1. Statistical Analysis and Data Exploration

- Number of data points for price – 506
- Number of data points for features – 6578
- Number of features for housing price – 1
- Number of feature for housing features – 2
- Minimum housing price – 5
- Maximum housing price – 50
- Mean Boston housing price – 22.53
- Median Boston housing price – 21.2
- Standard deviation for housing price – 9.188
- Standard deviation for housing features – 145.156

2. Evaluating Model Performance

2.1.1. Which measure of model performance is best to use for regression and predicting Boston housing data?

The best measure of model performance is mean squared error

2.1.2. Why is this measurement most appropriate?

This measure is most appropriate because it becomes optimal faster

2.1.3. Why might the other measurements not be appropriate here?

The other measurements might not be appropriate here because we are looking for how far off the model's prediction is from the real true value.

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

It is important to split date into training and testing data because:

1) Gives estimate performance on an unseen data
2) Serves as check on overfitting

If the training and test datasets are not partition we run into issues evaluating a model because it has already seen all the data.

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

I think that the most appropriate cross validation technique is K-Fold Validation because we can use all the data both for training and for testing. However, if we need to minimize training time than we have to use Train\Test validation instead.

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

Grid search determines which tune gives the best performance. We might want to use it because it can work through many combinations by writing several lines of code.

3. Analyzing Model Performance

3.1. What is the general trend of training and testing error as training size increases?

The general trend of training error as training size increases is going down to zero

The general trend of testing error as training size increases is slightly going down and then stay at some value.

3.2. When the model is fully trained does it suffer from either high bias/underfitting or high variance/overfitting?

Max-depth = 1

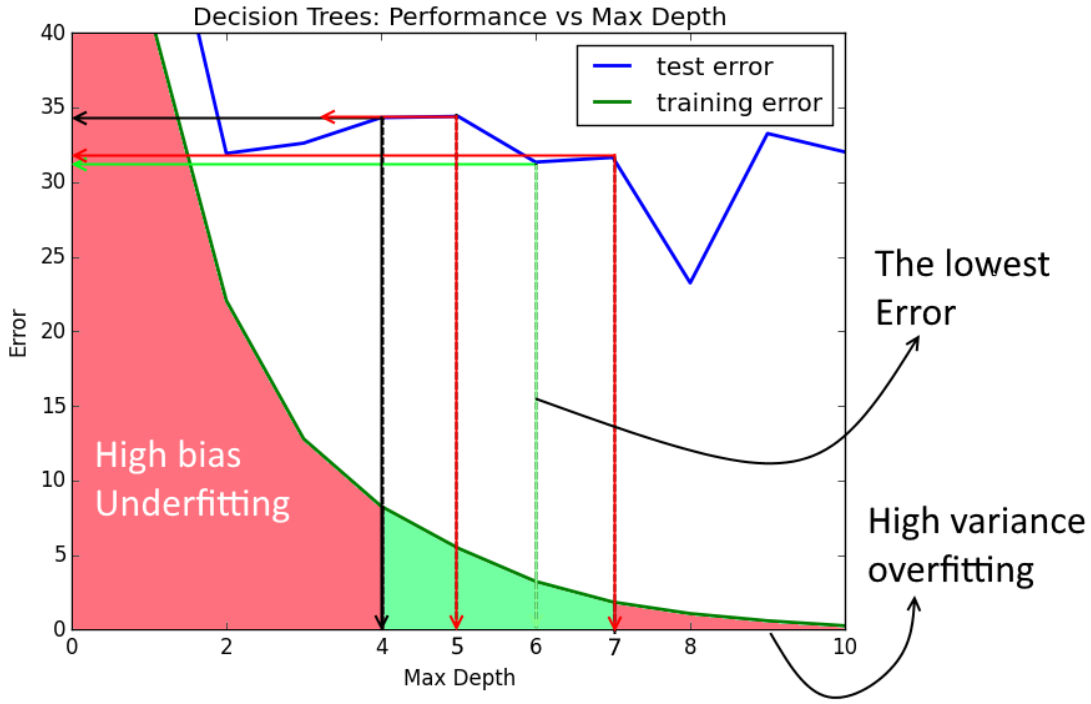
Supposed to the picture (Max-depth = 1) where both test and training errors are high, the model suffers from high bias/underfitting.

Max depth = 10

Supposed to the picture (Max-depth = 10) where test error is high but training error is very low, the model suffers from high variance/overfitting

3.3. 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?

If we choose more complex model, then we will get the less training error. As it's shown on the picture below (Model complexity), at the start we have high bias underfitting because the model isn't trained. However, while training, the model becomes more complex and its training error goes to 0 which leads to high variance overfitting.



Model Complexity

Supposed to picture (Model complexity) the 6th model best generalizes the dataset because it has the lowest generalization error on distance between high bias and high variance.

4. Model Prediction

The house with following parameters:

[11.95, 0.00, 18.100, 0, 0.6590, 5.6090, 90.00, 1.385, 24, 680.0, 20.20, 332.09, 12.13]

Predicted 21.62 price

Predicted price is close to mean and median.

$Mean = \overline{x} = 22.53$
$Standard\ deviation = \sigma = 9.188$
$Predicted\ price = x = 21.62$
$\overline{x} - \sigma < x < \overline{x}$
$13.34 < 21.62 < 22.53$

This result fit to our system.