Question 1

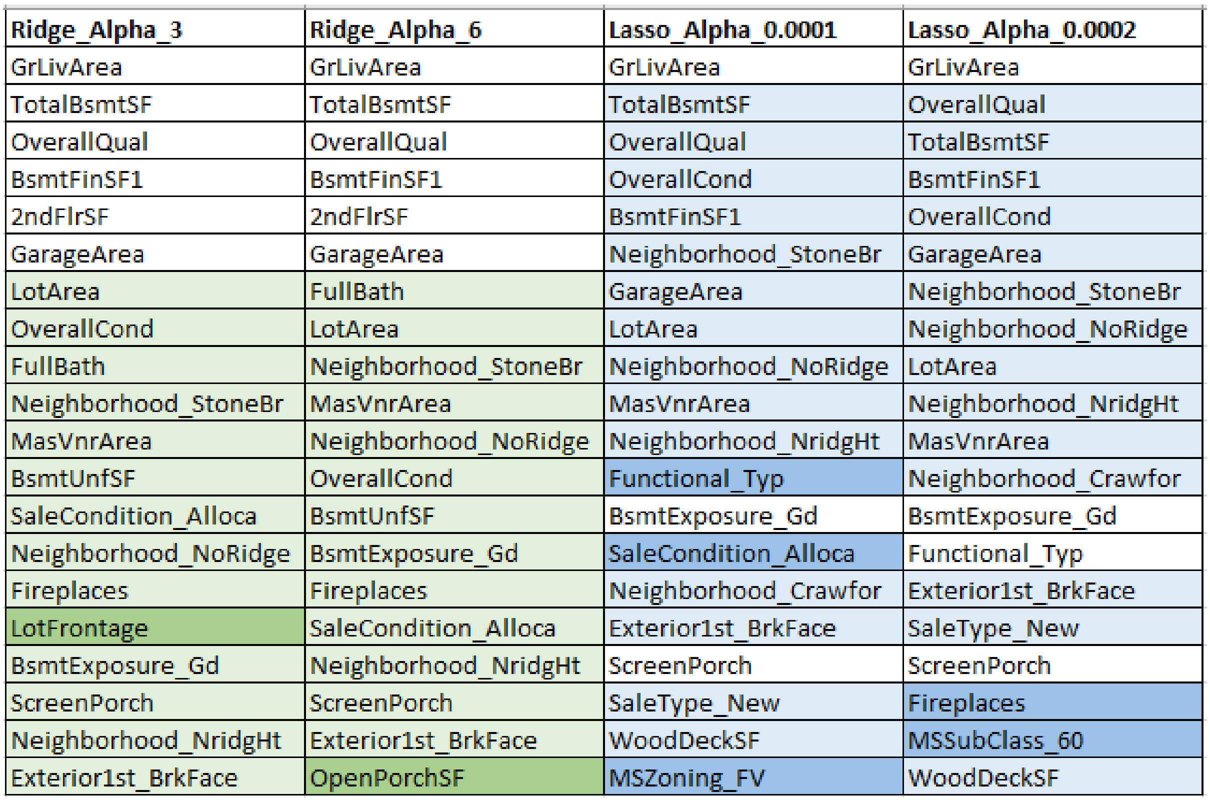
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer:

The optimal value of alpha obtained for the Ridge regression model is 3, and the optimal value of alpha obtained for the Lasso regression model is 0.0001.

When the alpha for the Ridge and the Lasso changed to double the optimal values, the coefficients of both Ridge and Lasso trend towards zero. In addition, for Lasso, with an optimal value of 0.0001, there were 60 features with nonzero coefficients. After double the coefficient value the number of nonzero coefficients reduced to 41.

Listed below are the most significant 20 predictors with optimal value for Ridge and Lasso as well as Ridge and Lasso with double the optimal value.



Ridge_Alpha_3	**Ridge_Alpha_6**	**Lasso_Alpha_0.0001**	**Lasso_Alpha_0.0002**
GrLivArea	GrLivArea	GrLivArea	GrLivArea
TotalBsmtSF	TotalBsmtSF	TotalBsmtSF	OverallQual
OverallQual	OverallQual	OverallQual	TotalBsmtSF
BsmtFinSF1	BsmtFinSF1	OverallCond	BsmtFinSF1
2ndFlrSF	2ndFlrSF	BsmtFinSF1	OverallCond
GarageArea	GarageArea	Neighborhood_StoneBr	GarageArea
LotArea	FullBath	GarageArea	Neighborhood_StoneBr
OverallCond	LotArea	LotArea	Neighborhood_NoRidge
FullBath	Neighborhood_StoneBr	Neighborhood_NoRidge	LotArea
Neighborhood_StoneBr	MasVnrArea	MasVnrArea	Neighborhood_NridgHt
MasVnrArea	Neighborhood_NoRidge	Neighborhood_NridgHt	MasVnrArea
BsmtUnfSF	OverallCond	Functional_Typ	Neighborhood_Crawfor
SaleCondition_Alloca	BsmtUnfSF	BsmtExposure_Gd	BsmtExposure_Gd
Neighborhood_NoRidge	BsmtExposure_Gd	SaleCondition_Alloca	Functional_Typ
Fireplaces	Fireplaces	Neighborhood_Crawfor	Exterior1st_BrkFace
LotFrontage	SaleCondition_Alloca	Exterior1st_BrkFace	SaleType_New
BsmtExposure_Gd	Neighborhood_NridgHt	ScreenPorch	ScreenPorch
ScreenPorch	ScreenPorch	SaleType_New	Fireplaces
Neighborhood_NridgHt	Exterior1st_BrkFace	WoodDeckSF	MSSubClass_60
Exterior1st_BrkFace	OpenPorchSF	MSZoning_FV	WoodDeckSF

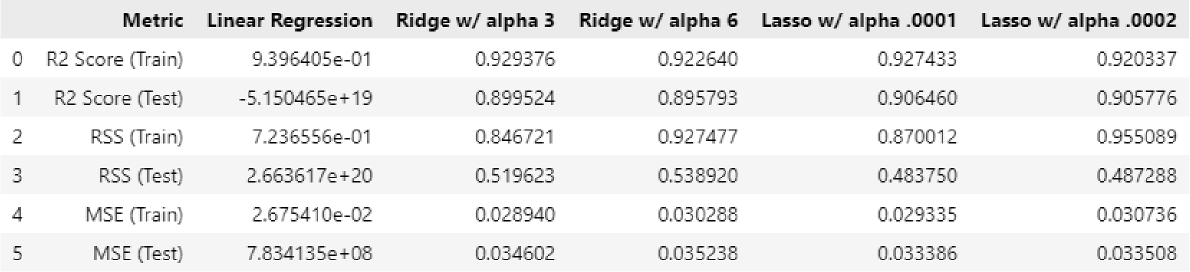
There are changes in the order of significance and some variables did not make it in the first 20 after applying double the value of alpha.

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:

Given below are the R2, RSS, and MSE scores for the models built for the regression.



	Metric	Linear Regression	Ridge w/ alpha 3	Ridge w/ alpha 6	Lasso w/ alpha .0001	Lasso w/ alpha .0002
0	R2 Score (Train)	9.396405e-01	0.929376	0.922640	0.927433	0.920337
1	R2 Score (Test)	-5.150465e+19	0.899524	0.895793	0.906460	0.905776
2	RSS (Train)	7.236556e-01	0.846721	0.927477	0.870012	0.955089
3	RSS (Test)	2.663617e+20	0.519623	0.538920	0.483750	0.487288
4	MSE (Train)	2.675410e-02	0.028940	0.030288	0.029335	0.030736
5	MSE (Test)	7.834135e+08	0.034602	0.035238	0.033386	0.033508

Out of all models, Lasso gives better performance on test data, and test and train differences are minimal.

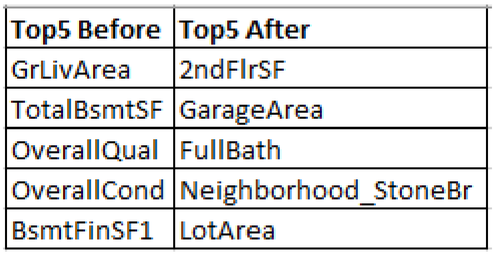
With the optimal Lasso alpha value of 0.0001, the model has 60 variables, and with double the optimal alpha value of 0.0002, the model can be explained with 41 variables. The difference in test and train R2 score and MSE values is very negligible. Hence, my preference is the model with Lasso with an alpha value 0.0002.

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 top five feature with the original, and the top five feature after dropping the top five from the original model is given below.



Top5 Before	Top5 After
GrLivArea	2ndFlrSF
TotalBsmtSF	GarageArea
OverallQual	FullBath
OverallCond	Neighborhood_StoneBr
BsmtFinSF1	LotArea

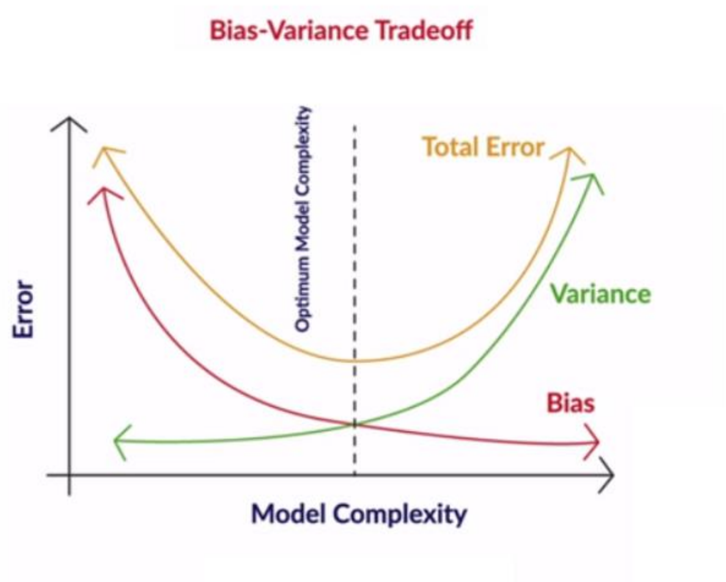
Question 4

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 can be generalizable by significantly reducing the variance in the model with a small reduction in bias. This can be done by increasing the alpha value in Ridge or Lasso.

By generalizing the model the accuracy of prediction will be reduced. Also, the model complexity will be reduced. For example, in the model assignment, when generalized the Lasso model by changing the alpha value from 0.0001 to 0.0002, the reduction in R2 score is 0.0071 and MSE increased by 0.0014 on the train set. With the change in alpha, the number of variables in the model reduced from 60 to 41 and the variables are trending towards zero as compared to the model with an alpha of 0.0001.



The model complexity will be much reduced if we generalize the model much further. The implication with this is model will not predict well with both training data as well as test data since the prediction error is much higher.