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?

Answers:

The optimal value of alpha for ridge and lasso Regression 100 and 0.001 respectively.

Changes observed in Models are as follows.

Linear Regression Model : House price prediction

R2 train Score is: 0.9169815891967786

R2 test Score is : 0.9062430917996098

RSS for train data is : 12.178278639970507

RSS for test data is : 6.924258792108792

MSE for train data is : 0.012057701623733175

MSE for test data is : 0.015954513345872792

Ridge Regression analysis for optimal alpha (100) value:

R2 score for training data: 0.9142758648298612

R2 score for test data: 0.9063673040871204

RSS for training data: 12.57519138431804

RSS for test data: 6.915085302491956

MSE for training data: 0.012450684538928752

MSE for test data: 0.01593337627302294

Ridge Regression analysis for double optimal alpha (200) value:

R2 score for training data : 0.911898919681178

R2 score for test data : 0.9048303770025571

RSS for training data : 12.923874285526601

RSS for test data : 7.028592467802612

MSE for training data : 0.012795915134184753

MSE for test data : 0.016194913520282517

It is observed that R2 Score got dropped down and RSS has increased.

Hence, it is not the right fit as higher RSS cannot fit the model better.

Lasso Regression analysis for optimal value (0.001) of alpha:

R2 score for training data : 0.9156975785970994

R2 score for test data : 0.9072346798235644

RSS for training data : 12.366634917912675

RSS for test data : 6.851026726069003

MSE for training data : 0.012244192988032351

MSE for test data : 0.015785775866518442

Lasso Regression analysis for double the optimal value (0.001*2) of alpha:

R2 score for training data : 0.9131831189751465

R2 score for test data : 0.9062499509543072

RSS for training data : 12.735490327319168

RSS for test data : 6.923752220773079

MSE for training data : 0.012609396363682345

MSE for test data : 0.015953346130813548

R2 Score has gone down and RSS value has been increased after doubling the optimal alpha 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: Choosing the Lasso Regression for the best model fit as the RSS value plays an important role. Here the RSS value is lower in Lasso than Ridge. Also, R2 Score is good in Lasso.

Question 3

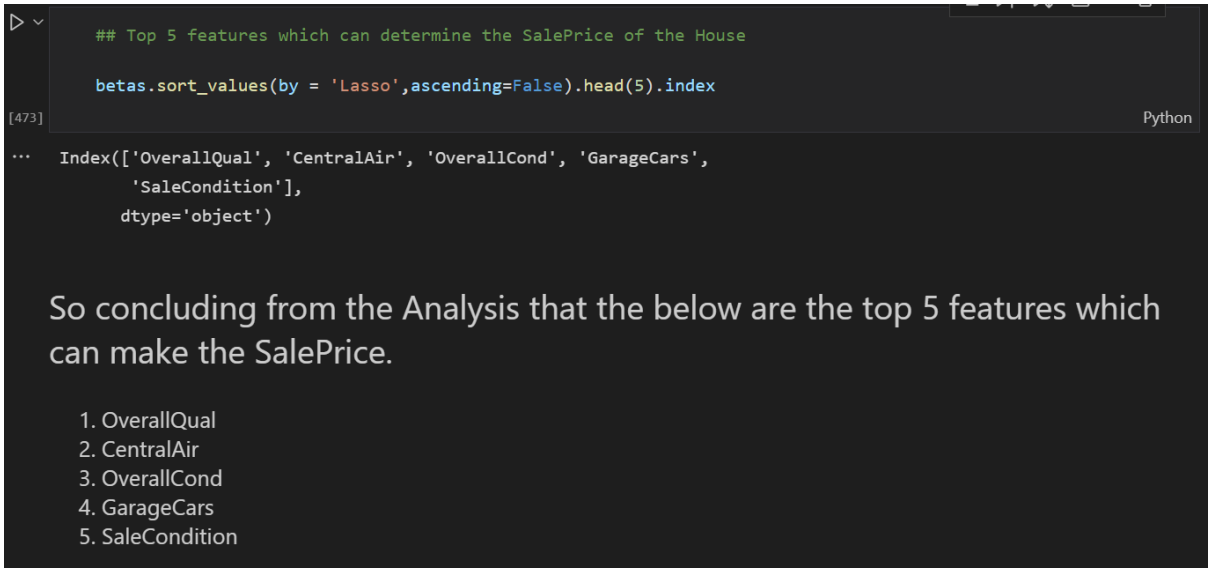
After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding

the five most important predictor variables. Which are the five most important predictor variables now?

Answer:

Below are the top Five features

1. OverallQual
2. CentralAir
3. OverallCond
4. GarageCars
5. SaleCondition



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?

To make sure the model is robust and generalisable,

1. The model should be more general and not so simple. For example, simpler model ends to underfitting of data (high variance and low bias) and complex model leads to Overfitting (low variance and high bias). So, the model should have low variance and low bias and that will be a robust model.
2. And the model should not pay more attention on removing the features if it is showing very less change. For example, outliers should not be removed completely as it may add values to the predictions.
3. The model should have the valid data. (For example, Data cleaning on the features having improper datatypes and null values).