ADVANCED REGRESSION ASSIGNMENT

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?

Based on the Model execution completed, my model gave the below Best params.
Best alpha value for Lasso: {'alpha': 0.001}
Best alpha value for Ridge: {'alpha': 10.0}
Based on the Doubling factor of the alpha value the score remains same in both Lasso and Ridge whereas, there is some change in the coefficient values.

Lasso Regression

	Featuere	double Coef
27	Neighborhood_OldTown	0.480347
26	Neighborhood_NridgHt	0.441699
32	Condition1_Norm	0.431969
7	GarageArea	0.348412
72	Exterior2nd_Other	0.316468
57	Exterior1st_BrkComm	0.300106
1	BsmtQual	0.282245
46	HouseStyle_2.5Fin	0.235078
69	Exterior2nd_CmentBd	0.211573
14	LotConfig_CulDSac	0.208250

Ridge Regression

	Feaure	Double Coef
7	GarageArea	0.322574
27	Neighborhood_OldTown	0.283950
1	BsmtQual	0.283564
26	Neighborhood_NridgHt	0.269342
46	HouseStyle_2.5Fin	0.225132
32	Condition1_Norm	0.224319
2	BsmtExposure	0.173249
15	LotConfig_FR2	0.159973
33	Condition1_PosN	0.146956
14	LotConfig_CulDSac	0.145791

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?

Best Parameters are listed below
Best alpha value for Lasso: {'alpha': 0.001}
Best alpha value for Ridge: {'alpha': 10.0}
R2Score for Lasso Training: 0.8595646392358296
R2Score for Lasso Test: 0.8618444610437768
R2Score for Ridge Training: 0.8484512863885689
R2Score for Ridge Test: 0.8568740959515067
Since Lasso has a better R2 score and considers Feature reduction we can see that Lasso can give a better performance than Ridge.

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?

As per the model (Lasso Regression with best Parameter of 0.001) we have the below best predictors (Features)

1. 'HouseStyle_2.5Fin'
2. 'Condition1_PosN'
3. 'Condition1_Norm'
4. 'Neighborhood_OldTown'
5. 'Neighborhood_NridgHt'

R2 of the new model without the top 5 predictors drops to 0.8481997462369055

1. RoofMatl_CompShg
2. Condition2_PosN
3. Condition2_Norm
4. Neighborhood_SWISU
5. Neighborhood_SawyerW

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?

Simple models are more generalizable as they can be scalable and can be widely applicable. Simpler models require fewer training samples for effective training than the more complex ones and hence are easier to train.
Regularization can be used to make the model simpler. Regularization helps to strike the delicate balance between keeping the model simple and not making it too naive to be of any use.

Bias quantifies how accurate is the model likely to be on test data. A complex model can do an accurate job prediction provided there is enough training data. Models that are too naïve, for e.g., one that gives same answer to all test inputs and makes no discrimination whatsoever has a very large bias as its expected error across all test inputs are very high.
Variance refers to the degree of changes in the model itself with respect to changes in the training data.