Subjective Questions: Advanced Regression
Submitted by: Rakesh C

Q1. 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?

Ans) The optimum alpha for the ridge is 2 and for the Lasso, it is 0.0001
Doubling up the alpha value that is 4 and 0.0002 results to the following:

BEFORE
The optimum alpha is 2
The R2 Score of the model on the test dataset for optimum alpha is 0.8308526739483172
The MSE of the model on the test dataset for optimum alpha is 0.0026219108677012896
AFTER
The R2 Score of the model on the test dataset for doubled alpha is 0.8401367121286467
The MSE of the model on the test dataset for doubled alpha is 0.0024780012879913383

RIDGE CO-EFFICIENT DIFFERENCE

	Ridge Co-Efficient
Total_sqr_footage	0.177934
TotRmsAbvGrd	0.157682
GarageArea	0.140742
RoofMatl_WdShngl	0.054566
OverallCond	0.050681
Total_porch_sf	0.045530
LotArea	0.044130
Condition2_PosA	0.038999
Neighborhood_StoneBr	0.036737
KitchenQual_Ex	0.036166
SaleType_ConLD	0.035953
CentralAir_Y	0.035617
Condition2_Norm	0.034374
HouseStyle_2.5Unf	0.033392
LandContour_HLS	0.032599
BsmtFullBath	0.032266
Neighborhood_Veenker	0.032239
SaleType_CWD	0.031380
BsmtQual_Ex	0.031072
Condition1_PosA	0.029804

	Ridge Doubled Alpha Co-Efficient
TotRmsAbvGrd	0.144170
Total_sqr_footage	0.132364
GarageArea	0.129064
Total_porch_sf	0.047310
OverallCond	0.044595
RoofMatl_WdShngl	0.043249
CentralAir_Y	0.037361
KitchenQual_Ex	0.036792
BsmtFullBath	0.036714
LotArea	0.035046
Neighborhood_StoneBr	0.033668
BsmtQual_Ex	0.033296
HouseStyle_2.5Unf	0.030123
Neighborhood_Veenker	0.029981
SaleType_ConLD	0.028258
Condition2_Norm	0.027573
LandContour_HLS	0.027504
Heating_GasW	0.026646
LandContour_Low	0.025887
Alley_Pave	0.025798

Lasso CO-EFFICIENT DIFFERENCE

	Lasso Co-Efficient
Total_sqr_footage	0.330670
GarageArea	0.149267
TotRmsAbvGrd	0.145800
OverallCond	0.052302
RoofMatl_WdShngl	0.046925
Total_porch_sf	0.041547
CentralAir_Y	0.035731
KitchenQual_Ex	0.031611
LandContour_HLS	0.030953
LandContour_Low	0.030330
HouseStyle_2.5Unf	0.029290
Neighborhood_StoneBr	0.029133
Heating_GasW	0.027140
BsmtQual_Ex	0.026466
Neighborhood_Veenker	0.026176
Alley_Pave	0.026042
Condition1_Norm	0.023493
MSSubClass_70	0.023208
Condition1_PosA	0.022783
LotArea	0.022001

	Lasso Doubled Alpha Co-Efficient
Total_sqr_footage	0.307323
GarageArea	0.149592
TotRmsAbvGrd	0.145473
OverallCond	0.045897
Total_porch_sf	0.042764
CentralAir_Y	0.036888
KitchenQual_Ex	0.031137
BsmtQual_Ex	0.027808
RoofMatl_WdShngl	0.026721
LandContour_Low	0.024908
LandContour_HLS	0.023481
Neighborhood_StoneBr	0.021890
Alley_Pave	0.021609
Condition1_Norm	0.021155
MSSubClass_70	0.020351
Heating_GasW	0.019444
BsmtCond_TA	0.017616
Neighborhood_Veenker	0.016854
HouseStyle_2.5Unf	0.016512
PavedDrive_Y	0.013547

The change in alpha value was very minimal, the change in R2 and MSE remained almost the same which did not have much impact. the predictor variables also hardly changed, only a few variables have minimal increase than the other variables.

Q2. 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?

Ans) The optimum value of lambda for ridge and lasso are as follows:
Ridge - 2
Lasso - 0.0001
R2
Ridge-0.8308526739483172
Lasso-0.8356403844911493
Mean Square Error
Ridge - 0.0026219108677012896
Lasso - 0.00254769775067085
The mean square error difference is minimal, it doesn't have much impact.

since lasso has the top edge in featuring few of the variables near zero it has quite a higher weightage compared to the ridge model

Q.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?

Ans) The five most important predictors are

1) total square footage area 2) Garage area 3) Total rooms above grade 4) overall condition 5) lot area

therefore a lasso model was built by removing these attributes from the dataset The R2 Score of the model on the test dataset is 0.7607336355362793 (without top5 predictors) The MSE of the model on the test dataset is 0.0037088087403233626 (without top5 predictors)

the new top5 predictor is shown in the below table

	Lasso Co-Efficient
LotFrontage	0.122682
HouseStyle_2.5Fin	0.094535
Total_porch_sf	0.083869
HouseStyle_2.5Unf	0.073230
RoofMatl_WdShngl	0.070518

Q4) 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?

Ans) The model is robust and generalizable:

1) if it is not impacted by the outliers

2) it should be simple and generalized so that more training is not required and it's widely used.

3) simple models are robust in nature but are more prone to errors it is more biased but has less variance (Bias- variance trade-off) more regularization is required.

4) the model should be accurate in the datasets and confidence analysis (standard deviation) uncertainty has to be determined for the prediction of the model's robustness.