Requires Changes

6 SPECIFICATIONS REQUIRE CHANGES

This is a very impressive submission. Just need some minor adjustments and you will be golden, but also check out some of the other ideas presented in this review. One tip here would be that some of these topics are extremely important as you embark on your journey throughout your Machine Learning career and it will be well worth your time to get a great grasp on these topics before you dive deeper in. Keep up the hard work!!

Data Exploration

All requested statistics for the Boston Housing dataset are accurately calculated. Student correctly leverages NumPy functionality to obtain these results.
Good job utilizing the power of Numpy!! Always important to get a basic understanding of our dataset before diving in. As we now know that a "dumb" classifier that only predicts the mean would predict $454,342.94 for all houses.

Student correctly justifies how each feature correlates with an increase or decrease in the target variable.
Although your ideas are correct, please provide some justification for why these would be true. For example why would you "*expect an increase of the value of the property*"? etc... A good idea would be to reference the feature descriptions here

- 'RM' is the average number of rooms among homes in the neighborhood.
- 'LSTAT' is the percentage of homeowners in the neighborhood considered "lower class" (working poor).
- 'PTRATIO' is the ratio of students to teachers in primary and secondary schools in the neighborhood.

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.

> "Depends of the context of the problem. Comparing the median absolute error with the standard deviation of the target variable, it gives a 20% rate of a standard deviation which in some domains it is not tolerated."

I agree that it does depends of the context of the problem, but this is just a hypothetical model. Thus, solely based on the R2 score of 0.923 would you consider this model to have successfully captured the variation of the target variable?Why or why not? How does this compare to the optimal score? How do the 'true values' and 'predictions' compare? etc...

(https://en.wikipedia.org/wiki/Coefficient_of_determination)

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 reasons! We need a way to determine how well our model is doing! As we can get a good estimate of our generalization accuracy on this testing dataset. Since our main goal is to accurately predict on new unseen data. And correct that we can try and protect against overfitting with this independent dataset.

If you would like to learn some more ideas in why we need to split our data and what to avoid, such as data leakage, check out these lectures

- https://classroom.udacity.com/courses/ud730/lessons/6370362152/concepts/63798118300923
- https://classroom.udacity.com/courses/ud730/lessons/6370362152/concepts/63798118310923

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.
You are correct with your comment of

> "If the curves converge, there is no need of adding more training points."

As in the beginning it is beneficial, but at the end if we look at the testing curve here, we can clearly see that it has converged to its optimal score, so more data is not necessary.

Therefore lastly for this section make sure you also describe the initial trends of the training and testing curves for your chosen max depth of 3, before they converge(increasing or decreasing?). As this is key in terms of how models initially scale to more 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.
You are correct that a max_depth of 1 suffers from high bias and a max_depth of 10 suffers from high variance. And good visual justification for high variance(large gap between the training and validation scores). But please also provide some specific visual justification for a max depth of 1 and high bias. Therefore make sure you mention what the training and validation scores are. High? Low? etc... The reason for this is so you can quickly identify these instances in the future.

Links

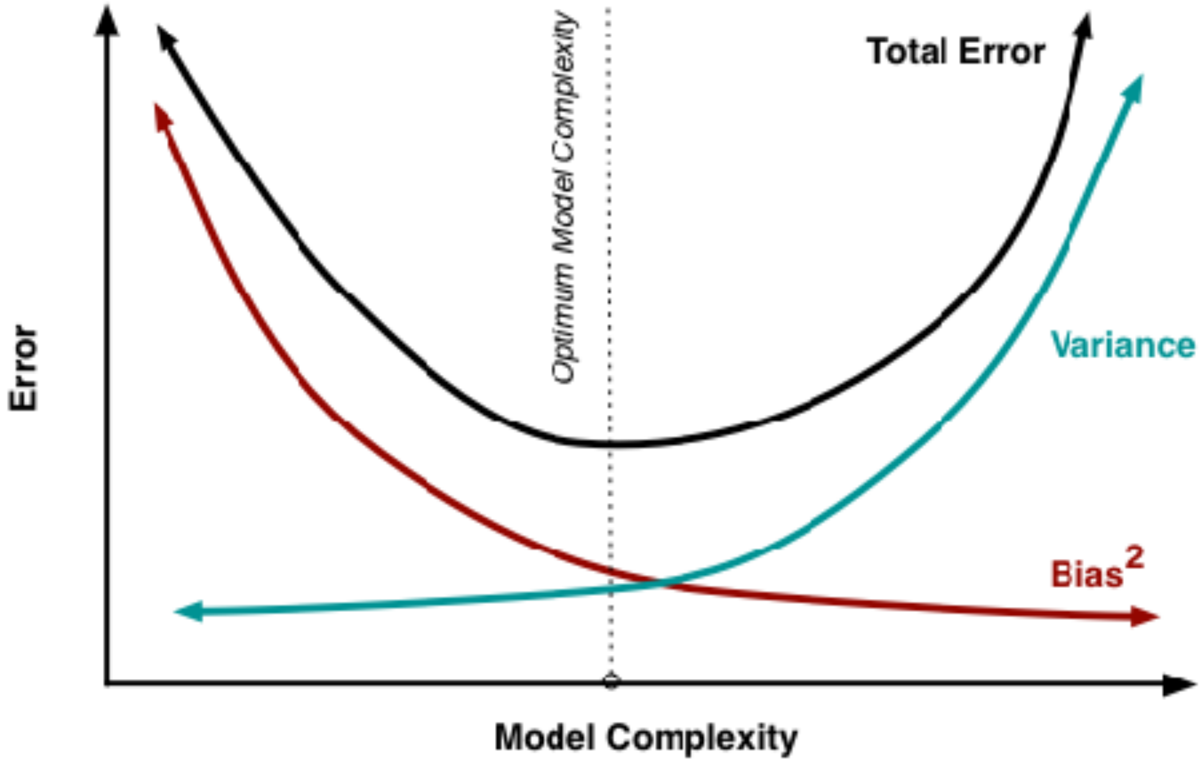
- http://scott.fortmann-roe.com/docs/BiasVariance.html
- https://insidebigdata.com/2014/10/22/ask-data-scientist-bias-vs-variance-tradeoff/
- http://machinelearningmastery.com/gentle-introduction-to-the-bias-variance-trade-off-in-machine-learning/

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

> "It is the depth with best score on the validation."

Exactly! As we are definitely looking for the highest validation score(which is what gridSearch searches for). And we are also looking for a good bias / variance tradeoff(with close training and validation scores while the training score is high).

Check out this visual, it refers to error, but same can be applied to accuracy(just flipped)



Evaluating Model Performance

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

"compute the F1 score. The cell with the best F1 score would be the best ML Model for the problem."

First off, just want to make you aware, that we are not restricted to using F1 score as our evaluation metric when optimizing a model with gridSearch, as we can use any evaluation metric we please. As you can see that we are using max depth, a decision tree and **r squared** score in this project. Here might be a list of common evaluation metrics.

Also please mention which hyper-parameter value combinations does it test (i.e. a random sample of them, every other combination, all of the exhaustively)? As there are many different search types. What does `GridSearchCV` use?

Links

- (http://scikit-learn.org/stable/modules/grid_search.html)
- (https://en.wikipedia.org/wiki/Hyperparameter_optimization#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 description of the k-fold cross-validation technique, probably the most use CV method in practice.

"The benefit is that does not waste data to be used to validate the result and avoid the use of the test dataset for overfitting."

This is an extremely important concept in machine learning, as this allows for multiple **validation**datasets and is not just reliant on the particular subset of partitioned data. For example, if we use single validation set and perform grid search then it is the chance that we just select the best parameters for that specific validation set. But using k-fold we perform grid search on various validation set so we select best parameter for generalize case. Thus cross-validation better estimates the volatility by giving you the average error rate and will better represent generalization error.

Student correctly implements the `fit_model` function in code.

Nice implementation! Good idea to set a `random_state` in your DecisionTreeRegressor for reproducible results. This is a great habit to get into!

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

Congrats! Can note that GridSearch searches for the highest validation score on the different data splits in this ShuffleSplit.

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.

" Using the logic and the intuition of the question 1, the values look reasonable and inside of the range of the data."

Can you please give some specific justification for why they are "reasonable"? As a good idea would be to go client by client and give an idea for each feature of `RM`, `LSTAT`, and `PTRATIO` and descriptive stats of the housing prices.

Optional: A more advanced and great idea would be to compare these to the descriptive stats of the features. We can compute the five number summary of the descriptive stats of the features with

```
features.describe()
```

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

Would agree. This dataset is quite old, probably doesn't capture enough about housing features, and the range in predictions are quite large. Therefore this model would not be considered robust!

If you would like to learn more about how to analyze the uncertainty in the output of a mathematical model, can check out these ideas (https://en.wikipedia.org/wiki/Sensitivity_analysis)