Predicting Boston Housing Prices

1. Statistical Analysis and Data Exploration

- Number of data points (houses)?

  > 506

- Number of features?

  > 13

- Minimum and maximum housing prices?

  > 5.0 - 50.0 (value in 000's)

- Mean and median Boston housing prices?

  > mean 22.533 and median 21.2

- Standard deviation?

  > 9.188

2. Evaluating Model Performance

- Which measure of model performance is best to use for predicting Boston housing data and analyzing the errors? Why do you think this measurement most appropriate? Why might the other measurements not be appropriate here?

  > Mean squared error (MSE) is the most appropriate measure for regression errors in predicting Boston housing prices. Absolute mean error could also work but puts less weight on the existence of large errors compared to MSE. Median absolute error would place less weight on outliers.

- Why is it important to split the Boston housing data into training and testing data? What happens if you do not do this?

  > Splitting the data is important in order to prevent overfitting of the model to the data used to train the model. If the data is not split then the model will be overfitted with low bias and high variance, resulting in poorer performance when used to predict on novel data.

- What does grid search do and why might you want to use it?

  > Grid search helps to tune the learning model by working through multiple combinations of parameter tunes and cross-validating to find the tune that gives the best performance. Grid search will help to find the best model given the limited amount of data in the dataset.

- Why is cross validation useful and why might we use it with grid search?

> CV is useful to tune the learning model to minimize both training and test error. It's used with grid search in order to find the best parameter tunes that result in the best learning model.

3. Analyzing Model Performance

- Look at all learning curve graphs provided. What is the general trend of training and testing error as training size increases?

  > Training error increases and testing error decreases as training size increases. Training error rises slowly and steadily whereas testing error decreases sharply at first and then fluctuates while still decreasing slightly.

- Look at the learning curves for the decision tree regressor with max depth 1 and 10 (first and last learning curve graphs). When the model is fully trained does it suffer from either high bias/underfitting or high variance/overfitting?

  > The low max depth model is likely underfitting the data and suffering from high bias, resulting in both higher training and test error compared to the high max depth model. The high max depth is likely overfitting given the greater discrepancy between training and test error.

- Look at the model complexity graph. How do the training and test error relate to increasing model complexity? Based on this relationship, which model (max depth) best generalizes the dataset and why?

  > Increasing model complexity reduces training error toward zero, but test error fluctuates in a range around 30-40. A model with depth 5 best generalizes the dataset because it generates test error at (or near to) the low point in the model complexity graph, and doesn't overfit the data.

4. Model Prediction

- Model makes predicted housing price with detailed model parameters (max depth) reported using grid search. Note due to the small randomization of the code it is recommended to run the program several times to identify the most common/reasonable price/model complexity.

- Compare prediction to earlier statistics and make a case if you think it is a valid model.

  > The model generally gives a prediction in the 19.0 - 22.0 range, which is within the dataset's range of prices and wihtin 1 sd of the mean price. The model appears to be valid.