and high variance for different `max_depth` parameters.

In this link you can read more about **high bias** and **high variance** of data and boost your understanding about this topic!

Also in Wikipedia you may find a nice article about the tradeoff regarding to a model with high bias and high variance.

Recently I've been studying this subject and this website brought me a lot of light about this issue. Note in the picture below that we can have an optimal point and have a model without the overfitting or underfitting problem.



✓

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

Great job! 😄

Nice guess! Let's check what your code will tell us. 😏

Evaluating Model Performance

✓

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

Correct! 😄

You show really a very good understanding about the Grid Search.

I'd like just to point out that Grid Search will test EVERY combination of hyperparameters. So it can be very computationally expensive if you want a model with many hyperparameters or many sets of them.

Usually I have some tips and links about Grid Search and I'd like to share with you. It can be good and complete your studies and I hope so!

1. Official sklearn page on gridSearch
2. How gridSearch works
3. Specifying multiple metrics for evaluation
4. The scoring parameter: defining model evaluation rules
5. Defining your scoring strategy from metric functions

✓

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.

Correct! 😄

You summarized well the K-fold concepts. I really liked your explanation and it shows that you understood this important machine learning tool.

I'd like to share some extra readings about K-fold:

- What is Cross Validation?
- K-fold and Cross Validation

✓

Student correctly implements the `fit_model` function in code.

Correct! 😄

Very nice implementation. 👍

✓

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

Correct! 😄

You have done an awesome job coding your project and explained how your answer compares with question 6.

✓

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.

Correct! 😄

The selling prices you evaluated are correct and your discussion shows your very good understanding about this project. Also, it shows that you can extract insights from data using your intuition and abilities.

Good job!

✓

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

Correct! 😄

Nice answers regarding to all questions in this last section. I highlight your answer about the quality of the features:

> To get a rough idea only, Yes. But to get better predictions, we need more features to train our model on. So, final ans as No only.

I agree with you. We should have more details or features about the houses, city and other important things to get better predictions.

Nice job! 👏