ADVANCED REGRESSION ASSIGNMENT

SUBJECTIVE QUESTIONS AND ANSWERS

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:

- **Ridge Optimal Alpha value: 3**
- **Lasso Optimal Alpha value: 0.001.**
- **We have R2 on Train as 83% and test R2 is 77%**
- ***After doubling the alpha values in the Ridge and Lasso, the prediction accuracy increases***
  We have R2 on Train as 94.5% and R2 Test is 87.9%

Lasso: Important Predictor variables (in alphabetical order):

Predictor
1stFlrSF
2ndFlrSF
Age
BsmtFinSF1
BsmtUnfSF
LogLotArea
Neighborhood_MeadowV
OverallCond
OverallQual_FA
OverallQual_VP

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 optimum lambda value in case of Ridge and Lasso is as follows:-

- **Ridge – 3**
- **Lasso – 0.0001**

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.834451	0.943874	0.945984
1	R2 Score (Test)	0.766122	0.877722	0.879035
2	RSS (Train)	23.958079	8.122481	7.817114
3	RSS (Test)	15.138004	7.914552	7.829625
4	MSE (Train)	0.155017	0.090260	0.088547
5	MSE (Test)	0.188067	0.135985	0.135254

Lasso is preferred over Simple Multiple linear regression or Ridge regression as the coefficient values get reduced to nearly zero).

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 5 most important predictor variables in the current lasso model are:

- **Total_sqr_footage**
- **Age**
- **LogLotArea**
- **OverallCond**
- **OverallQual_FA**

If we build a Lasso model after removing these attributes from the dataset the R2 of the new model drops significantly and the Mean Squared Error increases to 0.0028575670906482538

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: When comparing two models that show similar 'performance' in the finite training or test data, it is suggested that we should pick the one that makes fewer error on the test data due to following reasons:

- **Simpler models are normally 'generic' and are widely applicable**
- **Simpler models in general require fewer training samples**
- **Higher errors are observer in Simpler models make the training set.**
- **Complex models lead to overfitting and require more tarining**
- **Simple models**
    a. **low variance**
    b. **high bias**
- **Complex models**
    a. **Low bias,**
    b. **High Variance**
- **Complex models - Require extensive training and the can lead to overfitting.**
- **Frequent and unexpected changes are often observed in the Complex models variance**
- **Regularization can be used to make the model simpler. Regularization helps to strike the delicate balance between keeping the model simple and not making it too naive to be of any use. For regression, regularization involves adding a regularization term to the cost that adds up the absolute values or the squares of the parameters of the model.**
- **Also, Making a model simple leads to Bias-Variance Trade-off, since a complex model will need to change for every little change in the dataset in contrast a simpler model will create an abstract pattern based on the training data set.**