ADVANCED REGRESSION ASSIGNMENT – SUBJECTIVE QUESTIONS

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 REGRESSION:

The optimal value of alpha comes out to be 3. The top 5 most significant variables are :

	Weightage	Impact
GrLivArea	0.348239	Positive
1stFlrSF	0.273066	Positive
LotArea	0.218278	Positive
Condition2_PosN	0.205354	Negative
OverallQual_Excellent	0.169183	Positive

Root Mean Square Error train = 0.104
Root Mean Square Error test = 0.126
R2 score on train data = 0.932
R2 score on test data = 0.898

If alpha value is doubled i.e. alpha = 6;

The top 5 most significant variables comes out to be:

	Weightage	Impact
GrLivArea	0.268813	Positive
1stFlrSF	0.233188	Positive
OverallQual_Excellent	0.168135	Positive
LotArea	0.166431	Positive
FullBath	0.160829	Positive

Root Mean Square Error train = 0.109
Root Mean Square Error test = 0.126
R2 score on train data = 0.924
R2 score on test data = 0.898

The R2 score on train data has slightly reduced. However, for test data, it is the same.

LASSO REGRESSION:

The optimal value of alpha comes out to be 0.0001. The top 5 most significant variables are :

	Weightage	Impact
GrLivArea	1.042659	Positive
Condition2_PosN	1.038558	Negative
RoofMatl_WdShngl	0.663887	Positive
RoofMatl_CompShg	0.582522	Positive
RoofMatl_WdShake	0.577456	Positive

Root Mean Square Error train = 0.09
Root Mean Square Error test = 0.139
R2 score on train data = 0.947
R2 score on test data = 0.876

If alpha value is doubled i.e. alpha = 0.0002;

The top 5 most significant variables comes out to be:

	Weightage	Impact
GrLivArea	1.017767	Positive
Condition2_PosN	0.857384	Negative
LotArea	0.382956	Positive
MSZoning_RH	0.366281	Positive
MSZoning_FV	0.362100	Positive

Root Mean Square Error train = 0.097
Root Mean Square Error test = 0.130
R2 score on train data = 0.94
R2 score on test data = 0.891

The R2 score on train data has remained almost same. However, for test data, it has increased.

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 –

Ridge would be the appropriate choice. Reasons for this are as described below:

Comparing the Ridge and Lasso models:

	RIDGE MODEL	LASSO MODEL
Root Mean Square Error train	0.104	0.09
Root Mean Square Error test	**0.126**	0.139
R2 score on train data	0.932	0.947
R2 score on test data	**0.898**	0.876

As is evident form the above table;

a. RMSE on test data is better for Ridge model
b. R2 score of Ridge model is significantly higher
c. The R2 scores on test and train data are closer for Ridge model, indicating that the model has generalized well.

Also, if we analyse the top 5 significant variables prescribed by the two models:

RIDGE MODEL

	Weightage	Impact
GrLivArea	0.348239	Positive
1stFlrSF	0.273066	Positive
LotArea	0.218278	Positive
Condition2_PosN	0.205354	Negative
OverallQual_Excellent	0.169183	Positive

LASSO MODEL

	Weightage	Impact
GrLivArea	1.042659	Positive
Condition2_PosN	1.038558	Negative
RoofMatl_WdShngl	0.663887	Positive
RoofMatl_CompShg	0.582522	Positive
RoofMatl_WdShake	0.577456	Positive

Lasso model puts 3 different values of RoofMatl in the top 5 whereas Ridge model prescribes all different variables, which seems more realistic intuitively.

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 excluding the top 5 variables and re-building the Lasso model, following variables come out to be most significant:

	Weightage	Impact
1stFlrSF	0.652521	Positive
LotArea	0.303309	Positive
GarageCars	0.261508	Positive
OverallQual_Excellent	0.248108	Positive
FullBath	0.164275	Positive

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 –

A model can be made more robust and generalizable if it is ensured that the model does not overfit the training data. i.e. the model has low variance. Overfitting can be kept in check by the following methods:

1. The dataset should be large enough in comparison to the number of features.
   Ideally, dataset size should be at least 10 times the no. of features.

2. The model should not be made too complex by including all the variables available. The variables should be judiciously selected through any one or a mix of the below given methods:
   a. Business domain knowledge
   b. Manual Feature Selection (based on p-value)
   c. Recursive Feature Elimination technique
   d. Regularization – Ridge , Lasso, Elastic Net

   <u>Implication on Accuracy:</u>
   A model that overfits shall have high accuracy on training data as it would have learnt the complete data set to bring forth the model. However, such a model shall perform poorly on unseen data in most cases.

   Vice-versa, a robust and generalized model shall have a moderate accuracy on training data and a good enough accuracy on unseen data.
   Lesser the difference between train and test accuracy, more generalized the model is.