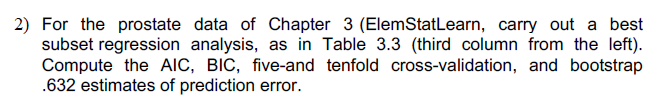
Question 2.

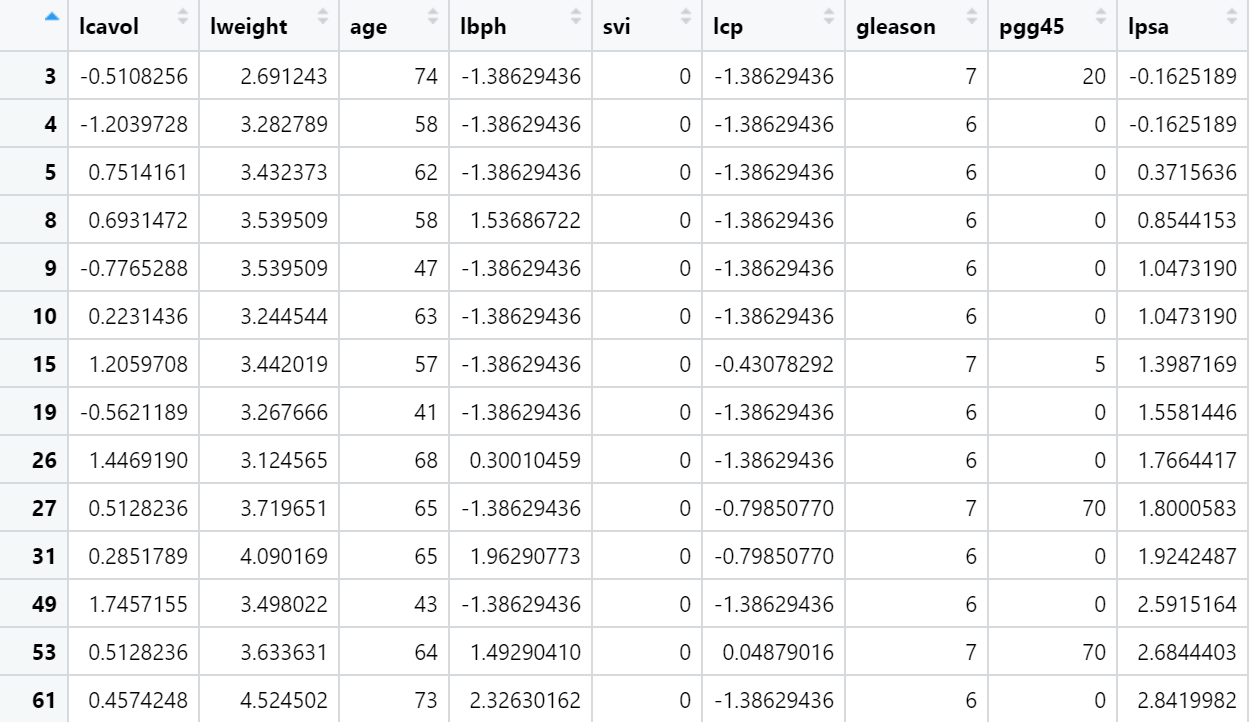


Process:

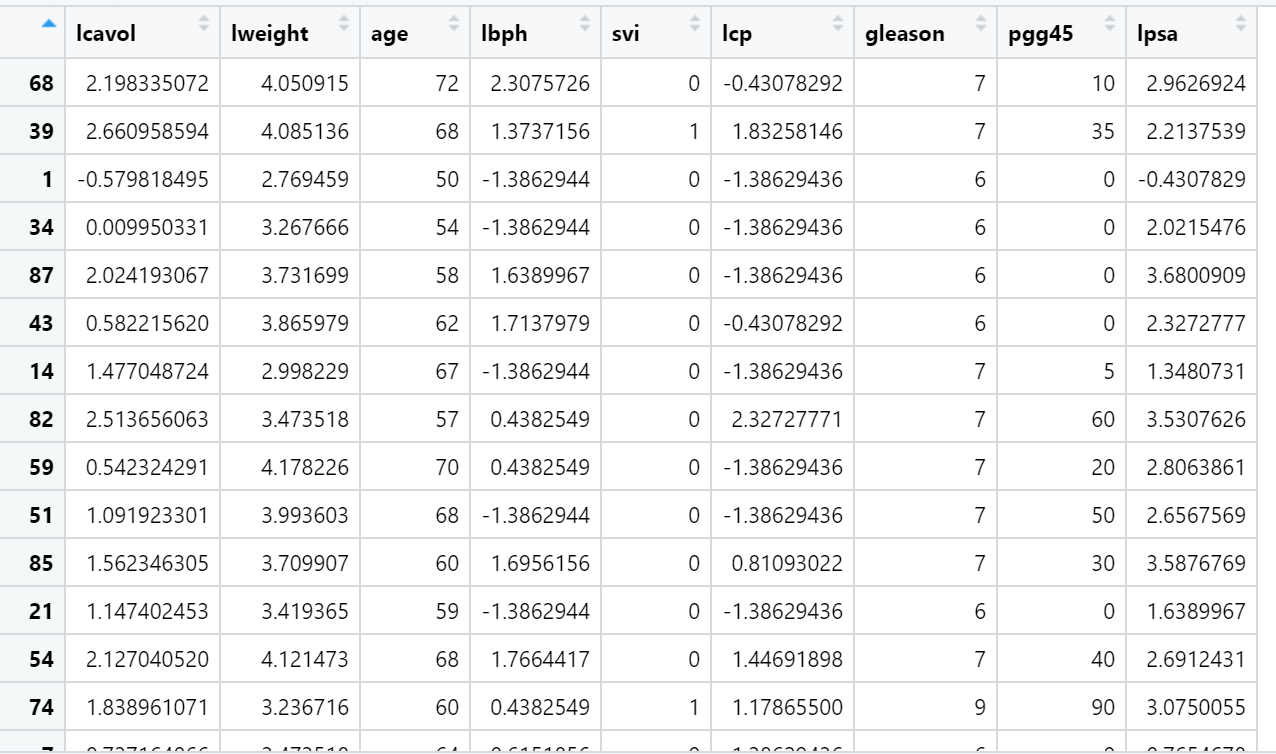
1. **Load data**
2. **Split the Train and Test data**: I used the method ‘sample’ to split the data into the train and test data.
3. **Fitting the exhaustive subset selection model with the data**: By using exhaustive subset selection, find out important features among all the features.
4. **Plot CP and BIC values by the number of variables**
5. **Calculate the train and test MSE**: By the number of variables, I put the train data to make the models and made the predicted data with each model. Also, I made MSE list by the number of variables with the predicted data and test data.
6. **Get values of AIC and BIC**: Using the selected model with CP and BIC, get values of AIC and BIC.
7. **Fit two Cross validation model to get MSEs:**  Using whole data, made crossvalidation models by the number of variables (ordered by importance). Calculate MSE by the variables. (CV10 and CV5)
8. **Get MSEs with Bootstrap.:** Using train and test data, made bootstrap models by the number of variables (ordered by importance). Calculate MSE by the variables.

Outputs:

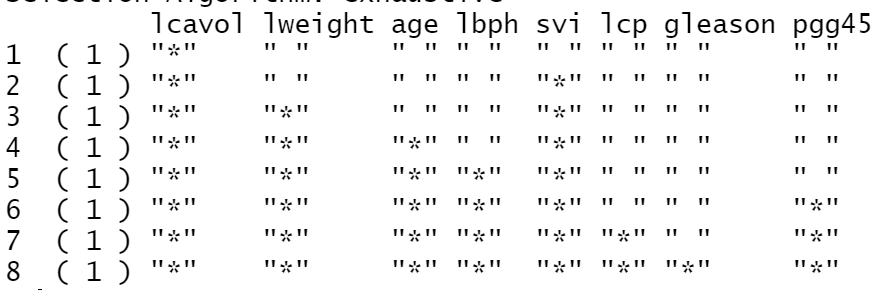
**Test data (80%)**



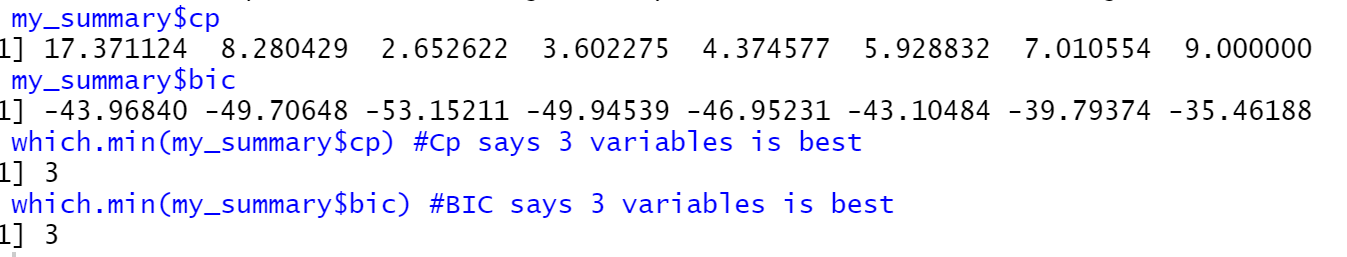
**Train data (20%)**



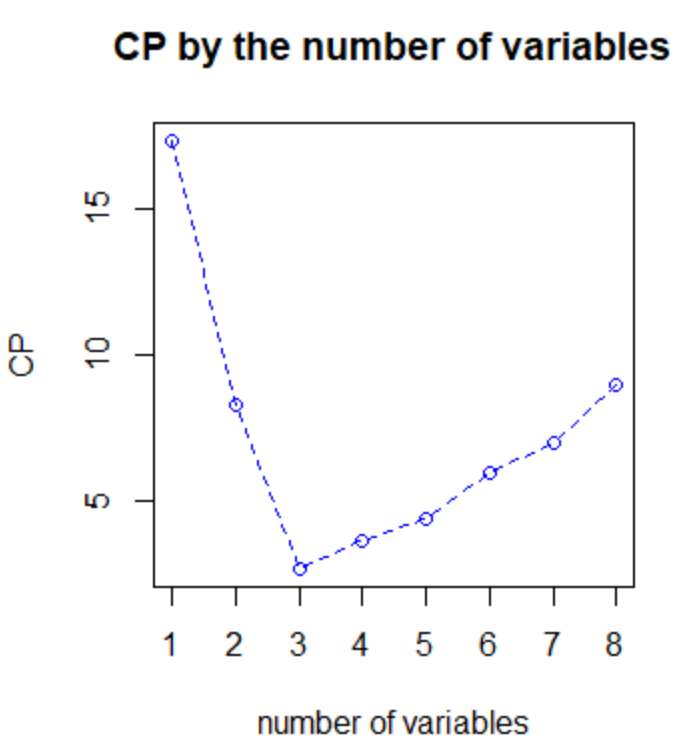
**Summary of Exhaustive subset selection**



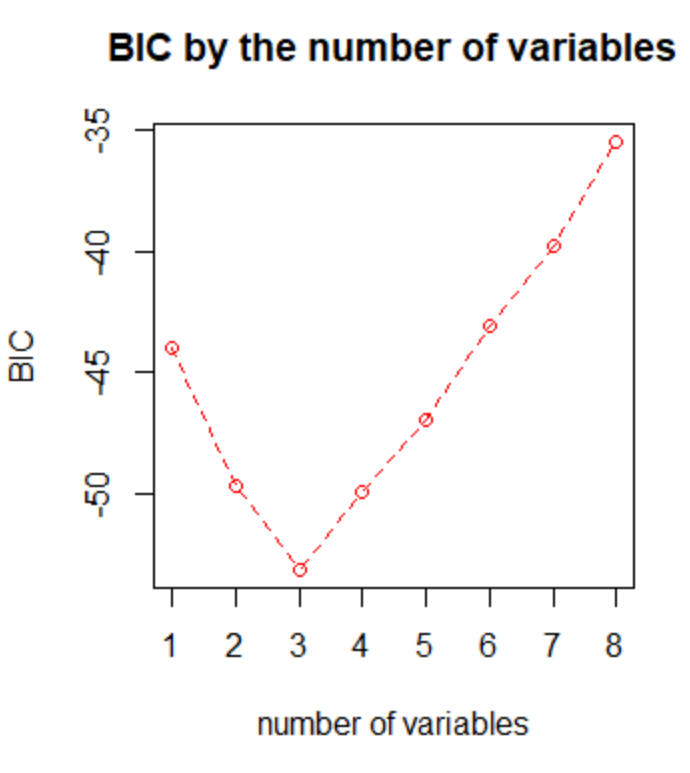
**The number of model’s variables(minimum cp value)**



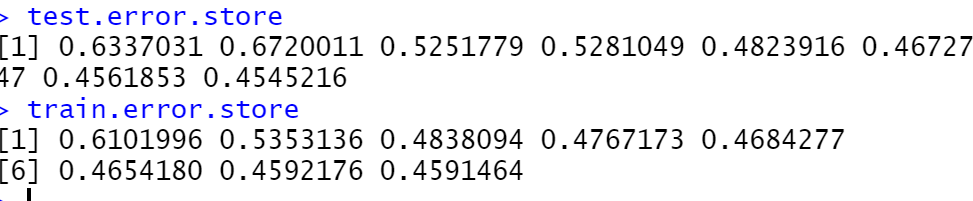
**The Best subset selection CP value by the number of variables in the model.**

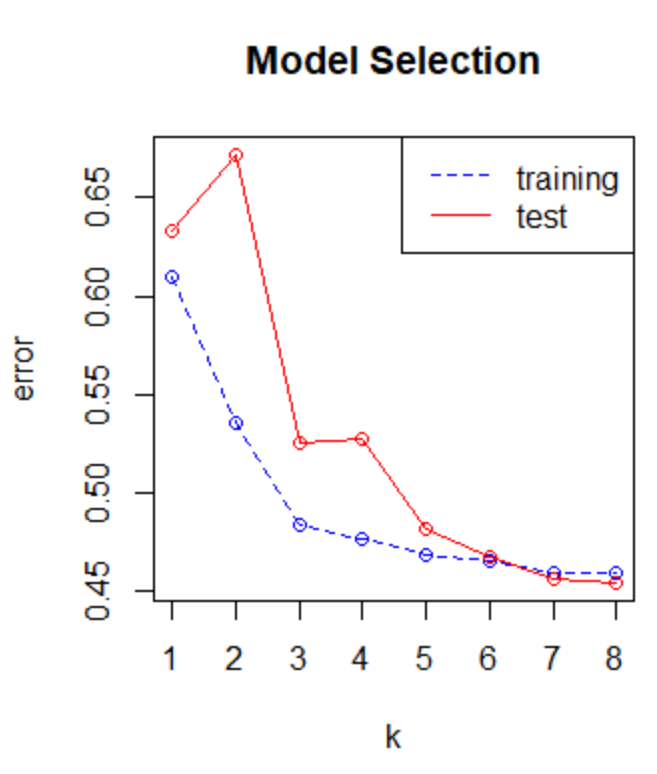


**The Best subset selection CP value by the number of variables in the model.**



**Train and Test MSE by the number of variables( the best subset regression model)**

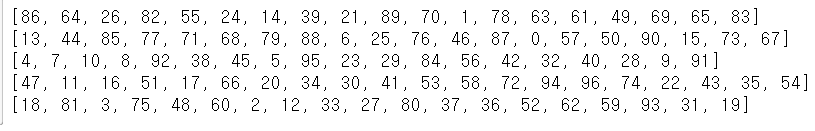




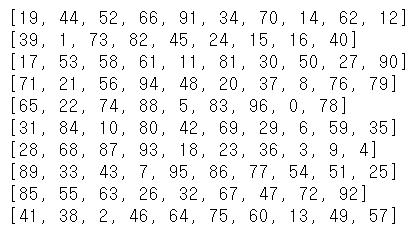
**BIC and AIC values with model of 3 features**



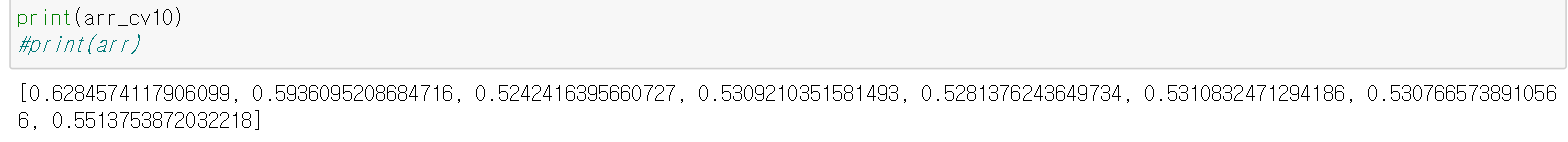
**Index for the Cross Validateion 5**



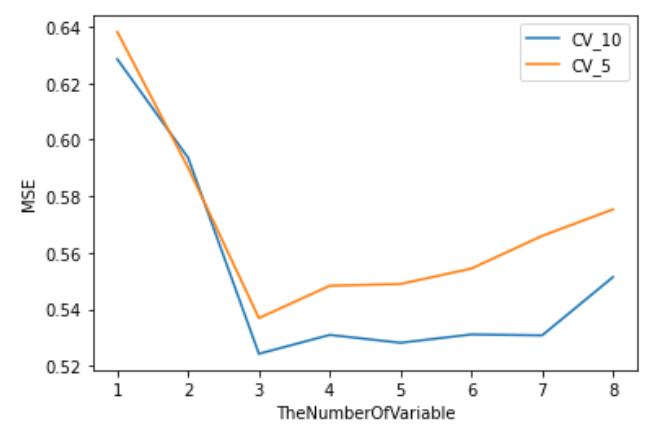
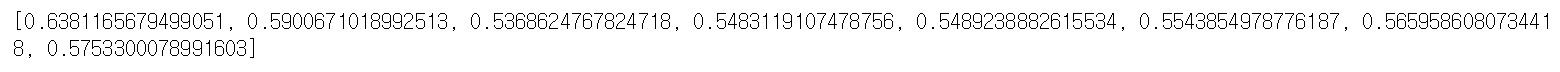
**Index for the Cross Validateion 10**



**CV\_10 MSE by the number of variables**



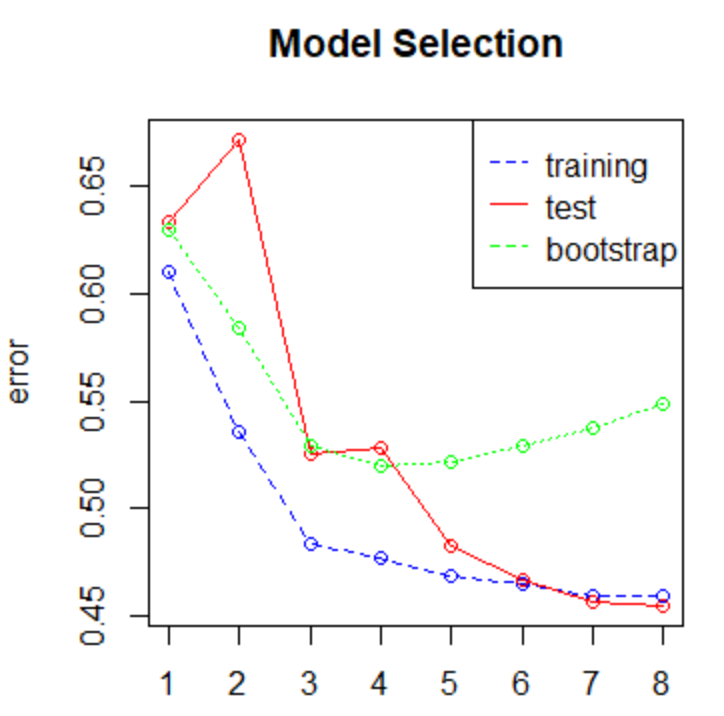
**CV\_5 MSE by the number of variables**



**MSE with bootstrap by the number of variables.**



**Test data, train, and the bootstrap of MSE by variables**



Discussion:

***The number of variables for the Best subset selection***

To find out the number of variables for the model, I calculated CP and BIC values by the model for the number of variables.

*AIC = -2log(likelihood) + 2p*

*BIC = -2log(likelihood) + plog(n)*

*P: variable n: data num*

The result indicated that the model with three variables shows minimum CP and BIC values. This means the model is parsimonious and explainable that maximize likelihood and minimize variables

Cp value with 3 variable model in the model summery: 2.652622

Bic value with 3 variable model in the model summery: -53.15211

AIC values with the selected model : 172.6096

BIC values with the selected model : 184.3286

***Prediction MSE(the model with 3 variables)***

*The best Subset regression(train): 0.4838*

*The best Subset regression(test): 0.525217*

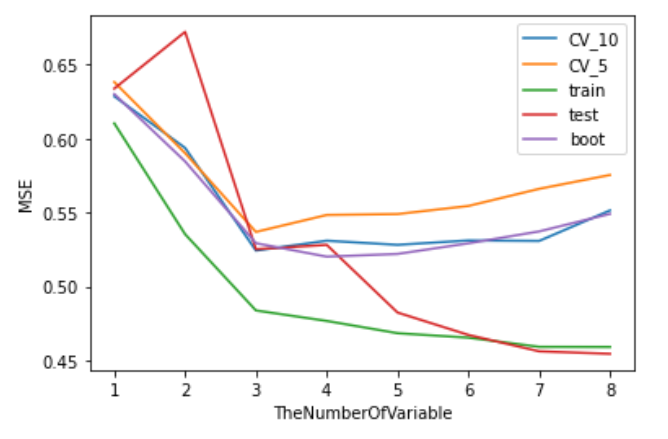
*Cv5 0.5368*

*Cv10 0.5242*

*Boosting 0.529247*

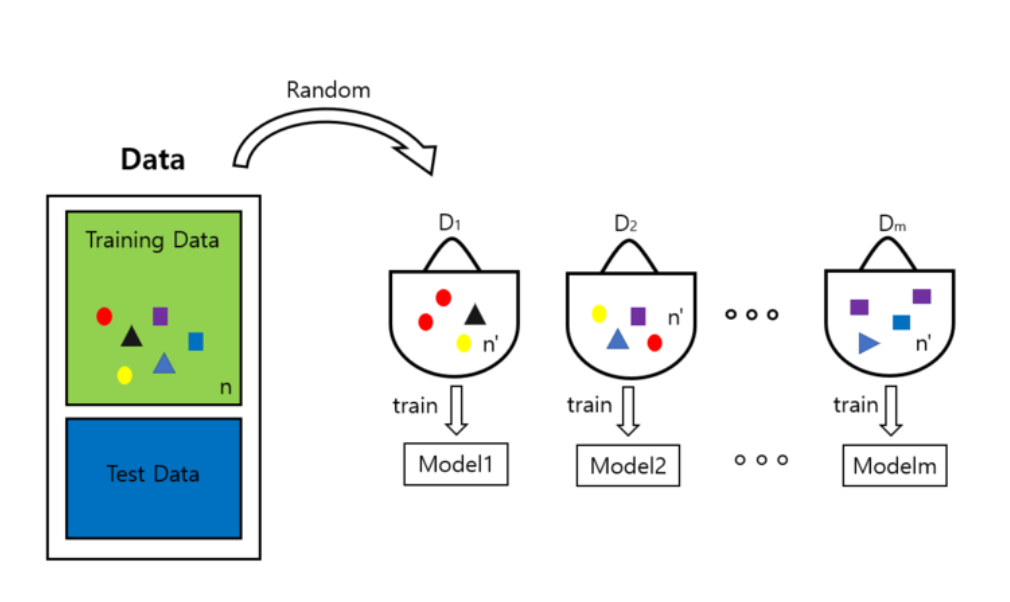
Except for the train MSE, the other 4 model has similar MSE for the result. The reason for the lowest train MSE is overfitting.

***Trend for the MSE by the number of variable(Train, test VS CV\_10, CV\_5, and boot)***

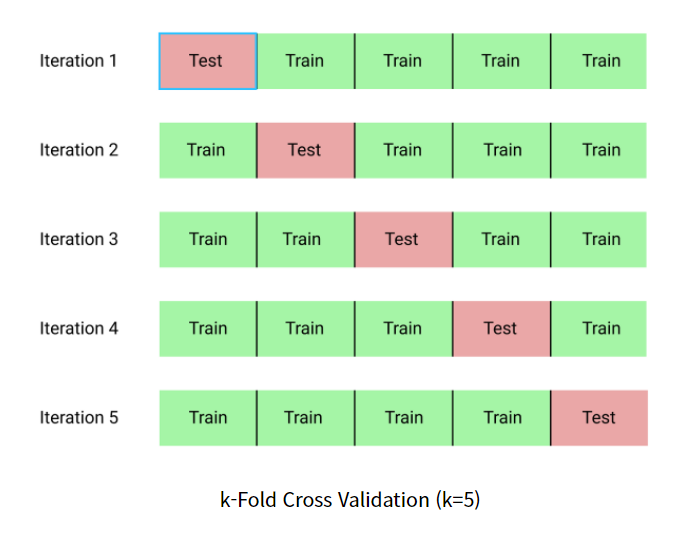


In the picture above, we can see the trend for the MSE by the number of variables at once. All 5 results show that reduce their MSE until 3 variable models. However, after 3 values of ‘TheNumberOfVariable’, the MSE of CV\_10, CV\_5, and Bootstrap model is going up while the Test and Train MSE is going down.

*Bootstrap*



*Cross validation*



When we see the process for bootstrap, the bootstrap model sampling train data several times. Likewise, the crossvalidation model also split randomly and repeat train and test process. These processes enable the model to avoid unfortunate split and overfitting. This can be one of the explanations of the trend in the graph after the value 3 of‘TheNumberOfVariable’ (Train, Test MSE **VS** bootstrap, CV\_10, CV\_5)

Reference :

<https://blog.naver.com/euleekwon/221377799792>

https://neocarus.tistory.com/entry/%EB%AA%A8%ED%98%95%EC%9D%98-%EC%A0%81%ED%95%A9%EC%84%B1%EC%97%90-%EB%8C%80%ED%95%9C-%EA%B5%90%EC%B0%A8%EA%B2%80%EC%A6%9D%EC%9D%84-%EC%88%98%ED%96%89%ED%95%98%EB%8A%94-%EB%B0%A9%EB%B2%95