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:

a. The Optimal value of alpha for ridge is 2 and for lasso it is 0.0009
b. After doubling the alpha values in the Ridge and Lasso, the prediction accuracy remains around 0.82 but there is a small change in the co-efficient values.
c. The new model is created and demonstrated in the Jupiter notebook. Below are the changes in the co-efficients.

<u>Ridge regression:</u>

	Ridge Co-Efficient
Total_sqr_footage	1.775099
GarageArea	1.022466
TotRmsAbvGrd	0.658284
LotArea	0.519446
OverallCond	0.460275
LotFrontage	0.385151
CentralAir_Y	0.340345
Total_porch_sf	0.322288
Alley_Pave	0.283545
Neighborhood_StoneBr	0.260305
MSSubClass_70	0.228895
HouseStyle_2.5Unf	0.220205
OpenPorchSF	0.208467
SaleType_Con	0.201785
Neighborhood_Veenker	0.201116
PavedDrive_P	0.198609
RoofMatl_WdShngl	0.186879
SaleType_Oth	0.178074
ExterCond_Ex	0.177688
Condition1_PosN	0.177110

	Ridge Doubled Alpha Co-Efficient
Total_sqr_footage	1.559966
GarageArea	0.927756
TotRmsAbvGrd	0.674250
LotArea	0.443779
OverallCond	0.419170
Total_porch_sf	0.344978
CentralAir_Y	0.336165
LotFrontage	0.322887
Alley_Pave	0.252418
Neighborhood_StoneBr	0.238455
MSSubClass_70	0.220241
OpenPorchSF	0.208703
BsmtQual_Ex	0.194574
HouseStyle_2.5Unf	0.187861
Neighborhood_Veenker	0.182717
PavedDrive_P	0.175108
MasVnrType_Stone	0.168838
KitchenQual_Ex	0.166826
Condition1_Norm	0.157282
PavedDrive_Y	0.157277

Lasso Regression:

	Lasso Co-Efficient
Total_sqr_footage	2.123419
GarageArea	1.119248
TotRmsAbvGrd	0.593445
LotArea	0.556863
OverallCond	0.460292
CentralAir_Y	0.349808
Total_porch_sf	0.292163
LotFrontage	0.273539
Alley_Pave	0.256738
Neighborhood_StoneBr	0.205235
MSSubClass_70	0.195613
OpenPorchSF	0.185743
Condition1_Norm	0.157461
BsmtQual_Ex	0.156083
HouseStyle_2.5Unf	0.152288
LandContour_HLS	0.149418
KitchenQual_Ex	0.140137
PavedDrive_P	0.130056
LotConfig_CulDSac	0.129984
Condition1_PosN	0.129700

	Lasso Doubled Alpha Co-Efficient
Total_sqr_footage	2.144308
GarageArea	1.067935
TotRmsAbvGrd	0.629239
OverallCond	0.418216
LotArea	0.390744
CentralAir_Y	0.342906
Total_porch_sf	0.304615
Alley_Pave	0.213485
BsmtQual_Ex	0.170352
MSSubClass_70	0.153329
OpenPorchSF	0.147276
Neighborhood_StoneBr	0.143520
KitchenQual_Ex	0.134287
LandContour_HLS	0.130258
Condition1_Norm	0.129846
BsmtCond_TA	0.121997
LotConfig_CulDSac	0.119421
MasVnrType_Stone	0.117413
SaleCondition_Partial	0.116304
PavedDrive_Y	0.077874

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:

1. The optimum lambda value in case of Ridge and Lasso is as follows:-
   - Ridge – 2
   - Lasso – 0.0009
2. The Mean Squared Error in case of Ridge and Lasso are:
   - Ridge - 0.1857491033294977
   - Lasso - 0.18787007533042074
3. The Mean Squared Error of both the models are almost same.
4. Since Lasso helps in feature reduction (as the coefficient value of some of the features become zero), Lasso has a better edge over Ridge and should be used as the final model.

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 five most important predictor variables in the current lasso model is:-

1. Total_sqr_footage
2. GarageArea
3. TotRmsAbvGrd
4. OverallCond
5. LotArea

We build a Lasso model in the Jupiter notebook after removing these attributes from the dataset.

The R2 of the new model without the top 5 predictors drops to 0.73

The Mean Squared Error increases to 0.2917218039744371

The new Top 5 predictors are:-

	Lasso Co-Efficient
LotFrontage	1.842349
HouseStyle_2.5Fin	1.164699
HouseStyle_2.5Unf	0.796963
Total_porch_sf	0.769369
Condition2_PosA	0.654268

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:

As Per, Occam's Razor— given two models that show similar 'performance' in the finite training or test data, we should pick the one that makes fewer on the test data due to following reasons:-

- Simpler models are usually more 'generic' and are more widely applicable
- Simpler models require fewer training samples for effective training than the more complex ones and hence are easier to train.
- Simpler models are more robust.
    - Complex models tend to change wildly with changes in the training data set

- Simple models have low variance, high bias and complex models have low bias, high variance
- Simpler models make more errors in the training set. Complex models lead to overfitting —they work very well for the training samples, fail miserably when applied to other test samples

Therefore, to make the model more robust and generalizable, make the model simple but not simpler which will not be of any use.

Regularization can be used to make the model simpler. Regularization helps to strike the delicate balance between keeping the model simple and not making it too naive to be of any use. For regression, regularization involves adding a regularization term to the cost that adds up the absolute values or the squares of the parameters of the model.

Also, Making a model simple leads to Bias-Variance Trade-off:

- A complex model will need to change for every little change in the dataset and hence is very unstable and extremely sensitive to any changes in the training data.
- A simpler model that abstracts out some pattern followed by the data points given is unlikely to change wildly even if more points are added or removed.

Bias quantifies how accurate is the model likely to be on test data. A complex model can do an accurate job prediction provided there is enough training data. Models that are too naïve, for e.g., one that gives same answer to all test inputs and makes no discrimination whatsoever has a very large bias as its expected error across all test inputs are very high.

Variance refers to the degree of changes in the model itself with respect to changes in the training data. Thus accuracy of the model can be maintained by keeping the balance between Bias and Variance as it minimizes the total error as shown in the below graph: