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 to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer:

Optimal value of alpha for ridge and lasso regression in my model comes out to be 500.

There are some changes in the model when we double the value of alpha for ridge and lasso regression models . The value of coefficients and the order of importance of variables changes after alpha is doubled .

In case of Lasso , the number of variables with zero coefficient increases after we double the value of alpha.

The 5 most predictor variables

Ridge model , alpha = 500 **Variable Name: Coefficient Value**	**Ridge model , alpha = 1000** **Variable Name: Coefficient Value alpha= 1000**
OverallQual: 8750.53976330277	OverallQual: 7273.680253670489
GrLivArea: 8299.910026024645	GrLivArea: 6874.079319201152
Neighborhood_NoRidge: 6122.602041956645	Neighborhood_NoRidge: 5183.671232762423
Condition2_PosN: -5921.51585197518	1stFlrSF: 4983.640317134498
Neighborhood_NridgHt: 5597.383661925816	Neighborhood_NridgHt: 4679.52583742666

Lasso model , alpha = 500 **Variable Name: Coefficient Value**	**Lasso model , alpha = 1000** **Variable Name: Coefficient Value**
RoofMatl_CompShg: 31006.8414673564	GrLivArea: 27978.80457117819
GrLivArea: 30986.94383238589	OverallQual: 16292.857329010067
RoofMatl_WdShngl: 19406.088440452804	Condition2_PosN: -10548.557766438924
RoofMatl_Tar&Grv: 18926.234034870933	BsmtQual_Gd: -8362.74731894178
OverallQual: 12997.649672874928	Neighborhood_NridgHt: 7377.990176331924

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:

Metric	Ridge Regression	Lasso Regression
R2 Score (Train)	8.835432e-01	9.231018e-01
R2 Score (Test)	8.566113e-01	8.454503e-01
RSS (Train)	7.430758e+11	4.906644e+11
RSS (Test)	4.041714e+11	4.356310e+11
MSE (Train)	2.697762e+04	2.192196e+04
MSE (Test)	3.037706e+04	3.153714e+04

We have done the above metric comparison in Jupyter Notebook, and we can observe that Ridge regression performs slightly better than lasso regression on the test dataset.

Despite that, I will prefer to apply lasso as it keeps the model simpler and performance difference is very small.

Lasso reduces the no. of significant predictor variables to almost half of ridge 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:

The next 5 important predictor variables are :
RoofMatl_WdShake, Condition2_PosN, BsmtQual_Gd, KitchenQual_Gd, KitchenQual_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:

We need to balance the bias and variance in the model . On one hand , model should not overfit and on the other hand , the model should not underfit . This can be done by regularization . When we apply regularization such as Lasso and Ridge Regression on a dataset to build a model , we perform the following steps in sequence to create a model which is more robust and generalizable .

1) EDA which includes steps such as Data Cleaning, removing outliers , imputation
2) Convert categorical variables to dummy variables.
3) Scaling
4) Split Dataset into Test and Train

5) Applying regularization techniques such as Lasso and Ridge
6) Cross validation.
7) Hyperparameter tuning.

When we try to make a model more robust and generalizable , the accuracy of the model may slightly go down on the training dataset but the performance of such a model will be better on new data as compared to a overfitting model.