1) Statistical Analysis and Data Exploration

- Number of data points (houses)?

  506
- Number of features?

  13
- Minimum and maximum housing prices?

  Minimum is 5.0. Maximum is 50.0
- Mean and median Boston housing prices?

  Mean is 22.5328063241. Median is 21.2.
- Standard deviation?

  Standard deviation is 9.18801154528.

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. Since the data is continuous data, It is appropriate to measure regression performance, not classification performance. There is a similar metric called Mean Absolute Error, but MSE is more appropriate since it can emphasize larger errors than smaller errors. Larger errors are evil. It means massive loss in the real-estate agent scenario.
- Why is it important to split the Boston housing data into training and testing data? What happens if you do not do this?

  If I don't split data, I need to use same data both for fitting model, and for predicting outputs. It will probably fit well on the specific dataset, but will fail to generalize to other datasets.
- What does grid search do and why might you want to use it?

  Grid search try all combinations of parameters given, and look for a best combination which help to improve model performance.
- Why is cross validation useful and why might we use it with grid search?

  There is a trade-off relationship on splitting training data and test data. More and more training data we prepare test data is reduced, and vice versa.

  Cross validation is useful because it run experiments(split data, train model on training data and validate the model on test data) multiple times. It generates multiple results, so we can use average of the results(values), which can reduce randomness on the values.

Using cross validation with grid search, we can automate parameter tuning and choose best combination of parameters without preparing additional data points.

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 increase gradually, and converge at one point. Testing error decrease sharply, and converge at one point.

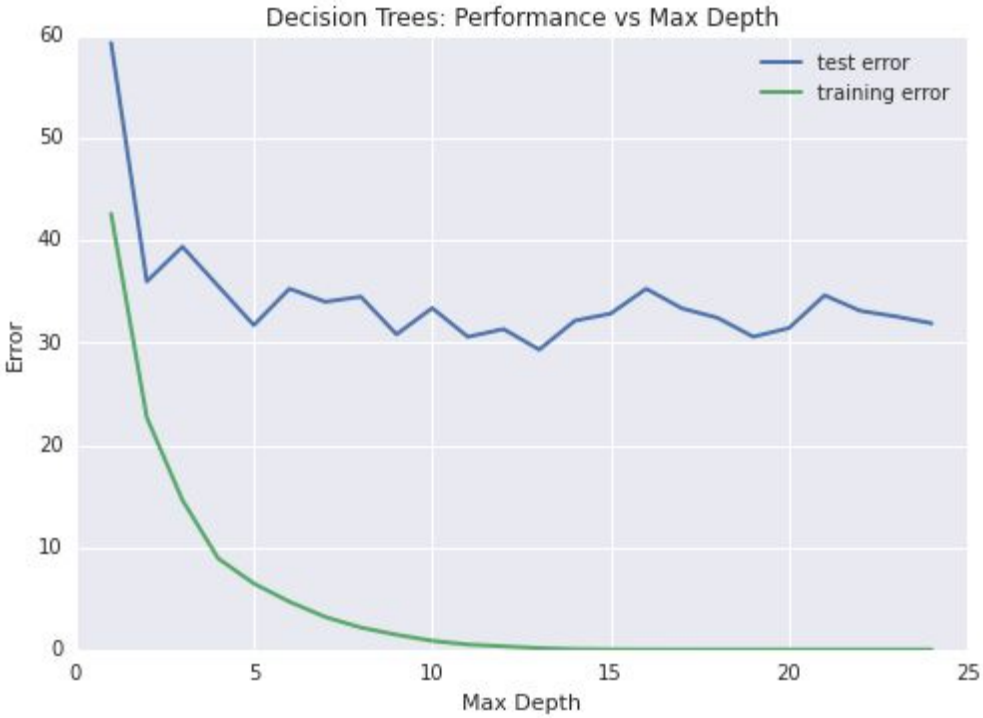
- Look at the learning curves for the decision tree regressor with max depth 1 and 10 (first and last learning curve graphs). When the model is fully trained does it suffer from either high bias/underfitting or high variance/overfitting?

  Regressor with max depth 1 suffer from high bias/underfitting. The model lacks complexity. It`s too simple to represent underlying relationships, so the increase of training size didn't help improve the model performance. As for regressor depth 10, it suffer from high variance/overfitting. Even though the test error decrease as training size increased, the test error is much higher than test error regardless of the training size. The model failed to generalize the dataset.

Graph 1: Performance of regressor with depth 1.

Graph2: Performance of regressor with depth 10.

- 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? **Increase of the model complexity reduce training error, while it stop reducing test error around max depth between 5 and 10. Model with max depth around 5 best generalizes the dataset. Even though training error decrease as the model complexity increase more, test error doesn't. It means the models with max depth > 5 failed to generalize the dataset well. In another word, it overfit the data.**



Graph 3: Performance of regressors differ with max depth.

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.

  Run the (grid search) program several times, and got several predicted prices and model parameters. Median of the prediction is 21.63. The most common model complexity are 4 and 6 (max depth).
- Compare prediction to earlier statistics and make a case if you think it is a valid model.

  I think it's a valid model, but I can't say it's a valid model by just comparing earlier statistics and the prediction. The actual price mean is 22.53, the max value is 50, the min value is 5, and the standard deviation is 9.18. From the statistics, I can say the prediction, 21.63, is not far from the actual prices. May be the model fit well on the dataset.