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 alpha value for Ridge- 0.9
Optimal alpha value for Lasso- 4

Effects of doubling the value of alpha:

Ridge regression: (alpha=1.8)

1) R2 score of train- 0.9015
2) R2 score of test- 0.8717
3) MSE (train) – 21126.88
4) MSE (test)- 23250.11

The train R2 score has slightly decreased and the test R2 score has increased. The train MSE has increased and the test MSE has decreased

Lasso regression: (alpha=8)

1) R2 score of train- 0.9055
2) R2 score of test- 0.8224
3) MSE (train) – 20689.28
4) MSE (test)- 27362.66

The train R2 score has slightly decreased and the test R2 score has increased. The train MSE has increased and the test MSE has decreased

The predictors remain the same, but the value of the coefficients are slightly changed. LotArea, OverallQual, OverallCond, YearBuilt, TotalBsmtSF, GrLivArea, TotRmsAbvGrd, Street_Pave, RoofMatl_Metal etc are some important predictors.

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:

Optimal alpha value for Ridge- 0.9
Optimal alpha value for Lasso- 4
R2 score -Ridge (train)= 0.9041704261373944
R2 score – Ridge (test)= 0.8571814420323907
R2 score-Lasso (test)= 0.8571814420323907
R2 score- Lasso (test)= 0.8185965198175262

The difference between test and train R2 score is less in Ridge regression is less when compared to lasso regression. This shows that the model performs well even on the test data. So, Ridge regression would be the choice.

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:

LotArea', 'OverallQual', 'OverallCond', 'YearBuilt', and 'BsmtFinSF1' are the top 5 important Predictor variables.

R2 score -Ridge (train)= 0.8597
R2 score – Ridge (test)= 0.5857
R2 score-Lasso (test)= 0.8622
R2 score- Lasso (test)= 0.4130

The scores have decreased a lot after removing the 5 most important variables.

Top 5 variables after removing-
'GrLivArea', 'KitchenAbvGr', 'ExterQual_Fa', 'BsmtQual_Fa', 'BsmtQual_TA'

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:

Firstly, the accuracy of the model should be very high. The model should perform well even on test data and even on unseen data. If it performs well only on the training data, this means that the model is overfitting. Also, it should have good R2 scores and less error. P value should be less than 0.05 and VIF should be less than 5. In general, the model should perform well on any dataset.