Name: Yongchao Tang
Email: yongchaotang@gmail.com

Questions and Report Structure

1) Statistical Analysis and Data Exploration

Number of data points (houses)?

Answer: Number of houses is: 506

Number of features?

Answer: Number of features is: 13

Minimum and maximum housing prices?

Answer: Minimum price is: 5.0
Maximum price is: 50.0

Mean and median Boston housing prices?

Answer: Mean price is: 22.5328063241
Median price is: 21.2

Standard deviation?

Answer: Stand deviation is: 9.19710408738

2) Evaluating Model Performance

Which measure of model performance is best to use for predicting Boston housing data and analyzing the errors?

Answer: Mean squared error

Why do you think this measurement most appropriate?

Answer: It automatically converts all the errors as positives, emphasizes larger errors rather than smaller errors. For houses prices, larger errors should be paid more attention to than small errors because large errors may result in big loss of money. And it is differentiable which allows us to find the minimum error.

Why might the other measurements not be appropriate here?

Answer: Other measurements such as mean absolute error does not reflect the larger errors as the mean squared error does.

Why is it important to split the Boston housing data into training and testing data?

Answer: By splitting the dataset into two parts, we have an independent set of data to verify that the model can generalize well rather than just to the training example. It also serves as check on overfitting.

What happens if you do not do this?

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

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

Answer: The grid search can systematically work through multiple combinations of parameter tunes and find the parameters that generate the best training performance. We use the grid search because we often need to optimize the parameters that are not directly learnt within estimators.

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

Answer: The cross validation can make use of whole dataset for training and whole dataset for testing. And then the assessment of learning algorithm will be more accurate.

When evaluating different settings for estimators, there is still a risk of overfitting on the test set because the parameters can be tweaked until the estimator performs optimally. To solve this problem, yet another part of the dataset can be held out as a so-called “validation set”. However, by partitioning the available data into three sets, we drastically reduce the number of samples which can be used for learning the model, and the results can depend on a particular random choice for the pair of (train, validation) sets. When doing with cross-validation, the validation set is no longer needed and does not waste too much data.

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 training error grows and approaches a certain value as the training size increase. On the contrary, the testing error decreases

and levels off as the training size increases.

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/under-fitting or high variance/overfitting?

Answer: In the graph for the model with max depth 1, both the training error and test error approach the same value as the training size grows, while the training error and test error have large difference in the graph for the model with max depth 10. The deviations between test error and training error account for the high variance/ overfitting suffered by the model with max depth 10. The test and training error for the model with max depth 1 is obviously larger than that for the model with max depth 10, which means the model with max depth 1 suffers from the high bias/under-fitting.

After all, the model with max depth 1 suffers from high bias/under-fitting, while the model with max depth 10 suffers from high variance/ overfitting.

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: The training and test error both decrease as the model complexity increases. Based on this relationship, the model with max depth 4 best generalizes the dataset because the test error has already levelled off and the deviation between test and training error is still small, which means a low variance.

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:
best estimator is: DecisionTreeRegressor(criterion='mse', max_depth=4, max_features=None,
           max_leaf_nodes=None, min_samples_leaf=1,
min_samples_split=2,
           min_weight_fraction_leaf=0.0, presort=False,
random_state=None,
           splitter='best')

```
best params is: {'max_depth': 4}
House: [11.95, 0.0, 18.1, 0, 0.659, 5.609, 90.0, 1.385, 24, 680.0,
20.2, 332.09, 12.13]
Prediction: [ 21.62974359]
```

Compare prediction to earlier statistics and make a case if you think it is a valid model.

Answer: The mean price is 22.5 and median price is 21.2. The predicted price 21.6 is close to the mean and median price. Furthermore, we can find the 10 nearest neighbour of the feature vector in the dataset and calculate the mean value of their prices. Since the mean price is 21.52 which is close to the prediction of the trained model, I think the predicted value is included in one standard deviation range and the trained model is a valid model.