Assignment Part -II - Subjective Questions and Answers

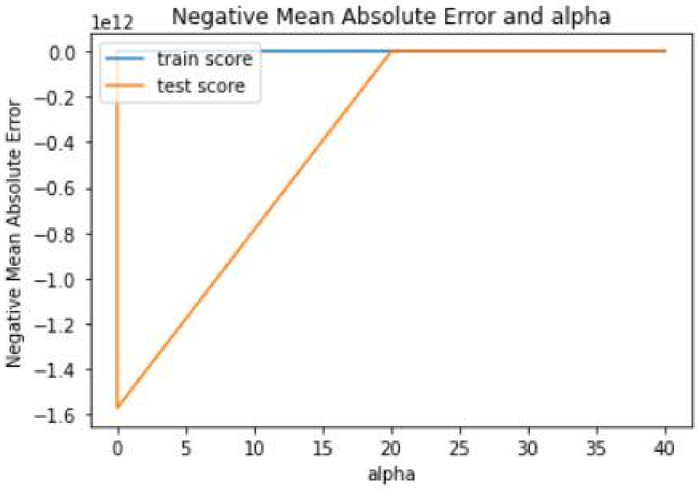
Question 1

What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

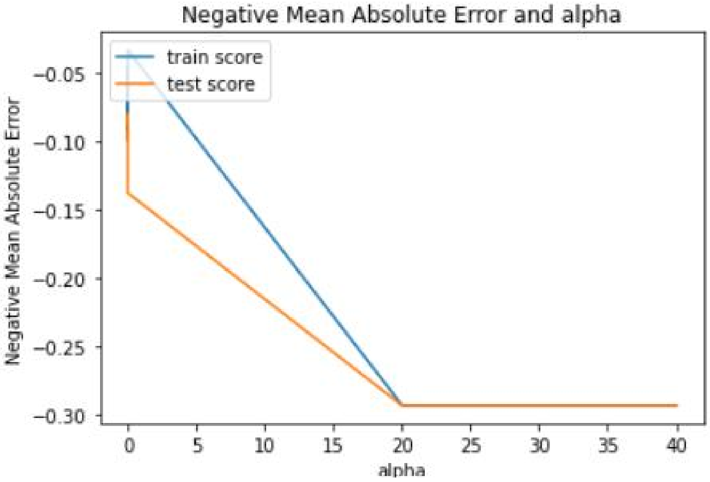
Answer.

Optimal value of alpha for Ridge: 20
Optimal value of alpha for Lasso : 0.0005

For Ridge:



For Lasso:



If we choose the double the value of alpha for ridge and lasso, the, we observed following changes:

1. For Ridge, the R2 for train drops slightly whereas R2 for test drops significantly by ~2%. Also the RMSE value also increases

2. For Lasso as well, we observe similar things, the RMSE value increases from 0.077 to 0.094 and test R2 value drops from 96% to 94%

3. Conceptually, as we increase the alpha in both, Ridge and Lasso, model reduces its complexity and hence move towards underfitting. The same we had observed if we double the alpha values.

After the change is implemented, the important variables are:

For Ridge:
1. Neighborhood_Crawfor
2. CentralAir_Y
3. Neighborhood_Edwards (negative impact)
4. ExterQual_TA
5. OverallCond_3 (negative impact)
6. Foundation_PConc

For Lasso:
1. Neighborhood_Crawfor
2. OverallCond_3 (negative impact)
3. OverallQual_9
4. ExterCond_Fa (nagative impact)
5. CentralAir_Y
6. SaleType_New

Question 2.
You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?
Answer.
For Ridge and Lasso, we have determined the optimal value of lamba as below:
Ridge: 20
Lasso: 0.0005

The RMSE in both the cases are:
Ridge: 0.097
LAsso: 0.077

R2 for Ridge:
Train: 94.31%
Test: 93.78

R2 for Lasso:
Train: 94.21
Test: 96.02

Hence we observe that R2 for Lasso is better than Ridge and also due to feature selection it is more simple model. Hence we will choose Lasso model to apply.

Question 3.
After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?
Answer.
For Ridge:
1. Neighborhood_IDOTRR

2. OverallCond_8
3. OverallQual_9
4. HeatingQC_Fa
5. Age_52

For Lasso:

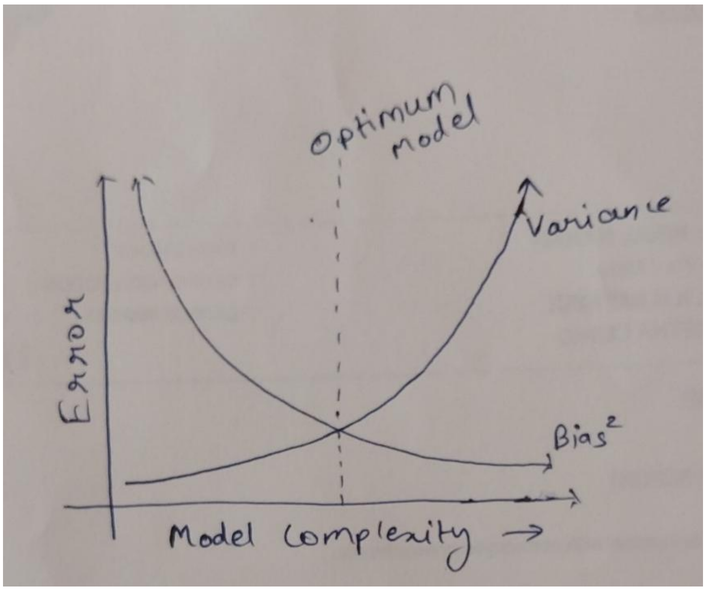
1. Neighborhood_IDOTRR
2. OverallCond_4
3. HeatingQC_Fa
4. Fence_GdWo
5. Age_52

Question 4.

How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

Answer.

- Model should be more and more simpler as simple models are more generic and robust but it decreases its accuracy.
- More complex models are overfitting, they perform good in training data but on unseen data its accuracy is very low.
- Hence there is a trade-off between bias-variance. As we increase the Bias, Variance decreases and vice versa.
- For overfitting model bias is very low and variance is very high because it is tested on unseen data and is not able to perform well.
- It is important to have balance between Bias-Variance and it is possible through Regularization techniques.
- Regularization helps in shrinking the coefficients values towards zero.
- This methods keep the model simple by penalizing the complex model
- Regaularization methods help to get the Bias-Variance tradeoff. It decreases the Bias so that Variance negligibly increases to achieve the optimum position where total error is minimum.


optimum model
Variance
Error
Bias^2
Model Complexity →