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:

<u>Optimal value of alpha for ridge and lasso regression is as below:-</u>
Ridge = 20
Lasso = 0.001

<u>Changes on the model if value of alpha is double for both ridge and lasso</u>

Metrics values when alpha is original (Ridge =20, Lasso = 0.001) :-

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.208343e-01	0.908259	0.920449
1	R2 Score (Test)	-2.658633e+22	0.816608	0.812931
2	RSS (Train)	8.082814e+01	93.667408	81.222024
3	RSS (Test)	1.201896e+25	82.906378	84.568870
4	MSE (Train)	2.813639e-01	0.302888	0.282049
5	MSE (Test)	1.654631e+11	0.434572	0.438908

Metrics value when alpha is doubled (Ridge = 40, Lasso = 0.002):-

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.208343e-01	0.897581	0.919282
1	R2 Score (Test)	-2.658633e+22	0.817581	0.814952
2	RSS (Train)	8.082814e+01	104.569500	82.413268
3	RSS (Test)	1.201896e+25	82.466816	83.655380
4	MSE (Train)	2.813639e-01	0.320029	0.284109
5	MSE (Test)	1.654631e+11	0.433418	0.436531

Observations:

- There has been very marginal change in the test scores of R2 and MSE for both Ridge and Lasso Regression, however in case of Train data set, R2 score has decreased for both model.
- Also MSE train score has marginally increased for both Ridge and Lasso. So this makes the original model based on original alpha to perform better.

Most important predictor variables after the change is implemented

Ridge Regression Model

Features	Coef
RoofMatl_CompShg	0.3799
GrLivArea	0.2360
RoofMatl_WdShngl	0.2277
RoofMatl_Tar&Grv	0.2046
OverallQual	0.1855
YearBuilt	0.1817
1stFlrSF	0.1811
KitchenQual_TA	-0.1723
RoofMatl_WdShake	0.1614
KitchenQual_Gd	-0.1515

Lasso Regression Model

Features	Coef
RoofMatl_CompShg	0.9393
RoofMatl_Tar&Grv	0.5353
RoofMatl_WdShngl	0.4915
GrLivArea	0.4560
RoofMatl_WdShake	0.4265
RoofMatl_Membran	0.2449
RoofMatl_Metal	0.2229
RoofMatl_Roll	0.2206
KitchenQual_TA	-0.2069
YearBuilt	0.1941

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 alpha for both ridge and Lasso is 20 and 0.001 respectively.

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.208343e-01	0.908259	0.920449
1	R2 Score (Test)	-2.658633e+22	0.816608	0.812931
2	RSS (Train)	8.082814e+01	93.667408	81.222024
3	RSS (Test)	1.201896e+25	82.906378	84.568870
4	MSE (Train)	2.813639e-01	0.302888	0.282049
5	MSE (Test)	1.654631e+11	0.434572	0.438908

- R2 test score for Ridge Regression is marginally better than Lasso regression, however R2 train score for Ridge has decreased as compared to Lasso. So this makes Ridge model better model to perform on unseen data.
- MSE test score for Ridge regression is slightly less than Lasso regression, so this makes Ridge perform better on unseen data.

So , Ridge model looks better based on the metrics to perform as compared to Lasso regression.

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:

Top 5 Features before removing variables are :-

Features
RoofMatl_CompShg
RoofMatl_Tar&Grv
RoofMatl_WdShngl
RoofMatl_WdShake
GrLivArea

Top 5 features after removing the important top 5 features are :-

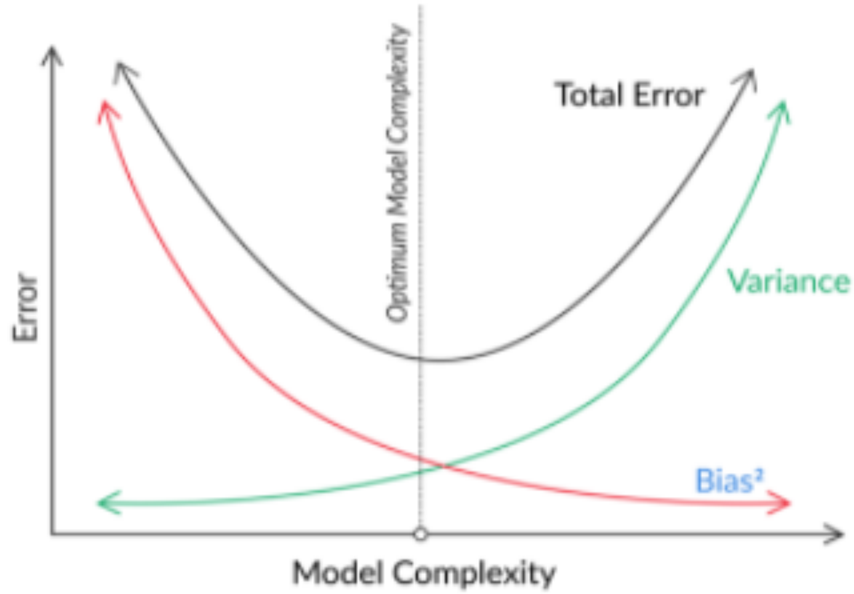
Features
2ndFlrSF
1stFlrSF
OverallQual
BsmtQual_Gd
BsmtQual_TA

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:

- A robust model performs consistently well on new data.
- Regularization helps to control the trade-off between model complexity and bias, which is important for model robustness.
- To make a model more robust and generalizable, we need to balance between keeping the model simple and not making it too naive.
- There is a bias-variance trade-off where a complex model is sensitive to changes in training data, while a simpler model is less sensitive.
- Bias measures how accurate the model is likely to be on test data, while variance measures the degree of changes in the model itself with respect to changes in the training data.
- Maintaining a balance between bias and variance helps to minimize the total error and maintain model accuracy.



What we need is lowest total error, i.e., low bias and low variance, such that the model identifies all the patterns that it should and is also able to perform well with unseen data.