Recursive Partitioning & Random Forest

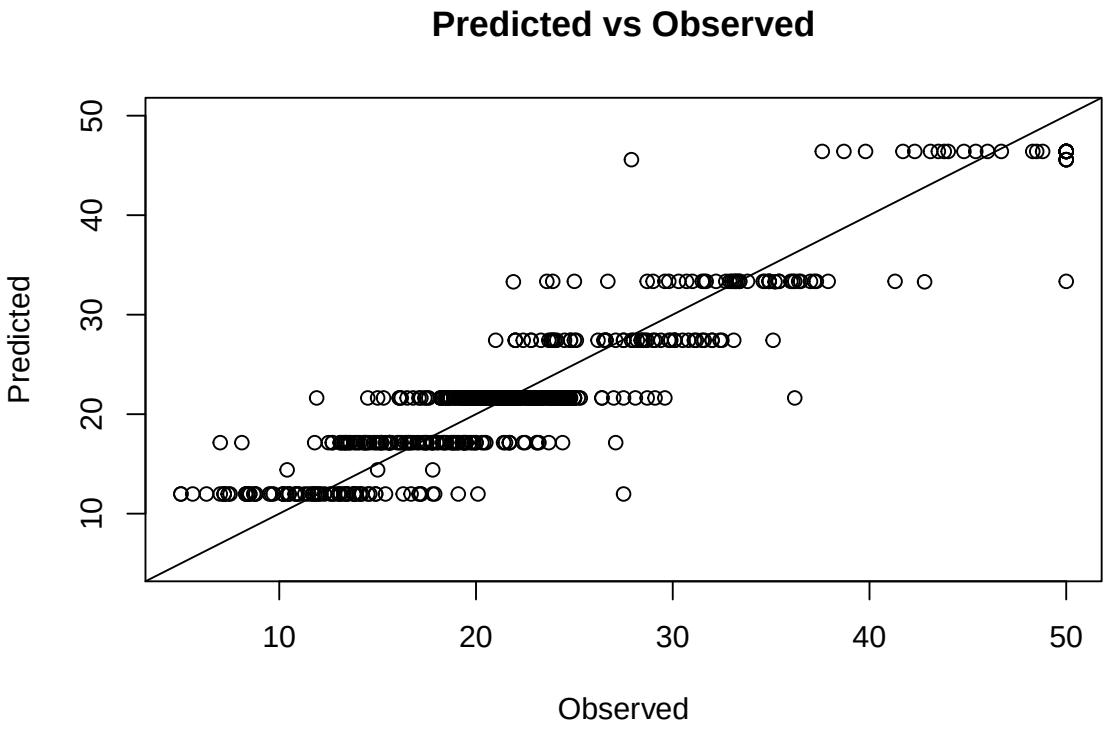
Naga Pakalapati

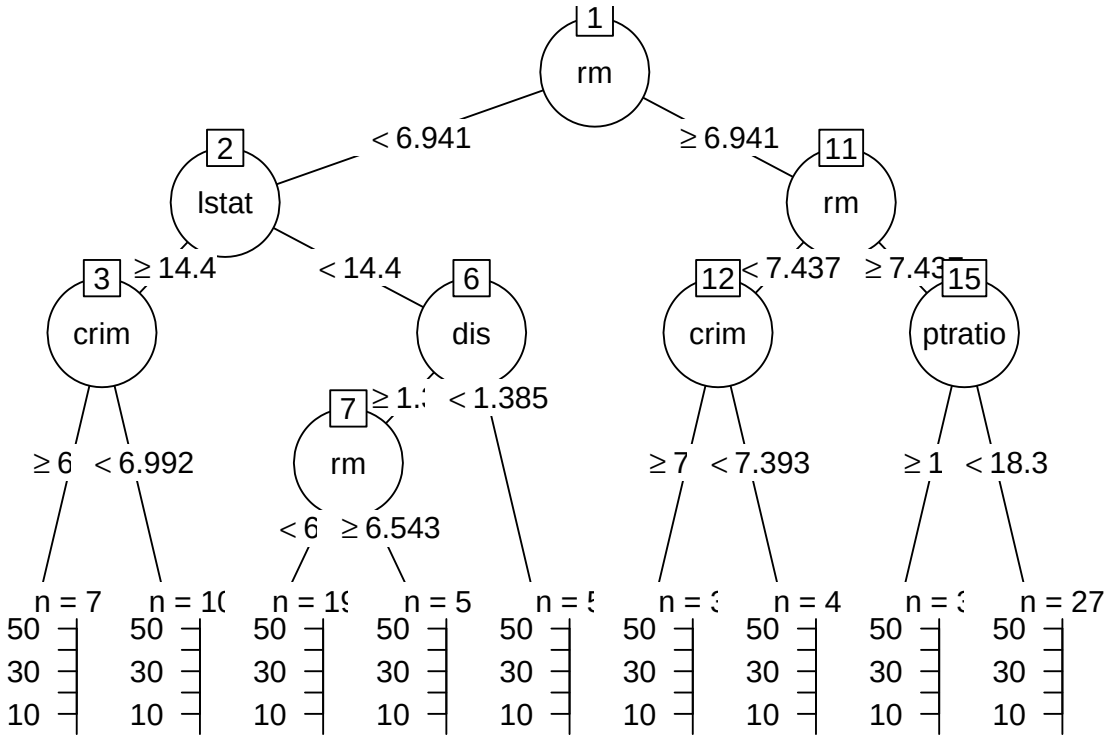
Please do the following problems from the text book R Handbook and stated.

1. The **BostonHousing** dataset reported by Harrison and Rubinfeld (1978) is available as data.frame package **mlbench** (Leisch and Dimitriadou, 2009). The goal here is to predict the median value of owner-occupied homes (medv variable, in 1000s USD) based on other predictors in the dataset. Use this dataset to do the following

   a.) Construct a regression tree using rpart(). The following need to be included in your discussion. How many nodes did your tree have? Did you prune the tree? Did it decrease the number of nodes? What is the prediction error (calculate MSE)? Provide a plot of the predicted vs. observed values. Plot the final tree.

```
## Prediction error:  12.71556
```





Regression tree is constructed using all variabels resulted in 9 nodes with rm (average number of rooms per dwelling) as the root node and lstat, crim, dis, ptratio, rm variables used in the others decison test nodes.

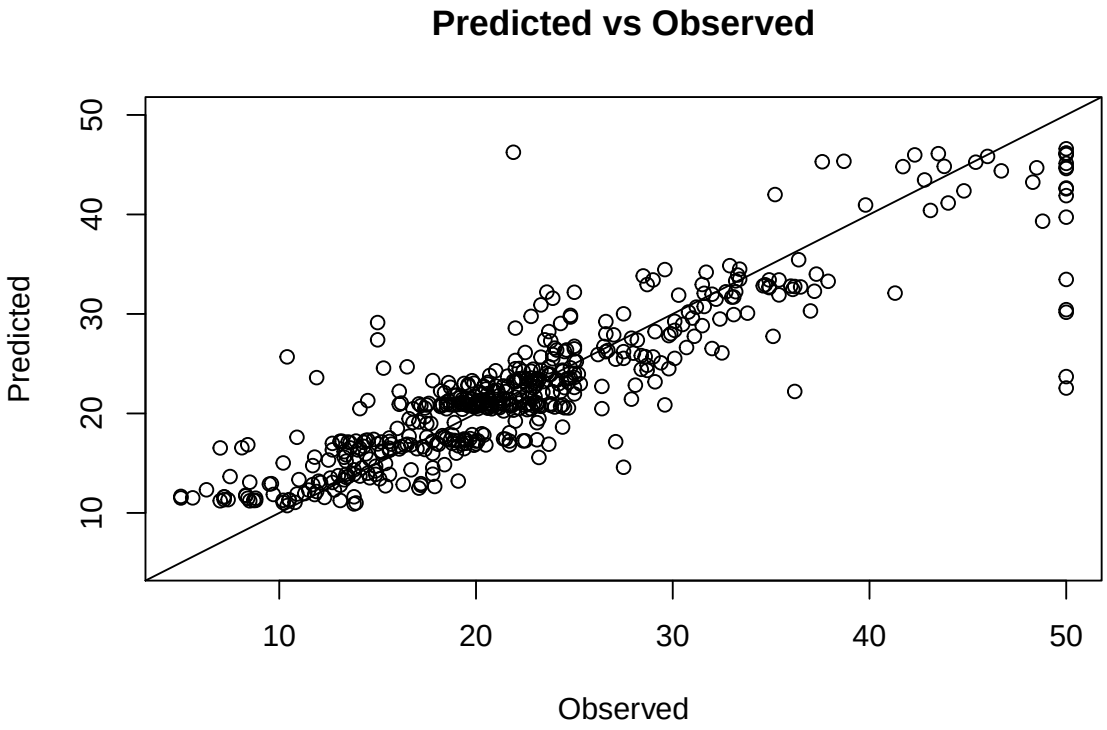
After pruning the tree there is no change in the number of nodes.

Using the original tree we have made predicition and plotted observed vs fitted plot. The prediction error we achieved is 12.7.

b) Perform bagging with 50 trees. Report the prediction error (MSE). Provide the predicted vs observed plot.

```
##
## lstat    rm
##    25    25
```

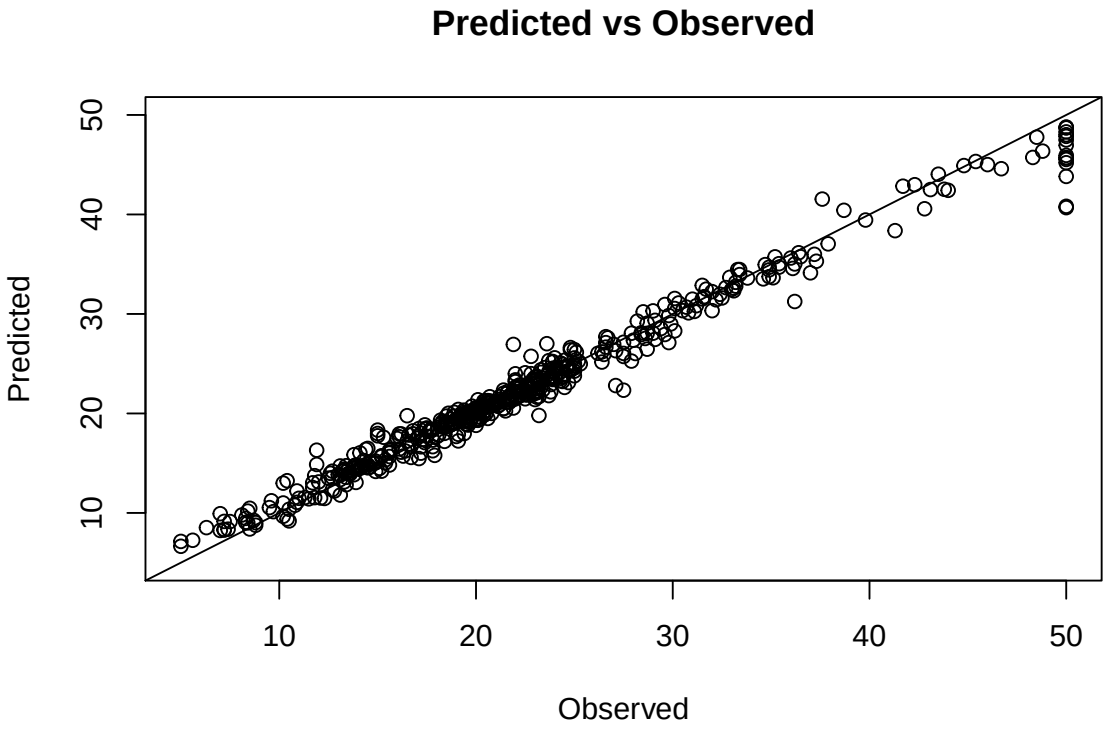
```
## Prediction error:  18.23437
```



We can see that the prediction error increased after performing bagging.

c) Use randomForest() function in R to perform bagging. Report the prediction error (MSE). Was it the same as (b)? If they are different what do you think caused it? Provide a plot of the predicted vs. observed values.

```
## Prediction error:  1.947255
```

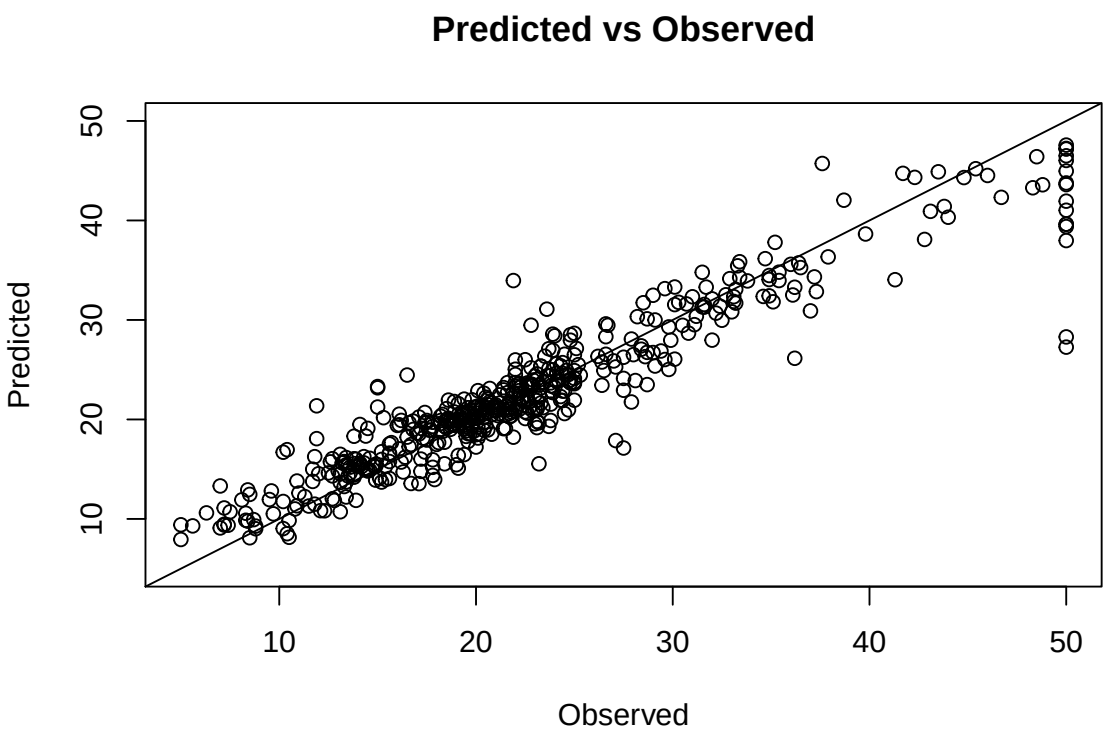


Using the **randomForest** function for bagging the prediction error (MSE) is minimized significantly compared to earlier model. Primary reason could be the number of trees grown in each iteration is **500** in this model compared to only **1** for each iteration usign **rpart** model.

d) Use randomForest() function in R to perform random forest. Report the prediction error (MSE). Provide a plot of the predicted vs. observed values.

```
## Prediction error:  10.50507
```

```
## Number of trees grown:  500
```



e) Provide a table containing each method and associated MSE. Which method is more accurate?

	mse
rpart_mse	12.715559
rpart_bg_mse	18.234375
rf_mse	10.505066
rf_bg_mse	1.947256

2. Consider the glacoma data (data = "**GlaucomaM**", package = "**TH.data**").

a) Build a logistic regression model. Note that most of the predictor variables are highly correlated. Hence, a logistic regression model using the whole set of variables will not work here as it is sensitive to correlation.

```
glac_glm <- glm(Class ~., data = GlaucomaM, family = "binomial")
#warning messages  -- variable selection needed
```

The solution is to select variables that seem to be important for predicting the response and using those in the modeling process using GLM. One way to do this is by looking at the relationship between the response variable and predictor variables using graphical or numerical summaries - this tends to be a tedious process. Secondly, we can use a formal variable selection approach. The *step*() function will do this in R. Using the *step* function, choose any direction for variable selection and fit logistic regression model. Discuss the model and error rate.

```
#use of step() function in R
?step
glm.step <- step(glac_glm)
```

Do not print out the summaries of every single model built using variable selection. That will end up being dozens of pages long and not worth reading through. Your discussion needs to include the direction you chose. You may only report on the final model, the summary of that model, and the error rate associated with that model.

```
##
## Call:
## glm(formula = Class ~ ag + at + as + eag + eat + eas + ean +
##     eai + abrs + mhcg + mhct + mhcn + phci + hvc + vbsg + vbst +
##     vbss + vbsn + vbsi + vasg + vasn + vasi + vbrg + vbrt + vbrs +
##     vbrn + vbri + mdg + mdt + tmg + tmt + tmn + rnf, family = "binomial",
##     data = GlaucomaM)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -3.286e-04  -2.000e-08   0.000e+00   2.000e-08   3.323e-04
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -7281     678525  -0.011    0.991
## ag             -7449     898506  -0.008    0.993
## at             11545    3178876   0.004    0.997
## as             19828    2797565   0.007    0.994
## eag          -267543   51263505  -0.005    0.996
## eat           288950   50720919   0.006    0.995
## eas           269866   50731884   0.005    0.996
## ean           264468   50442954   0.005    0.996
## eai           267976   52197620   0.005    0.996
## abrs           -9067     600973  -0.015    0.988
## mhcg           53348    2347442   0.023    0.982
## mhct          -39550    1641947  -0.024    0.981
## mhcn          -10319     652212  -0.016    0.987
## phci           -6055     787294  -0.008    0.994
## hvc             4580     623930   0.007    0.994
## vbsg          349255   34265139   0.010    0.992
## vbst         -487395   35691929  -0.014    0.989
## vbss         -331822   35209583  -0.009    0.992
## vbsn         -350225   33654952  -0.010    0.992
## vbsi         -310476   35049443  -0.009    0.993
## vasg           82428    6018140   0.014    0.989
## vasn          -98307   11989243  -0.008    0.993
## vasi          -80706    8160662  -0.010    0.992
## vbrg         -259478   18191860  -0.014    0.989
## vbrt          406283   22274158   0.018    0.985
## vbrs          241335   17695898   0.014    0.989
## vbrn          269489   18700673   0.014    0.989
## vbri          199868   16455453   0.012    0.990
## mdg            -4730     229696  -0.021    0.984
## mdt             6051     374384   0.016    0.987
## tmg           -10080     746012  -0.014    0.989
## tmt             5871     340626   0.017    0.986
## tmn             1504     343608   0.004    0.997
## rnf            32460    1527690   0.021    0.983
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2.7171e+02  on 195  degrees of freedom
## Residual deviance: 1.6180e-06  on 162  degrees of freedom
## AIC: 68
##
## Number of Fisher Scoring iterations: 25
```

b) Build a logistic regression model with K-fold cross validation (k = 10). Report the error rate.

```
## Error rate using 10-fold CV: 0.3050847
```

c) Find a function (package in R) that can conduct the "adaboost" ensemble modeling. Use it to predict glaucoma and report error rate. Be sure to mention the package you used.

```
##                Observed Class
## Predicted Class glaucoma normal
##        glaucoma       32      2
##        normal          4     21
```

```
## Error rate using adaboost: 0.1016949
```

d) Report the error rates based on single tree, bagging and random forest. (A table would be great for this).

err	or_rates
single_tree	0.5000000
bagging	0.1683673
randomForest	0.1428571

e) Write a conclusion comparing the above results (use a table to report models and corresponding error rates). Which one is the best model?

err	or_rates
glm	0.3050847
adaboost	0.1016949
single_tree	0.5000000
bagging	0.1683673
randomForest	0.1428571

The best performing model of all for this problem is adaboost with an error rate of 0.08.

f) From the above analysis, which variables seem to be important in predicting Glaucoma?

```
## Important variables contributing the most that are common in above models are:
##  hvc vars vari abrg vbst mdt mhci vass vart tms varg varn rnf vbri NA mhcg mhcn eas mhcs tmi pho
```