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 alpha for lasso and ridge model:

- **Ridge Model:** Optimal value of alpha = 0.5
- **Lasso Model:** Optimal value of alpha = 0.0001

Ridge Model: Top 5 predictors with their coefficients:

Predictors	Coefficients
GrLivArea	0.3222
OverallQual	0.2054
OverallCond	0.1056
TotalBsmtSF	0.1056
Functional	0.0672

Lasso Model: Top 5 predictors with their coefficients:

Predictors	Coefficients
GrLivArea	0.3308
OverallQual	0.2108
OverallCond	0.1057
TotalBsmtSF	0.1049
Functional	0.0637

After doubling the value of alpha for both ridge and lasso i.e Ridge = 1 & Lasso = 0.0002

- Lasso regression: shrinks coefficients of predictors all the way to zero, hence removing them from the model.
- Ridge regression: shrinks coefficients toward zero, but they rarely reach zero. Coefficients of predictors decrease, and then their effect in the model decreases. And hence the flexibility of the model shall decrease.

Ridge Model: Top 5 predictors with their coefficients:

Predictors	Coefficients
GrLivArea	0.305171
OverallQual	0.204071
YearBuilt	0.137284
TotalBsmtSF	0.105974
OverallCond	0.100251

Lasso Model: Top 5 predictors with their coefficients:

Predictors	Coefficients
GrLivArea	0.318643
OverallQual	0.216155
YearBuilt	0.138691
TotalBsmtSF	0.104584
OverallCond	0.099586

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 Regularization Model Can be chosen depending on the below purpose:

- Lasso Regularization: If there are too many variables and reducing the complexity is the goal the Lasso model is the solution.
- Ridge Regularization: If the variables are important for business and needed in the model then use Ridge model.

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.919211	0.918959	0.918884
1	R2 Score (Test)	0.889109	0.889099	0.890041
2	RSS (Train)	1.316858	1.320959	1.322187
3	RSS (Test)	0.802669	0.802741	0.795921
4	MSE (Train)	0.038640	0.038700	0.038718
5	MSE (Test)	0.046081	0.046083	0.045887

It is observed that R2_scores are almost same for both Ridge and Lasso but Lasso Regularization will penalize more on the dataset and can also help in feature elimination hence Lasso Regularization will be chosen as final model. Also the given dataset is large, where we have around 130 features + (some of which are obsolete and needs to be removed)

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 excluding the five most important predictor variables. Five most important predictor variables are :

Predictors	Coefficients
GarageCars	0.259207
OpenPorchSF	0.149553
BsmtFinSF1	0.144440
YearRemodAdd	0.137517
LotArea	0.132951

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 :

- The model should be generalizable so that the test accuracy is not lesser than the training score. i.e the model should be accurate for datasets other than the ones which were used during training.
- The model needs to be made robust and generalizable so that it is not affected by outliers in the training data.
- The outliers which does not make sense to keep must be removed from the dataset. This would help increase the accuracy of the predictions made by the model which help standardize the predictions made by the model. If the model is not robust, it cannot be trusted for predictive analysis.

- To make a model robust, model must be built using Median or Median Absolute Deviation which are unaffected by outliers in the data. And will not be affected while predicting on unseen or altering dataset.

Hence, we should always select a model which is just complex enough to understand the variance in the data without much inaccuracy at the same time not too complex that it overfits.

This can be achieved using regularization. Regularization is the process of deliberately simplifying models to achieve the correct balance between keeping the model simple and yet not too Complex.