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:
In the case of ridge regression:- When we plot the curve between negative mean absolute error and alpha we see that as the value of alpha increase from 0 the error term decrease and the train error is showing increasing trend when value of alpha increases .when the value of alpha is 2 the test error is minimum so we decided to go with value of alpha equal to 2 for our ridge regression.

For lasso regression I have decided to keep very small value that is 0.01, when we increase the value of alpha the model try to penalize more and try to make most of the coefficient value zero. Initially it came as 0.4 in negative mean absolute error and alpha. When we double the value of alpha for our ridge regression no we will take the value of alpha equal to 10 the model will apply more penalty on the curve and try to make the model more generalized that is making model more simpler and no thinking to fit every data of the data set .from the graph we can see that when alpha is 10 we get more error for both test and train. Similarly when we increase the value of alpha for lasso we try to penalize more our model and more coefficient of the variable will reduced to zero, when we increase the value of our r2 square also decreases.

The most important variable after the changes has been implemented for ridge regression are as follows:-
1. MSZoning_FV
2. MSZoning_RL
3. Neighborhood_Crawfor
4. MSZoning_RH
5. MSZoning_RM
6. SaleCondition_Partial
7. Neighborhood_StoneBr
8. GrLivArea
9. SaleCondition_Normal
10. Exterior1st_BrkFace

The most important variable after the changes has been implemented for lasso regression are as follows:-
1. GrLivArea
2. OverallQual3. OverallCond
4. TotalBsmtSF
5. BsmtFinSF1
6. GarageArea

7. Fireplaces
8. LotArea
9. LotArea
10. LotFrontage

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:

Ridge regression is generally preferred when:

1. There are many predictors with small to moderate effect sizes.
2. The goal is to reduce the impact of multicollinearity among predictors.
3. Predictors are believed to contribute somewhat equally to the outcome.

Lasso regression is preferred when:

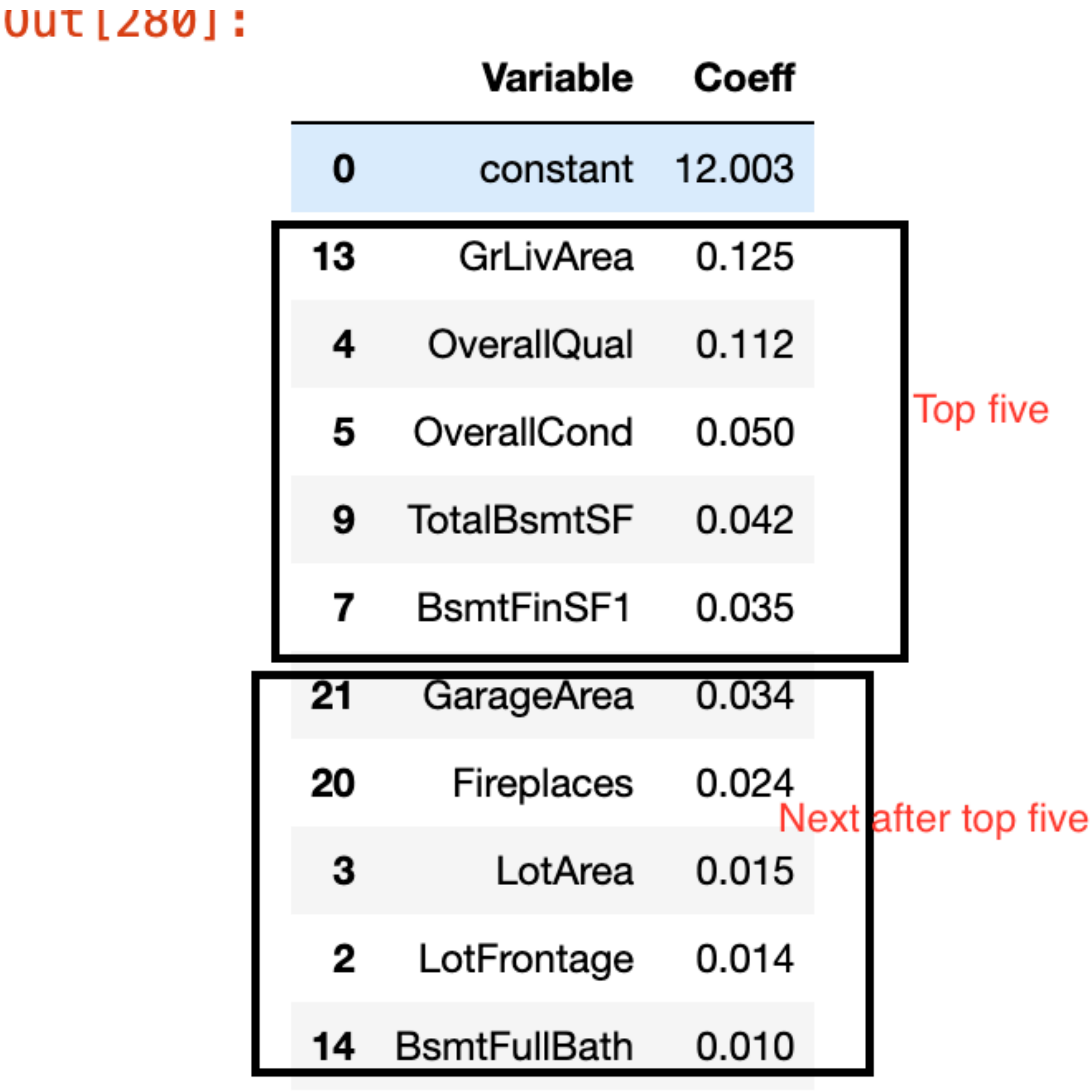
1. There are many predictors, but only a subset of them are expected to have a significant effect on the outcome (sparse model).
2. Feature selection is desired, as lasso tends to shrink some coefficients to exactly zero, effectively removing them from the model.
3. There is a need for a more interpretable model with a smaller number of predictors.

Ultimately, the choice between ridge and lasso regression depends on the balance between bias and variance in the model, the interpretability of the results, and the specific goals of the analysis. If the goal is to select a more parsimonious model with feature selection, lasso regression might be preferred. If multicollinearity is a concern and the goal is to retain all predictors but reduce their impact, ridge regression might be more appropriate.

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, the next five most important predictor variables would be those with the largest non-zero coefficients in the updated lasso model. These variables would have the strongest influence on the outcome after accounting for the exclusion of the original five variables.



Out[280]:

	Variable	Coeff
0	constant	12.003
13	GrLivArea	0.125
4	OverallQual	0.112
5	OverallCond	0.050
9	TotalBsmtSF	0.042
7	BsmtFinSF1	0.035
21	GarageArea	0.034
20	Fireplaces	0.024
3	LotArea	0.015
2	LotFrontage	0.014
14	BsmtFullBath	0.010

Top five

Next after top five

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:
The model should be as simple as possible, though its accuracy will decrease but it will be more robust and generalisable. It can be also understood using the Bias-Variance trade-off. The simpler the model the more the bias but less variance and more generalizable. Its implication in terms of accuracy is that a robust and

generalisable model will perform equally well on both training and test data i.e. the accuracy does not change much for training and test data.

Bias: Bias is error in model, when the model is weak to learn from the data. High bias means model is unable to learn details in the data. Model performs poor on training and testing data.

Variance: Variance is error in model, when model tries to over learn from the data. High variance means model performs exceptionally well on training data as it has very well trained on this of data but performs very poor on testing data as it was unseen data for the model.

It is important to have balance in Bias and Variance to avoid overfitting and under-fitting of data.