1 Statistical Analysis and Data Exploration

- Number of data points: 506
- Number of features: 13
- Max and min housing prices: 5, 50
- Mean and median prices, respectively: 22.53, 21.20
- Standard deviation: 9.19

2 Evaluating Model Performance

Citation #5

- Since the Boston housing data is a regression, we can narrow down the model performance into two error measuring functions: mean squared error and mean absolute error. Mean squared error is used to magnify the error of points that are farther away from the regression line. I'll use the mean squared error metric for that reason. However, mean absolute error would also be appropriate.

Citation #4

- The Boston housing data needs to be split into training data, to learn from the data, and testing data to test its model on data that it has never seen before. There is no way of knowing if the model is predicting correctly if there's no testing data to check the accuracy of the model.

Citation #7

- Grid search is necessary to test a combination (in this case, a range) of the depths of the decision tree. Using the parameters, we can find the optimal depth of the decision tree to use for this model.

Citation #4

- Cross validation is useful for checking the validity of the model when the dataset is potentially limited. Using cross validation, as opposed to using simple validation, uses the average of combinations and assigns the testing set to each bin; this ensures that all the data is used for both testing and training. Grid search, combined with cross validation, gives a more accurate result for each parameter combination when finding the optimum.

3 Analyzing Model Performance

- We see a trend as the model is given more data. At each max depth graph, the training data slightly increases. At each max depth, the testing data decreases when given more data.

Citation #9

- For max depth of 1, the model runs into the problem of underfitting. After increasing the training size past 100 the test error no longer decreases. It doesn't matter how much data we feed it past a certain point. The test error approaches ~20%, and the general slope of the testing error is nearing zero. The model runs into overfitting when the gap between the training and testing error increases.
- Both the training and testing error decrease as max depth approaches 5. The training error continues to decrease as the model learns from data it has already seen. The testing error reaches a trough at the max depth of ~5. Increasing model complexity past a max depth of 5 doesn't improve the model. This means that the variance has gone up slightly. Increasing the max depth of the decision tree won't make the model too complex.
- In the model complexity graph, the model that best generalizes the dataset uses the parameter with the decision tree max depth of 6. At this point, the model doesn't improve the its accuracy with the testing data.

4 **Model Prediction**

- The optimal max depth and housing price reported by grid search best estimator is 4 and 21.630 (in thousands).
- The prediction aligns very close to the mean and median price of houses in Boston. This house price is right between the mean and median house prices. This is reasonable: given that the standard deviation of this dataset is 9.190, which gives this prediction a z-score of -0.098. When finding the ten nearest neighbors and computing their average, we see that it returns 21.52.

Citations

Citation #1:
- https://scikit-learn.org/

Citation #2:
- http://docs.scipy.org/doc/numpy-1.10.0/index.html#

Citation #2:
- https://www.udacity.com/course/viewer#!/c-ud725-nd/l-5454078888/m-3515108966

Citation #3:
- https://www.udacity.com/course/viewer#!/c-ud725-nd/l-5454838999/e-3019448561/m-2951438859

Citation #4:
- https://www.udacity.com/course/viewer#!/c-ud725-nd/l-5453289163/m-5500579243
- http://scikit-learn.org/stable/modules/cross_validation.html
- http://www.anc.ed.ac.uk/rbf/intro/node16.html

Citation #5:
- http://scikit-learn.org/stable/auto_examples/model_selection/grid_search_digits.html
- https://en.wikipedia.org/wiki/Hyperparameter_optimization
- http://stats.stackexchange.com/questions/886/what-is-the-fundamental-idea-of-machine-learning-for-estimating-parameters
- https://www.youtube.com/watch?v=GoI_qOgRqfA

Citation #6:
- https://discussions.udacity.com/t/model-complexity/38112/2
- http://scikit-learn.org/stable/modules/generated/sklearn.metrics.make_scorer.html

Citation #7:
- http://scikit-learn.org/stable/modules/grid_search.html
- https://www.udacity.com/course/viewer#!/c-ud725-nd/l-5406799334/m-3056108547
- https://discussions.udacity.com/t/differences-in-model-predictions-with-gridsearchcv-parameters/38350/3

Citation #8:
- https://discussions.udacity.com/t/gridsearch-guidance-needed/40356/6

Citation #9:
- http://scott.fortmann-roe.com/docs/BiasVariance.html
- https://www.udacity.com/course/viewer#!/c-ud725-nd/l-5450289429/m-5479332868
- https://www.udacity.com/course/viewer#!/c-ud725-nd/l-5450289429/m-5465878606