2017

Boston Housing Data Analysis

SRIHARSHA SURINENI

Summary:

Boston Housing dataset from the R library "MASS", is analyzed using various techniques such as generalized linear model, Regression Tree, Generalized additive model and Neural Network and their performance are compared. This report starts with basic exploratory plots of all the variables and correlation among them. There were no missing values or any other abnormalities observed in the data. Significant correlations were found for the response variable with the predictor variables. Models were built using the above mentioned techniques and their performances were observed as follows:

Mean squared error:

```
##                glm     tree       gam  gamfinal         nnet
## train mse 21.91861 13.55013  7.488375  7.490935 2.977359e-07
## test mse  23.94790 18.18418 12.818108 12.817956 1.550811e+01
```

It has been found that neural network performed better both in case of in sample and out of sample predictions. But, this improvement does come at the cost of intense computational requirement in both learning the weights and finding right parameters for the neural network. Next best approach is generalized additive model, though its performance is not too different to that of the regression tree in this specific case. Glms performed worse among all the techniques

Boston Housing data:

The following analysis is carried out on popular data set from the "MASS" library - "Boston Housing Data", comprising of various metrics likely affecting the housing prices of a region like per capita crime rate, average number of rooms, nitrogen oxides concentration etc., and the median value of owner - occupied homes in thousands of dollars. The major objective of this analysis is to model the effects of these factors on the median of house prices using different techniques like generalized linear regression, tree, generalized additive models and neural network, arrive at the best possible models using these techniques and compare their in-sample and test sample performances. Performance metric used to compare the performance – Mean Squared Error.

The following analysis starts with exploratory analysis of the data to find major abnormalities like missing values, the distributions of all the variables and correlations among the variables. Next step is to sample the data into 75% train and 25% test samples, build models using the above mentioned techniques and selecting the best possible model for each technique and compare the performance of these models both in-sample and out of sample (test) .Deduce major conclusions from the observations.

Methodology:

- Summary and exploratory analysis of the data
- Random sample the data - 75% train and 25% test sets
- Modeling of the data using above mentioned techniques
- Comparison of their performance
- Conclusion

Exploratory analysis of the data:

Summary:

```
##       crim                zn             indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
##       nox               rm             age              dis        
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
##       rad              tax           ptratio          black       
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38  
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44  
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67  
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23  
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90  
##      lstat            medv
##  Min.   : 1.73   Min.   : 5.00
##  1st Qu.: 6.95   1st Qu.:17.02
##  Median :11.36   Median :21.20
##  Mean   :12.65   Mean   :22.53
##  3rd Qu.:16.95   3rd Qu.:25.00
##  Max.   :37.97   Max.   :50.00
```

Correlation :

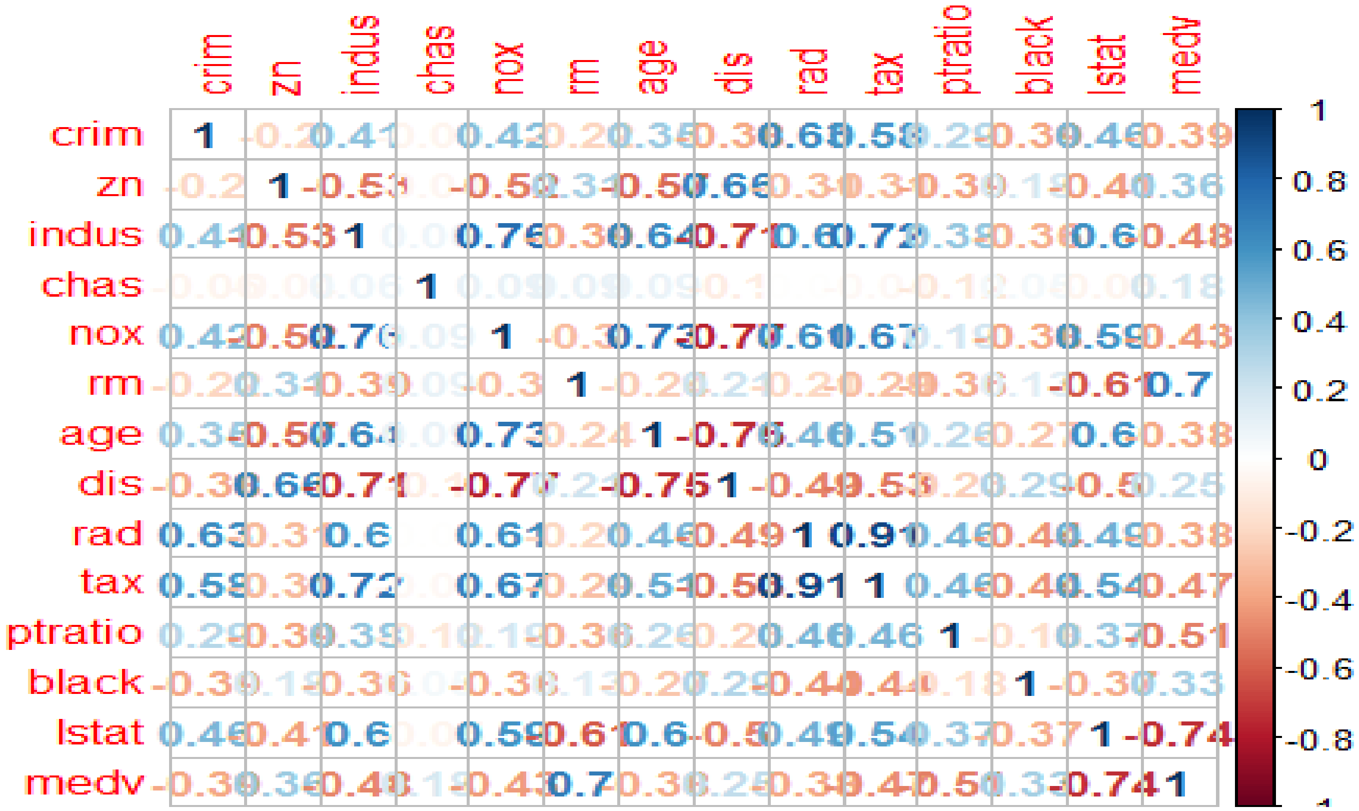


Figure 1: Correlation Plot

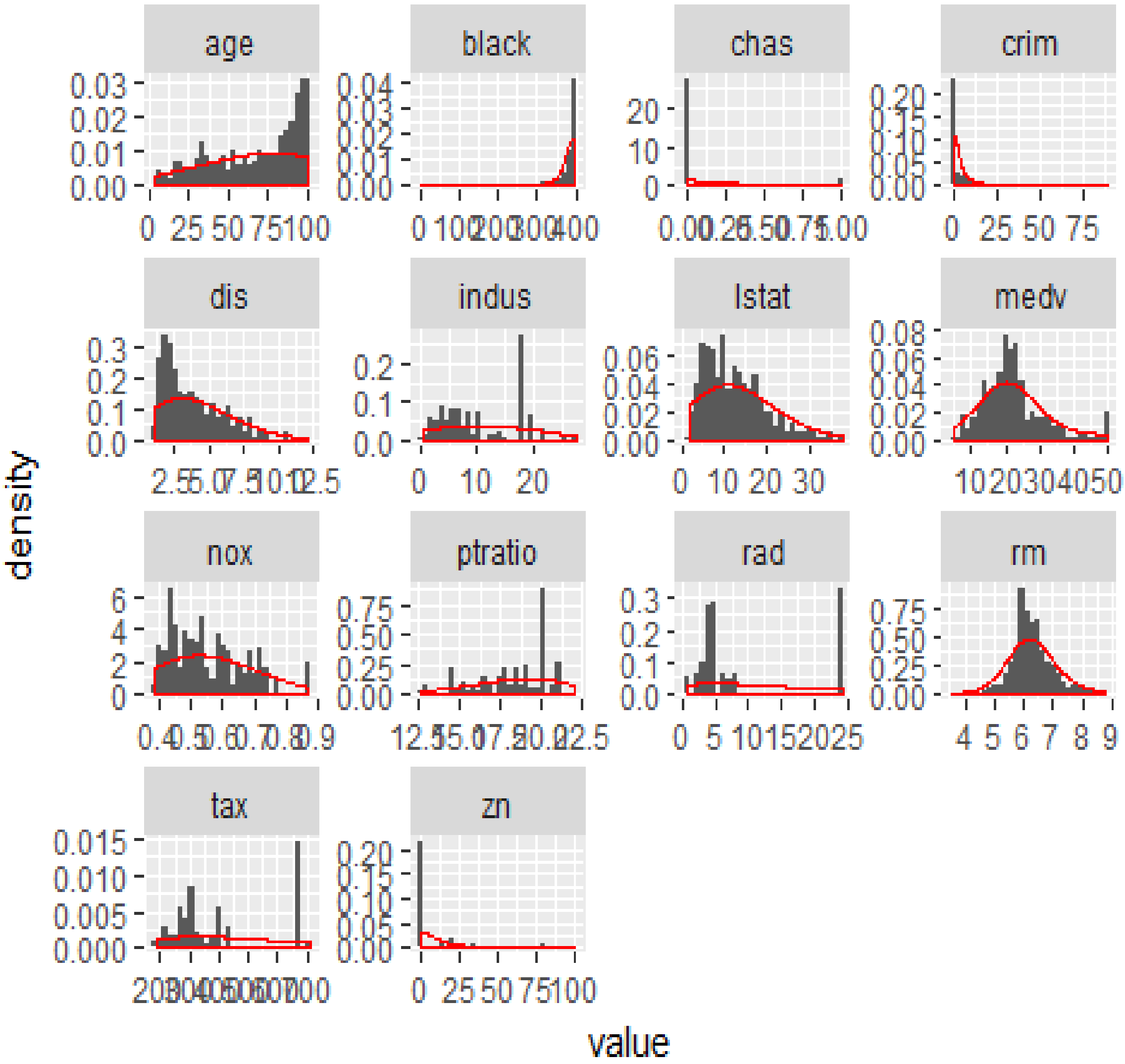
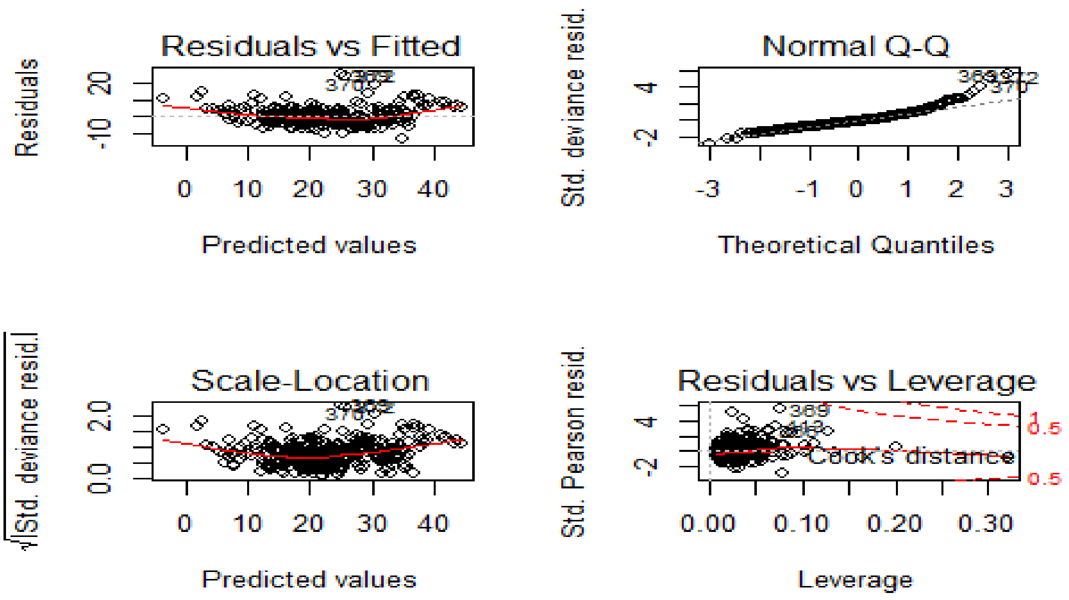


Figure 2: Histogram and density pots

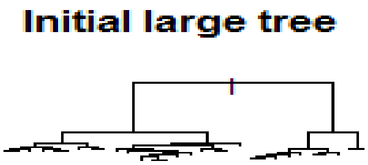
- There are no missing values in the data.
- Response variable is clearly correlated with most of the predictor variables, especially 'lstat', 'rm', 'ptrratio'
- There is not much variation in some of the variables such as black, crim, zn etc., chas is a factor variable

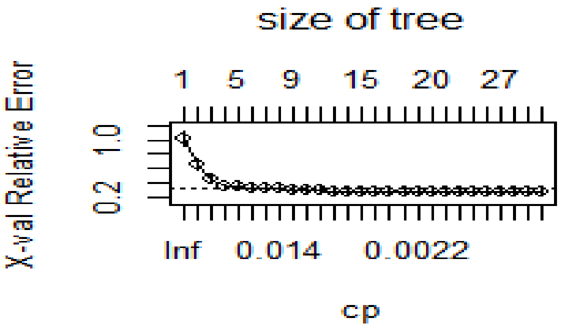
Generalized Linear Model:

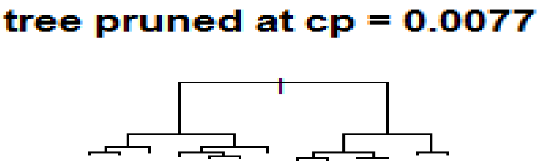


There is an obvious trend in the residual vs fitted value plot. Transforming response variable might help.

Tree:



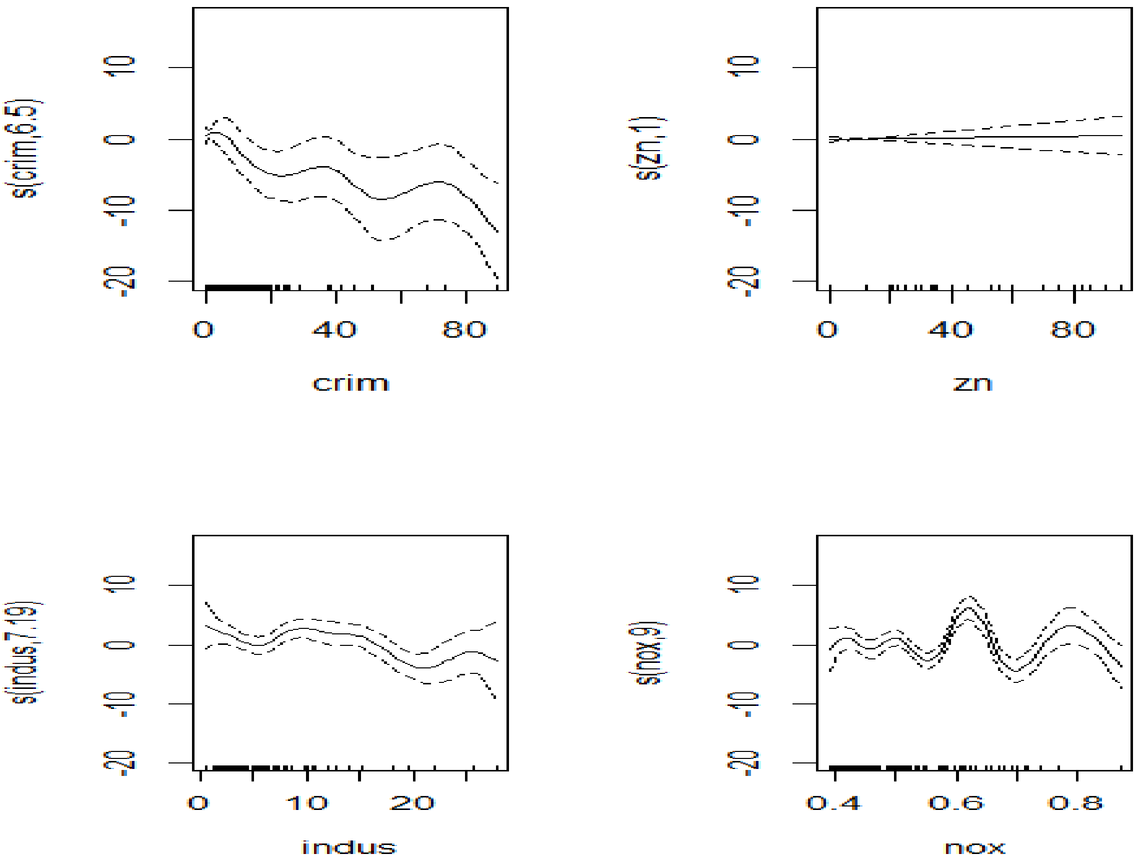


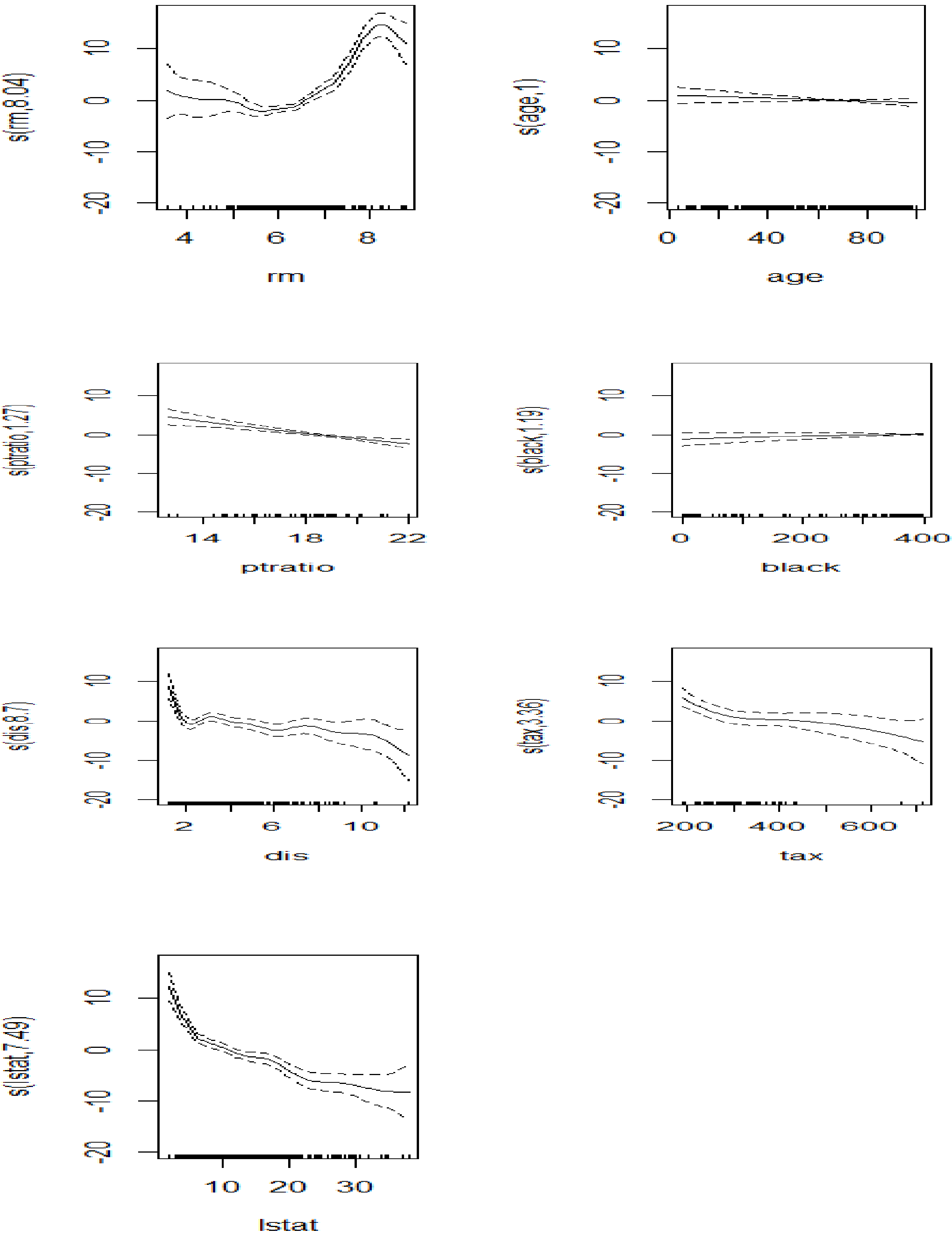


Initial tree is pruned at cp of 0.0077 to obtain final tree model. There is not much difference in performance of both the models as such. So, the pruning of the tree is justified.

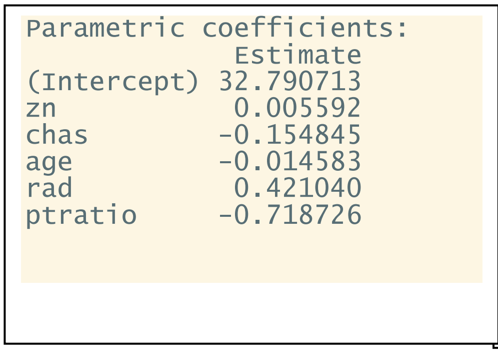
GAM model:

Generalized additive model, modeled using splines of quantitative predictor variables.

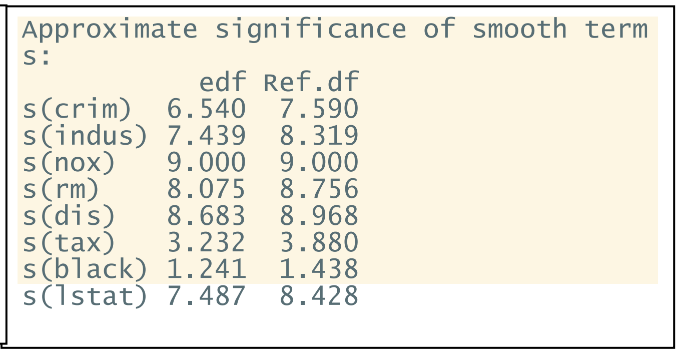



s(rm,8.04)
-20
-10
0
10
4
6
8
rm
s(age,1)
-20
-10
0
10
0
40
80
age
s(ptratio,1.27)
-20
-10
0
10
14
18
22
ptratio
s(black,1.19)
-20
-10
0
10
0
200
400
black
s(dis,8.7)
-20
-10
0
10
2
6
10
dis
s(tax,3.36)
-20
-10
0
10
200
400
600
tax
s(lstat,7.49)
-20
-10
0
10
10
20
30
lstat

Modeling gam, considering spline function for some of the variables is not justified as the plots of the spline functions are linear. After removing such variables (ptratio, age and zn), final model obtained:



```
Parametric coefficients:
             Estimate
(Intercept) 32.790713
zn           0.005592
chas        -0.154845
age         -0.014583
rad          0.421040
ptratio     -0.718726
```



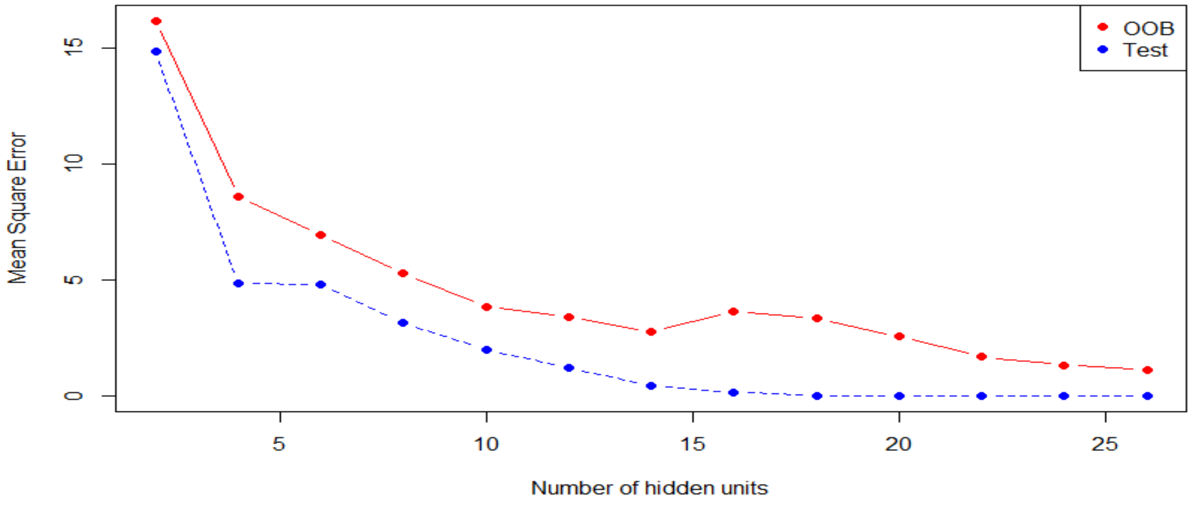
```
Approximate significance of smooth term
s:
            edf Ref.df
s(crim)   6.540  7.590
s(indus) 7.439  8.319
s(nox)    9.000  9.000
s(rm)     8.075  8.756
s(dis)    8.683  8.968
s(tax)    3.232  3.880
s(black) 1.241  1.438
s(lstat) 7.487  8.428
```

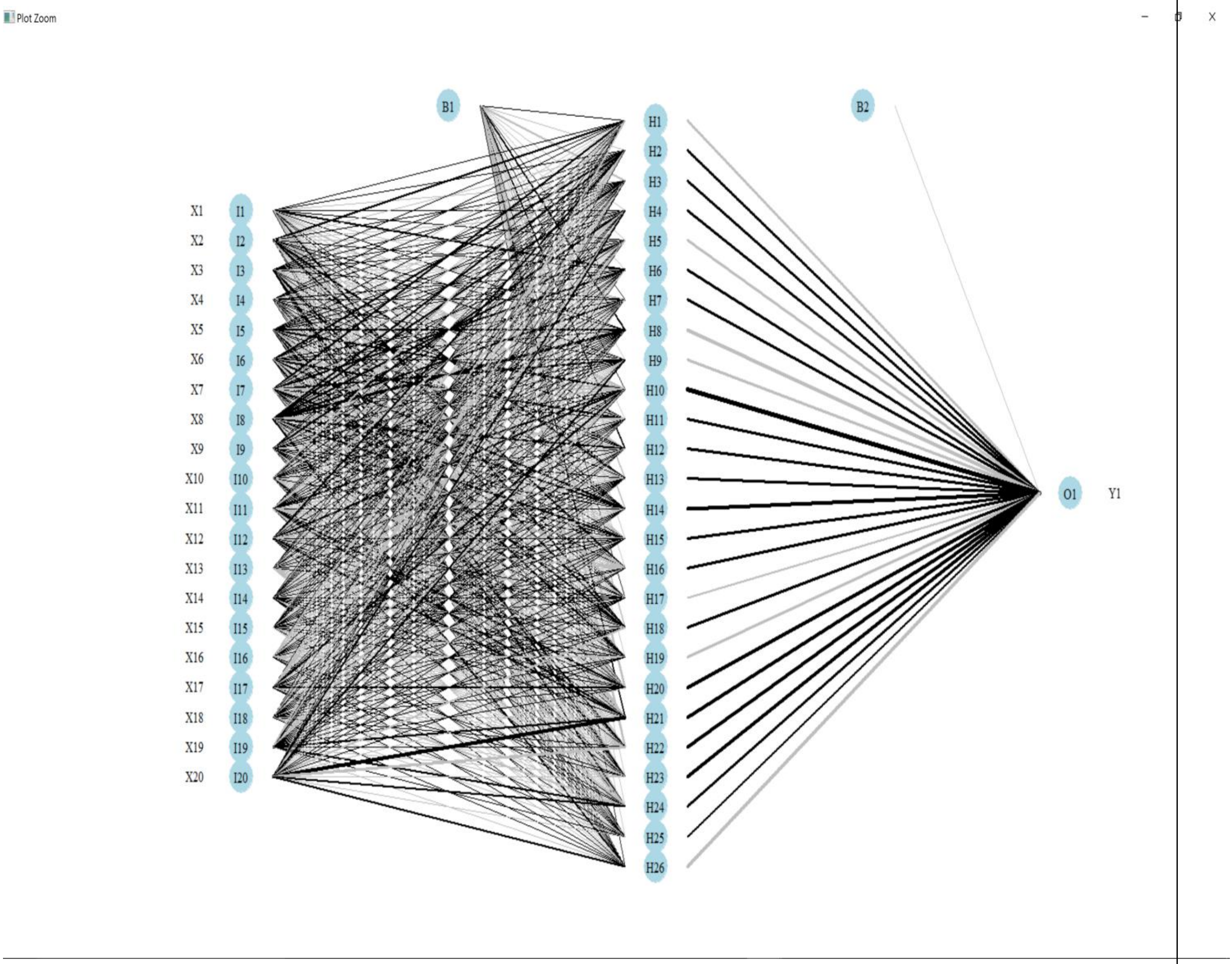
Neural network:

Neural network parameters:

- Hidden layer size – 15
- Learning rate – 0.0001
- Maximum iterations – 8000

For the given learning rate and number of maximum iterations, hidden layer size of 15 is obtained by plotting average mean squared error of training set and cross validation set (20% random sample of training data set).





Mean squared error:

```
##                   glm     tree       gam  gamfinal         nnet
## train mse 21.91861 13.55013  7.488375  7.490935 2.977359e-07
## test mse  23.94790 18.18418 12.818108 12.817956 1.550811e+01
```

Conclusion:

- Neural network with one hidden layer of size 19 and learning rate of 0.0001 seems to perform better both in-sample and out sample
- Next to neural network, generalized additive model is slightly better than pruned regression tree(at cp = 0.0071)
- Generalized linear model (Gaussian) is the worst performer among all four techniques
- Mean squared error reduced considerably from 23 to 1.55 on out of sample data