1) Statistical Analysis and Data Exploration

Number of data points?

- 506 houses

Number of features?

- 6578/506 = 13 features per house

Minimum and maximum housing prices?

- Min = 5
- Max = 50

Mean and median Boston housing prices?

- Mean = 22.53
- 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?

- We are dealing with continuous data, so we will need a regression metric.
- We want to care about large errors more than small ones. Under selling a house will make clients upset, but over selling a house may mean it doesn't get sold. We want values close to the mean and median.
- I choose mean squared error for the above reasons.

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

- It's important to split the data so that the algorithm has two sets of independent data to work with. The first trains the algorithm and the second tests the algorithm.
- If we didn't do this, we wouldn't know if the algorithm overfit the data, which may lead to inaccurate future house prices.

Which cross validation technique do you think is most appropriate and why?

- I choose a 10 K-Fold cross validation technique because of the lack of data. This will allow all data to be used for training and testing purposes.
- “Cross validation is an iterative process where train/test sets are randomly generated multiples times in order to evaluate the algorithm at each split, the results are then averaged over the splits.” - Udacity Review Feedback

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

- “GridCV is a way of systematically working through multiple combinations of parameter tunes, cross-validating as it goes to determine which tune gives the best performance. The beauty is that it can work through many combinations in only a couple extra lines of code.“ - Udacity https://www.udacity.com/course/viewer#!/c-ud9013-nd/l-5406799334/m-3056108547
- We want to use this because we don't know which depth will be the best to use, and this can find a good estimate for us.

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?

- When the max depth is 1, both training and testing error asymptotes to an error of about 50.
- When the max depth is 10, the training error asymptotes to about 0 error while the testing error asymptotes to about 15-20 error.

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?

- When the max depth is 1, the model suffers from high bias/underfitting. This can be seen by the large training error (about 50), but a small difference from the testing error (about 50).
- When the max depth is 10. the model suffers from high variance/overfitting. This can be seen by the small training error (about 0), but a moderate difference from the testing error (about 15-20).

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?

- The test error gets better until around 4-6 and then starts to get slightly worse.
- I would choose 4 since we don't want to overfit the data and choosing a lower value is more likely to avoid this.

4) Model Prediction

Model makes predicted housing price with detailed model parameters. 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.

- After running the program several times, I got that the predicted house price is about 21.63
- The max depth was 4 in this case.

Compare prediction to earlier statistics

- The predicted house price seems reasonable since it is close to the Mean (22.53) and the Median (21.2).