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 value for alpha for Ridge is **4** and Lasso is **0.0001**.
- Doubling the value of alpha does not lead to a significant change in the r2_score but small change in the value of the features coefficients.
- Below are the top 10 important predictor variables after the change:

Ridge:

	index	Ridge
0	GrLivArea	0.261517
1	OverallCond	0.195735
2	1stFlrSF	0.194324
3	TotalBsmtSF	0.161718
4	BsmtFinSF1	0.157150
5	OverallQual_EX	0.140760
6	LogLotArea	0.139931
7	Age	-0.128913
8	OverallQual_PO	-0.128421
9	MSZoning_FV	0.124611

Lasso:

	index	Lasso
0	GrLivArea	0.516172
1	MSZoning_FV	0.348760
2	OverallCond	0.289277
3	Age	-0.275433
4	TotalBsmtSF	0.272770
5	MSZoning_RH	0.269544
6	MSZoning_RL	0.267364
7	OverallQual_PO	-0.254681
8	MSZoning_RM	0.224446
9	LogLotArea	0.219857

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:

Lasso is preferred over Ridge as it helped to reduce the values of multiple coefficients to nearly zero thus helping to build a simple model. Below is the result for all the 3 models.

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.878990	0.933688	0.939439
1	R2 Score (Test)	0.877056	0.906006	0.913980
2	RSS (Train)	17.269558	9.463527	8.642711
3	RSS (Test)	8.007925	6.122285	5.602873
4	MSE (Train)	0.132009	0.097721	0.093387
5	MSE (Test)	0.137267	0.120022	0.114818

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 important 5 features obtained from Lasso model with alpha = 0.0001 are:

- GrLivArea
- MSZoning_FV
- OverallCond
- TotalBsmtSF
- MSZoning_RH

If we drop the above features and re-train the model, then we got below top 5 features:

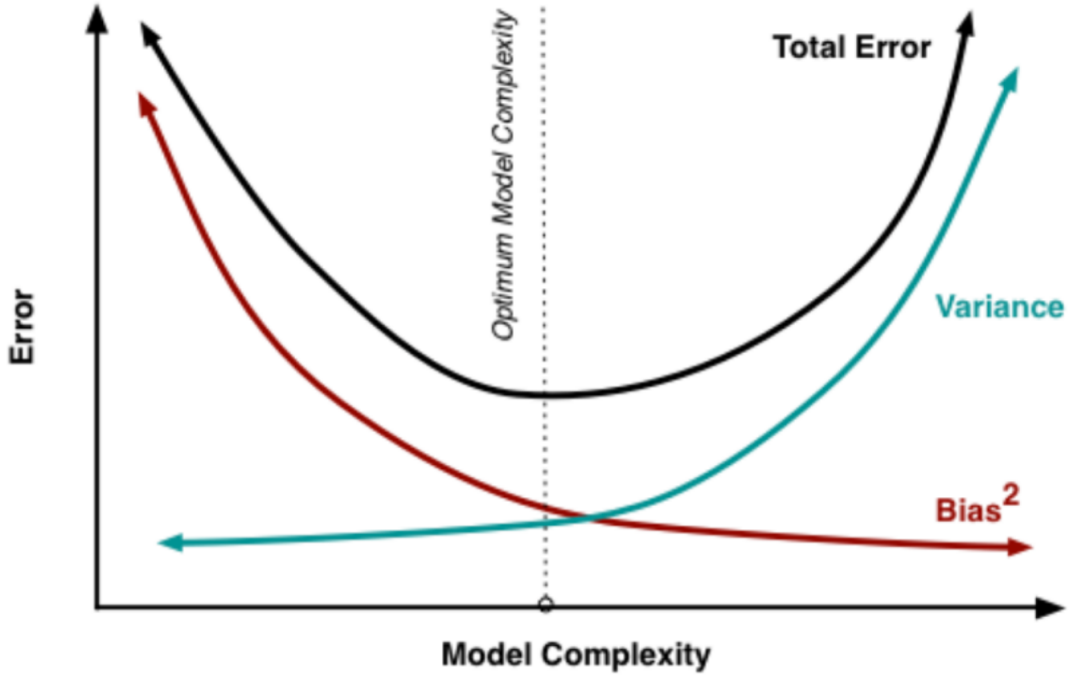
- 1stFlrSF
- OverallQual_PO
- 2ndFlrSF
- BsmtFinSF1
- LogLotArea

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 Model will be robust and generalisable when the performance is not degraded on any variation in data and performs well new/unseen data.
- In general, model should not be too complex, If complexity increases there is a fair chance of overfitting.
- Complex model will have high accuracy on train data but not on test data, due to high variance.
- We have to reduce the variance which can introduce bias.



- We have to find the balance between variance, bias and complexity to make our model robust and generalisable. Which can be achieved by Regularization techniques like Ridge Regression and Lasso.