Question 1

What is the optimal value of alpha for ridge and lasso regression?

Optimal value of alpha:

Lasso 0.001

Ridge 5.0

What will be the changes in the model if you choose double the value of alpha for both ridge and lasso?

For Lasso:

r2_score drops, also more number of columns have coefficients 0 and can be eliminated

For Ridge:

r2_score slightly drops.

What will be the most important predictor variables after the change is implemented?

Lasso:

Significant Feature	Coef
OverallQual	0.792631
TotRmsAbvGrd	0.359551
GarageArea	0.344564
BsmtFullBath	0.184843
Fireplaces	0.178119

Ridge:

Significant Feature	Coef
OverallQual	0.363363
GarageArea	0.239136
TotRmsAbvGrd	0.220802
FullBath	0.180535
2ndFlrSF	0.169874

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:

Lasso regression has r2_score which is more generalizable, and also it helps to eliminate the independent variables by reducing coefficients to 0.

Greater alpha, the greater is the regularization.

Regularization penalizes model complexity, therefore the higher the alpha, the lesser is model complexity, decreasing the error due to variance (overfit).

At the same time higher alphas increase the error due to bias (underfit).

So, we need to choose an optimal Alpha such that the error is minimized in both direction**s. Hence will choose and apply Lasso regression with optimal alpha 0.001**

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

Answer:

1. MasVnrArea

2. OverallCond

3. YearRemodAdd_Age

4. MSZoning_RH

5. MSZoning_RM

6. HeatingQC

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:

A model is robust and generalizable when it performs with almost same accuracy on the unseen data as on trained data.

 Model should not be impacted by outliers in training data.

 Example - Data is within the variable boundary. Giving weightage to outliers can give high accuracy for the model with train data but may not be so with unseen test data. (i.e. Overfitting)

So, to make the model robust – only relevant outliers should be retained. This implies – low accuracy of model with train data as the predictions for outliers but a better accuracy with unseen or test data.