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:

Optimum Value for alpha - Ridge Regression: 3.0

Optimum Value for alpha - Lasso Regression: 0.0008

After doubling the optimal value for alpha, we see that the coefficients for 9 features are reduced to 0 causing the model to be simpler.

For Ridge regression we saw that the coefficient values for a few features have reduced a little.

The change in alpha does not seems to have significantly impacted the accuracy between the models.

Below are the 10 most important predictor variables after the change.

Ridge Regression: -

	feature_name	coeff	abs_coeff
7	MSZoning_Residentail	0.492	0.492
5	GrLivArea	0.357	0.357
22	Neighborhood_MeadowV	-0.356	0.356
21	Neighborhood_IDOTRR	-0.328	0.328
28	Neighborhood_NridgHt	0.326	0.326
56	SaleType_New	0.319	0.319
37	Exterior1st_BrkFace	0.285	0.285
35	BldgType_Twnhs	-0.279	0.279
1	OverallQual	0.266	0.266
29	Neighborhood_OldTown	-0.263	0.263

Lasso Regression: -

	feature_name	coeff	abs_coeff
7	GarageCars	0.721	0.721
5	CentralAir	0.356	0.356
22	Neighborhood_IDOTRR	-0.350	0.350
28	Neighborhood_NoRidge	0.331	0.331
35	Condition1_RRAn	-0.313	0.313
56	GarageType_Detchd	0.297	0.297
21	Neighborhood_Edwards	-0.282	0.282
37	HouseStyle_Unf	0.269	0.269
29	Neighborhood_NridgHt	-0.249	0.249
4	BsmtFinType1	0.225	0.225

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:

After finalizing the model and the R2 scores on both train and test data for both the models do not vary much, but as lasso will penalize more on the dataset and has eliminated 4 more features will be selecting and applying Lasso regression model.

Question 3

After building the model, you realized 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 creating the new model below are the new top five important predictor variables based on the absolute value of their coefficients.

	feature_name	coeff	abs_coeff
1	OverallQual	0.510	0.510
23	Neighborhood_NoRidge	0.400	0.400
16	Neighborhood_ClearCr	0.386	0.386
51	SaleType_New	0.381	0.381
32	Exterior1st_BrkFace	0.369	0.369

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:

A model can be considered generalizable when it doesn't overfits the training data and performs equally well on the test data set as well.

A model can be considered robust if it works for broad range of input data set i.e. is does not drastically change its behavior on changing of input data. Ideally speaking accuracy should not vary much for training and test datasets.