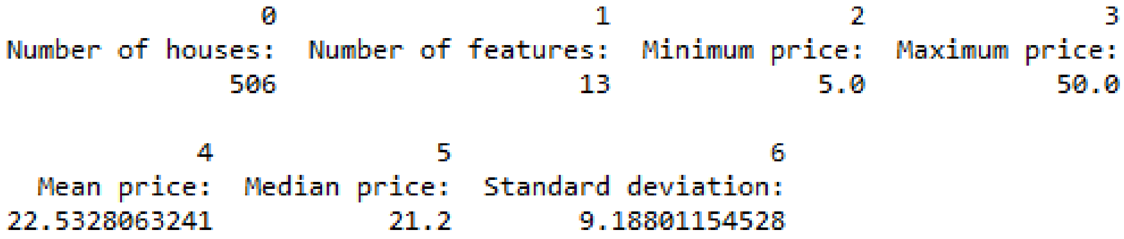
Data Exploration

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.

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. After we specify a number k, the dataset will be divided into k different buckets, and we will run the analysis for k times – an iterative process. For each time, we choose different buckets to be our training and test data (at the same training/test split ratio) so that all of our data is used to training the model and test the performance.

We can use cross validation in combination with grid search for model selection – to find the parameters that optimize the score function we specify.

Gridsearch

Gridsearch is an exhaustive algorithm to find the best parameters set by searching through all the possible choices within the range we give. In this case, the best parameters set will optimize the score function we define.

The gridsearch function I set up is very straightforward:

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. I leave other parameters as their default values.

Analyzing Model Performance

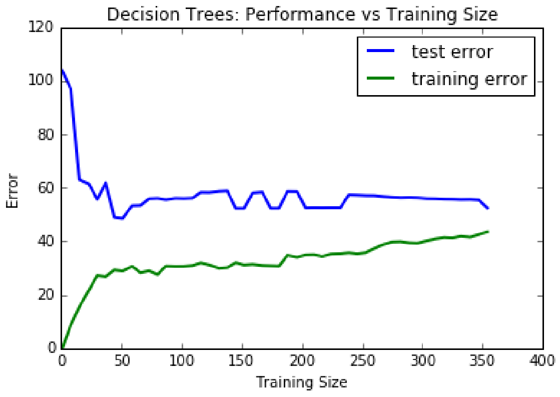
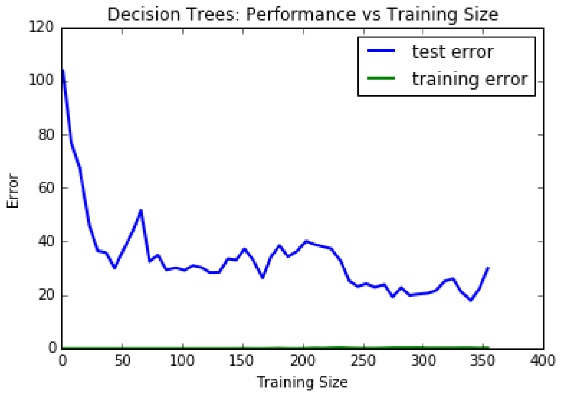
Learning Curves and Training Analysis

In general, the test error is decreasing while the training error is increasing as training size increases.

Learning Curves and Bias & Variance Analysis

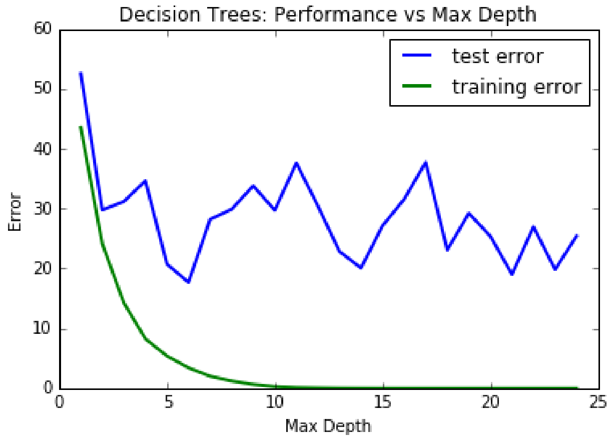
When the Max depth = 1, the model has high bias/underfitting (Not complex enough to explaining the data). The training and test are converging to certain amount of error that can't be reduced, which indicates high bias.

When the Max depth = 10, the model has high variance/overfitting (Too complex that the model fits perfectly with the training data but not the test). The training error is almost zero while the test error is not converging to zero, which indicates high variance or overfitting.

Decision Tree with Max Depth: 1	Decision Tree with Max Depth: 10
	

Error Curves and Model Complexity

As the model complexity increases, to the dataset as the training error goes to 0 but testing error fluctuating. As we can see from the following graph, the model becomes overfitting when the Max Depth is greater than 7 – test error goes up then fluctuates while training error goes to zero.



Picking the Optimal Model

best parameters: Max Depth = 4

best score: -37.09966579 (MSE, loss function)

best estimator: 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')

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

Other than optimizing the score function we can see from the complexity graph that at Max Depth = 4 both training and test error are relatively low.

Model Prediction

Predicted Housing Price: 21.38 (The mean value over 10 runs)

Run	Prediction	Best Score	Max Depth
1	20.76598639	-37.13990024	6
2	20.76598639	-36.09699181	6
3	21.62974359	-36.23962562	4
4	21.62974359	-35.11201693	4
5	21.62974359	-39.83476154	4
6	21.62974359	-35.43960286	4
7	21.62974359	-39.83476154	4
8	21.62974359	-36.16994047	4
9	21.62974359	-36.16994047	4

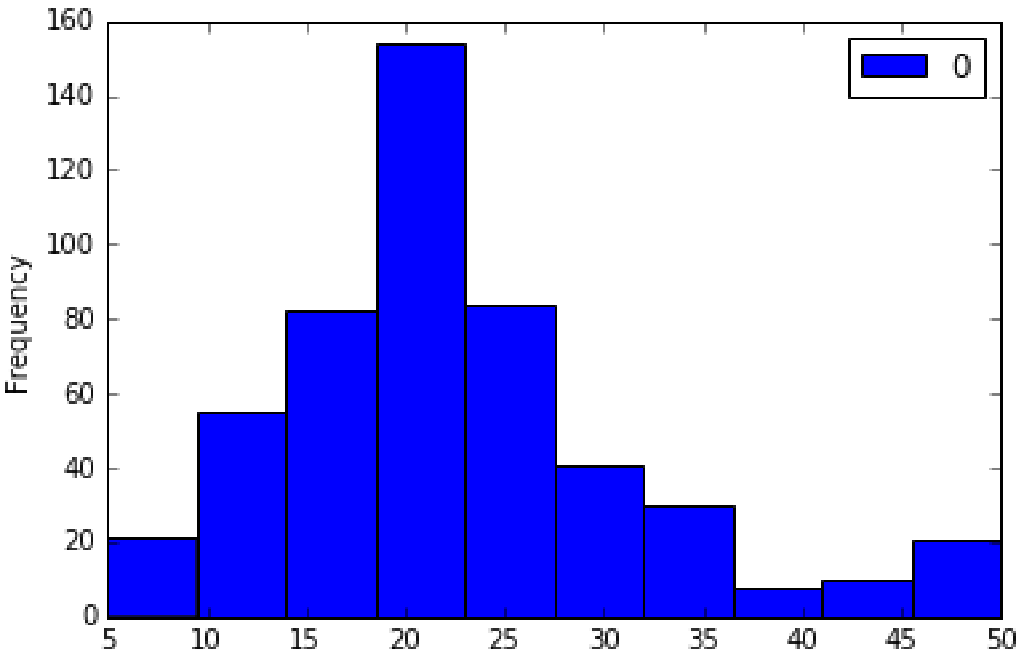
10	19.32727273	-38.48070211	9
	Avg Prediction	Mean Score	Median Max Depth
	21.38295582	-37.09966579	4

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.

It also falls into the middle 50% range.

Histogram of housing price distribution



25th percentile: 17.025, 75th percentile: 25.0