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 are :

- For Ridge Regression : 20.0
- For Lasso Regression : 0.001

1) looks like after double the value of alpha, R2 score for training dropped slightly and R2 score went up very slightly.

2) model_coefficients also decreased while double the value of alpha.

Below are the most important predictor variables after double the value of alpha.

['1stFlrSF', '2ndFlrSF', 'OverallQual', 'OverallCond', 'MSZoning_RL']

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 :

We will choose Lasso regression as it is giving the feature selection option also with 89% accuracy on train data and 87% accuracy on test dataset. It has removed unwanted features from the model without affecting model accuracy which makes the model simple and accurate.

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 :

The top five features with actual model were :

1stFlrSF	0.093158
2ndFlrSF	0.092032
OverallQual	0.087899
OverallCond	0.037571
MSZoning_RL	0.035395

After removing the above feature, new top 5 features are :

TotRmsAbvGrd	0.088227
FireplaceQu	0.061937
GarageArea	0.049930
KitchenQual	0.047112
FullBath	0.044832

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?

Answer:

To make the model generalised and robust, here are the top 3 features :

- Model Accuracy should be > 70-75%, in our case it is coming 89% for train and 87% for test dataset which is considered as the good model.
- P-value should be < 0.05
- VIF should be < 5