Advanced Linear Regression Subjective Questions

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
In the given dataset the optimal value for alpha for Lasso and Ridge regression is '0.0004'
If we double the alpha value, then train R2 to decrease and test R2 to increase.

Alpha: 0.0004
Training R2
0.892052186581573
Testing R2
0.8265210310687388

Alpha: 0.0008
Training R2
0.8670772664783504
Testing R2
0.836063024168417

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 is more useful with this dataset since after dummy encoding the no of features reached 256 and with Lasso, we can reduce the no of features.

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?

Top 5 features

OverallQual	0.245500
TotRmsAbvGrd	0.120509
GarageCars	0.108760
YearRemodAdd	0.043824
CentralAir_Y	0.030067

top5 = ['OverallQual', 'TotRmsAbvGrd', 'GarageCars', 'YearRemodAdd', 'CentralAir_Y']

The new optimal alpha after dropping the top 5 columns is : 0.00031

Now the top 5 most important features are

GrLivArea	0.413085
OverallCond	0.110985
GarageArea	0.072033
Neighborhood_StoneBr	0.064200
YearBuilt	0.057666

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?

A model needs to be made robust and generalizable so that they are not impacted by outliers in the training data. The model should also be generalisable so that the test accuracy is not lesser than the training score. The model should be accurate for datasets other than the ones which were used during training. Too much weightage should not be given to the outliers so that the accuracy predicted by the model is high. To ensure that this is not the case, the outlier analysis needs to be done and only those which are relevant to the dataset need to be retained. Those outliers which it does not make sense to keep must be removed from the dataset. This would help increase the accuracy of the predictions made by the model. This would help standardize the predictions made by the model. If the model is not robust, it cannot be trusted for predictive analysis.