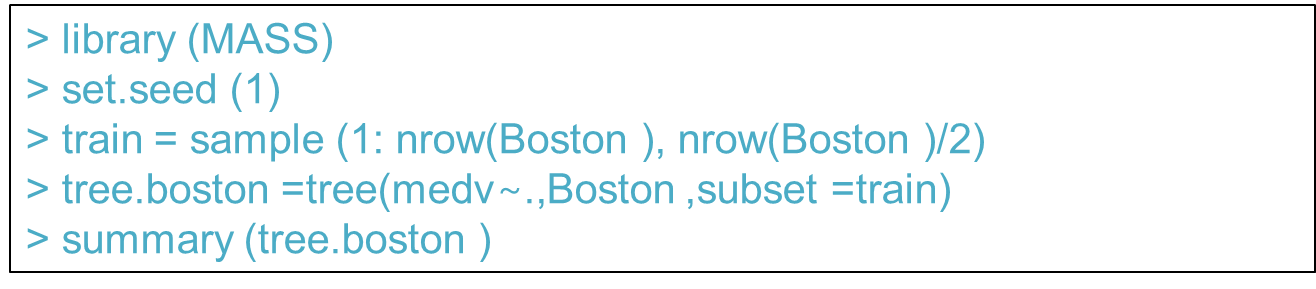
|  |  |
| --- | --- |
| Big Data and Machine Learning in Logistics | 2nd semester 2021 |
| Student number: 502115907 | Student name: 놀웬 피지언 Nolwenn PIGEON |

**R Self-work submission 4**

1. Regression tree with Boston dataset

Please follow the instruction in R-lab session in Week 12 and present the output from the r-work below



Answer)

Regression tree:

tree(formula = medv ~ ., data = Boston, subset = train)

Variables actually used in tree construction:

[1] "rm" "lstat" "crim" "age"

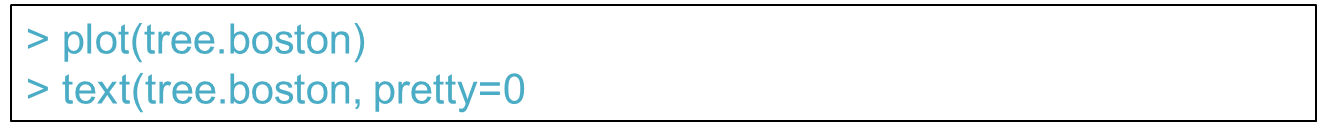
Number of terminal nodes: 7

Residual mean deviance: 10.38 = 2555 / 246

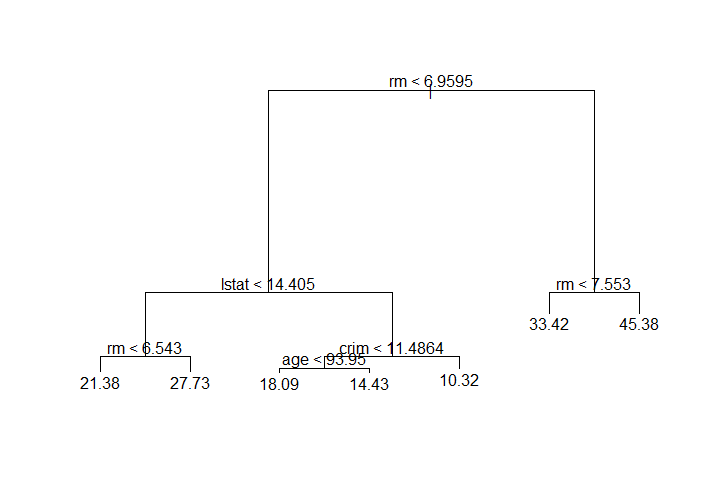
Distribution of residuals:

Min. 1st Qu. Median Mean 3rd Qu. Max.

-10.1800 -1.7770 -0.1775 0.0000 1.9230 16.5800



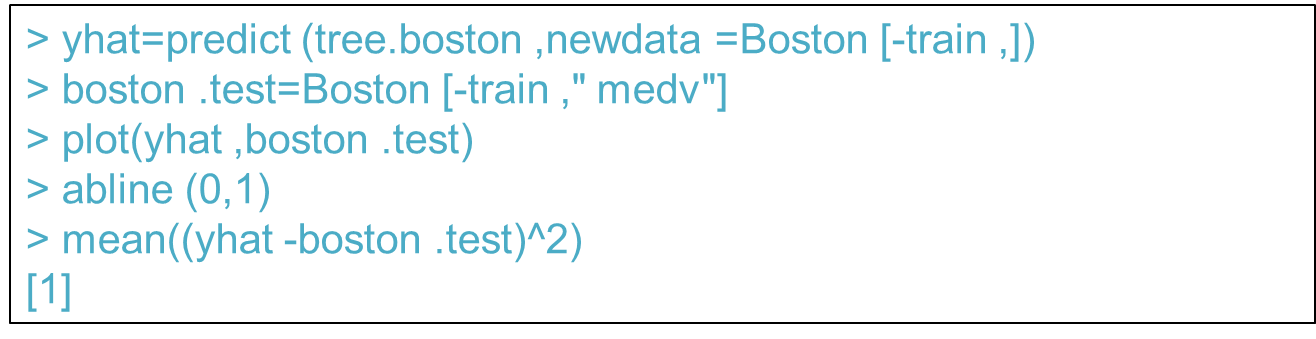
Answer)



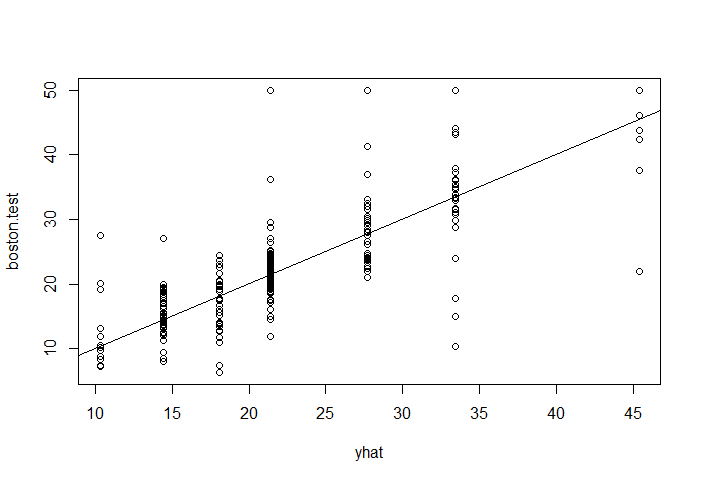
Also Include possible interpretation of the tree

the variable lstat measures the percentage of individuals with lower socioeconomic status. The tree indicates that lower values of lstat correspond to more expensive houses.

The tree predicts a median house price of $45.38 for larger home in suburbs in which residents have high socioeconomic status (rm<6.96 and rm<7.553)



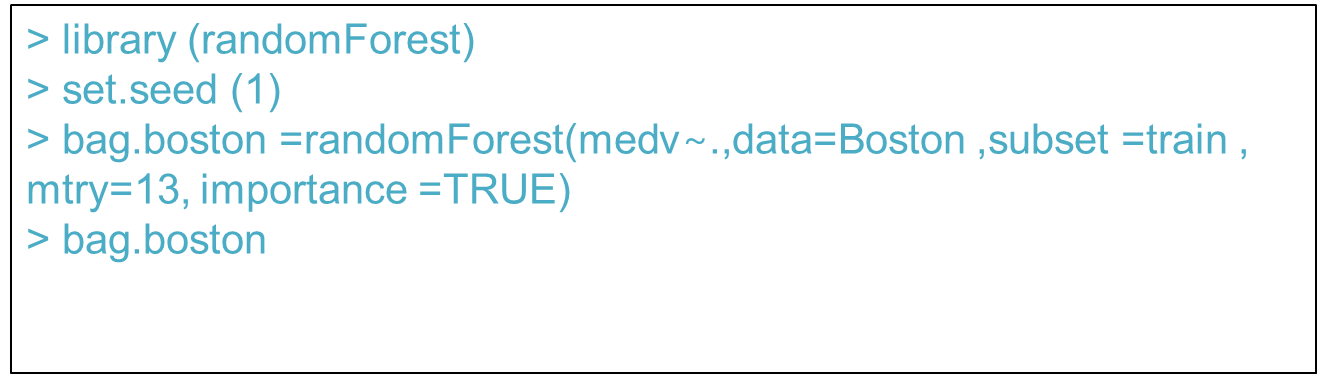
Answer)



Mean = 35.28688

In other words, the test set MSE associated with the regression tree is 35.28. The square root of the MSE is therefore around 5.94, indicating that this model leads to test predictions that are within around $5.94 of the true median home value for the suburb.

2. Bagging and Random forest with Boston dataset



Answer)

Call:

randomForest(formula = medv ~ ., data = Boston, mtry = 13, importance = TRUE, subset = train)

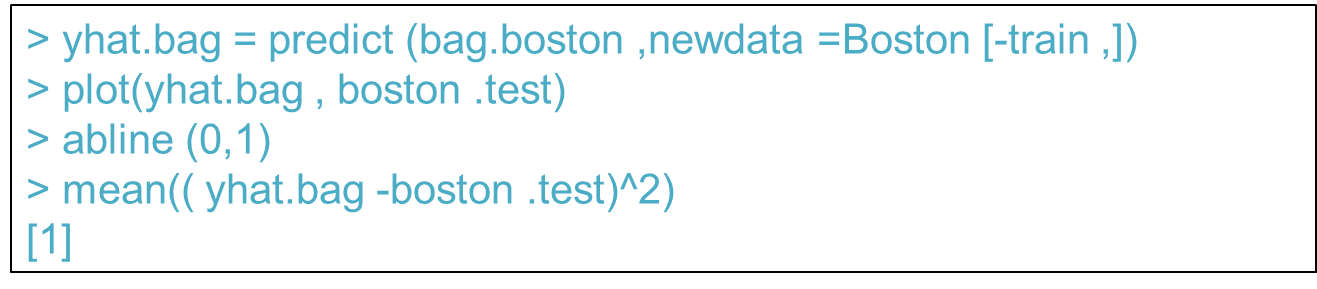
Type of random forest: regression

Number of trees: 500

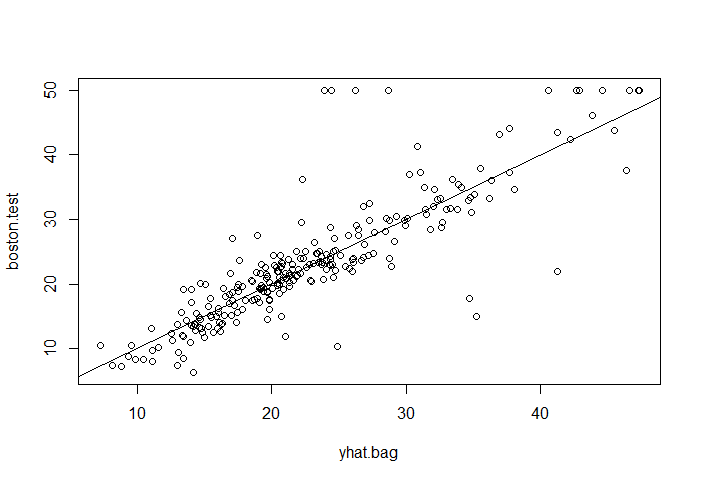
No. of variables tried at each split: 13

Mean of squared residuals: 11.39601

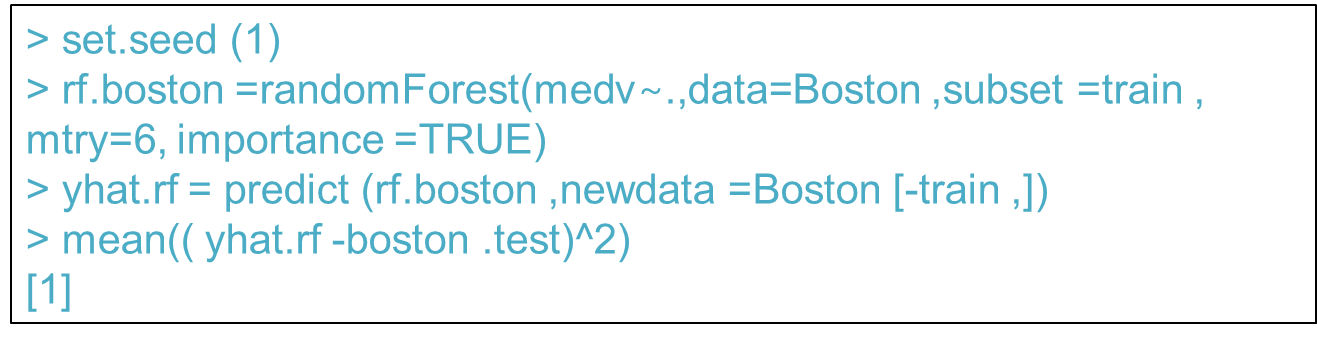
% Var explained: 85.17



Answer)

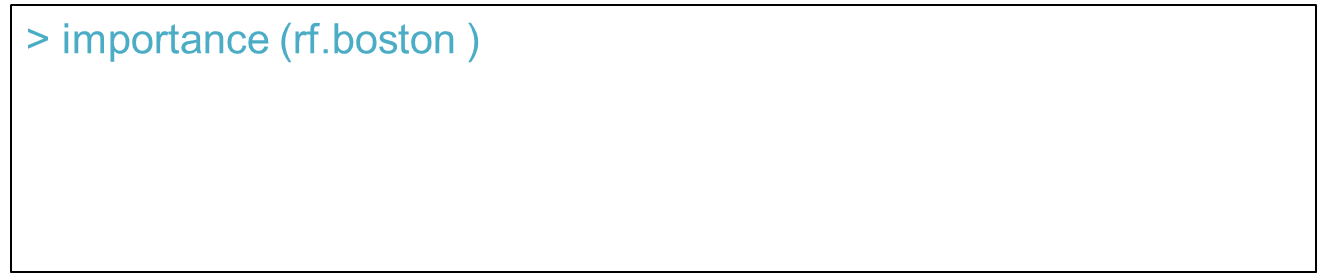


Mean =23.59273



Answer)

Mean = 19.62021



Answer)

%IncMSE IncNodePurity

crim 16.697017 1076.08786

zn 3.625784 88.35342

indus 4.968621 609.53356

chas 1.061432 52.21793

nox 13.518179 709.87339

rm 32.343305 7857.65451

age 13.272498 612.21424

dis 9.032477 714.94674

rad 2.878434 95.80598

tax 9.118801 364.92479

ptratio 8.467062 823.93341

black 7.579482 275.62272

lstat 27.129817 6027.63740 The 2 most important predictors are rm and lstat, if we leave them out then the accuracy decreases of 32% and 27% respectively.

3. Model comparison

Make comparison of the three methods in terms of prediction accuracy.

Answer)

In terms of prediction accuracy, if we compare the mean of the three methods, we observe a decrease in the mean values. In the first method, using regression tree the mean is about 35% accuracy.

Using the bagging method, mean goes down to 24% of accuracy

Finally, in the random forest method the mean represents an accuracy of 20%.

2 measures of variable importance are reported. The former is based upon the mean decrease of accuracy in predictions on the out of bag samples when a given variable is excluded from the model. The latter is a measure of the total decrease in node impurity that results from splits over that variable, averaged over all trees. In the case of regression trees, the node impurity is measured by the training RSS, and for classification trees by the deviance.