Advance 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?

Ans.

The most optimal value are as follows:

For Lasso regression it is: **0.001**

For Ridge regression it is: **20**

If we double the value of alpha the model will be more generalize and its training error may increase:

For lasso the r2 score for train and test will get affected and will reduce to **0.8867** for train and **0.885** for test.

For Ridge the r2 score for train will be **0.903** and **0.892** for test.

The most important predictor variable will be as follows:

For Lasso regression:

	Feature	Coefficient
0	const	11.895641
3	OverallQual	0.109049
10	2ndFlrSF	0.066971
9	1stFlrSF	0.066962
20	GarageCars	0.059945
55	Neighborhood_Crawfor	0.053281
75	Condition1_Norm	0.051178
114	Exterior1st_BrkFace	0.049545
152	Foundation_PConc	0.042213
65	Neighborhood_NridgHt	0.042059

For Ridge Regression:

	Feature	Coefficient
0	constant	11.863700
3	OverallQual	0.090310
10	2ndFlrSF	0.061026
9	1stFlrSF	0.052205
55	Neighborhood_Crawfor	0.050153
114	Exterior1st_BrkFace	0.046533
20	GarageCars	0.045900
75	Condition1_Norm	0.044510
71	Neighborhood_StoneBr	0.043872
65	Neighborhood_NridgHt	0.040025

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?

Ans.

Lasso Regression eliminate some variables by making their coefficient equal to zero so lasso regression proves to be better than ridge regression. Moreover, if we talk about their model prediction and accuracy, both tend to give almost same result.

Eliminating the variables gives lasso regression an edge over ridge regression as it makes the model simpler to interpret.

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?

Ans.

After removing the top five important predictor variables, the variable important in Lasso are as follows:

	Feature	Coefficient
0	const	10.700239
107	RoofMatl_WdShngl	1.286724
106	RoofMatl_WdShake	1.180355
101	RoofMatl_CompShg	1.157105
105	RoofMatl_Tar&Grv	1.085711
104	RoofMatl_Roll	1.065561

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?

Ans.

We can make sure that the model is robust and generalised by implementing the techniques of advance regression as follows:

- Understanding the trade-off between Bias and Variance of our model.
- Implementing the Cross-Validation technique with multiple folds.
- Using Ridge and Lasso regression techniques.
- Performing the Hyper Parameter Tuning to select the optimal value of alpha (Hyper-parameter).
- Rebuilding the model with tuned value of alpha.

The accuracy is not the only measure we should look on while model building because it can be biased towards our training data. We should always try to generalise our model to make it more robust. Getting high value of accuracy on training data set and low accuracy on test set implies that our model is getting overfitted and that's not we want from our model. So, we should always focus on generalising our model rather than increasing the accuracy of the model.