Boston Housing Price Analysis

Joe Sieber

Table of Contents

Abstract

With low interest rates the housing market is very competitive. Real estate agents need an edge over other agents to attract clients. One way to add value to the customer is to be able to better predict a house's value. For sellers this will help sell their house at the optimal value in the shortest time and for buyers this will be benchmark to see which houses are undervalued. A dataset was provided with 506 training examples and 13 features to help predict a house that we are currently an agent for. Using machine learning, a decision tree model was used to help facilitate the prediction of the house in question. Several models were tried with various depths used. Using our house's features with the final model it is predicted that the houses estimated value is $21,630. This is slightly lower than the average price for the dataset.

Introduction

The housing market is currently experiencing a surge in prices around the country due to low interest rates. This is especially true in Boston, shown in Figure 1 (2). Please note that the values in the figure are listed in today's dollars and do not relate directly to the dollar values listed in the dataset used for this report.

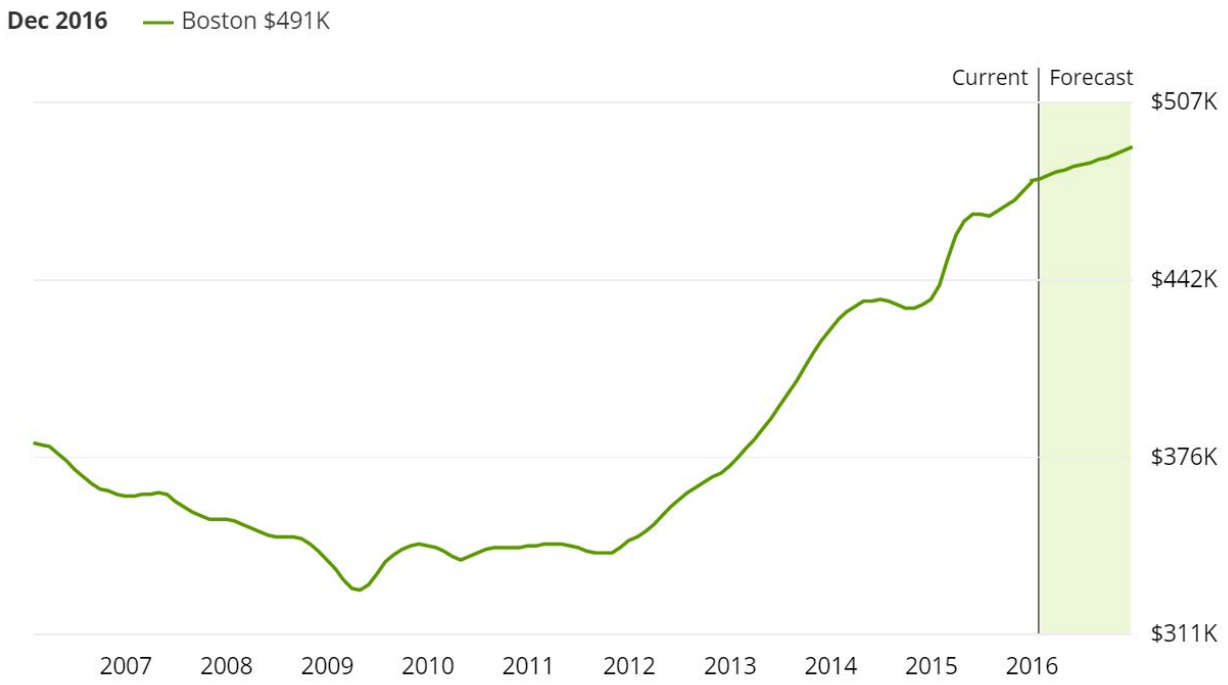

Figure 1 - Boston Housing Prices

To remain competitive in this market against other real estate agents and to get the customer the best price, new and innovative methods will need to employed. There is a wealth of data available and this leads to machine learning being a viable solution to predict house prices which in turn adds value to the customer.

Data Exploration

The data used for this analysis was provided from the scikit-learn datasets. This data set contains 13 features and 506 training examples. It also contains the housing price for those examples. The housing price statistics are shown in Table 1.

Max	$5000
Min	$50000
Mean	$22,533
Median	$21,200
Standard Deviation	$9,188

Table 1 - Housing Price Statistics

Figure 2 shows a histogram of the housing prices for this dataset.

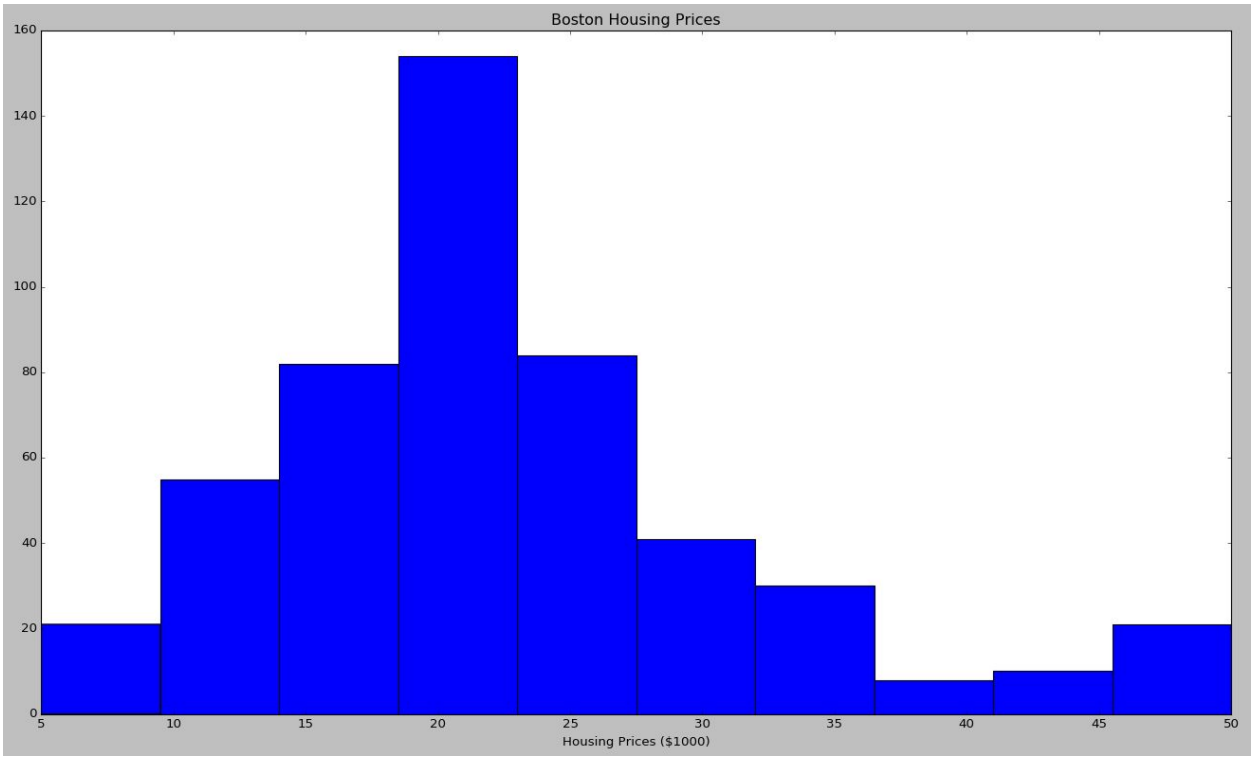

Figure 2 - Housing prices

The 13 features for this data set are shown in Table 2.

Feature	Description	Mean	Std. Dev.
CRIM	Per capita crime rate	3.59	8.59
ZN	Proportion of residential land zoned for lots over 25,000 sq. ft.	11.4	23.3
INDUS	Proportion of non-retail business acres	11.1	6.85
CHAS	Charles River dummy variable	0.0692	0.254
NOX	Nitric Oxides (parts 10 million)	0.555	0.116
RM	Average number of rooms per dwelling	6.28	0.702
AGE	proportion of owner-occupied units built prior to 1940	66.6	28.1
DIS	weighted distances to five Boston employment centres	3.80	2.10
RAD	index of accessibility to radial highways	9.55	8.70
TAX	full-value property-tax rate per $10,000	408	168
PTRATIO	pupil-teacher ratio by town	18.5	2.16
B	1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town	357	91.2
LSTAT	% lower status of the population	12.7	7.13

Table 2 - Features and Statistics

Evaluating Model Performance

For this report, a decision tree was used to predict the housing prices. This model can be evaluated using many different metrics. Some of the metrics include mean square error (MSE), root mean squared error (RMSE), and mean absolute error (MAE). Any of these metrics would work to help refine the model and determine the best fit but MAE was used for the model in this report. MAE was chosen because it is a linear score and it does not excessively penalize the model for outliers. There are many different features that can determine the outcome of the housing price and the 13 features in this dataset do not cover all of those. Since there will be outliers that can't be explained by the 13 features in the dataset it is best to not overly penalize the model with a metric such as MSE and RMSE.

The dataset was split into a training set and a test set. 70% of the data went to the training set and the remaining went to the test set. This was done in order to check the model to see if it is overfitting the data and to get a fair representation of how the model performs.

A grid search was used to find the optimal parameters for the decision tree. Specifically, the grid search is looking at various depths for the decision tree and determining the best depth. This rapidly speeds up the process of finding the ideal depth instead of manually changing the parameters. During the grid search process cross validation was used to determine how the model performed. Cross validation is the process of splitting the data up into several groups and training on a portion of that data and testing against the portion not trained on. There are several methods to do this but a common method is a X-fold pattern. Where X is the number of splits in the data. For example, if X where 10 then the data would be split into 10 groups and a model would be fitted to 9 of the groups and validated against the other group. This is done 9 more times with the validation set being different for each model and the average of all of those models is final model. With this method you get an out of sample scoring of the models and are able to train on all of the data for the averaged final model.

Analyzing Model Performance

Different models were generated with the grid search and their error rate, based on MAE, was plotted. Figures 1, 2, and 3 show the training and test error rates at depths of 1, 5, and 10. These figures show that as you increase the training set size the error rate quickly drops and plateaus with a training set size of around 100 examples. Figure 1, which has a depth of 1, starts with a test error around 10 and quickly drops to around 5 and then has almost no improvement as the training size increases. The training error does the opposite, starting at 0 error and going up to around 5 also. This is an example of underfitting because the model test error and training error are almost identical. To fix this the model should be increased in complexity, such as increasing the depth.

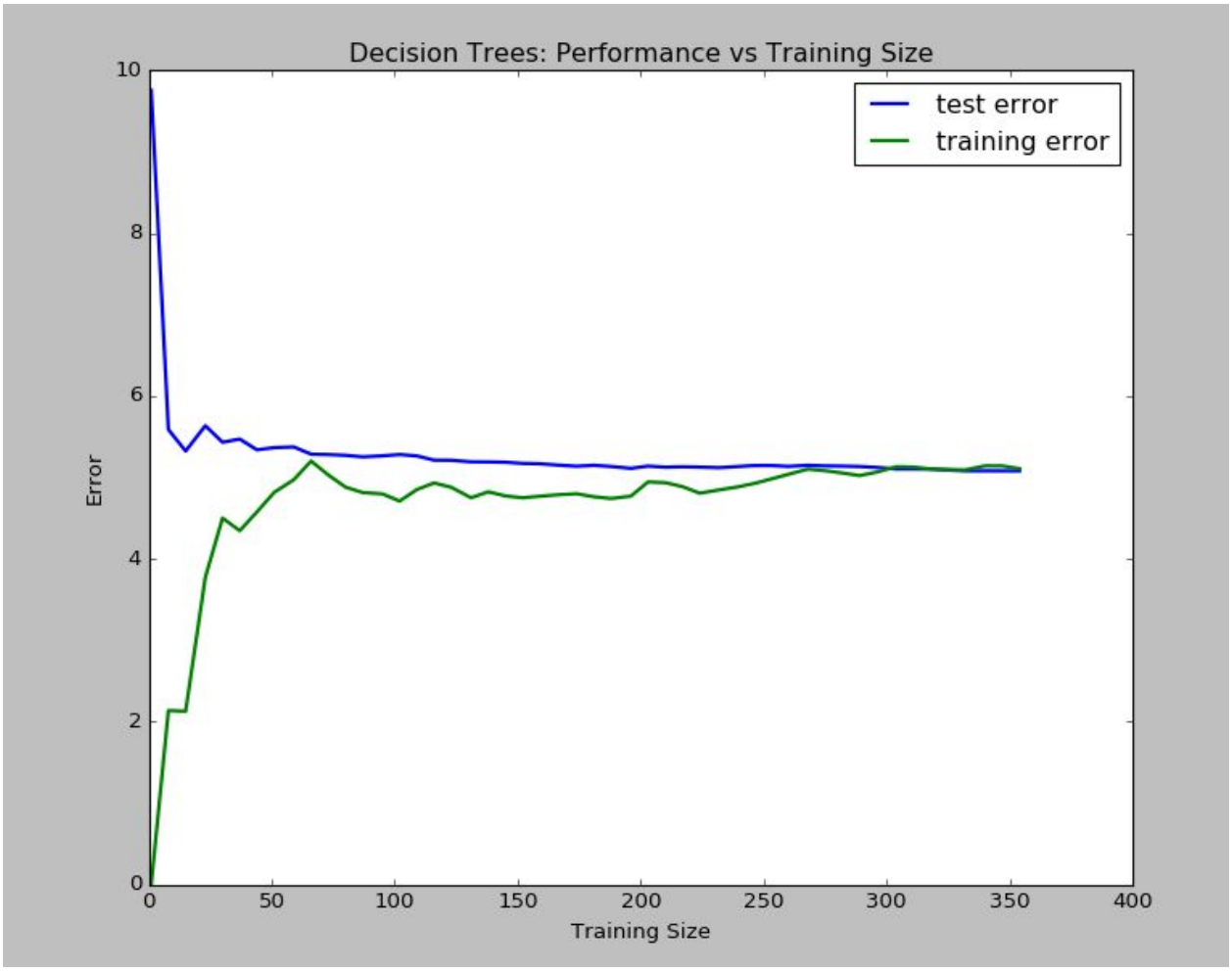

Figure 3 - Decision Tree with Depth of 1

Figure 4, which has a depth of 10, also starts with a test error around 10 and quickly drops to around 3 and then has almost no improvement as the training size increases. The training error starts at 0 error and slowly goes up to around 1. This is an example of overfitting because the model test error and training error are very different. To fix this more data should be given or reduce the model complexity.

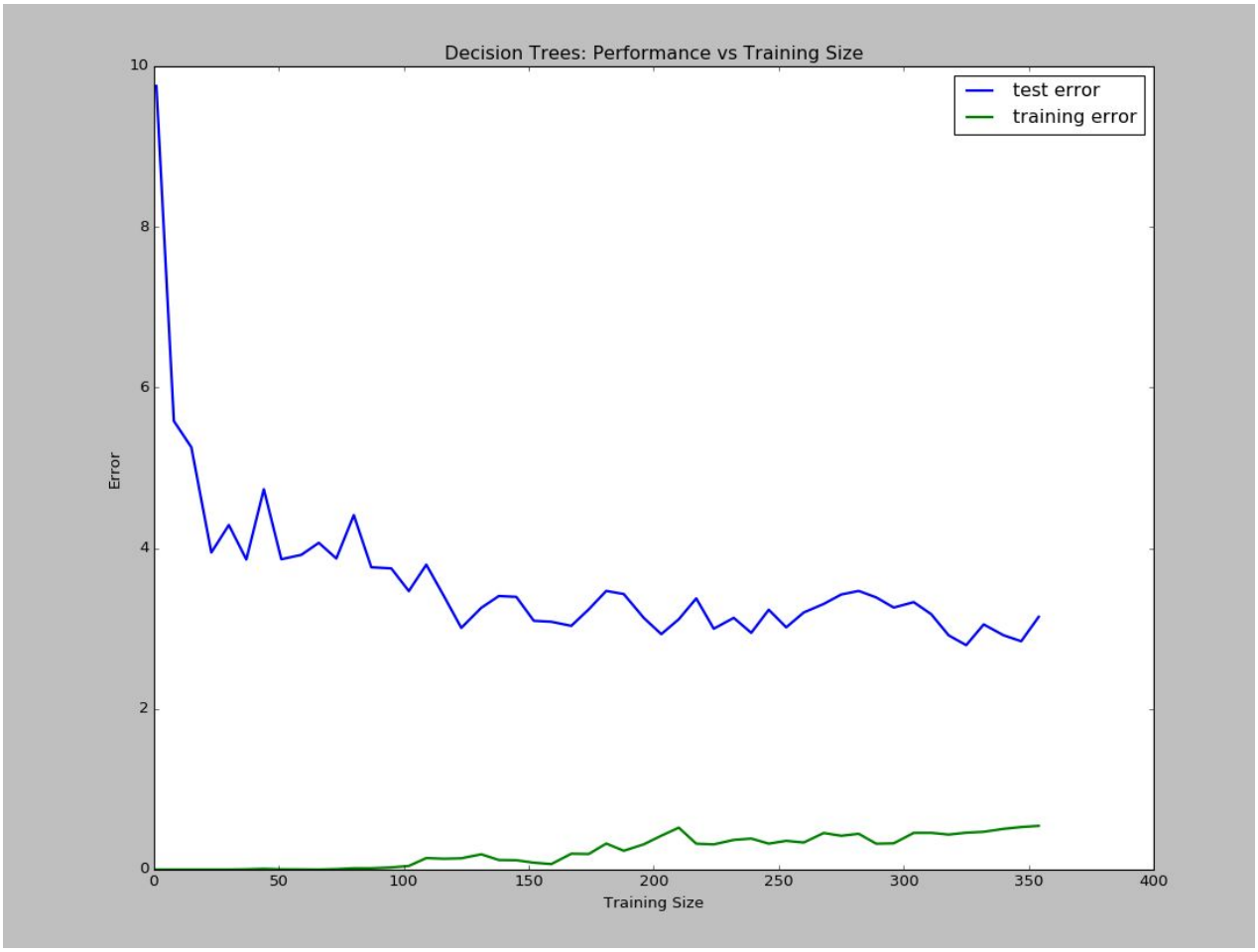

Figure 4 - Decision Tree with Depth of 10

Figure 5 shows the performance of the different decision tree models more clearly. The general trend of these different models is that as the depth of the decision tree increases, the training error goes down. However, the test error seems to stagnate with a depth of 5. This visually shows that the test error does not seem to improve as the depth of tree increases past 5. The model with a max depth of 5 should be used as to minimize overfitting.

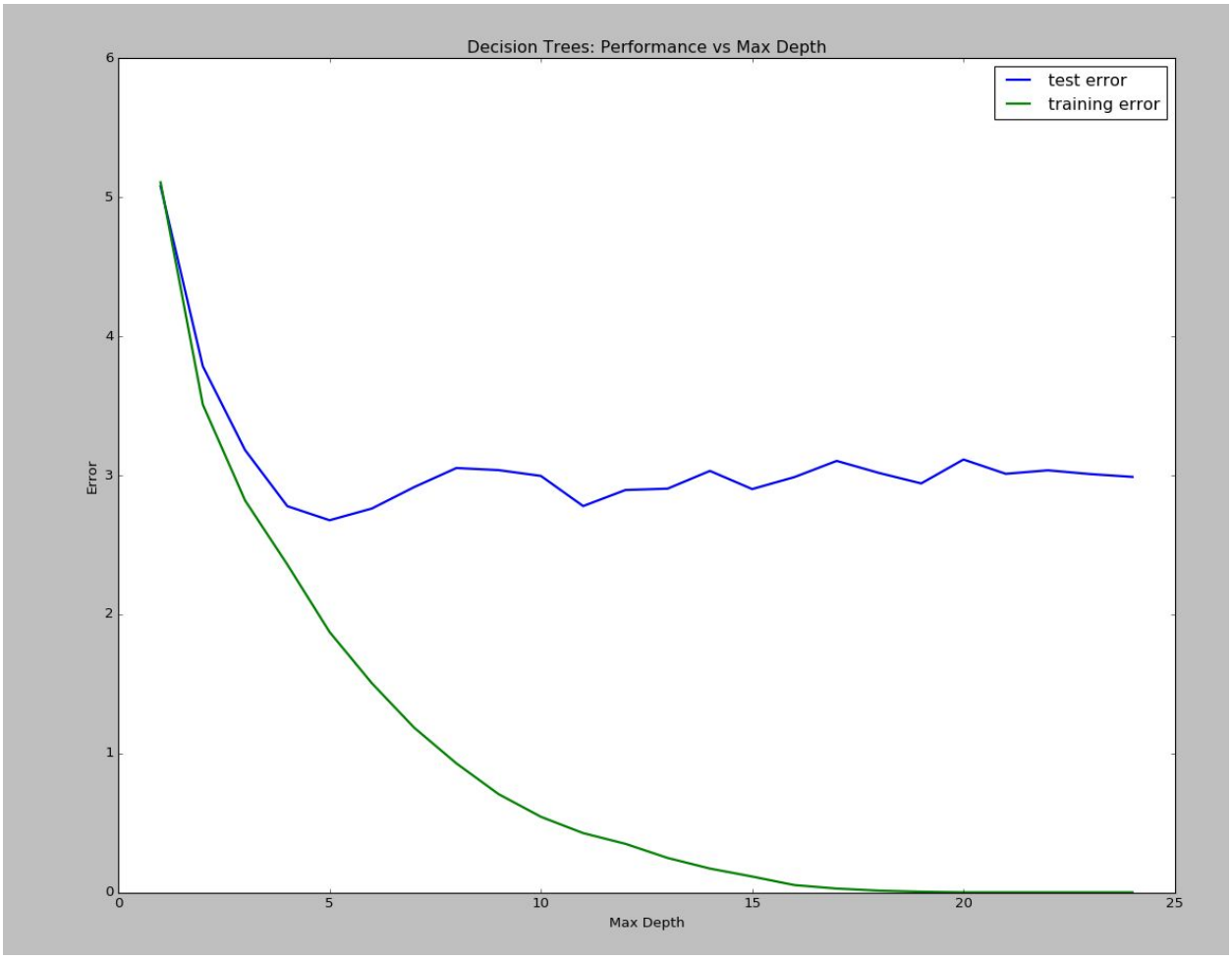

Figure 5 - Decision Tree Performance vs Max Depth

These results shows us where to focus energy on to create a better model. Just collecting more training examples will get us marginal improvement with this model. Effort should be focused on finding a different algorithm to improve these results.

Model Prediction

The process of creating a decision tree model has some variables that are randomly initialized. In order to get reproducible results the random seed was set to 1 using the numpy.random.seed() method. Using the grid search the best model found was with a depth of 4. With this model using our house with inputs to the features the predicted value for the house is $21,630. This seems to be a valid model since the mean house price in this dataset is $22,533 with a standard deviation of $9,188 and the house's features in question seem to be pretty close to the mean for that feature.

Feature	Prediction House Value
CRIM	11.95
ZN	0.00
INDUS	18.10
CHAS	0
NOX	0.6959
RM	5.6090
AGE	90.00
DIS	1.385
RAD	24
TAX	680.0
PTRATIO	20.20
B	332.09
LSTAT	12.13

Table 3 - Values for House to Predict

Appendix A - References

1. Http://Www.Bumc.Bu.Edu/Gms/Files/2010/06/Darmouth-Street.Jpg. 2016. Print.

2. Zillow, Inc. "Boston MA Home Prices & Home Values | Zillow". Zillow. N.p., 2016. Web. 31 Jan. 2016.