Report for Project 1 – Predicting Boston Housing Prices
by Mingming Guo

1) Statistical Analysis and Data Exploration

- Number of data points (houses)? Answer: 506
- Number of features? Answer: 13
- Minimum and maximum housing prices? Answer: 5.0, 50.0
- Mean and median Boston housing prices? Answer: 22.53, 21.2
- Standard deviation? Answer: 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? Answer: mean squared error (MSE) is best to use for this regression problem. Mean squared error is the average value of the squares of the errors (which is the difference between the predicted response and the real response). As an error metric, it measures the absolute fit of the model to the data. It incorporates both the variance of the estimator and its bias. By minimizing the mean squared error, we are minimizing the bias and the variance of the estimator. If the estimator is unbiased, mean squared error is just the variance of the estimator, and its root will be standard deviation. Other measurements like R-squared score is just a relative measure of fit because it mainly explains the proportion of variance in the response values that can be explained by using models. The mean absolute error (MAE) gives the expected value of the absolute error loss which will also minimize the bias and variance of the estimator. However, by squaring the errors, MSE gives more weights to large errors than small errors because large errors have even larger square values than small ones, and squaring doubles the different between large error and small error. This will make the data points with large errors more important in the learning model. Assuming we removed the outliers of the dataset, it is better to use MSE than MAE when we have many far points. The explained variance score is also a relative measure of fit by examining the explained variance versus overall variance, etc.
- Why is it important to split the Boston housing data into training and testing data? What happens if you do not do this? Answer: by splitting data into training and testing data, we can use training data to generalize the model, and use testing data to check if the generalization of the model is good or bad. If we do not do this, we will not know if the model we build is a good model or a bad model for unseen data.
- What does grid search do and why might you want to use it? Answer: grid search is a way to run the combination of different parameters, and then choose the best combination. This is an automatic way to choose the optimal parameters, and free us to do it manually.

- Why is cross validation useful and why might we use it with grid search? Answer: cross validation is useful because it can fully utilize the dataset for training and testing by dividing the dataset into pieces and iteratively choose different pieces of data for training and testing. We normally use it with gird search because parameters that cannot directly learnt within estimators can be set by searching a parameter space for the best cross-validation score.

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? Answer: the general trend of training and testing error as training size increases is that the training error increases, testing error decreases, and they show the trend to converge.
- 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? Answer: the model is fully trained when the training size is maximum. For the model with max depth 1, it suffer from high bias or underfitting because the training error and testing error are close to each other and both are high as the training size increases. For the model with max depth 10, it suffer from high variance or overfitting because the training error is really low and increasing very little as the training size increase, the testing error is high and there is a gap between testing error and training error.
- 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? Answer: as the model complexity increases, the training error continuously decreases while the testing error firstly decreases and then stop to decrease at some level and/or increase later. Base on this relationship, the model with max depth 5 best generalizes the dataset, because after max depth 5, the testing error stops to decrease and/or increase later as the max depth increases.

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. Answer: after running the program several times, the most reasonable price is 20.77 with the model complexity max depth 6.
- Compare prediction to earlier statistics and make a case if you think it is a valid model. Answer: the statistics shows that the mean is 22.53 with standard deviation 9.188. The result predicted price is 20.77 which falls into the 95% interval which is (4.154, 40.906). The prediction price is reasonable and I believe it is a valid model.