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 for Ridge and Lasso regression are **2** & **0.001** respectively. For these values the R2 of the model was **0.83** approx.

When we doubled the alpha value in Ridge & Lasso, the prediction accuracy remained around **0.82** but there was a small change in the coefficient values.

The new model was demonstrated in Jupyter notebook. The changes in the coefficient are listed below:

Ridge Regression Model

	Ridge Co-Efficient
Total_sqr_footage	0.169122
GarageArea	0.101585
TotRmsAbvGrd	0.067348
OverallCond	0.047652
LotArea	0.043941
CentralAir_Y	0.032034
LotFrontage	0.031772
Total_porch_sf	0.031639
Neighborhood_StoneBr	0.029093
Alley_Pave	0.024270
OpenPorchSF	0.023148
MSSubClass_70	0.022995
RoofMatl_WdShngl	0.022586
Neighborhood_Veenker	0.022410
SaleType_Con	0.022293
HouseStyle_2.5Unf	0.021873
PavedDrive_P	0.020160
KitchenQual_Ex	0.019378
LandContour_HLS	0.018595
SaleType_Oth	0.018123

	Ridge Doubled Alpha Co-Efficient
Total_sqr_footage	0.149028
GarageArea	0.091803
TotRmsAbvGrd	0.068283
OverallCond	0.043303
LotArea	0.038824
Total_porch_sf	0.033870
CentralAir_Y	0.031832
LotFrontage	0.027526
Neighborhood_StoneBr	0.026581
OpenPorchSF	0.022713
MSSubClass_70	0.022189
Alley_Pave	0.021672
Neighborhood_Veenker	0.020098
BsmtQual_Ex	0.019949
KitchenQual_Ex	0.019787
HouseStyle_2.5Unf	0.018952
MasVnrType_Stone	0.018388
PavedDrive_P	0.017973
RoofMatl_WdShngl	0.017856
PavedDrive_Y	0.016840

Lasso Regression Model

	Lasso Co-Efficient
Total_sqr_footage	0.202244
GarageArea	0.110863
TotRmsAbvGrd	0.063161
OverallCond	0.046686
LotArea	0.044597
CentralAir_Y	0.033294
Total_porch_sf	0.028923
Neighborhood_StoneBr	0.023370
Alley_Pave	0.020848
OpenPorchSF	0.020776
MSSubClass_70	0.018898
LandContour_HLS	0.017279
KitchenQual_Ex	0.016795
BsmtQual_Ex	0.016710
Condition1_Norm	0.015551
Neighborhood_Veenker	0.014707
MasVnrType_Stone	0.014389
PavedDrive_P	0.013578
LotFrontage	0.013377
PavedDrive_Y	0.012363

	Lasso Doubled Alpha Co-Efficient
Total_sqr_footage	0.204642
GarageArea	0.103822
TotRmsAbvGrd	0.064902
OverallCond	0.042168
CentralAir_Y	0.033113
Total_porch_sf	0.030659
LotArea	0.025909
BsmtQual_Ex	0.018128
Neighborhood_StoneBr	0.017152
Alley_Pave	0.016628
OpenPorchSF	0.016490
KitchenQual_Ex	0.016359
LandContour_HLS	0.014793
MSSubClass_70	0.014495
MasVnrType_Stone	0.013292
Condition1_Norm	0.012674
BsmtCond_TA	0.011677
SaleCondition_Partial	0.011236
LotConfig_CulDSac	0.008776
PavedDrive_Y	0.008685

Since the alpha values are small so we do not observe a huge change in the model after doubling the 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:

The optimum value of lambda for Ridge regression is **2** & for Lasso regression is **0.0001**.

The mean squared error for Ridge regression is **0.00183960907** & for Lasso regression is **0.00186341526** approx. So, the mean squared error for Ridge & Lasso are almost same. Hence, to decide the final model we will check the feature reduction feature & we know that Lasso helps in reducing the features as we can see the coefficient values of some of the features became 0 . Hence, we will choose Lasso 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 Lasso Model are as follows:

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

After removing above predictor variables , we again build a Lasso model. Now, the R2 of new model dropped to 0.73 & the mean squared error increases to 0.0028575670 approx. For more information, please refer the Jupyter notebook.

The top 5 predictors in new model are as follows:

	Lasso Co-Efficient
LotFrontage	0.146535
Total_porch_sf	0.072445
HouseStyle_2.5Unf	0.062900
HouseStyle_2.5Fin	0.050487
Neighborhood_Veenker	0.042532

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 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 more 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 widely with changes in the training dataset.
- Simpler models make more errors in the training set where as complex models leads to overfitting as they work very well on training samples & fail when apply to other test sample.
- Simple models have low variance, high bias whereas complex models have low bias & high variance.

Hence, to make a model robust & generalisable make it simple but not simpler which is of no use.

Regularization helps in making a model simple. It maintains a balance between a model to be simpler & not too naïve to be of no 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 lead to Bias-Variance Trade-off:

- A complex model will need to change for every little change in the dataset and hence is very unstable & extremely sensitive to 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 i.e., which gives same answer to all test inputs and makes no discrimination whatsoever has a very large bias as it's 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 a balance between Bias & Variance as it minimizes the total error as shown in the graph below.