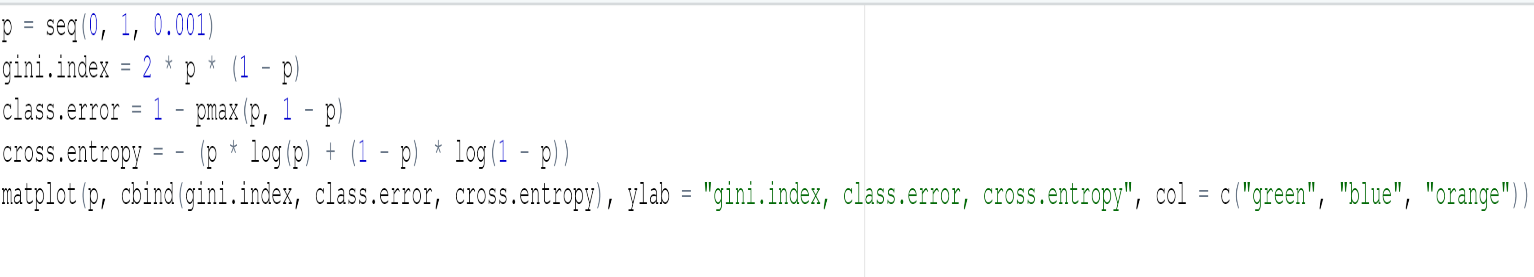
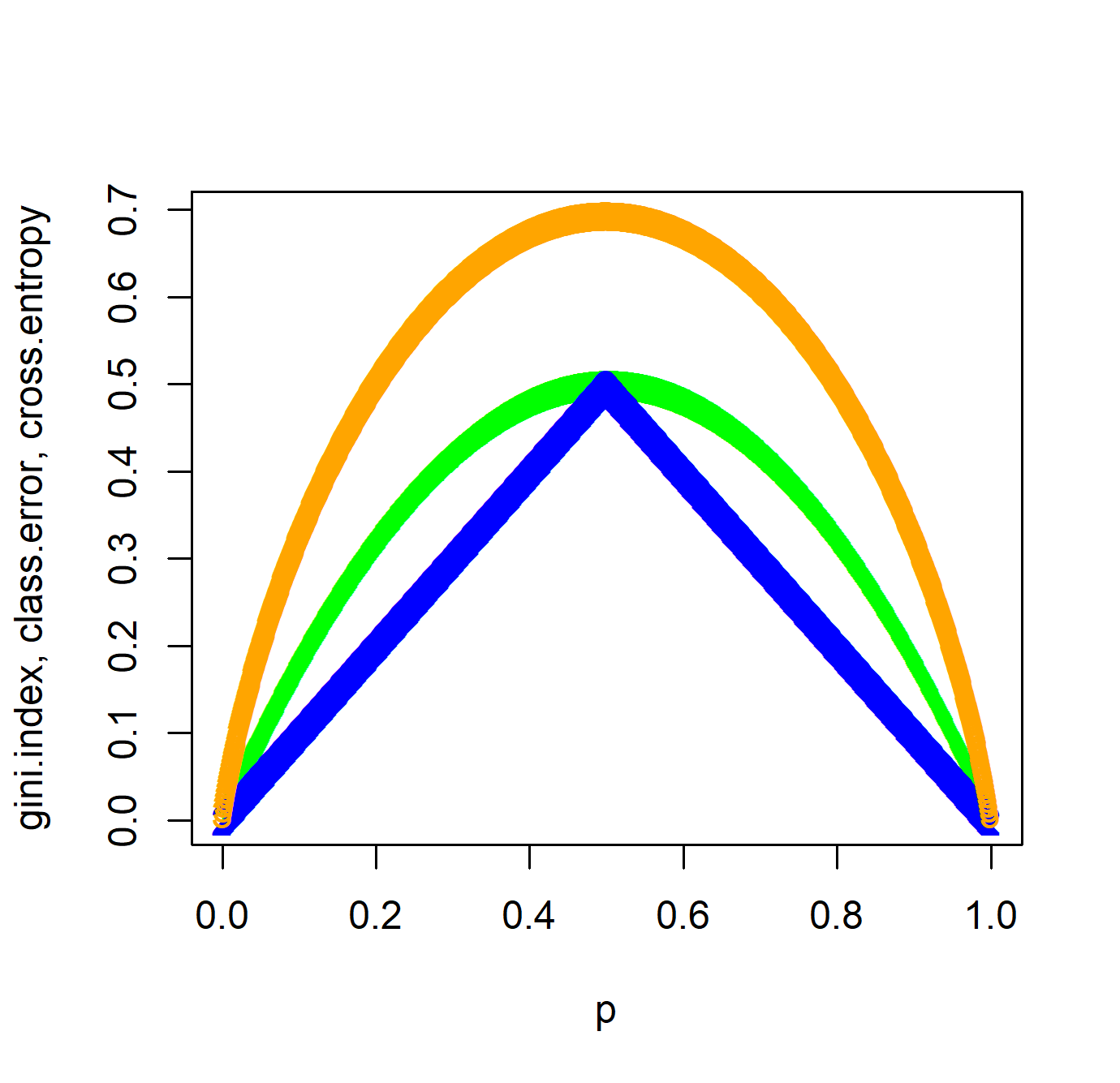
3)

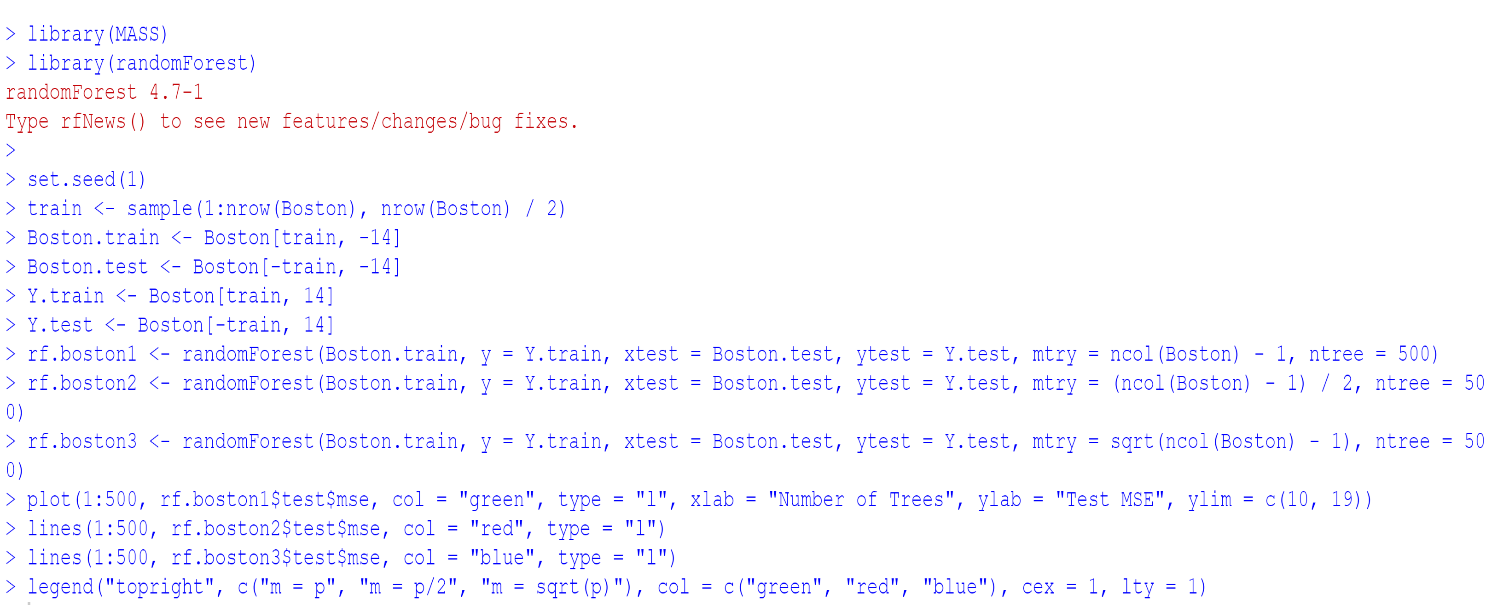


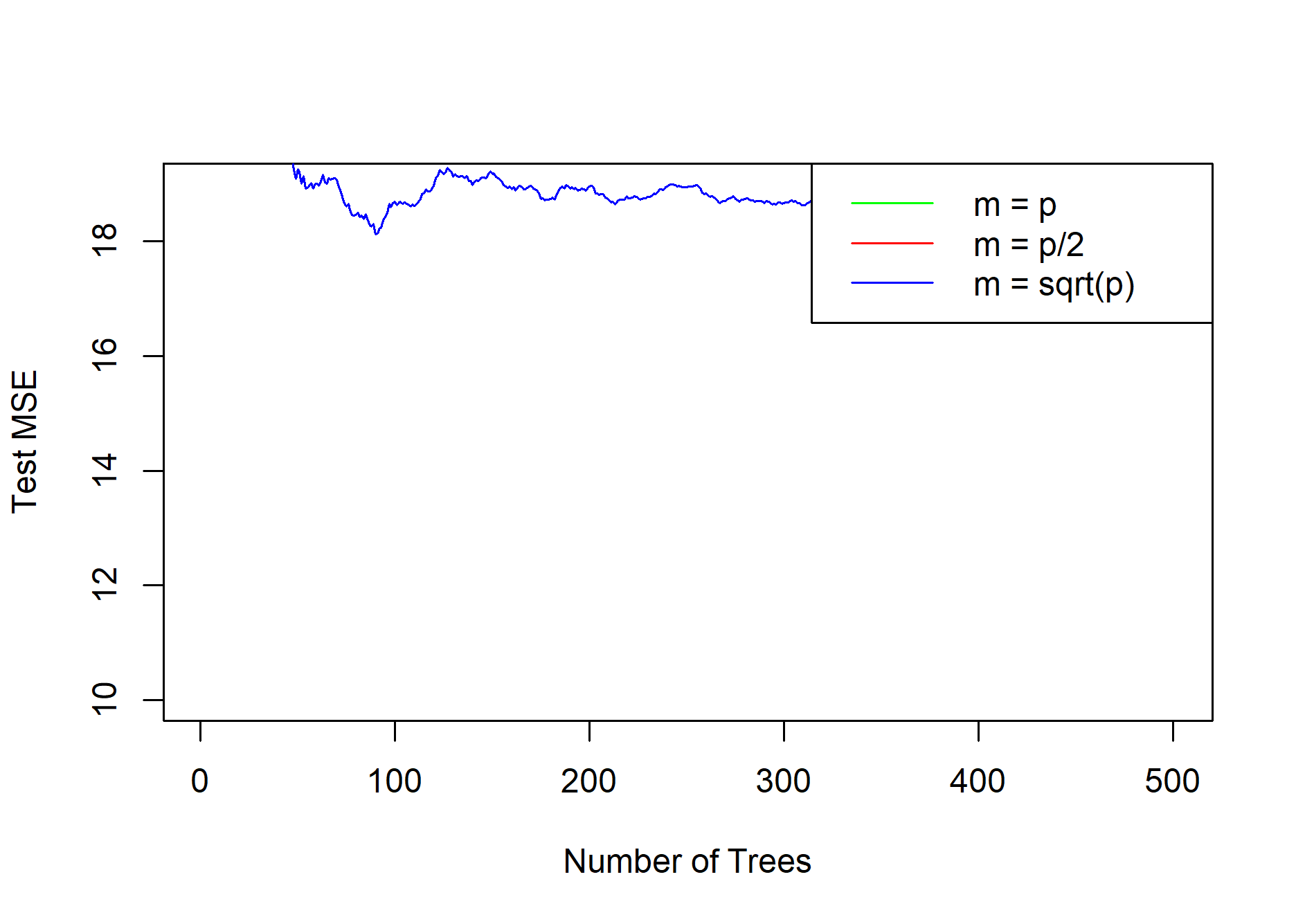


The Gini Index and the Cross-entropy are similar, as they both maximize at p = 1/2 and are concave. The misclassification rate is different since it is linear. It has similarities as it is maximized at p = 1/2, equals zero at p = 0,1 and it is concave.

7)

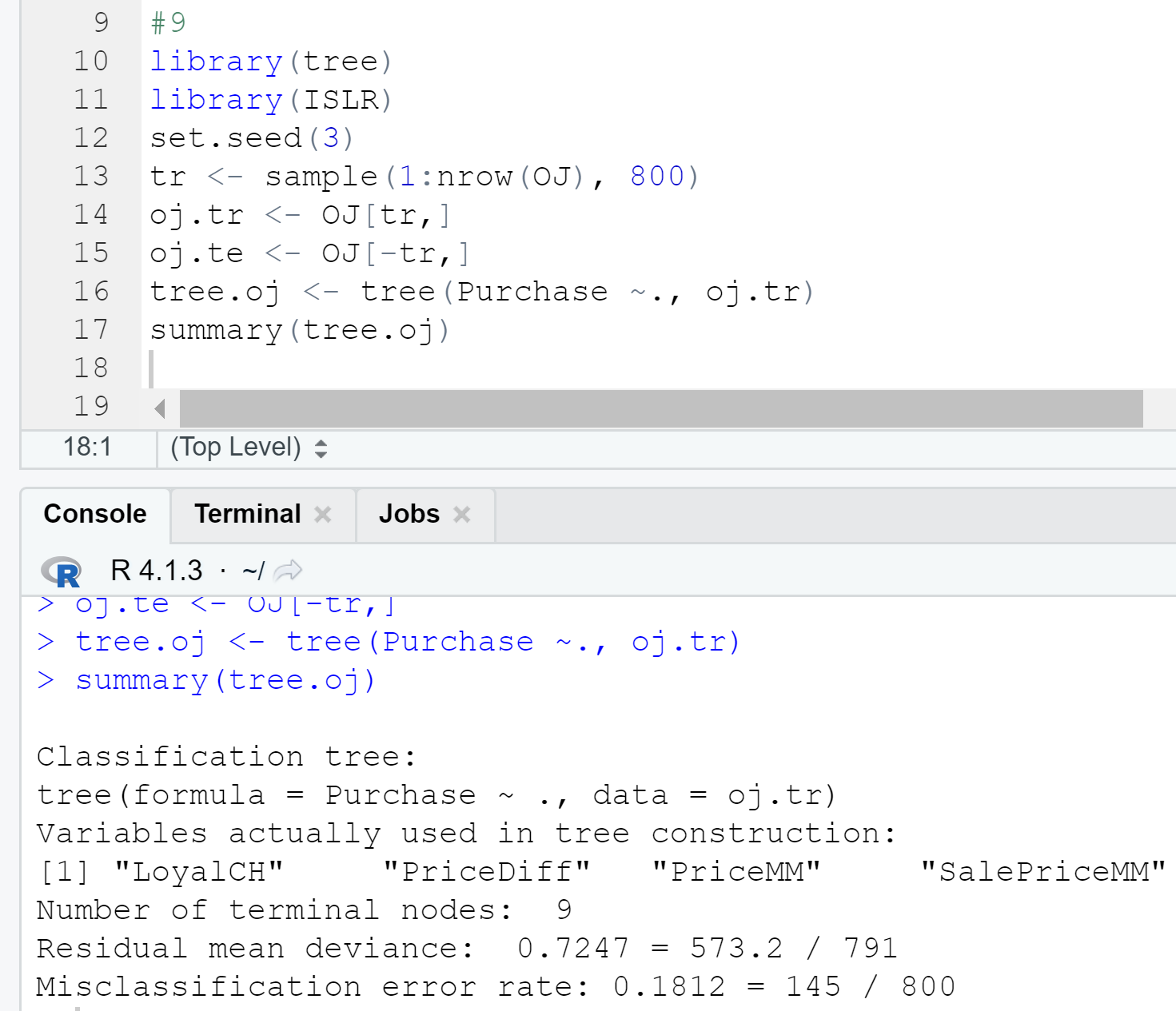
The Test MSE is very high for a single tree. As it decreases, the number of trees increases. Also, the Test MSE for all predictors is higher than half the predictors or the square root of the number of predictors.



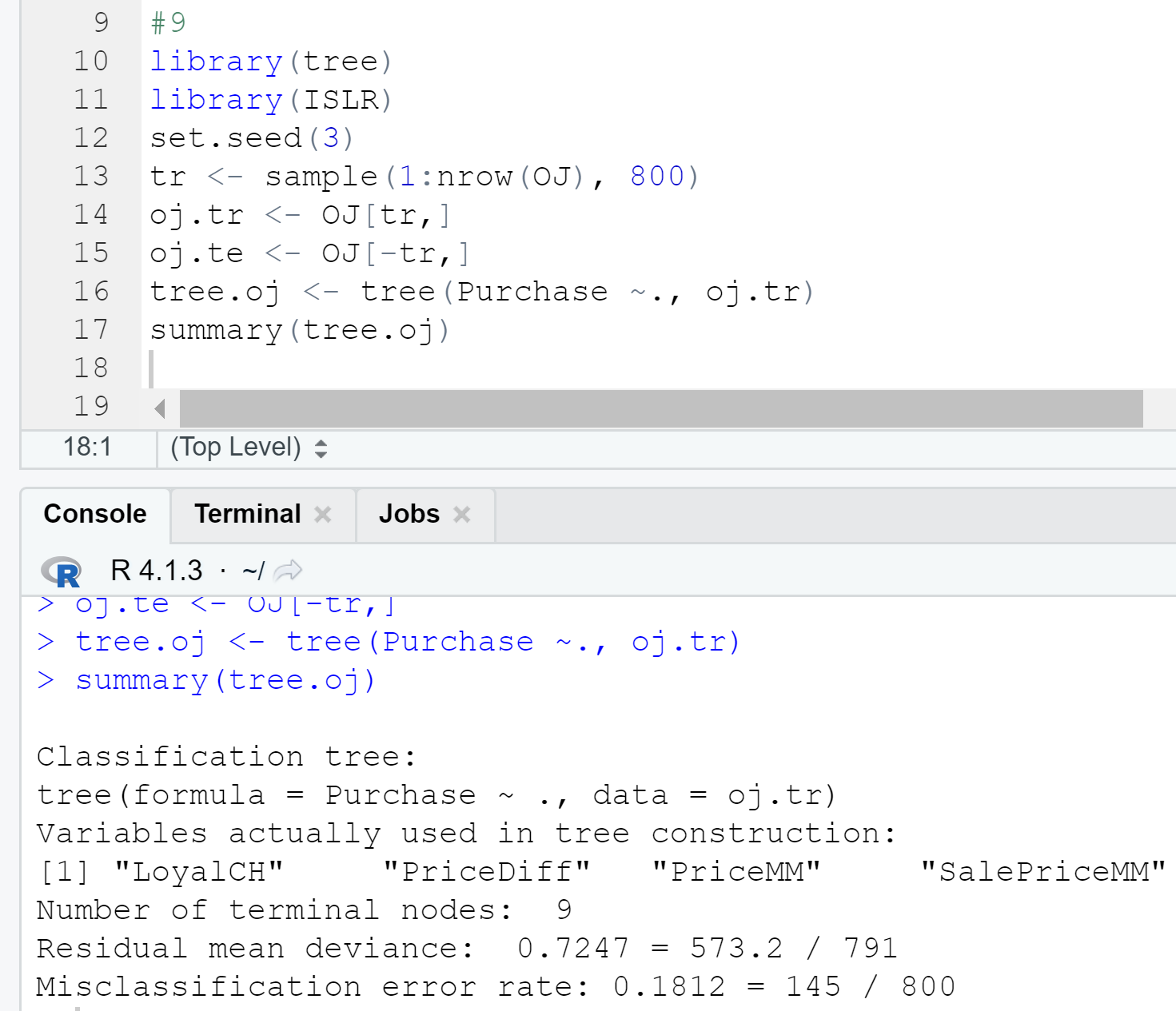


9)

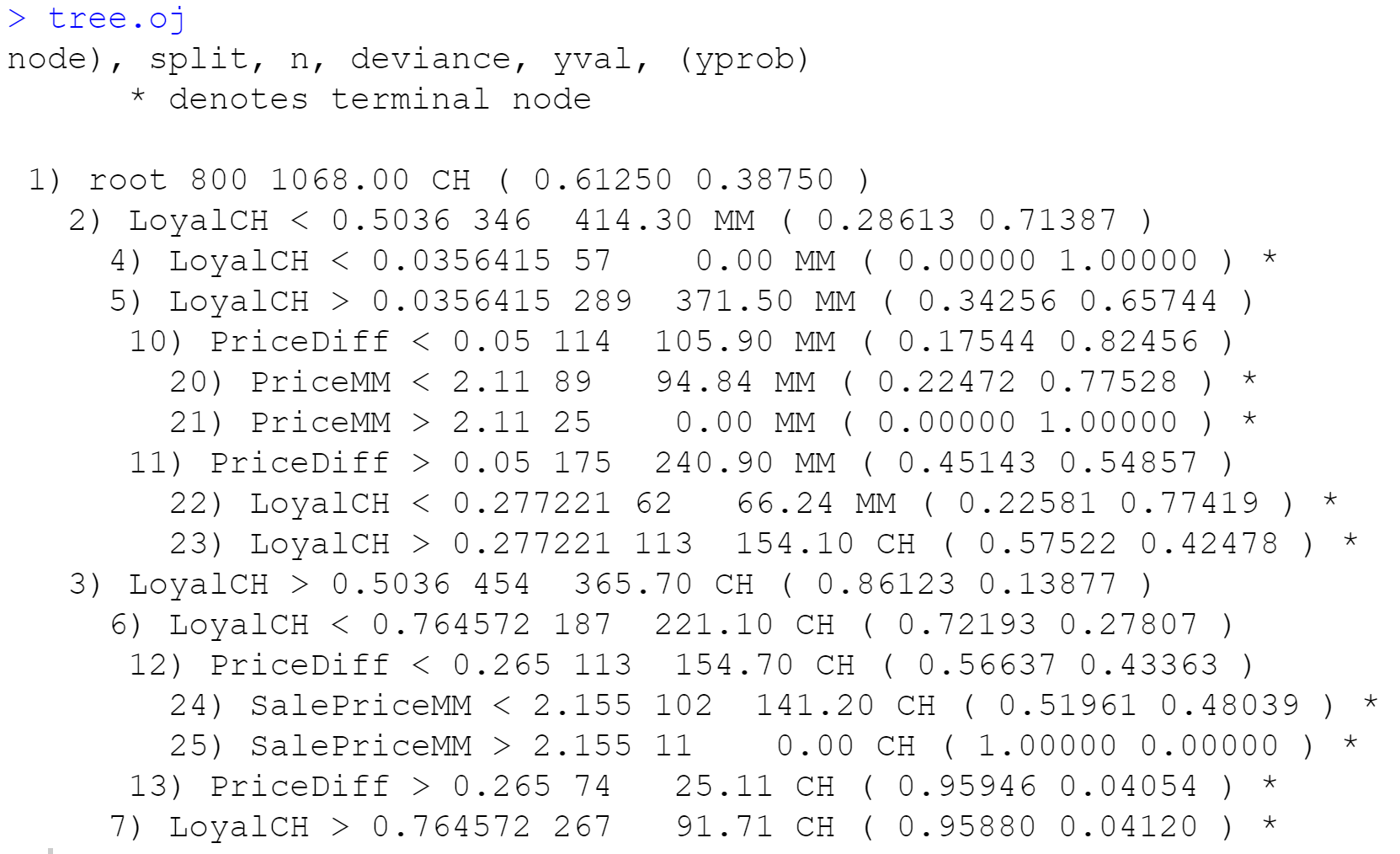
a)



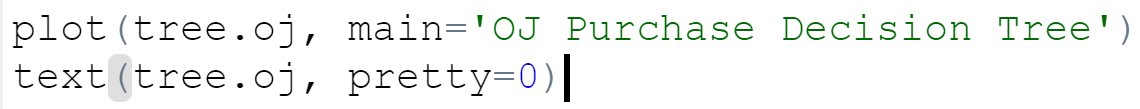
b)

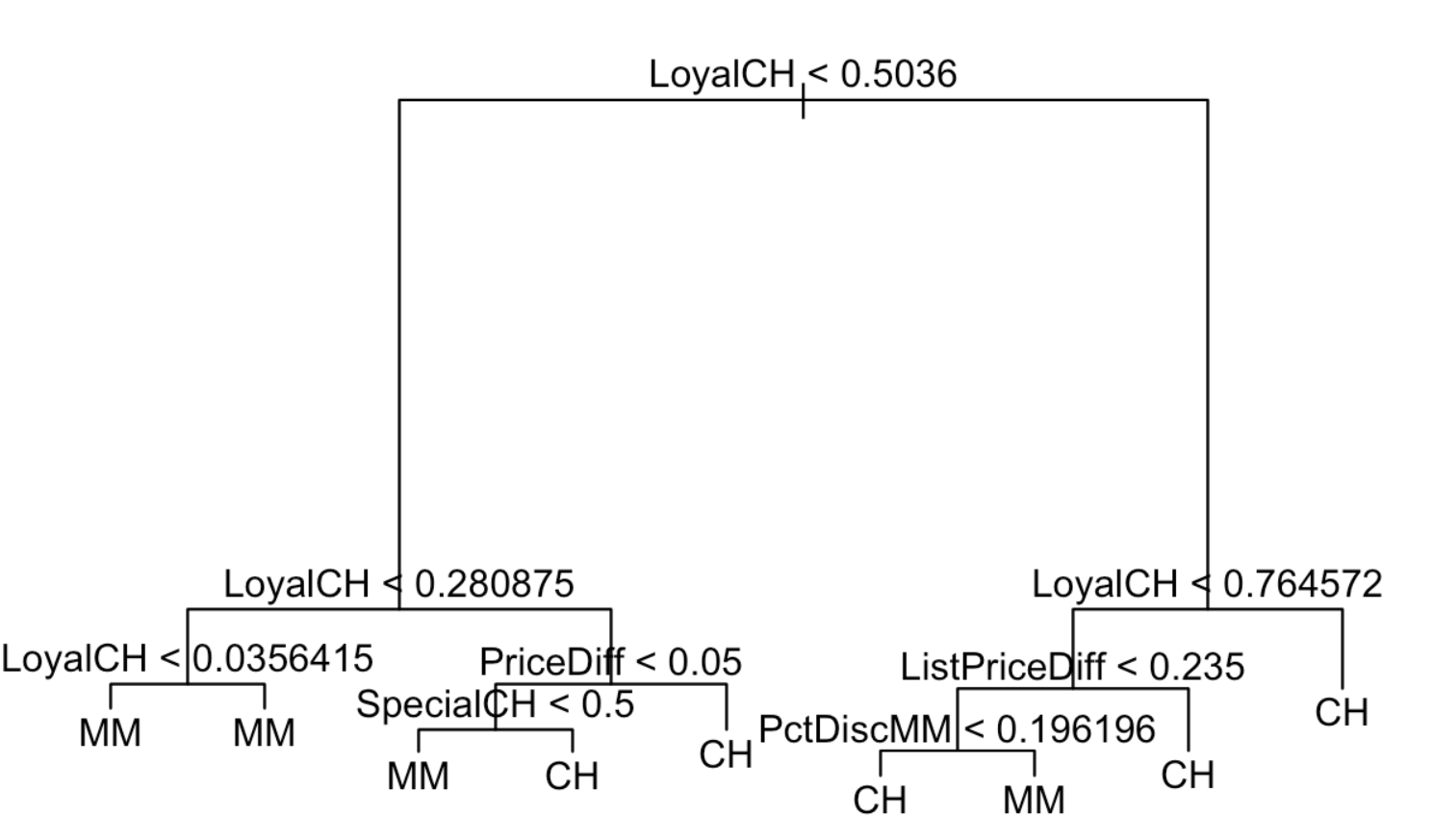


c)

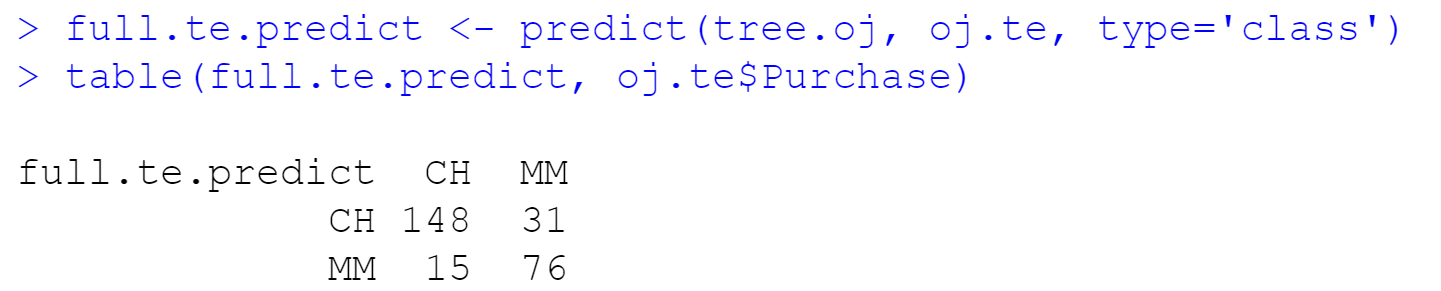


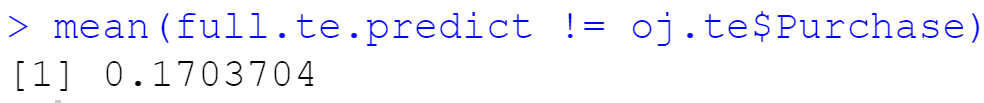
d)



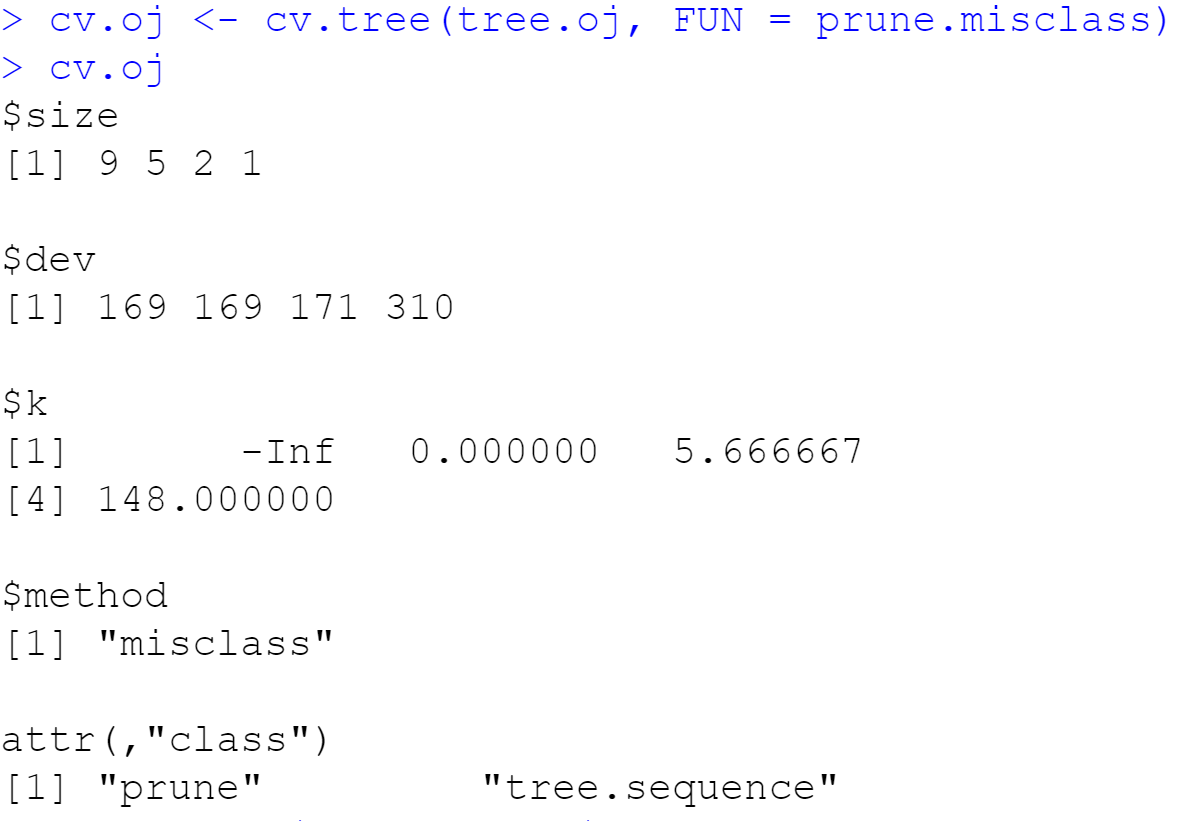


e)

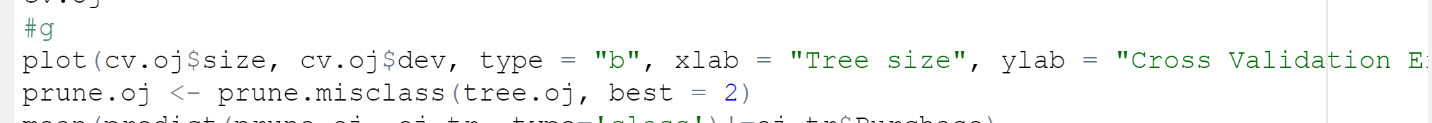


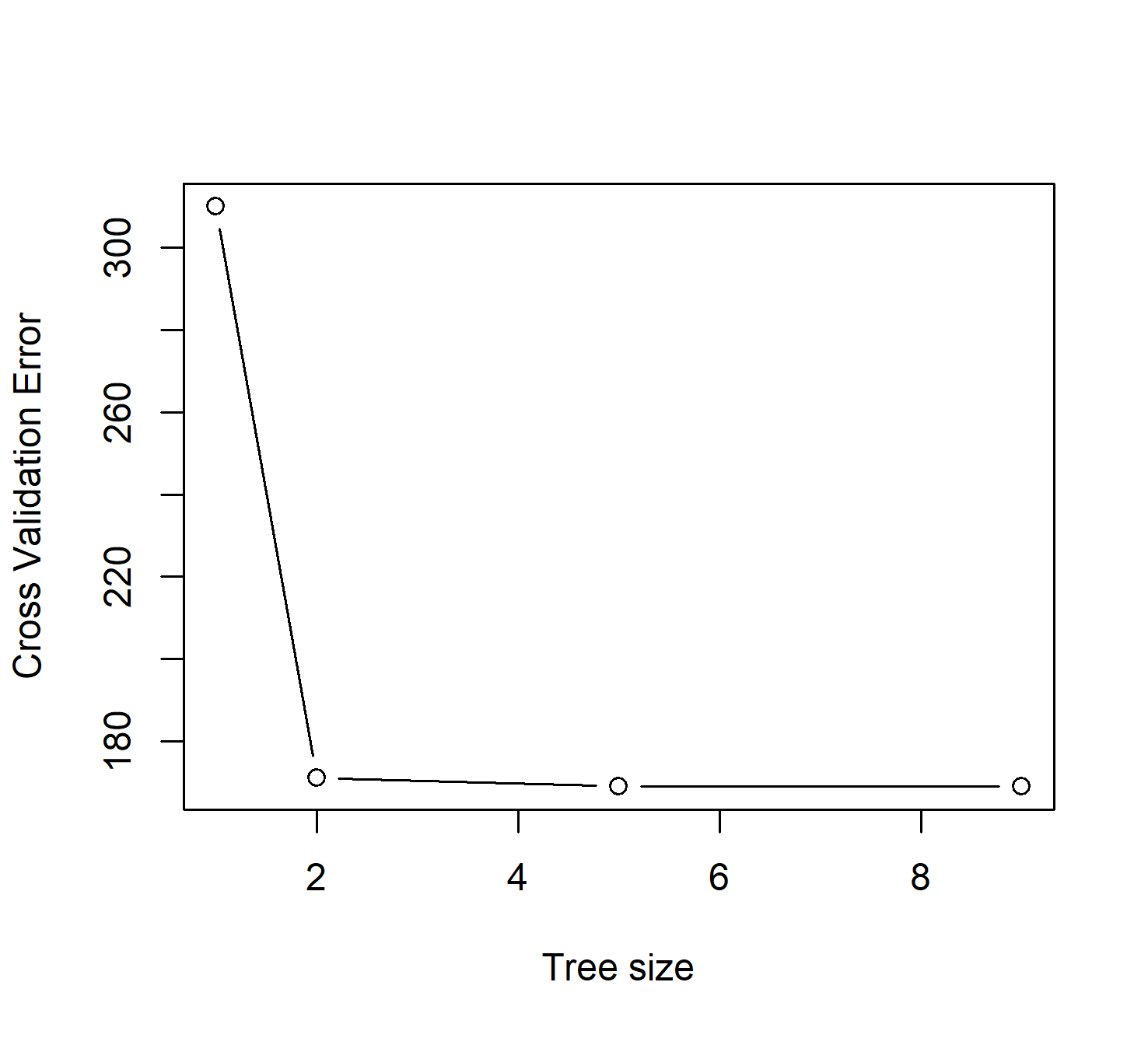


f)



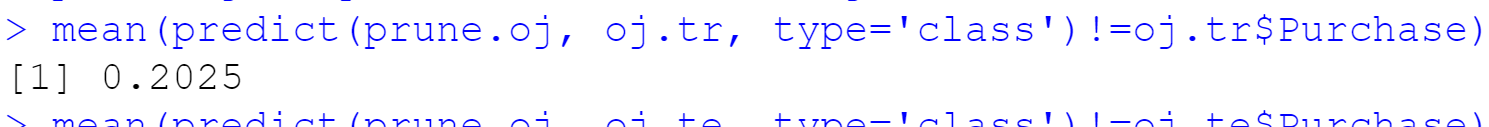
g)



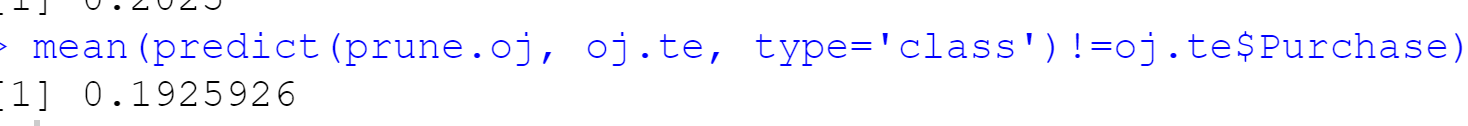


h) The tree size of 7 corresponds to the lowest cross-validated classification error rate. I choose the simple mode of size 2.

i)



j)



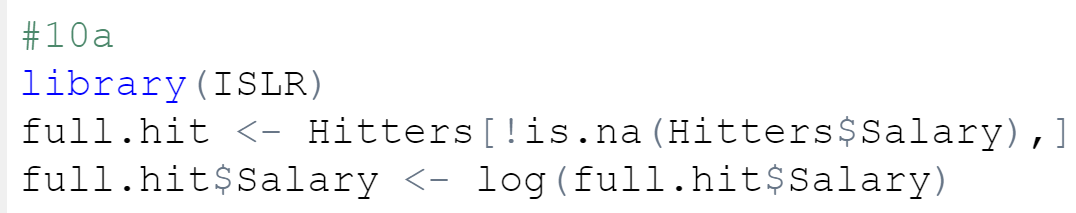
For Unpruned, the Training error rate is 15.8%. For Pruned, Training error rate is 18.3%. There may be some bias for a significantly lower variance model. It appears to be slightly over-pruned.

k)

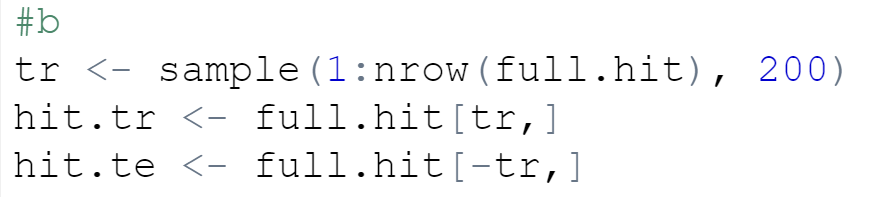
For Unpruned, the test error rate is 20%. For Pruned, the Test error rate is 21.5%. There may be some bias for a significantly lower variance model, it does appear to be slightly over-pruned however.

10)

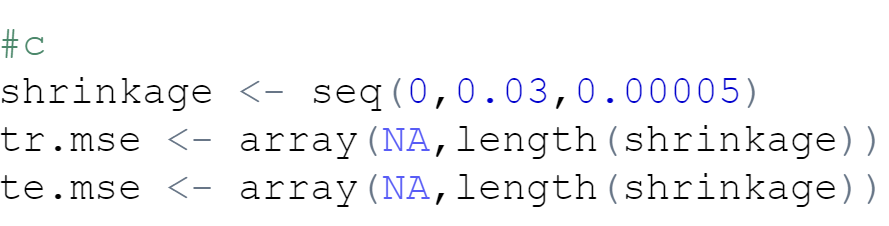
a)



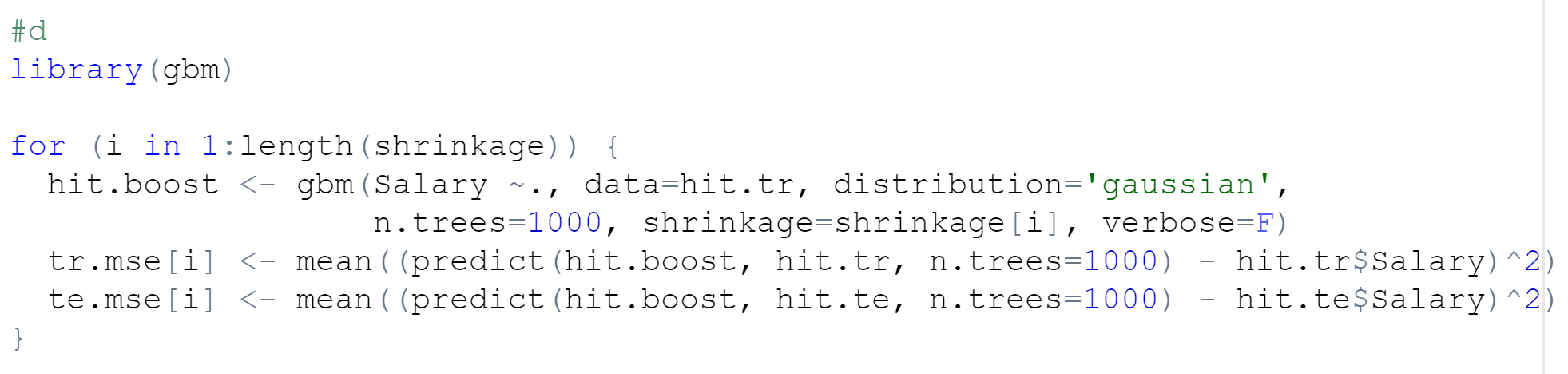
b)



c)

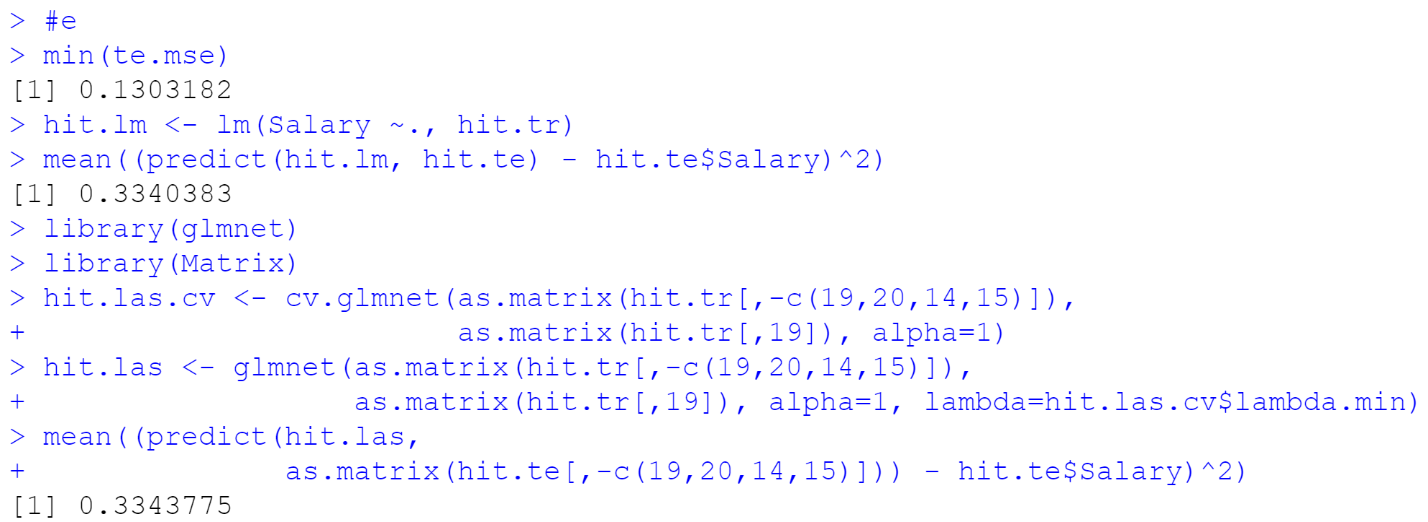


d)



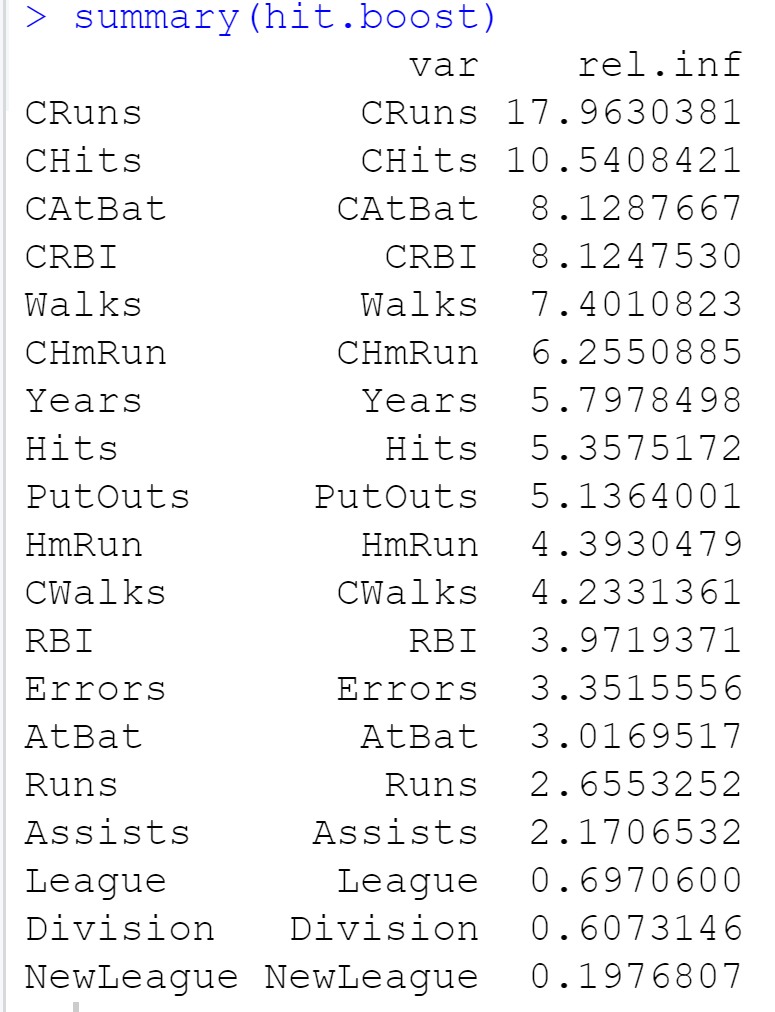
e)

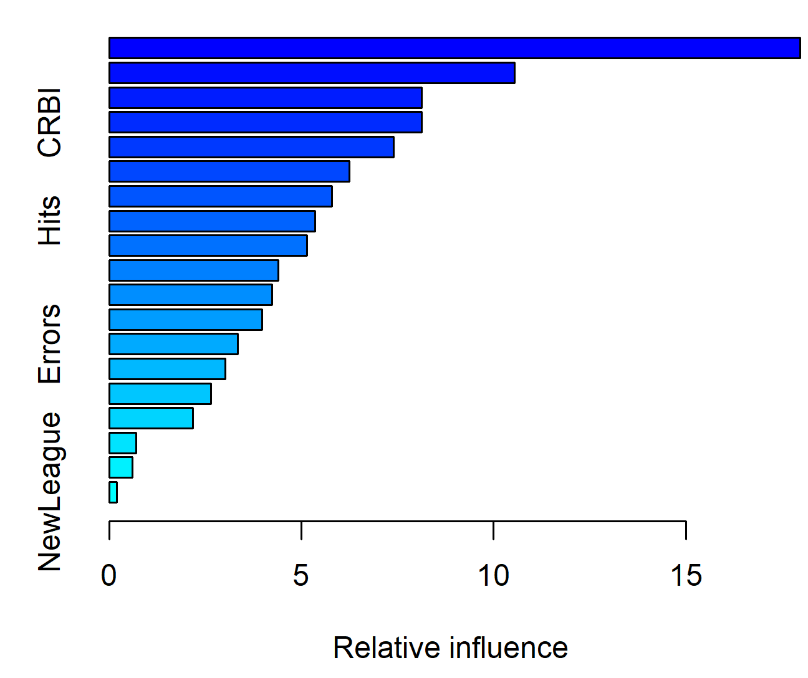
The boosting’s minimum test error of 0.456 with the test error is being compared with that of a simple linear model and lasso regression. Linear regression obtains a slightly better MSE than lasso, there is a much larger number of samples than covariates so we don’t really need to enforce sparsity. The MSE of linear regression, 0.62, is vastly greater than boosting’s 0.45. Boosting wins.



f)

The most influential variables are CAtBat, CRBI, CHits, CWalks, CRuns. “How many career bats”, “runs batted in”, “hits”, and “walks and runs the player has made is” are the most influential information affecting his salary, respectively. This looks like it is not based upon the current year’s statistics.





g)

The observed test MSE of 0.15 is better than that of linear or lasso regression, even though it is still slightly greater than boosting’s 0.47.

