1) Statistical Analysis and Data Exploration

- Number of data points (houses)?
- Number of features?
- Minimum and maximum housing prices?
- Mean and median Boston housing prices?
- Standard deviation?

- ✓ Number of houses: 506
- ✓ Number of features: 13
- ✓ Minimum House Price: 5.0
- ✓ Maximum House Price: 50.0
- ✓ Mean House Price: 22.532806324110677
- ✓ Median House Price: 21.2
- ✓ Standard Deviation of House Price: 9.188011545278203

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?

- ✓ Since this is a regression model, it is better to use one of the regression metrics such as mean square or absolute mean etc. Mean square error works better in these scenarios since it is second moment of the error and does cover the variance of the estimator as well as bias. In this regard, MSE can be expressed as the sum of variance and squared bias of the estimator and minimizing MSE will provide the optimum solution. Thus, I have used mean square error as the metric. The metrics for classifications are not appropriate since they calculate the error term by looking at the estimator classes which does not apply in this case.

The MSE can be written as the sum of the variance of the estimator and the squared bias of the estimator

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

- ✓ We need training data to train our model, if we use all our data for training and use the same data for testing, we will observe lower error rate than actual due to training bias

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

- ✓ Grid search uses the parameters we provide and evaluate the performance of the estimator using default k-fold cross validation. It is an easier to find the optimum estimator parameter

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

- ✓ Cross validation is important to evaluate the performance of estimator, since it averages over same data by changing the training and test sets and this reduces the bias of the expected error rate

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?

- ✓ It is obvious that test errors get smaller as we increase the size of the dataset for the estimator, while the training errors increases with the increasing number of sample

- 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?

- ✓ For the first learning curve graph, the model suffers from underfitting because of the size of the regression tree limited at 1. We see that both training and test errors are quite high in this case which indicates underfitting of the model. On the other hand, model overfit when we set the max-depth to 10. We see that the training error rate is low while test error rate shows variation. This behavior indicates that the model is overfit to training data and performance of test data shows a lot of variation over the graph. In other words, model width max_depth=10 uses some of the random errors in the data instead of underlying relation.

- 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?

- ✓ As we increase the depth, model complexity increases since we are adding more branches to our tree regressor. Up to certain depth value, we observe that the increasing model complexity does help to reduce the MSE; however as we increase the complexity after some point we do not observe any improvement and finally hit the overfitting problem. We observe that the best value is observed at depth=6, we see this after running the model with random data several times to minimize the training set bias.

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.

- ✓ This is reasonable model based on the first level statistics. The predicted result is 21.62974359 which is close to mean and median value of training data set. It is within one sigma of the mean value.