Assignment Part II

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: From our analysis done in Part I, we see the optimal value of alpha is 2 (for Ridge regression) and alpha is 0.0001 (for Lasso regression)
For Ridge regression, we will double value of alpha which becomes 4:
R2 Score of model on test dataset for doubled value of alpha = 0.7933990805211957
MSE of model on test dataset for doubled value of alpha = 0.002211210591542772

20 important features with alpha=2

	Ridge Co-Efficient
TotalSqrFoot	0.161724
GarageArea	0.116254
TotalNoOfBathrooms	0.066629
OverallCond	0.065650
LotArea	0.039030
TotRmsAbvGrd	0.032450
TotalPorchSf	0.031344
Neighborhood_Veenker	0.030141
Neighborhood_StoneBr	0.029971
MSSubClass_70	0.026807
LotFrontage	0.026113
BsmtQual_Ex	0.024106
OpenPorchSF	0.022561
MasVnrType_Stone	0.019371
KitchenQual_Ex	0.017596
HouseStyle_2.5Unf	0.014273
Neighborhood_Blmngtn	0.012583
ExterQual_Ex	0.012200
LotConfig_CulDSac	0.011707
RoofStyle_Mansard	0.010072

20 important features with alpha=4

	Ridge Model Doubled Alpha Co-Efficient
TotalSqrFoot	0.142311
GarageArea	0.104930
TotalNoOfBathrooms	0.062920
OverallCond	0.061542
TotRmsAbvGrd	0.039411
LotArea	0.035790
TotalPorchSf	0.032456
Neighborhood_StoneBr	0.027246
BsmtQual_Ex	0.025614
LotFrontage	0.025117
Neighborhood_Veenker	0.025107
MSSubClass_70	0.024882
OpenPorchSF	0.022264
MasVnrType_Stone	0.020136
KitchenQual_Ex	0.018274
ExterQual_Ex	0.012982
HouseStyle_2.5Unf	0.012251
LotConfig_CulDSac	0.012005
Neighborhood_Blmngtn	0.009457
RoofStyle_Mansard	0.009052

The above table shows the comparison of ridge coefficients for alpha value 2 & 4 respectively.

20 Important features for alpha=0.0001

	Lasso Co-Efficient
TotalSqrFoot	0.196702
GarageArea	0.128735
OverallCond	0.066294
TotalNoOfBathrooms	0.060259
LotArea	0.035915
TotalPorchSf	0.026834
Neighborhood_StoneBr	0.025067
TotRmsAbvGrd	0.024709
BsmtQual_Ex	0.023347
OpenPorchSF	0.020177
MSSubClass_70	0.019504
Neighborhood_Veenker	0.018665
MasVnrType_Stone	0.016484
KitchenQual_Ex	0.015966
ExterQual_Ex	0.009875
LotConfig_CulDSac	0.008327
Neighborhood_Blmngtn	0.007932
RoofStyle_Hip	0.006812
LotShape_IR2	0.005944
MasVnrType_BrkFace	0.005312

20 Important features for alpha=0.0002

	Lasso Model Doubled Alpha Co-Efficient
TotalSqrFoot	0.199039
GarageArea	0.123918
OverallCond	0.062796
TotalNoOfBathrooms	0.052147
TotRmsAbvGrd	0.028907
TotalPorchSf	0.027062
BsmtQual_Ex	0.024043
LotArea	0.022112
Neighborhood_StoneBr	0.018196
KitchenQual_Ex	0.015432
OpenPorchSF	0.015411
MasVnrType_Stone	0.015012
MSSubClass_70	0.013777
ExterQual_Ex	0.008771
LotConfig_CulDSac	0.007993
RoofStyle_Hip	0.007343
MasVnrType_BrkFace	0.005158
LotShape_IR2	0.002888
Neighborhood_Veenker	0.001508
GarageType_BuiltIn	0.000414

The above table shows the comparison of Lasso coefficients for alpha value 0.0001 & 0.0002 respectively.

Below is our conclusion based on above table values:
1) As the alpha value was small, doubling this value did not have much impact in both the models.
2) Both R2 and MSE remained almost same even after doubling the alpha value.
3) Also, the most important predictor variables remained same.
4) Ridge coefficients for alpha value 2 are slightly higher than the value for alpha value 4.
5) Lasso coefficients for alpha=0.0001 we see a slight difference than the value for alpha=0.0002.
6) However, for both Ridge and Lasso we see the feature “Total rooms above ground” has taken the 5th rank by replacing the feature “Lot Area”.

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: Lasso regression would be a better option because it helps in feature elimination. Also, the model will be more robust.

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: From our above analysis the 5 most important predictors are:

1) Total Square foot
2) Garage Area
3) Total No of Bathrooms
4) Overall Condition
5) Lot Area

After we delete the above top 5 predictors from incoming dataset and recalculate the lasso coefficients, we found the below 5 most important predictor variables:

	Lasso Co-Efficient
TotRmsAbvGrd	0.123240
LotFrontage	0.106755
TotalPorchSf	0.068439
Neighborhood_Veenker	0.050751
Neighborhood_StoneBr	0.036715

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 should be robust and generalizable so that they are not affected by outliers in the training data. The model also needs to be generalizable so that the test accuracy is not lesser than the training score. The model needs to perform accurately for any datasets other than the training dataset. Less weightage should be given to the outliers so that the accuracy of the model is high. To build a highly accurate model, the outlier analysis should be done and only the dataset which are relevant should be retained. The outliers which are less relevant needs to be removed. This will help in increasing the accuracy of the predictions made by the model. Also, confidence intervals can be used (typically 3-5 standard deviations). This will benefit in standardizing the predictions made by the model. The model needs to be robust so that it can be trusted for predictive analysis.