Problem Statement - Part II

Assignment Part-II

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?

Ridge (Optimal alpha value)-20

Lasso(Optimal alpha value)-20

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.058867e-01	9.041197e-01	9.056574e-01
1	R2 Score (Test)	-4.851794e+22	8.640544e-01	8.582805e-01
2	RSS (Train)	6.012594e+11	6.125482e+11	6.027243e+11
3	RSS (Test)	1.367581e+35	3.831915e+11	3.994666e+11
4	MSE (Train)	2.425524e+04	2.448187e+04	2.428476e+04
5	MSE (Test)	1.767012e+16	2.957814e+04	3.019974e+04

Top Feature- BsmtQual_Gd

After Doubling the alpha to ridge and lasso both, Here is the metrics

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.058867e-01	9.021148e-01	9.052711e-01
1	R2 Score (Test)	-4.851794e+22	8.676223e-01	8.608359e-01
2	RSS (Train)	6.012594e+11	6.253564e+11	6.051921e+11
3	RSS (Test)	1.367581e+35	3.731346e+11	3.922635e+11
4	MSE (Train)	2.425524e+04	2.473651e+04	2.433443e+04
5	MSE (Test)	1.767012e+16	2.918742e+04	2.992622e+04

Top Predictor variables

1. BsmtQual_Gd
2. BsmtQual_TA
3. KitchenQual_TA
4. KitchenQual_Gd
5. HouseAge

After comparing both the metrices, the model performed slightly better on test data by doubling the alpha.

Top Feature- BsmtQual_Gd (this remains same)

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?

I would prefer Ridge(alpha=40) as the performance on R2 Score on Test data is better than Laso.

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?

After deleting the 5 most important predictors., here are the 5 most important predictor variables

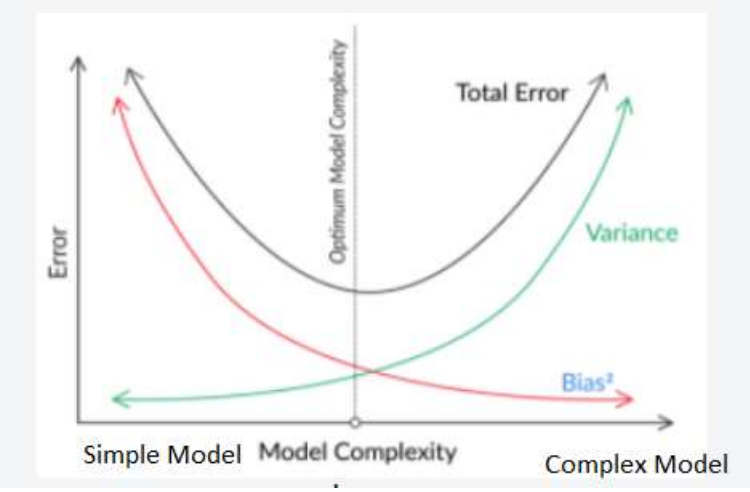
1. Neighborhood_Edwards
2. BsmtFinType1_Unf
3. KitchenAbvGr
4. BsmtExposure_No
5. OverallCond_Average

Question 4

How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

A simple model is more robust and generalizable but won't give accuracy as it majorl y fails on test data. In simple terms, A simple model would usually have high bias and low variance, whereas a complex model would have low bias and high variance. In ot her words, bias in a model is high when it does not perform well on the training data it self, and variance is high when the model does not perform well on the test data.

So, if a model is too simple the accuracy will be low as bias, and variance will be low as shown below: -



In order to make this robust and generalizable which neither overfit or underfit - comp lexity needs to be managed: It should neither be too high, which would lead to overfitt ing, nor too low, which would lead to a model with high bias (a biased model) that do es not even identify necessary patterns in the data. This can be achieved by using Reg ularization which helps with managing model complexity by essentially shrinking the model coefficient estimates towards 0. This discourages the model from becoming too complex, thus avoiding the risk of overfitting.