Executive Summary:

Boston Housing Data:

The objective of this report is to analyze the various models that can be fitted to the Boston Housing Data and to determine their average sum square errors for comparison. The Boston Housing Dataset contains information about the housing prices in a particular area of Boston, with the following variables - CRIM: per capita crime rate by town; ZN: proportion of residential land zoned for lots over 25,000 sq.ft.; INDUS: proportion of non-retail business acres per town; CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise); NOX: nitric oxides concentration (parts per 10 million); RM: average number of rooms per dwelling; AGE: proportion of owner-occupied units built prior to 1940; DIS: weighted distances to five Boston employment centres; RAD: index of accessibility to radial highways; TAX: full-value property-tax rate per $10,000; PTRATIO: pupil-teacher ratio by town; B: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town; LSTAT: % lower status of the population; MEDV: Median value of owner-occupied homes in $1000's.

The original dataset contains 506 observations with 14 variables (including the response variable). The dataset is split into training and testing using stratified random sampling, where 70% of the entire data is fixed as training data and the rest 30% is used for testing. (Training dataset = 354 observations, Testing dataset = 152 observations)

Models chosen:

Linear model- A linear model is fitted to the training data using step wise regression. (Direction used = both backward and forward)

Additive model- A generalized additive model is fitted to the training data using splines. (Continuous predictor variables are used in the model)

Neural Network- A model is generated using artificial neural networks for predicting the response variable.

Regression Tree- A regression tree is fitted to the training data set. The tree is populated, pruned and tested.

Important Results:

Model Type	Average SSE in-sample	Average SSE out-of-sample
Linear Model	21.60	25.69
Additive Model	14.13	18.86
Neural Networks	20.63	25.82
Regression Tree	21.92	27.56

From the above results, we see that the Additive Model produces the least average sum squared error, indicating that it is the best model out of the lot. This is because, as in most cases, the generalized additive model is more flexible to changes in the response variables as compared to the predictors.

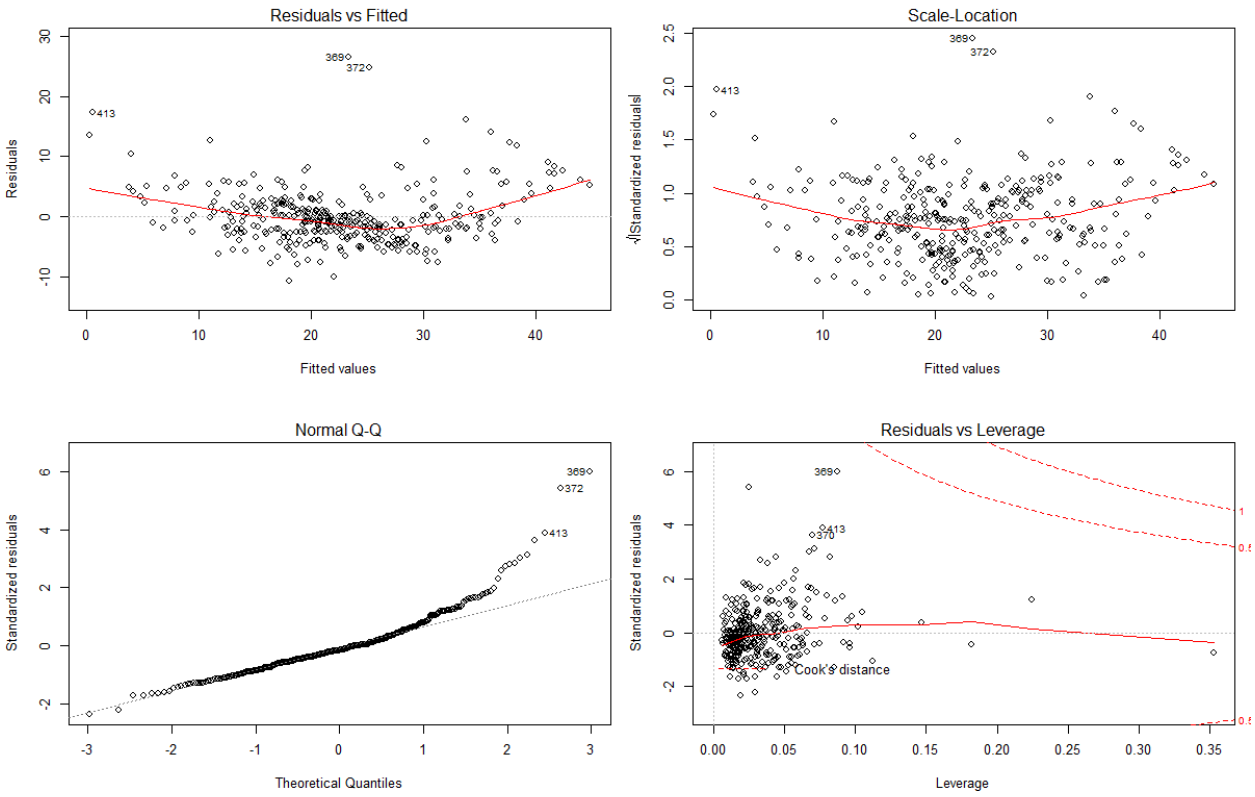
PROBLEM 1 – Part A: BOSTON HOUSING DATASET

Model 1: Linear Model

A linear model is fitted to the training dataset in a step-wise regression method. A combination of both forward and backward regression is used to arrive at the final model. The variables are added to the model based on their AIC scores. Initially, a null model is entered and variables are added/dropped at each step. The following table shows the entry and exit of variables into the model.

Step 1: lstat enters;	Step 2: rm enters;	Step 3: ptratio enters;	Step 4: black enters;
Step 5: dis enters;	Step 6: nox enters;	Step 7: chas enters;	Step 8: crim enters;
Step 9: rad enters;	Step 10: tax enters;	Step 11: zn enters	

The following graphs show the effectiveness of the fit by comparing the fitted values with residuals, scale, checking for normality of residuals, leverage vs residuals.



The model is then validated with the testing data and the average sum squared error is calculated.

The details of the model is are follows:

Model: medv = 34.31-0.505*lstat + 4.305*rm -1.01*ptratio + 0.011*black -1.58*dis -20.77*nox + 3.45*chas -0.131*crim + 0.255*rad -0.007*tax + 0.028*zn

Average SSE out of sample = 25.69 **Average SSE in sample** = 21.60

Model 2: Generalized additive model

Unlike the linear model, the generalized additive model can be considered as a non-linear model. Splines are fitted to each of the predictor variables and they are then used to predict the responses. The splines are applied only to the continuous predictor variables. The degrees of freedom of each spline depends on the combination of covariates within the variable.

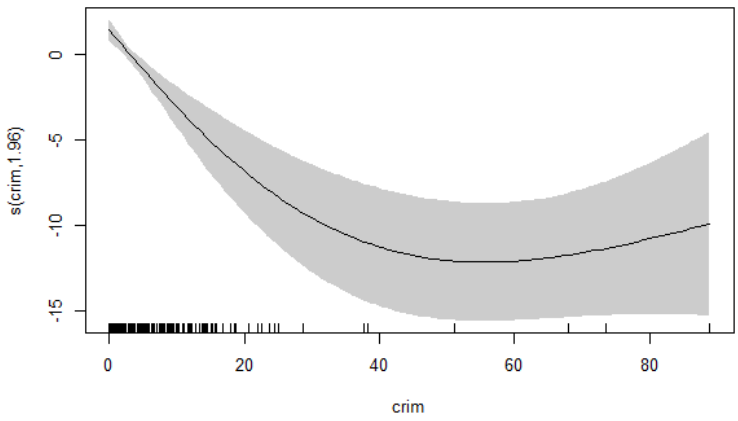
Variables used in the model:

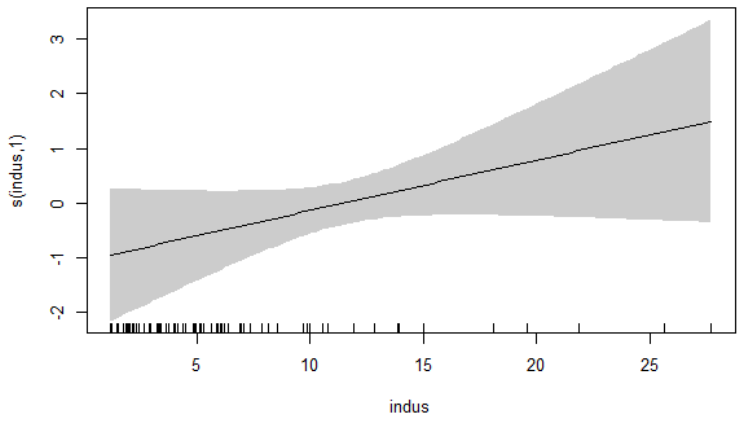
Continuous numerical predictors – crim, zn, indus, nox, rm, age, dis, rad, tax, ptratio, black, lstat

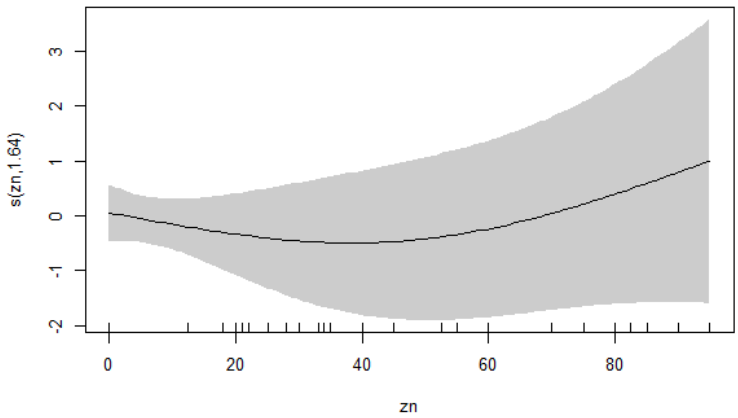
The model is trained using the training data and the testing data is used for validation of the additive model. For each of the variables, the spline plot is obtained as shown in figure a2. We see that the transformations have rendered the variables non-linear.

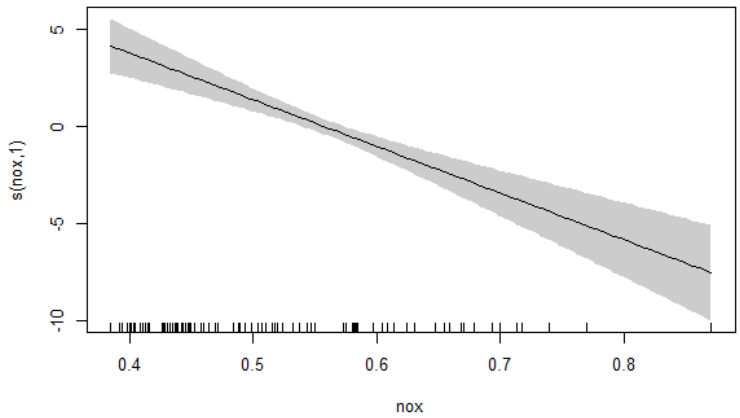
Upon generating the additive model, it is validated with the training dataset. The predicted medv values are then used to calculate the average sum squared error using the med values in the testing set.

Average sum squared error = 18.86 **Average sum squared error =** 14.13









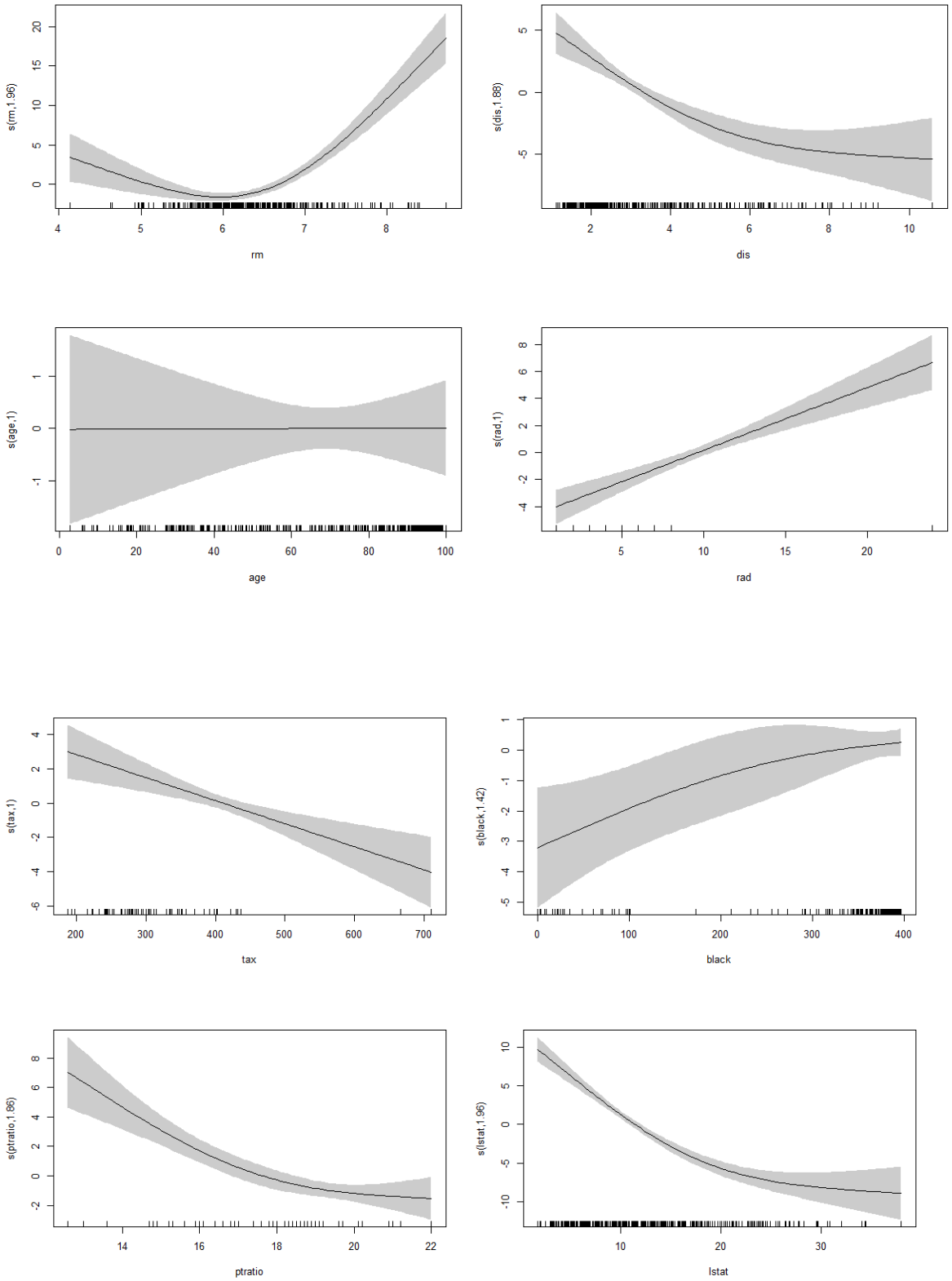


Figure a2: Spline plots of variables

Model 3: Neural Networks:

Artificial neural networks is a machine learning process that is used as a form of model fitting. They are inspired by the biological neural networks and are used to approximate functions that use a large number of inputs. They consist of an input layer, a set of hidden layers with a combination of hidden nodes and an output layer. The hidden layers and nodes are used to train the model with the given data. Similar to the previous models, artificial neural networks are also trained with the training data set from the Boston Housing data and validated using the testing dataset.

Specifications of the neural network used:

Input layer: Set of all input predictor variables

Hidden layer: Number of layers = 1, Maximum number of iterations = 500

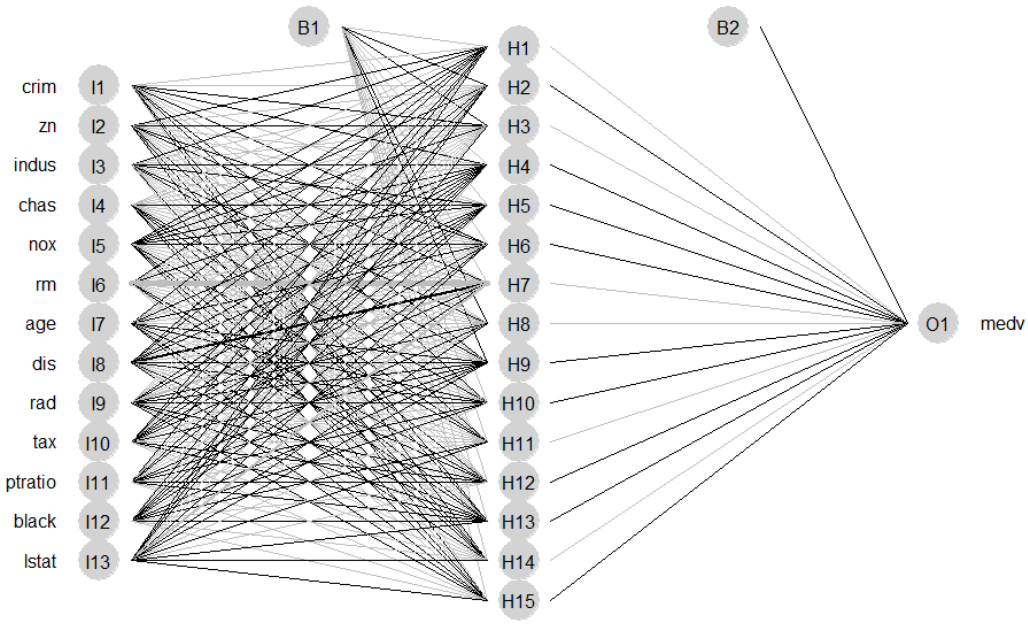


Figure a3: Artificial Neural Network

Upon creating the network model, the predicted medv values are generated using the predictor values from the testing dataset. The average sum squared error is then calculated.

Average SSE out of sample = 26.91 **Average SSE in sample =** 20.63

Model 4: Regression Tree

Another way of creating a model is by employing a regression tree. A regression tree is generated with the predictor variables (both continuous and categorical) as input with each tree node acting as a decision node. The terminals of the regression tree contains the predicted outputs.

Pruning the tree:

Pruning of the tree is necessary to avoid overfitting and to obtain a minimum average sum squared error. The initial tree is generated with a Cp value of 0.001 so as to obtain a large tree.

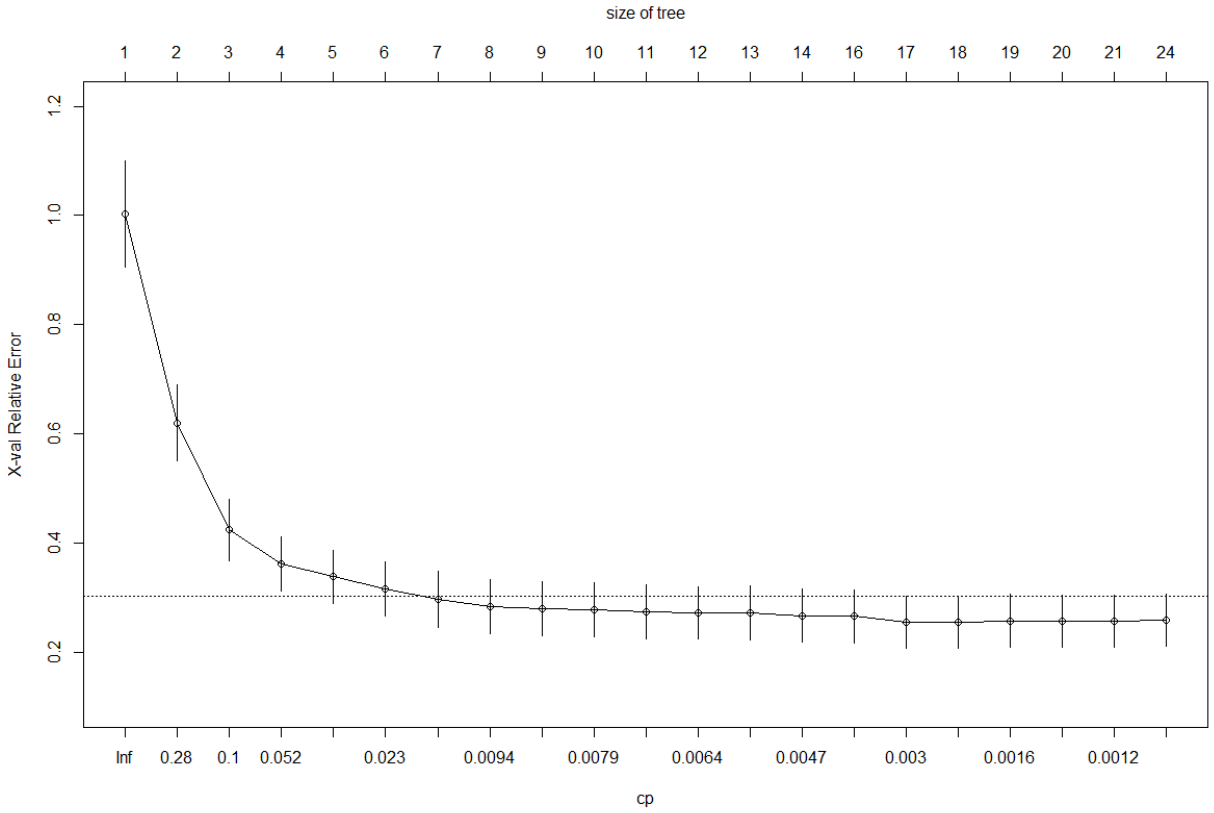


Figure a4: Plot of Cp values with Relative Error

The leftmost point of the graph below the horizontal line (one standard error above the most minimum value) is chosen as the optimum Cp value. In this case, the leftmost point is a Cp of 0.01

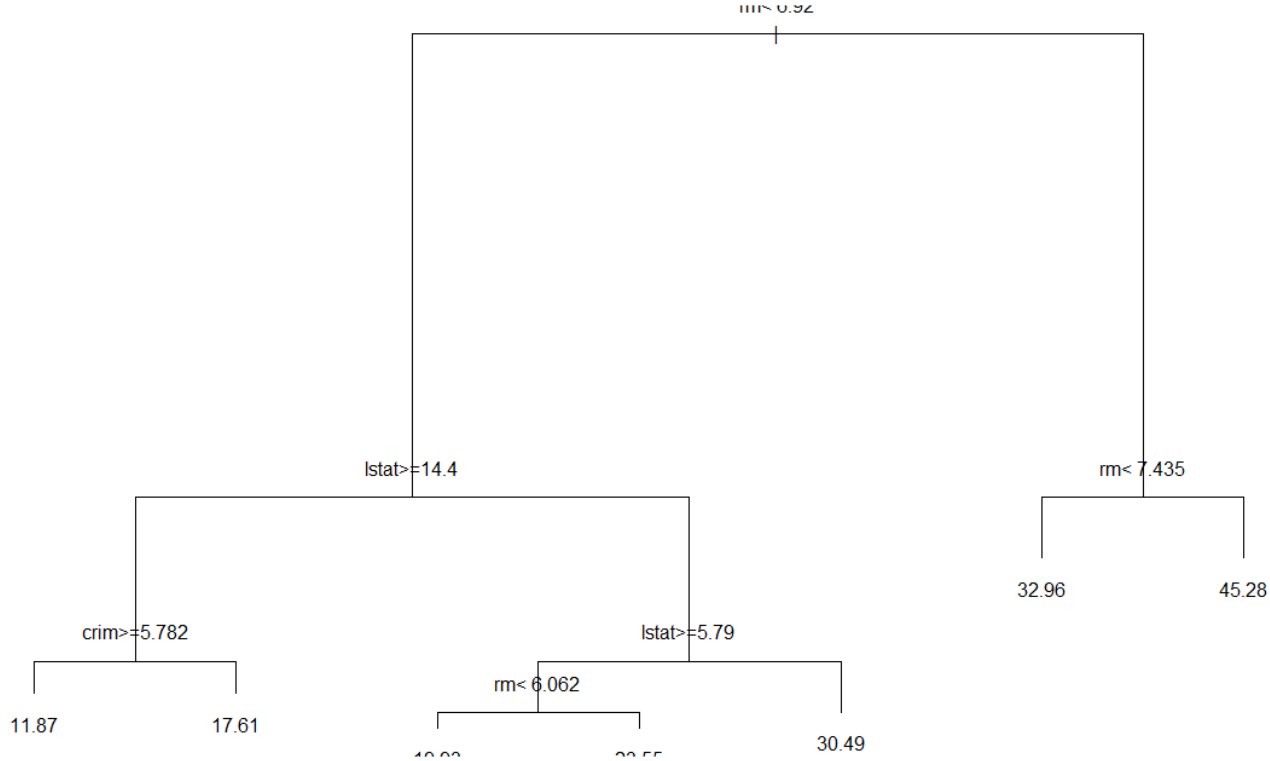
The regression tree is regenerated with this new Cp value, with the training data set.

Textual representation of the tree:

```
1) root 354 30376.0800 22.48051
   2) rm< 6.92 303 12489.2300 19.86271
     4) lstat>=14.4 123  2497.2060 14.90569
       8) crim>=5.7819 58   622.4159 11.87241 *
       9) crim< 5.7819 65   864.9702 17.61231 *
     5) lstat< 14.4 180  4904.3900 23.25000
      10) lstat>=5.79 158  2477.7660 22.24241
```

```
      20) rm< 6.0625 57   362.7467 19.93333 *
      21) rm>=6.0625 101  1639.5900 23.54554 *
   11) lstat< 5.79 22  1114.1860 30.48636 *
3) rm>=6.92 51  3474.0130 38.03333
  6) rm< 7.435 30  1178.2900 32.96333 *
  7) rm>=7.435 21   422.9381 45.27619 *
```

Visual representation of the Regression Tree:



Upon creating the regression tree, it is used to predict the medv values using the testing data set. The average sum squared values for the regression tree is obtained by comparing the predicted values with the observed values.

Average SSE out of sample = 27.56 **Average SSE in sample =** 21.92

Conclusion:

Various types of models were used to predict the medv values of the Boston dataset and their average sum squared errors were calculated in order to compare their effectiveness in prediction.

- The Generalized Additive Model provided the best prediction among the four models with an average sum squared error value of **18.86**