Data Exploration

Statistical Analysis

0	1	2	3
Number of houses:	Number of features:	Minimum price:	Maximum price:
506	13	5.0	50.0

4	5	6
Mean price:	Median price:	Standard deviation:
22.5328063241	21.2	9.18801154528

Evaluating Model Performance

Performance Metric

Mean Squared Error is chosen because (1) analyzing Boston housing dataset is a regression task; (2) Squaring never gives a zero sum value and emphasizes larger differences.

Testing/Training Split

One reason for 70:30 split is because the number of observations in this dataset is quite small. We need enough data for training the model, but also enough data for testing to simulate the unseen observations in the real world. In this case, a 70:30 split seems ideal.

Gridsearch

```
reg = GridSearchCV(estimator = regressor,
                   param_grid = parameters,
                   scoring = scorer,
                   cv = 10)
```

Regressor is the default decision tree regressor;

Scorer is the Mean Squared Error performance metric with the greater the better set to false because it's a loss function.

CV is changed to 10 for better accuracy.

I leave other parameters as default.

Cross Validation

Cross validation is very important because it allows you to train and test using all of the data points, solving the training-testing dilemma.

To get a more accurate result without sacrificing too much computing power I increase the K in CV to 10 fold.

Learning Curves and Training Analysis

Max depth = 1 high bias/underfitting => Model doesn't have enough explanatory power (not complex enough) => both types of error persist as training size increases.

Max depth = 10 high variance/overfitting => Model fits perfectly with training data while the test error fluctuates.

Learning Curves and Bias & Variance Analysis

As the model complexity increases, both training and test error goes down until the model becomes overfitting – almost zero training error but test error not going down.

Decision Tree with Max Depth: 1

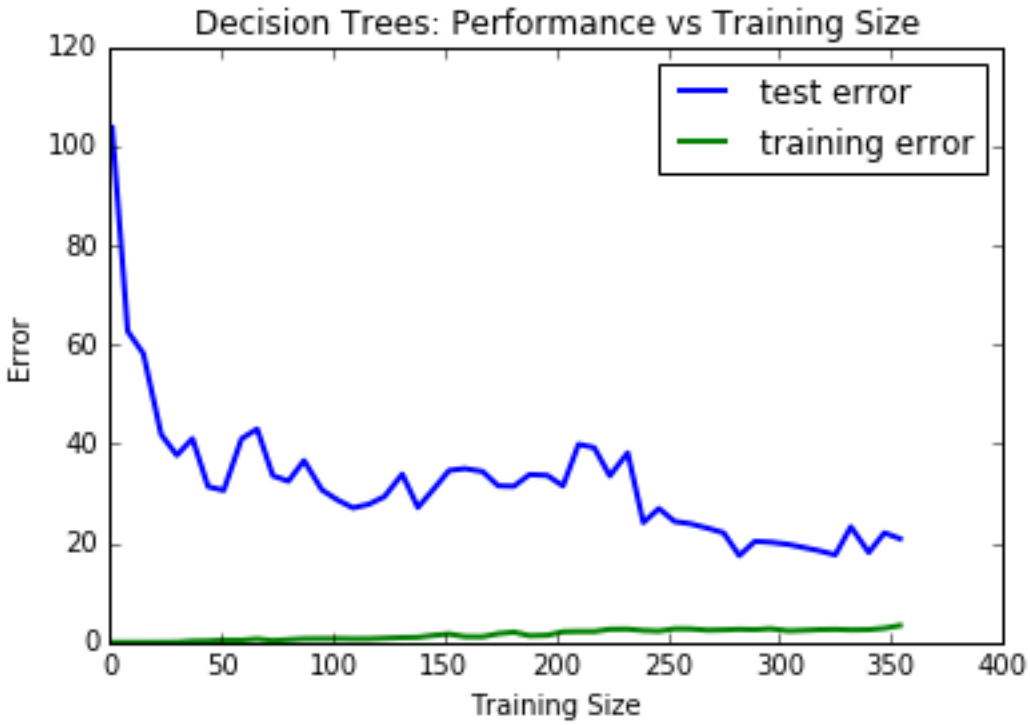
Decision Tree with Max Depth: 2

Decision Tree with Max Depth: 3

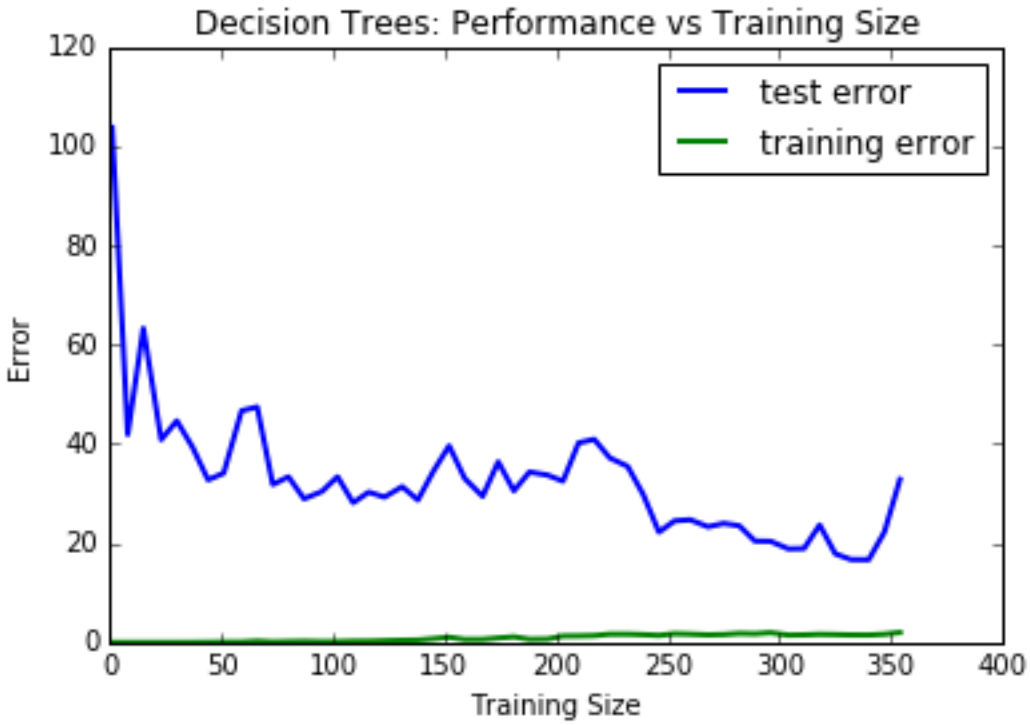
Decision Tree with Max Depth: 4

Decision Tree with Max Depth: 5

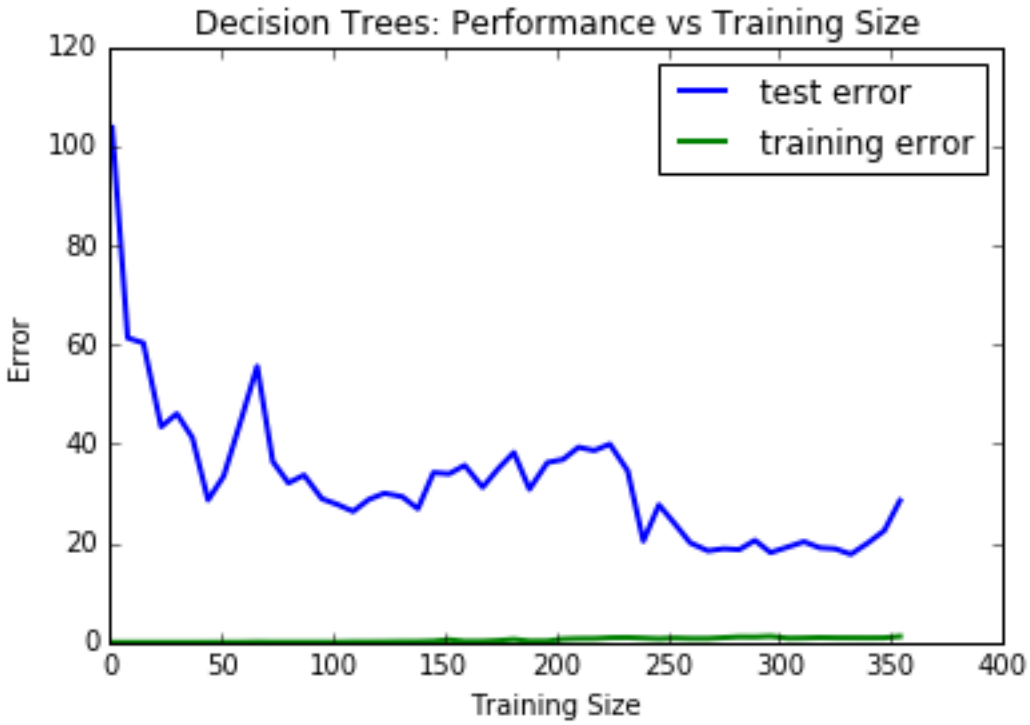
Decision Tree with Max Depth: 6



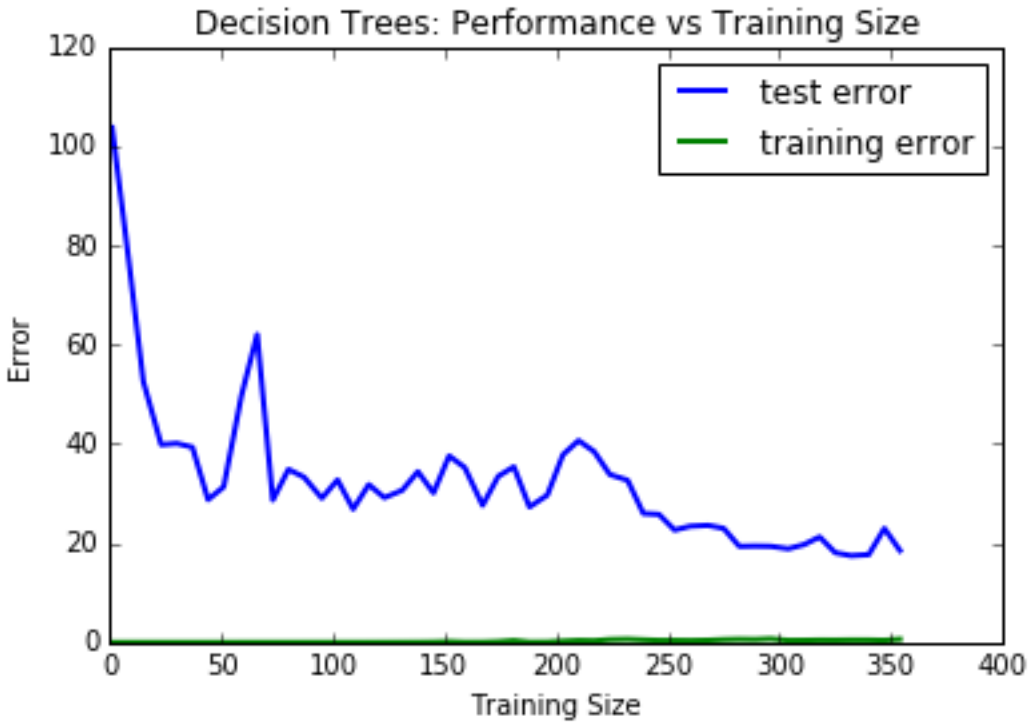
Decision Tree with Max Depth: 7



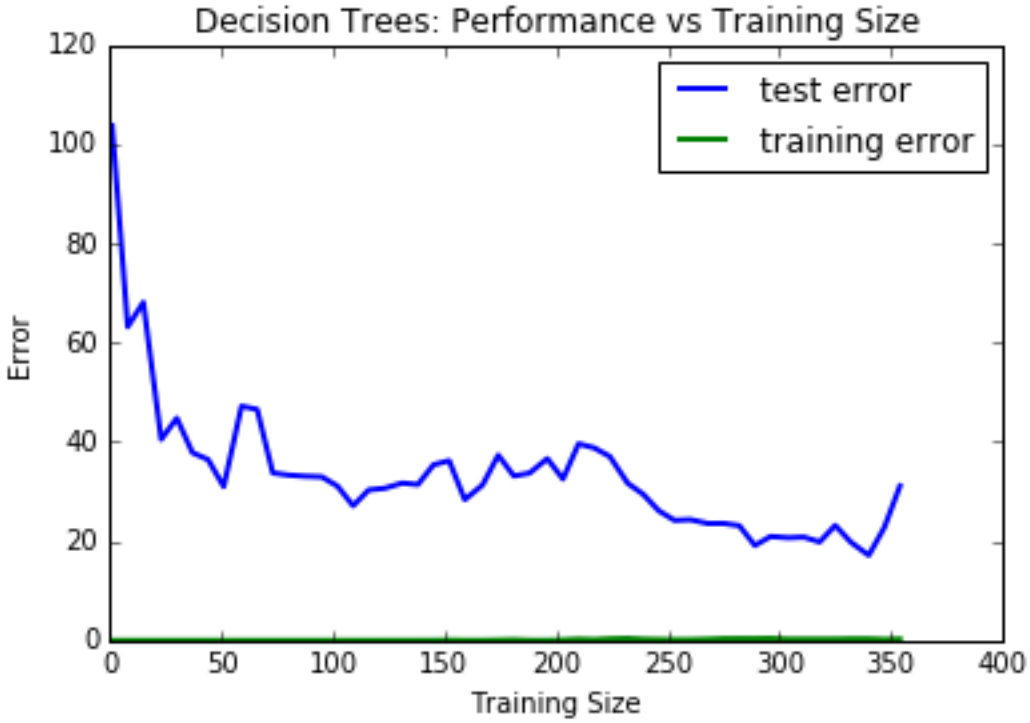
Decision Tree with Max Depth: 8



Decision Tree with Max Depth: 9



Decision Tree with Max Depth: 10



Error Curves and Model Complexity

Explanation: After max depth = 10, we can see from the graph that more complexity will cause overfitting to the dataset as the training error goes to 0 but testing error fluctuating.

Picking the Optimal Model

best parameters: {'max_features': 5, 'max_depth': 10}

best score: -28.6977337707 (MSE, loss function)

This model is picked by Gridsearch for the lowest MSE score.

Predicted Housing Price: 27.9

Comparing Model Price to Housing Statistics:

The price is reasonable as it falls into the range of 5 to 50 from the dataset, and it is fairly close to the mean 22.53.