Assignment Part II – 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:

➔ The optimal values of lambda are:

- Ridge: 10
- Lasso: 100

➔ The r2 score for the optimal values of lambda comes out to be as:

	R2 Score Train	R2 Score Test
Ridge (alpha=10)	0.9417942155185163	0.920382076055638
Lasso (alpha=0.001)	0.9308150158241388	0.9194788349631385

When the values of lambda are doubled:

	R2 Score Train	R2 Score Test
Ridge (alpha=20)	0.9376489425319786	0.9197129231744094
Lasso (alpha=0.002)	0.9192883582503254	0.913965377201741

- When the value of lambda is doubled, there is a slight difference in the r2 scores for both ridge and lasso.

- The important variables after the value of alpha is doubled are:

	Lasso new
Neighborhood_Crawfor	0.097230
SaleType_New	0.086747
Functional_Typ	0.070421
OverallQual	0.061974
MSZoning_FV	0.061197
MSZoning_RL	0.048202

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:

The optimal values of lambda for ridge is 10 and for lasso is 0.001. When building models are based on these values then there isn't a significant difference in the performance of both models. The r2_score for the test data shows no significant difference.

However as when you look at the model parameters, Ridge does not make the coefficients zero, while on the other hand Lasso does make the coefficients of quite a few variables zero, thus helping in feature selection. Hence it would be better to use Lasso regression with lambda set to 0.001.

Lasso helps in reducing the features in the model, helping to create a simpler final model. This is important for creating a robust and generalizable model.

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?

Answer:

After the top 5 features are dropped the next most important predictors become **MSZoning_RL, SaleCondition_Normal, Neighborhood_StoneBr, OverallCond, Condition1_Norm.**

	Lasso new
Neighborhood_Crawfor	0.097230
SaleType_New	0.086747
Functional_Typ	0.070421
OverallQual	0.061974
MSZoning_FV	0.061197
MSZoning_RL	0.048202
SaleCondition_Normal	0.047181
Neighborhood_StoneBr	0.047052
OverallCond	0.045083
Condition1_Norm	0.044492
Neighborhood_NridgHt	0.043187

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?

Answer:

A model can be considered robust and generalizable if it is shows no drastic change in performance when the training set is changed, i.e. the model should not overfit on the training data and should be able to handle new/unseen data properly.

When it comes to accuracy, a model which is robust and generalizable should perform equally well on both the training and test data.

It is also important to consider the values obtained for train and test, so that the model will perform well on unseen data. This means that the data should retain some outliers to help with predictions. As demonstrated in the assignment, accuracy of the model will vary, depending on the way data is processed and how features are selected. There may be no perfect model, but different steps are available to ensure that the model developed is fit for purpose for the specific context and the uniqueness of the business case.

This is in line with Occam's razor, that is, the model to be chosen should not be more complex than it needs to be