Ghorpade, Mrunal Mahesh

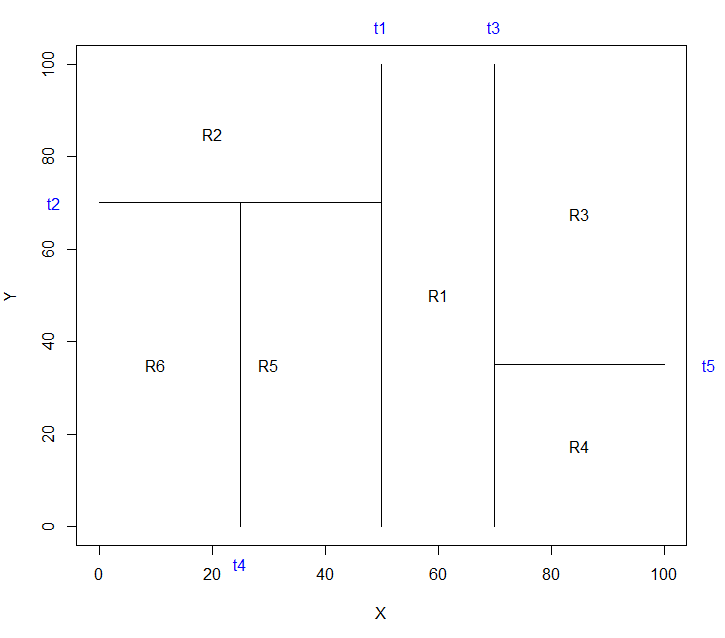
Mghorp2@uic.edu

UIN:677441117

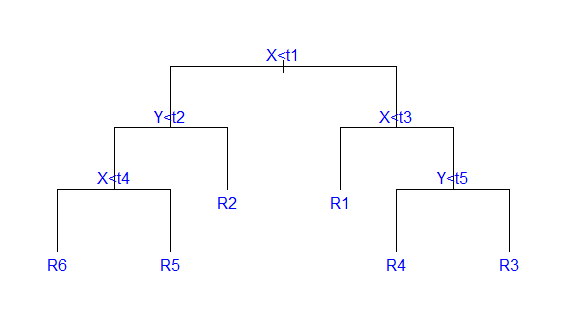
Assignment 3

1. Draw an example (of your own invention) of a partition of two-dimensional feature space that could result from recursive binary splitting. Your example should contain at least six regions. Draw a decision tree corresponding to this partition. Be sure to label all aspects of your figures, including the regions *R*1*,R*2*, . . .*, the cutpoints *t*1*, t*2*, . . .*, and so forth.

Answer: Following is the partition of 2-dimensional feature space that could result from recursive binary splitting:

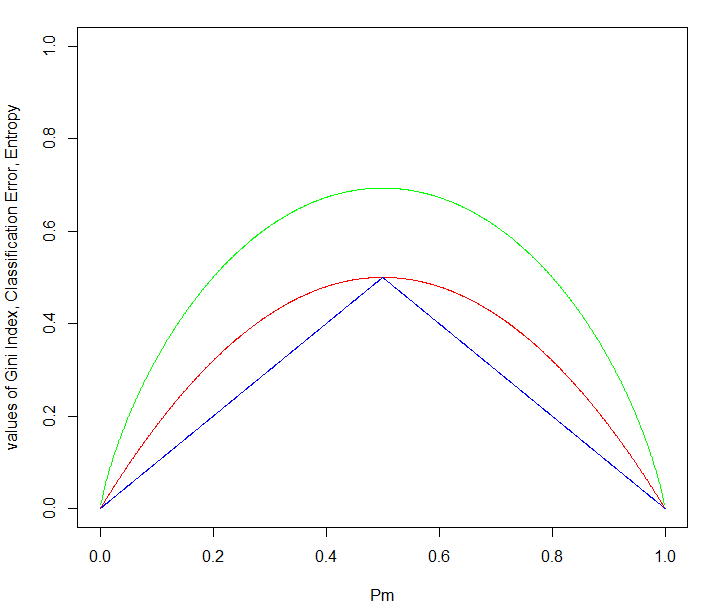


Decision tree corresponding to the above partition is given below;



1. Consider the Gini index, classification error, and entropy in a simple classification setting with two classes. Create a single plot that displays each of these quantities as a function of ˆ*pm*1. The *x* axis should display ˆ*pm*1, ranging from 0 to 1, and the *y*-axis should display the value of the Gini index, classification error, and entropy.

Answer: Below is the plot of Pm Vs Values of Gini index, classification error, and entropy



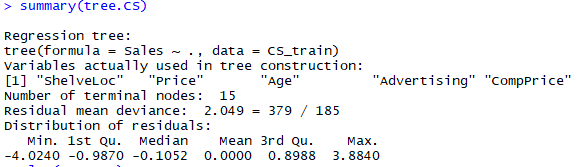
1. In the lab, a classification tree was applied to the Carseats data set after converting Sales into a qualitative response variable. Now we will seek to predict Sales using regression trees and related approaches, treating the response as a quantitative variable.

(a) Split the data set into a training set and a test set. (Please Refer to “Problem 8” R code)

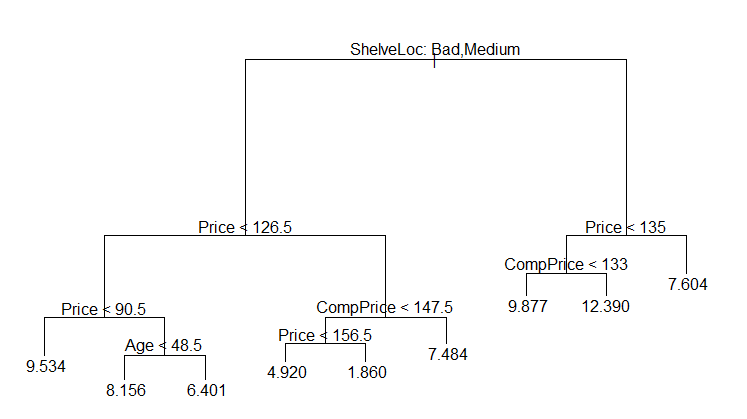
(b) Fit a regression tree to the training set. Plot the tree, and interpret the results. What test MSE do you obtain?

Observation:

Following is the summary of Regression Tree on Training Data;



Below in the Regression tree obtained:



For Carseats Data, a regression tree for predicting the log Sales of car seats at different locations, based on number of parameters i.e. predictor variables in the data e.g. Shelveloc: the quality of the shelving location for the car seats at each site, Price: Price charged by the company, Age: Average age of local population, CompPrice: price charged by competitors etc. At a given node the lable indicates the left-hand branch emanating from the split e.g. Split on the first node is on variable “ShelveLoc”, where ShelveLoc: Bad, Medium means the left-hand branch contaings the observations corresponding to ShelveLoc = Bad/Medium whereas the right-hand branch contains the data containing ShelveLoc = Good. The tree has 8 internal nodes and 9 leaf nodes. The number in each leaf is the mean of the Sales for the observations that fall there.

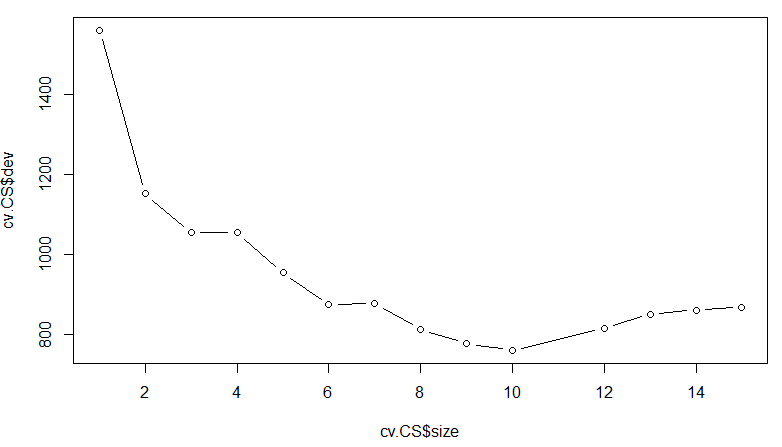
The Test MSE obtained is 4.325436



(c) Use cross-validation in order to determine the optimal level of tree complexity. Does pruning the tree improve the test MSE?

Observations:

After applying Cross Validation we get that optimal size of the tree should be 10 as shown in the below plot.



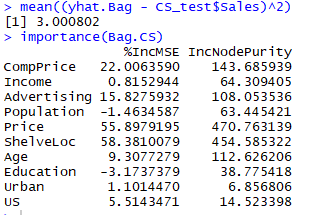
After applying pruning to get 10 node tree. We get MSE as 5.077. So, pruning increased the Test MSE.



(d) Use the bagging approach in order to analyze this data. What test MSE do you obtain? Use the importance() function to determine which variables are most important.

Observations:

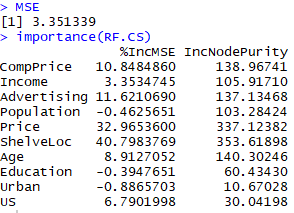
Bagging decreases Test MSE to 3.0008 . The two most important variables are “Price” and “ShelveLoc” as shown below;



(e) Use random forests to analyze this data. What test MSE do you obtain? Use the importance() function to determine which variables are most important. Describe the effect of *m*, the number of variables considered at each split, on the error rate obtained.

Random Forest decreases Test MSE to 3.35 but bagging decreases more. Here as well the two most important variables are “Price” and “ShelveLoc”.

Below results obtained are with m=sqrt(# of variables) which is approximately 3



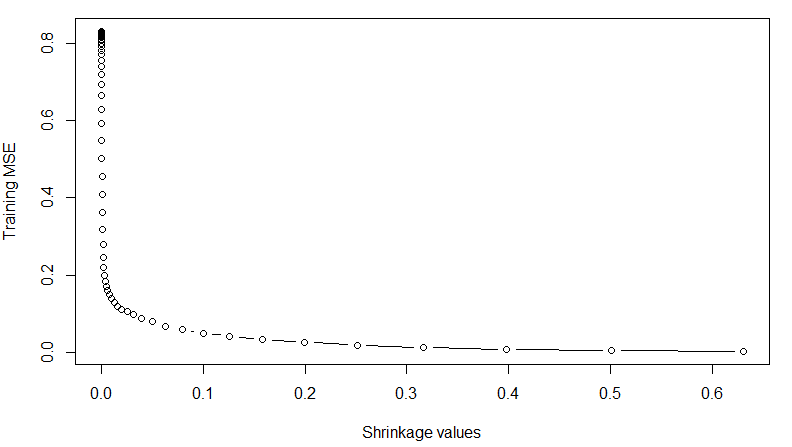
As m is increased the test error rate decreases till certain limit and it increases again.

|  |  |
| --- | --- |
| m | MSE |
| 2 | 3.835014 |
| 3 | 3.35133 |
| 5 | 3.00397 |
| 8 | 2.932019 |
| 9 | 2.940827 |
| 10 | 2.96737 |

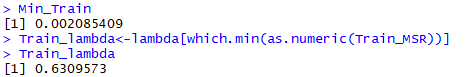
1. We now use boosting to predict Salary in the Hitters data set.
2. Remove the observations for whom the salary information is unknown, and then log-transform the salaries. (Please Refer to “Problem 10” R code)
3. Create a training set consisting of the first 200 observations, and a test set consisting of the remaining observations. (Please Refer to “Problem 10” R code)
4. Perform boosting on the training set with 1,000 trees for a range of values of the shrinkage parameter λ. Produce a plot with different shrinkage values on the x-axis and the corresponding training set MSE on the y-axis.

Observations:

Below is the Plot of Different Shrinkage values Vs Training MSE;



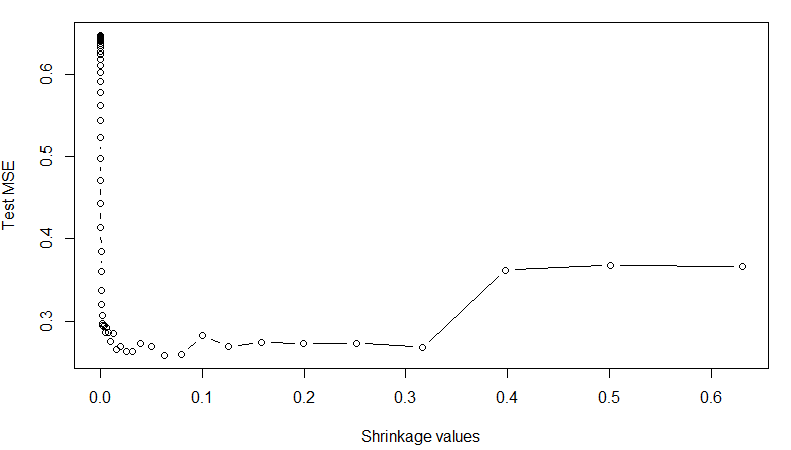
We get minimum Training Error 0.00208 at lambda = 0.63 as shown below;



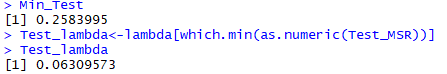
1. Produce a plot with different shrinkage values on the *x*-axis and the corresponding test set MSE on the *y*-axis.

Observations:

Below is the Plot of Different Shrinkage values Vs Test MSE;



We get minimum Test Error 0.258 at lambda = 0.063 as shown below;



1. Compare the test MSE of boosting to the test MSE that results from applying two of the regression approaches seen in Chapters 3 and 6.

Observations:

Applying Ridge Regression on the Test data we get MSE as follows;



Applying Linear Regression on the Test data we get MSE as follows;

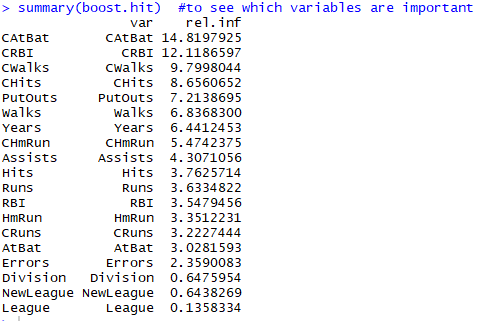


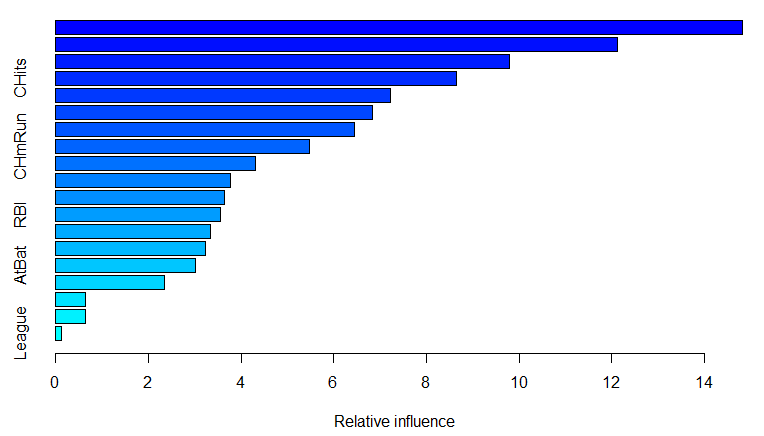
Both the approaches give higher MSE on Test Data as compared with Boosting

1. Which variables appear to be the most important predictors in the boosted model?

Observation:

From the below summary we can see that CAtBat, CRBI, CWalks, Chits are the most important variables in the boosted model





1. Now apply bagging to the training set. What is the test set MSE for this approach?

Observation:

After Bagging is applied we get following MSE;



Bagging gives lesser MSE than Boosting.