ADVANCED REGRESSION SUBJECTIVE QUESTIONS

Question 1

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

- For Ridge: 20
- For Lasso: 0.001

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

- Negative Mean squared error will decrease

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

- GrLivArea, OverallQual, BsmtFinSF1, GarageCars, SaleType_New, OverallQual, SaleCondition, TotalBsmtSF, MSZoning_FV, LotArea

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?

Optimal alpha for Ridge as 20 and for Lasso as 0.001.

In terms of train and test R2 and Adjusted R2 for both the models remains the same.

- Ridge_train_score 0.92
- Ridge_test_score 0.85
- Adjusted_R2_score_Ridge_Train 0.90
- Adjusted_R2_score_Ridge_Test 0.81

- Lasso_train_score 0.92
- Lasso_test_score 0.85
- Adjusted_R2_score_Lasso_Train 0.90
- Adjusted_R2_score_Lasso_Test 0.81

Theoretically, Lasso tends to do well if there are a small number of significant parameters and the others are close to zero and Ridge works well if there are many large parameters of about the same value.

In this case, we have many large parameters - so we can choose Ridge though we have similar results as things stand, however the results can change in the live data.

If we want the model to do the feature elimination by itself, then Lasso would be a good choice in this case, considering the large number of variables.

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?

- GarageCars, YearRemodAdd, TotalBsmtSF, OverallQual_5, BsmtFinSF1

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?

The robustness is the property that characterizes how effective an algorithm is while being tested on the new independent (but similar) dataset. The robust algorithm tends to have the testing error close to the training error.

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. Outlier analysis to be carried out properly so that the train and test produces the proper results.

If the model is not generalised then we will have a better accuracy in train and the model will not perform the same in the test, because we made the model to learn each and every aspect of the data and didn't generalise it.