Questions and Report Structure

1) Statistical Analysis and Data Exploration

- Number of data points (houses)?
- there are 506 houses in this dataset, as calculated by the size of the data target list which is the number of house prices, assuming one price per house

- Number of features?
- there are 6578 total features, categorized by 13 different attributes.

- Minimum and maximum housing prices?
- the minimum is $5000 and the maximum is $50,000

- Mean and median Boston housing prices?
- the mean of the prices is $22,532.8 and the median is $21,200

- Standard deviation?
- the standard deviation is 9.188 or $9188.01 from the mean

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?
- The best score attribute of the grid search CV object is the best to use for predicting data because the higher the accuracy the better. Other measurements are not as appropriate because they can be the same result from different models. I have ran some different models and they conclude to a small range of prices:

performance metric	prediction	best score	max_depth
median_absolute_error	19.93372093	5.131723517	1
median_absolute_error	19.93372093	5.131723517	1
mean_absolute_error	19.93372093	5.481653145	1
mean_absolute_error	19.93372093	5.481653145	1
mean_squared_error	19.93372093	56.82662232	1
mean_squared_error	19.93372093	56.82662232	1
r2_score	19.99746835	-2.149001939	
r2_score	20.76598639	-2.128119323	6
explained_variance_score	19.99746835	-0.447883282	
explained_variance_score	20.76598639	-0.6016504362	6

As we can see from the table most of the predictions close in on $19,933.72 for the house price.

- Why is it important to split the Boston housing data into training and testing data? What happens if you do not do this?
- It's important to use the test data to see if the model really works after fitting to the training data. If we do not do this then the model could be overfitting to the training data and unable to predict new data.

- What does grid search do and why might you want to use it?
- Grid search looks through all the parameters and finds the one with best performance. We want to use it so we can go through parameter combinations and optimize our models automatically.

- Why is cross validation useful and why might we use it with grid search?
- Cross validation is useful because it maximizes the usage of all given data by utilizing them for both training and testing data. We use it with grid search so we can use all the data to find the best parameter combinations.

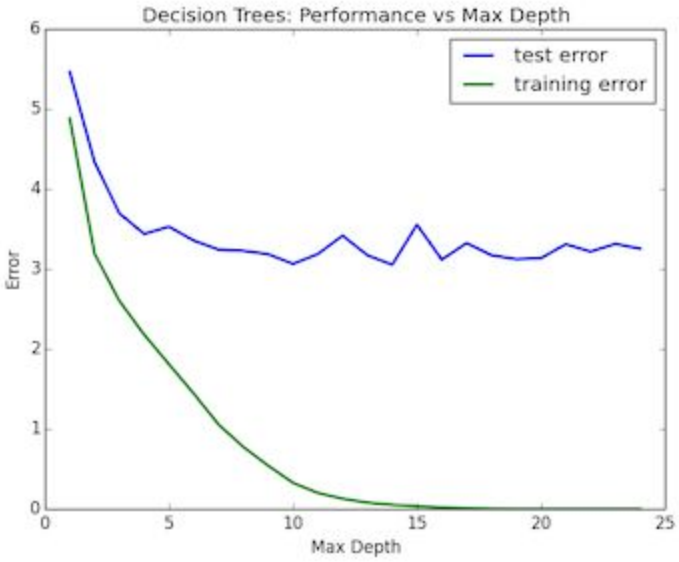
3) Analyzing Model Performance

- Look at all learning curve graphs provided. What is the general trend of training and testing error as training size increases?
- As training size increases the training error increases then stays approximately the same, and testing error decreases and stays approximately the same.

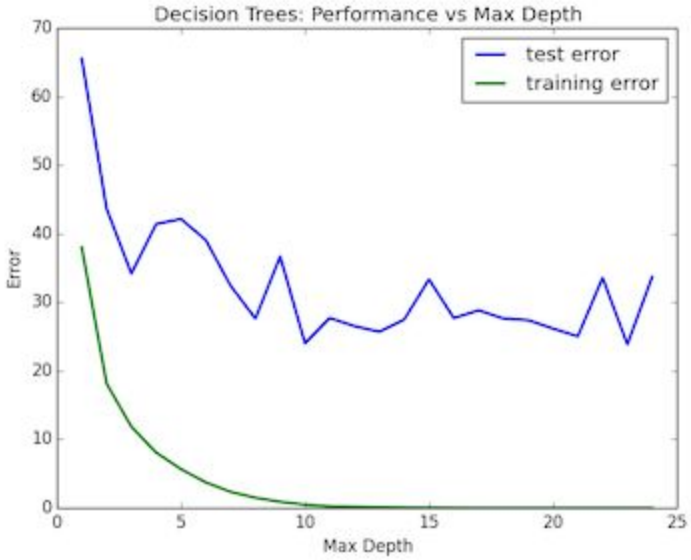
- Look at the learning curves for the decision tree regressor with max depth 1 and 10 (first and last learning curve graphs). When the model is fully trained does it suffer from either high bias/underfitting or high variance/overfitting?
- When the model is fully trained it suffers from high bias and underfitting because the general slope has a downward trend as the training size increases, and there is a gap between training error and 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?
- As the model complexity increases the training error decreases but the test error stays relatively flat. Based on this relationship a max depth of less than 5 is best. Grid search CV gives best parameter at max depth of 1 and from looking at the graph it seems a max depth of 2 or 3 would work as well. This way the two errors stay relatively close together and the model is less affected by underfitting. R2 score and explained variance score have positive slopes but they are relatively flat as well.

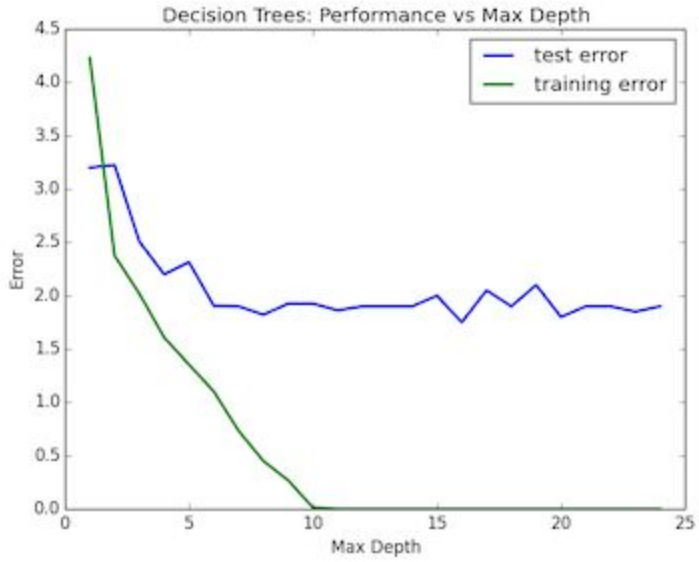
mean absolute error:



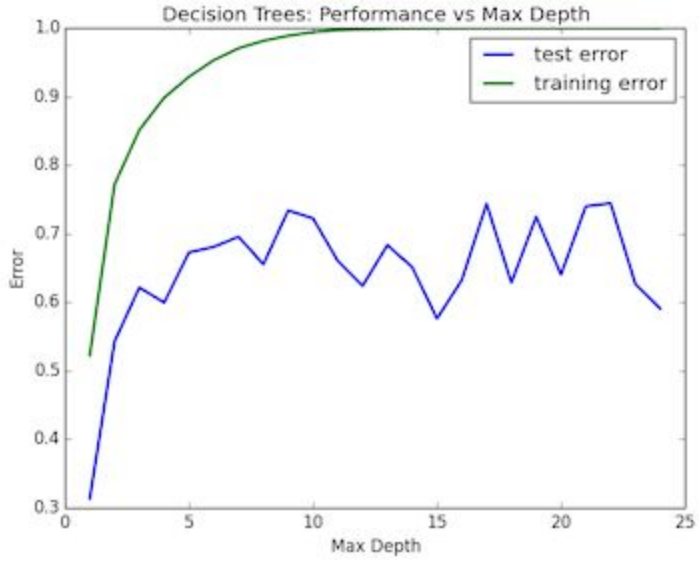
mean squared error:



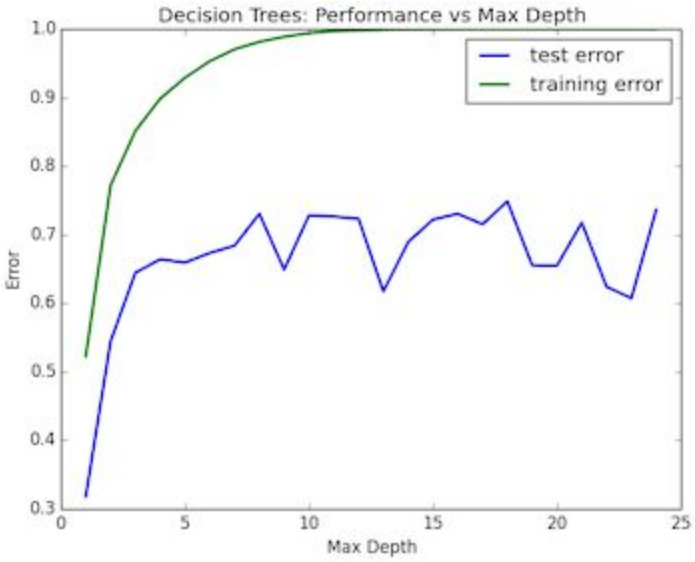
median absolute error:



r2 score:



explained variance score:



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.
- From grid search the predicted house value is from $19,933.72 to $20,765.98 with a range of $832.26 and a mean of $20,349.85

- Compare prediction to earlier statistics and make a case if you think it is a valid model.
- This is a valid model because the results produced by all the regression models are close together, and running the grid search multiple times produced same results.