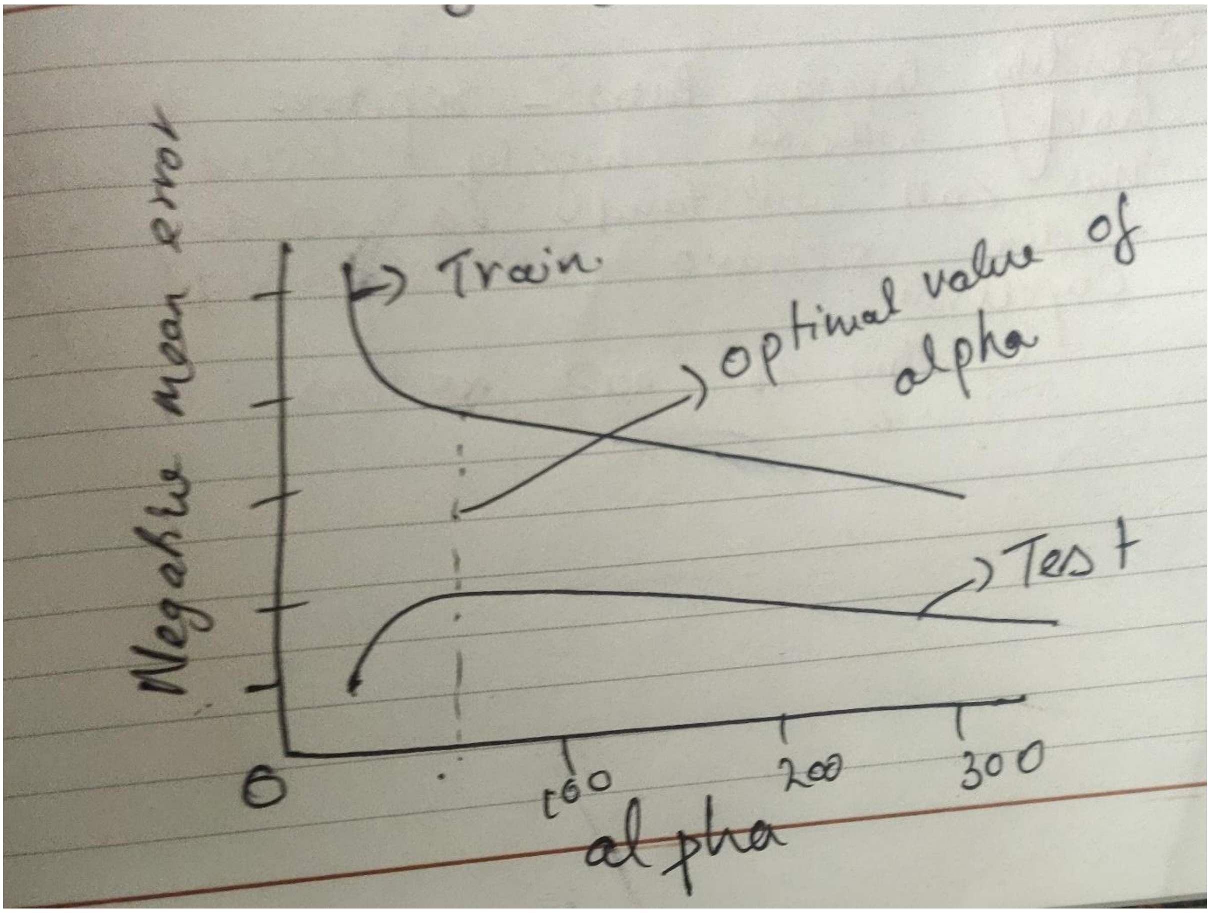
Question 1

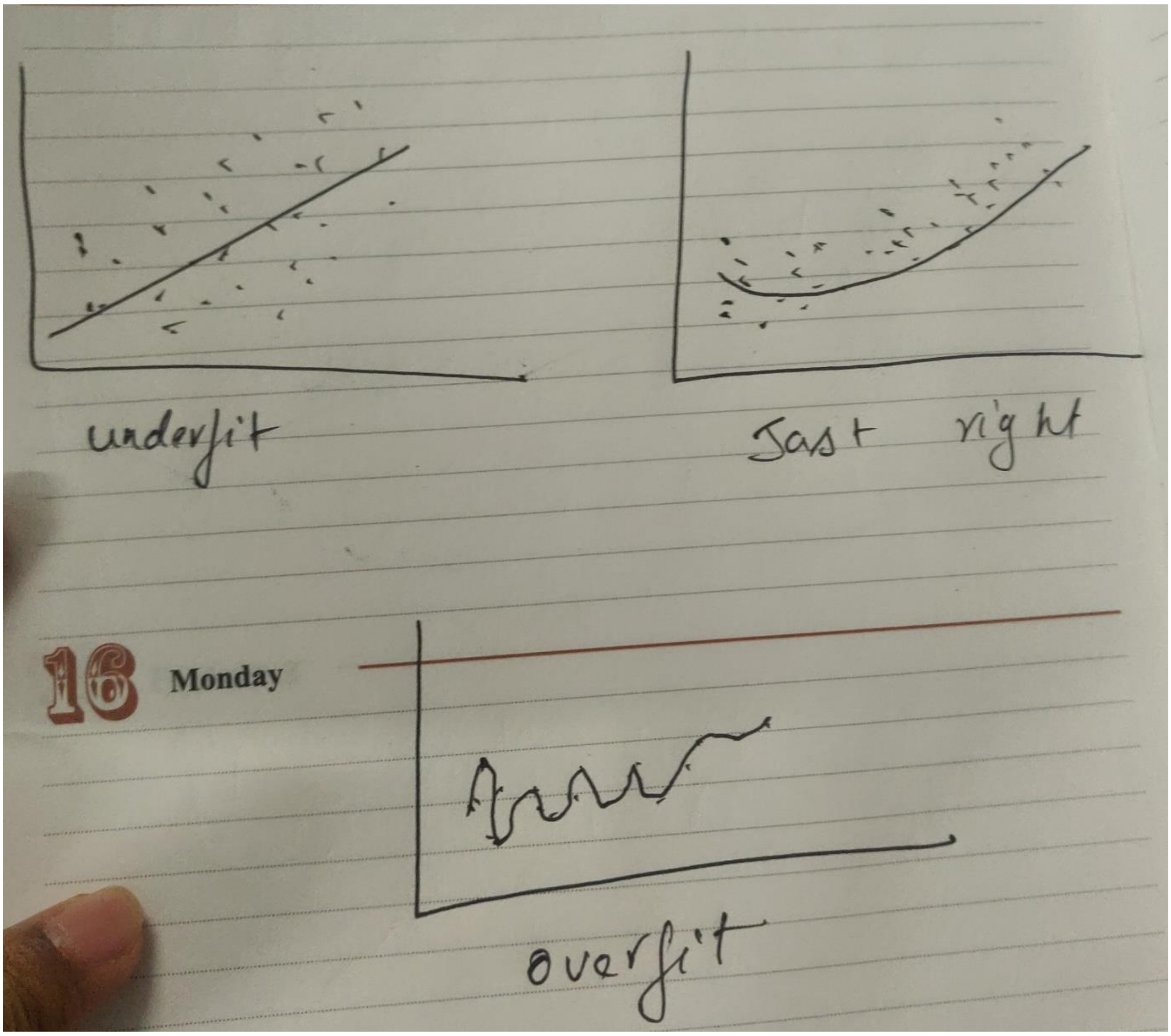
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer

The optimal value of the alpha is determined by hit and trail basis. We need to take a range of alpha value and then execute the cross validation to find the optimal value of alpha. After executing the cross validation with the range of alpha value then we plot a graph against the negative mean error and alpha value. We can see that after a particular value of alpha the graph decreases for both train and test dataset.



If we double the value of alpha then this might lead to overfitting of the model. As higher, the alpha model better the model performance as we are penalizing the features. But, this will lead to overfitting of the model. Therefore, we need to choose the value of alpha carefully to not to over fit the model and neither too small to under fit the model.



The important predictor are the coefficient value, adjusting R2 which helps in determining the predicting value.

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?

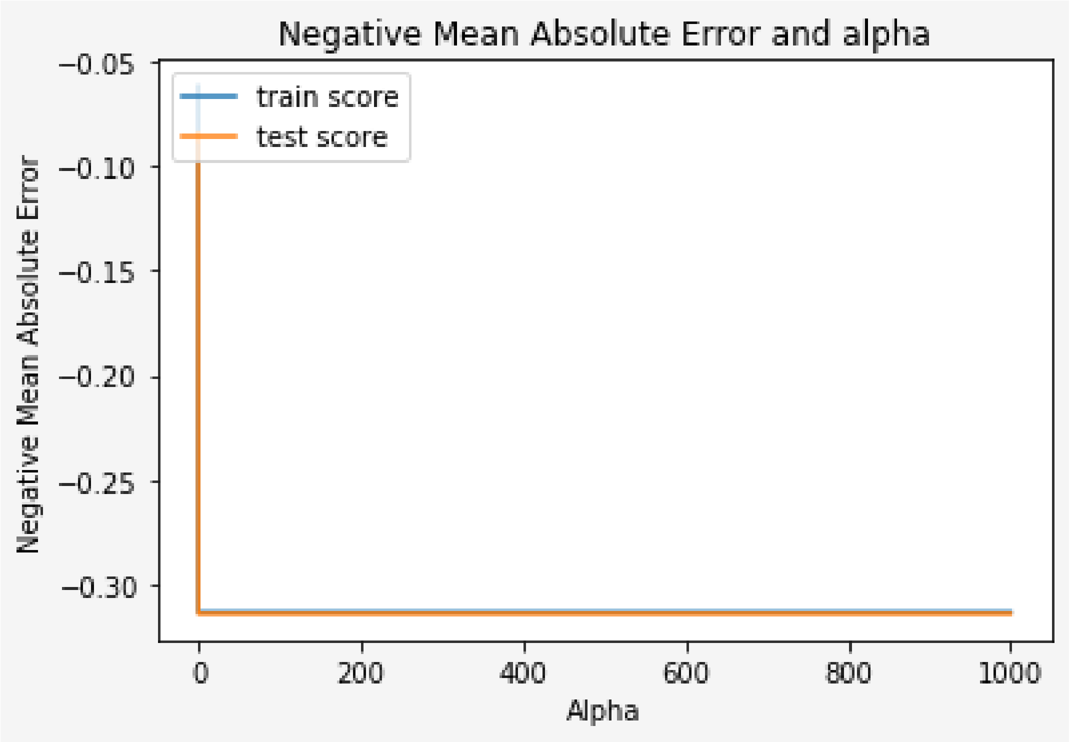
Answer:

We have seen that Ridge has the optimal value of 100 and that of Lasso is 0.001.

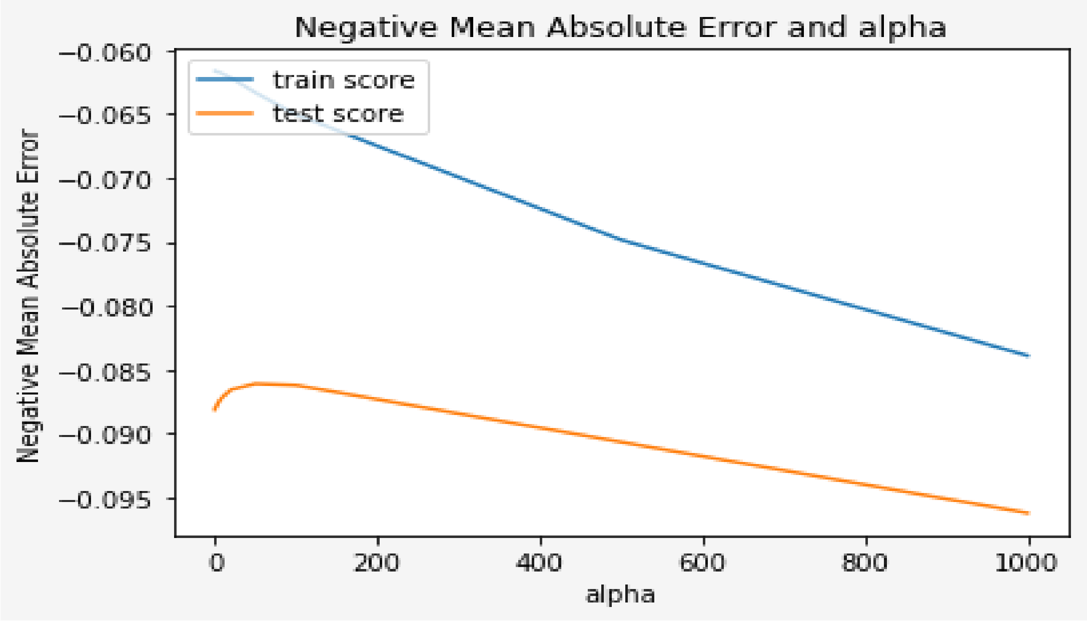
Ridge regression adds a regularization term to penalize the model which has high coefficient. But ridge does not make the value to zero, it is near to zero.

In Lasso regression also add the regularization term to punish high coefficient, but it also set them to zero of they are not relevant in predicting the target value. This can act as feature selection and as well as regularization term.

But for the model that we have prepared the model is performing better with Ridge regression. The model performs better with different values of alpha. As seen below:



Lasso Regression



Ridge Regression

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:

After removing the top five most important predictor from the model and the model is executed again and the next top five predictor model are:

1. MSZoning_C (all)
2. Functional_Typ
3. Neighborhood_crawfor
4. CentralAir_N
5. SaleCondition_Abnormal

Question 4

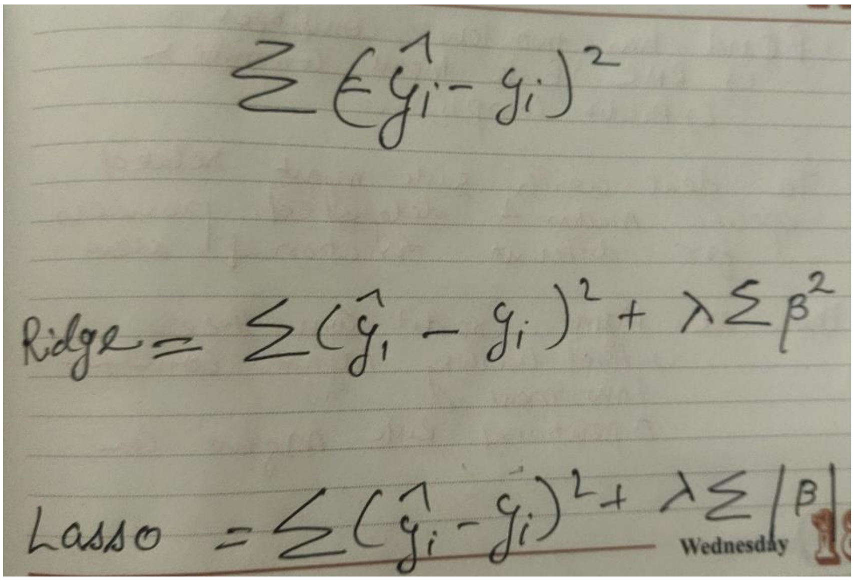
How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

Answer

The model is said to be robust when the model predicts the value correctly when there is slight change in the data. The model predicts the target value accurately when there is change in data, as this does not affect the model. We add the regularization term to make the is said to be robust when we add a regularization term to penalize the coefficient when the value high.

The normal cost function calculates the sum of square of difference between the actual and predicted value. And this is known as RSS(Residual sum of squares)

But, by adding the regularization term we try to penalize the value so that the coefficient value is not high.



$$\Sum(\hat{y}_i - y_i)^2$$

$$\text{Ridge} = \Sum(\hat{y}_1 - y_i)^2 + \lambda \Sum \beta^2$$

$$\text{Lasso} = \Sum(\hat{y}_i - y_i)^2 + \lambda \Sum |\beta|$$

Wednesday

In Ridge regression, we multiply alpha with sum of square of coefficient and in Lasso the sum of magnitude of coefficient.

The other factor, which helps in determining the accuracy of the model are:

- AIC – Akaike Information Criterion
- BIC – Bayes Information Criterion
- Adjusted R2