**Homework4**

**20200639 Chae Woojin**

**Chapter 8:** 2, 3, 4, 5, 8, 11

**Chapter 11:** 2, 3, 4, 7, 8, 11

**Chapter 8:**

**#2.**

A decision in which a single split occurs on the variable (at value ) can be simplified as follows.

With boosted stumps, following algorithm gives final decision as below.

b=1:

b=2:

b=B:

After Bth iteration, one could get final boosting model , given by

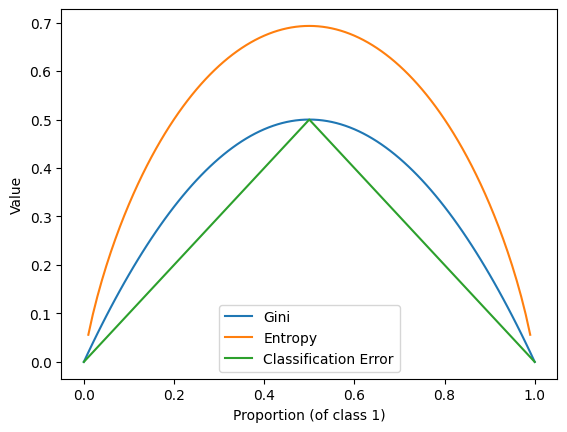
Since , we can see that this estimate is indeed an additive model of the form .

**#3.**

Gini index is defined by .

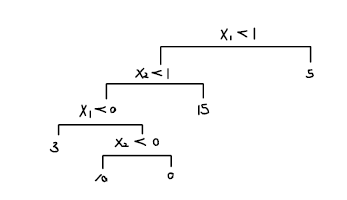
Entropy is defined by .

The classification error is .

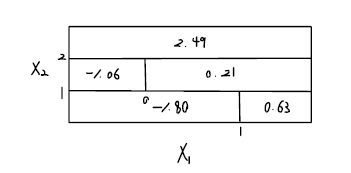


**#4.**

**(a)**



**(b)**

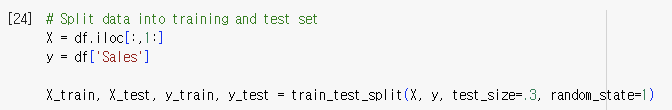


**#5.**

According to majority vote approach gives ‘red’ as a result since 6 out of 10 probabilities exceed 0.5, which means that class for given data is likely to be red. However, according to approach based on average probability, it results in ‘green’, since average probability for then estimates is 0.45, which allocates given data to class ‘green’.

**#8.**

(a)



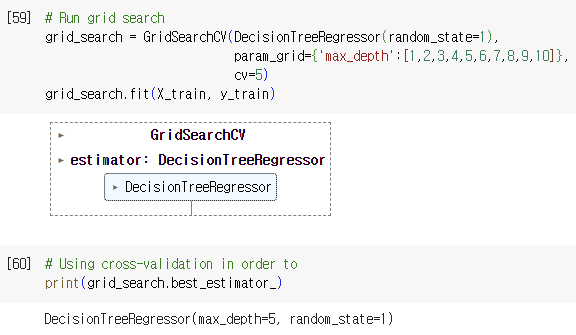
(1st column is ‘Sales’, so that X was constructed by excluding 1st column)

(b)

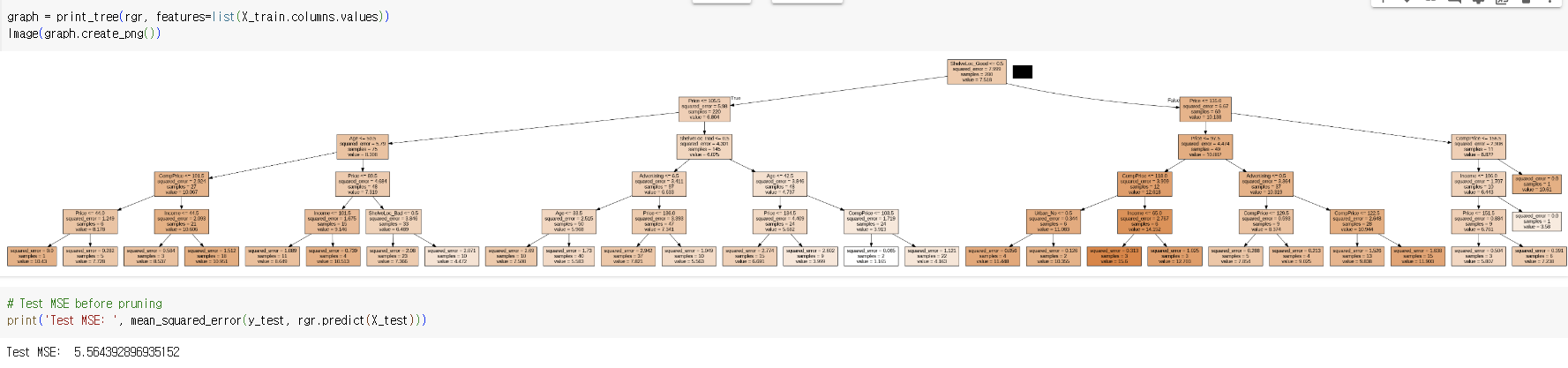


Test MSE obtained from this tree was 5.990749166666666. Although this tree seems too complex, it gives better performance than base test MSE, which was evaluated as residual sum of squares using test mean as predicted values.

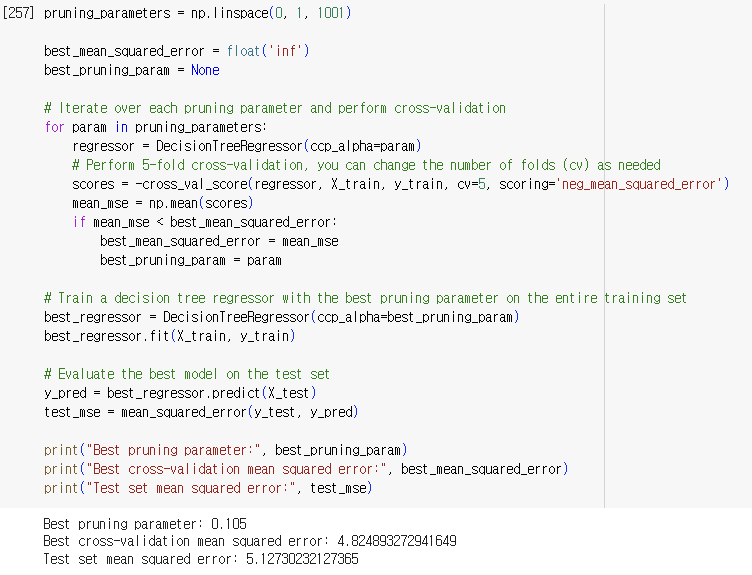
(c)

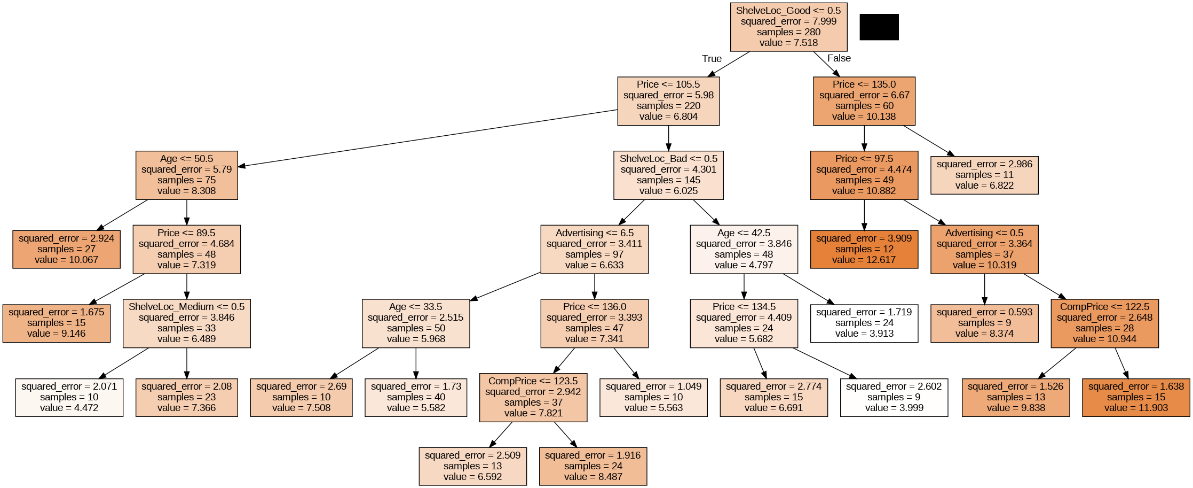


Using cross validation gives maximum depth 5 as the optimal level of convexity when we do not consider pruning. In this case, tree was obtained as follow, and test MSE was 5.564392896935152. It gives better test MSE than before.

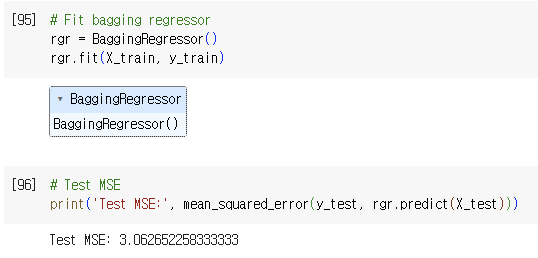


But when we additionally consider pruning, test MSE decreased further into 5.12730232127365. From this result, one can observe pruning improves the performance, as it prevents tree to be overly overfitted, which makes model less explainable in the new observations.

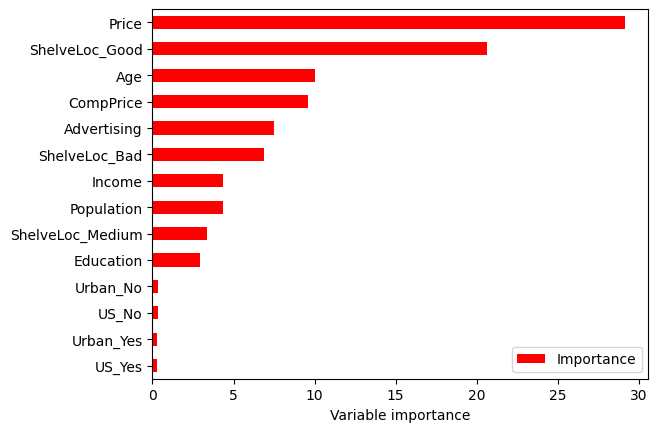




(d)

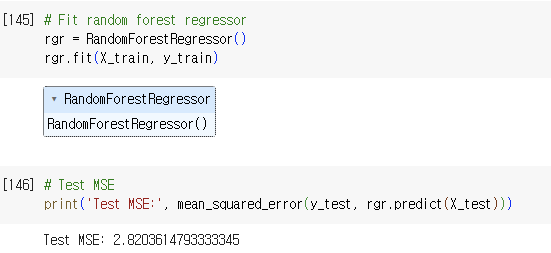


Using bagging approach gave 3.062652258333333 of test MSE.

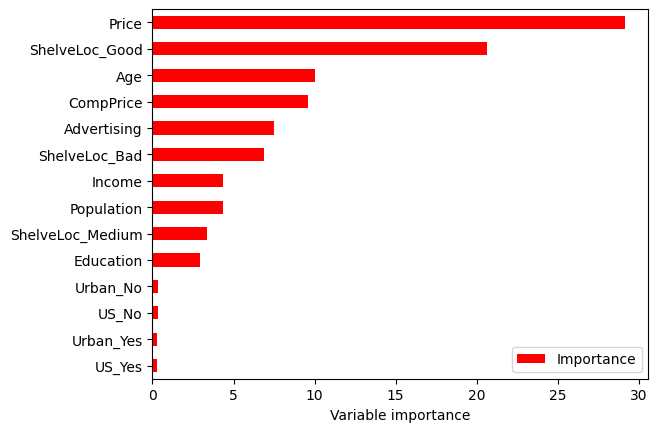


The figure above illustrates the importance of variable predicting response variable. From this result, one can conclude that price is the most important variable in predicting response variable.

(e)

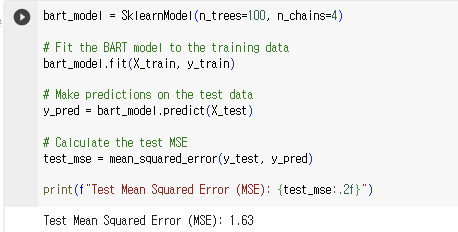


Using random forest approach gave 2.8203614793333345 of test MSE.



This figure describes the relative importance between predictors. From this result, one can conclude that price is the most important variable in predicting response variable.

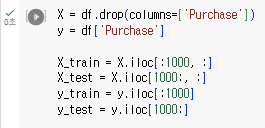
(f)



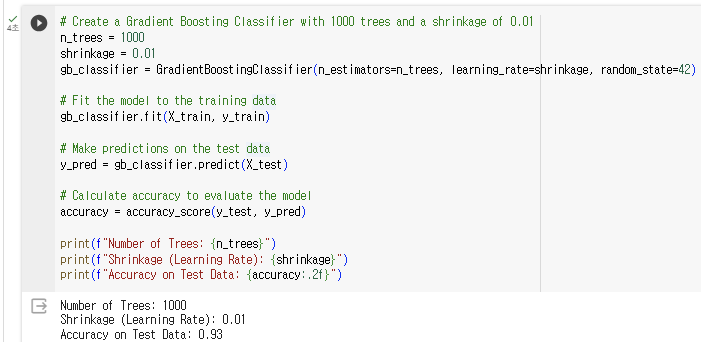
Using BART model, the lowest test MSE(about 1.63) was obtained.

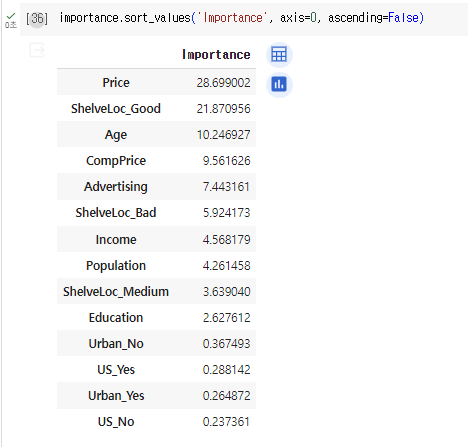
**#11.**

(a)



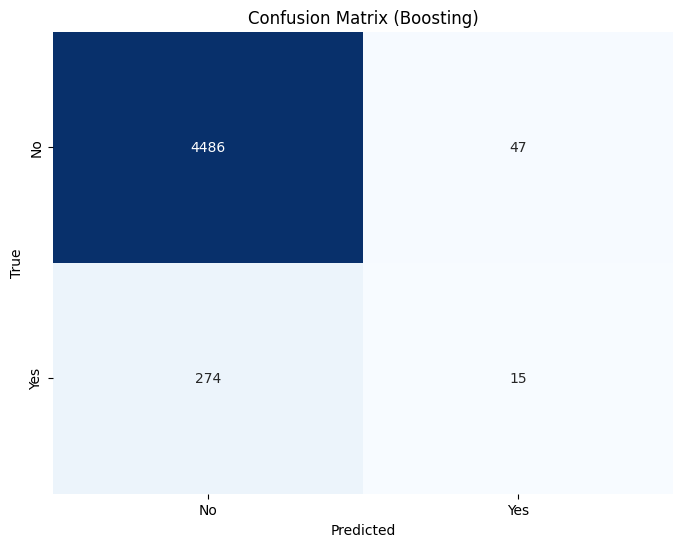
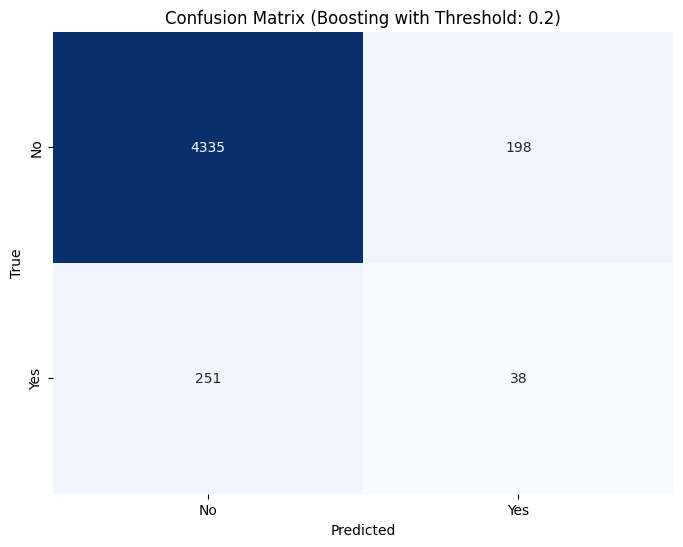
(b)

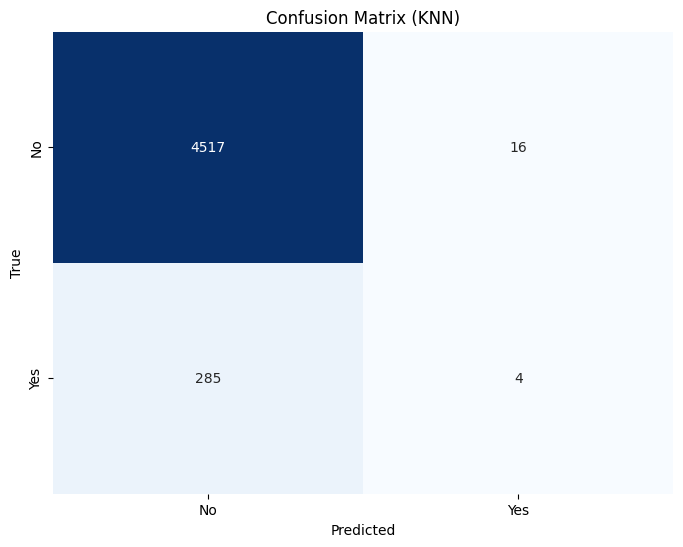
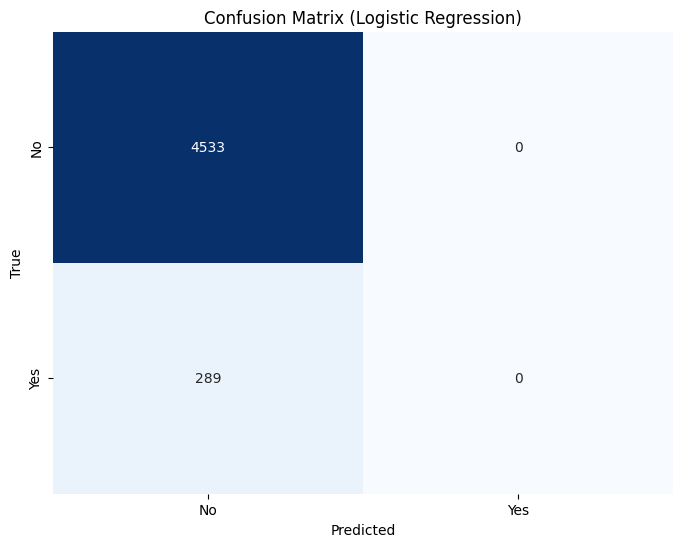




From a boosting model consists of 1000 trees, and a shrinkage value 0.01, ‘Price’ resulted into be the most important predictor.

(c)

These four figures illustrate classification result as a form of confusion matrix using Boosting, Boosting with acceptance threshold probability 0.2, KNN, and Logistic Regression respectively.

From this result, the fraction of the people predicted to make a purchase do in fact make from these models are evaluated as below.

* Boosting:
* Boosting with acceptance threshold probability 0.2:
* KNN:
* Logistic Regression:

In overall, every method did not perform well in terms of classification. I think this is because there are dominantly many ‘No’ in the given ‘Purchase’ data, while there is poor amount of ‘Yes’ in the data. Therefore, it cannot learn data well.

**Chapter11:**

**#2.**

(a) No, Yes, Yes and No. Second participant was the case that study has ended before event occurs, and the third participant was the case that one dropped out of the study.

(b)

(c)

(d)

(e)

**#3.**

: the number of unique replacements

: unique replacement times

: the number of participants who did not replaced phone until event happens

: the number of participants who replaced the phone

**#4.**

(a) There are two 2 events that happen before 50 days. There are 6 data, but as the earliest observation has censored, we start the number of risk to 5. Using this fact, the estimated probability of survival past 50 days can be evaluated as follows.

(b) As there are 3 events among uncensored data occurred at a unique time, estimated survival probability can be evaluated as follows, which consists of 4 intervals.

**#7.**

(a) can be considered as the random variable that chooses people of deaths occurred in group 1 among group 1( people) and group 2( people). Therefore, we can argue follows the hypergeometric distribution as follows.

(b) It is well known fact that a random variable which follows hypergeometric distribution has mean and variance as below.

Using this fact, we can evaluate the mean and variance of as follows.

It corresponds to the result of (11.5) and (11.6) in the textbook

**#8.**

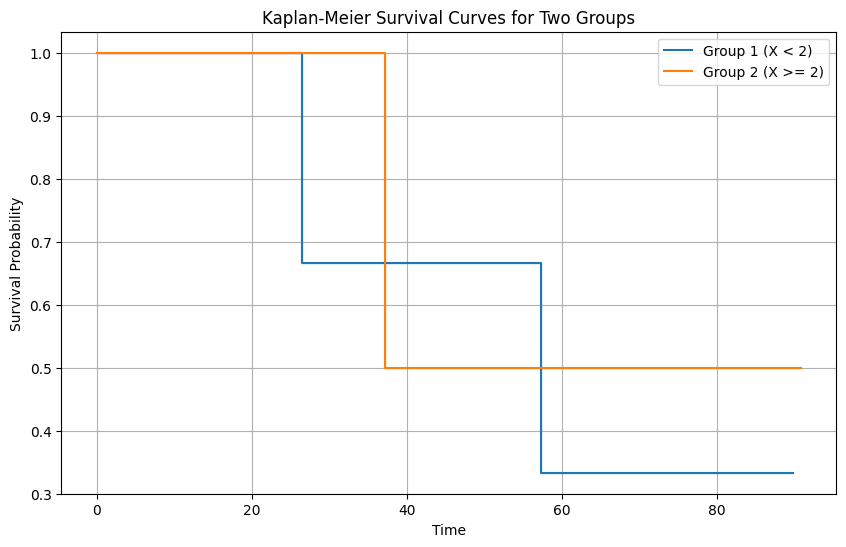
As , is a cdf of .

.

Then, . Integration from 0 to t gives .

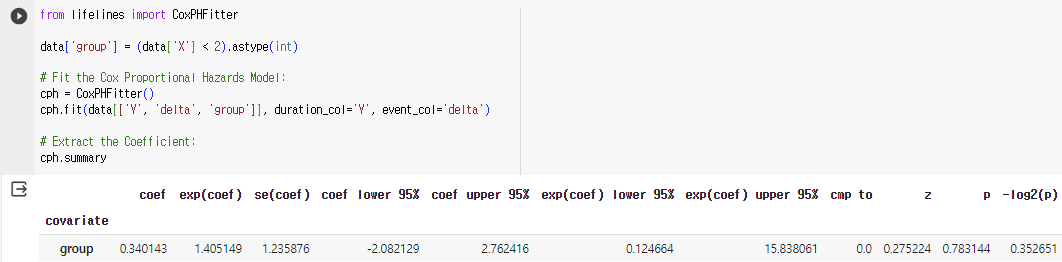
**#11.**

(a)



It does not seem to have similarity in terms of survival curves between two groups.

(b)



The coefficient is 0.34. This means that there exists an increased hazard with group having covariate ‘X’ less than 2. However, there is no evidence that the true coefficient value is non-zero, because the p-value of it is 0.78.

(c)



They are identical.