Questions and Report Structure

1) Statistical Analysis and Data Exploration

- Size of data: 506
- Number of features: 13
- Minimum Housing Price: 5.00
- Maximum Housing Price: 50.00
- Mean Housing Price: 22.53
- Median Housing Price: 21.20
- Standard Dev of Housing Prices: 9.19

2) Evaluating Model Performance

- Which measure of model performance is best to use for regression and predicting Boston housing data? Why is this measurement most appropriate? Why might the other measurements not be appropriate here?

  The predicted value in this case is the housing price which is a numerical data and we are using regression model. Accordingly, I chose Mean Square error (MSE) as the measure of model Performance. The other estimator is Mean Absolute Distance (MAD). Unless we plot the data and see if we have huge outliers, it is difficult to figure out if MSE is better than MAD. MSE seems to be more popular unless there is a specific insight and hence I chose MSE.

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

  The performance of a model needs to be evaluated on unseen data to see how well the model generalises. By splitting data into two sets and using training data to build the model and test data (unseen data while training) to evaluate the performance of a model, we get a better picture of model's ability to generalise.

  If you do not split the data into testing, there is no way to test the performance of the model at the time of building the model since all the data is used for training. The actual performance of such a model can only be gauged when the model is used for prediction and the predicted value is compared with the actual output value.

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

  There are two types of Cross validation techniques. A) Exhaustive like “Leave-p-out” cv or “Leave-one-out” cv; B) Non-exhaustive like “k-fold” cv. In k-fold, it is common to choose k=10. Leave-p-out is a good way to validate but the number of runs ($^{n}C_{r}$) required become very large as n increases.

  k-fold has the advantage of having each sample used once in validation and it allows to effectively train the model without sacrificing data for validation.

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

  Grid search is used to systematically search a parameter space with cross-validation to identify the best set of parameters. In out example we will use it to identify the best depth of decision tree that can be used for boston housing price estimator mdoel.

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?

  As the training size increases, the training error starts from zero and then slowly increases and soon stabilizes. The initial training error is zero as the number of training samples is small and it is possible to have perfect fit for this small size of training set. The fit worsens as we increase the training set and hence training error increases. The increase is usually logarithmic and hence it kind of stabilizes after the training size in increased beyond a point.

  The testing error is very high when the training size is small. This is due to the lack of generalization in the model. As the training size increases, for a given complexity and structure of model, the generalization improves and testing error drops. After a certain number of training samples, the testing error also stabilizes and usually will be higher than the training error of the model. This is also seen from the 10 learning curves that are plotted.

- 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?

At a tree depth of 1, the model suffers from high bias reflected in the generally high value of error in training and testing, that of around 40.

At a tree depth of 10, the model suffers from high variance/overfitting. This is reflected in very low training error and test error being high plus the learning curve (for depth of 10) showing a lot of jaggedness in the testing 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?

The model with max depth of 4 is the best model. Till the time model hits the max depth of 4, the training and testing errors both decrease. After this point, while the training error falls, the test set error starts to show an increase and for subsequent higher tree depths, the test error continues to fluctuate around the error value of 15 – same as the test set error for depth of 4.

However, gridsearchCV with 10-fold validation gives the best tree depth to be 5. I implemented the leave-one-out version of cross validation.

4) Model Prediction

- Model makes predicted housing price with detailed model parameters

The price predicted by the model is 20.96776

- Compare prediction to earlier statistics

The predicted price is fairly close to the median price. Accordingly, this house must be a very standard and commonly occurring type inside the given data.