Assignment Part-II

Question 1.

What is the optimal value of alpha for ridge and lasso regression?

1. optimal value of alpha for Ridge = 100
2. optimal value of alpha for Lasso = 0.01

What will be the changes in the model if you choose double the value of alpha for both ridge and lasso?

When I changed the alpha value from 100 to 200 and 0.01 to 0.02 respectively in Ridge and lasso Saw little bit dip in r2 score of model for ridge the values of coeff did not change but for lasso there was lot of change as you can see in below pic coeff of predictor like LotFrontage , half bath went to 0 and many of the other the coeff are reduced.

	Linear	Ridge	Lasso	Ridge_double	Lasso_double
LotFrontage	-0.046849	0.011008	-0.004351	0.011008	0.000000
LotArea	0.050607	0.053592	0.041139	0.053592	0.036371
MasVnrArea	0.073453	0.092969	0.071776	0.092969	0.071033
GrLivArea	0.455308	0.323298	0.429572	0.323298	0.415095
GarageArea	0.102480	0.130295	0.107424	0.130295	0.115554
Age	-0.207189	-0.151055	-0.196963	-0.151055	-0.189566
Condition1_Others	-0.064127	-0.054135	-0.056488	-0.054135	-0.048745
BldgType_Duplex	-0.061291	-0.042844	-0.052330	-0.042844	-0.044288
BldgType_Twnhs	-0.073115	-0.045907	-0.045756	-0.045907	-0.033139
HouseStyle_2Story	-0.129461	-0.058984	-0.088027	-0.058984	-0.059658
HouseStyle_Others	-0.087107	-0.059641	-0.061764	-0.059641	-0.043401
OverallQual_Good	0.201930	0.198760	0.202596	0.198760	0.202324
BsmtQual_Fa	0.001676	-0.007841	-0.000000	-0.007841	-0.000000
BsmtQual_Gd	-0.118706	-0.074176	-0.098363	-0.074176	-0.076514
BsmtQual_NA	-0.041487	-0.035982	-0.032039	-0.035982	-0.022905
BsmtExposure_No	-0.096115	-0.092407	-0.090377	-0.092407	-0.086991
smtFinType1_Other	-0.145149	-0.103037	-0.107985	-0.103037	-0.075407
BsmtFinType1_Unf	-0.141763	-0.103666	-0.119960	-0.103666	-0.100162
HalfBath_1	0.035462	0.036097	0.011652	0.036097	0.000000
KitchenQual_Fa	-0.008312	-0.014204	-0.001701	-0.014204	-0.000000
KitchenQual_Gd	-0.028950	-0.003248	-0.008542	-0.003248	-0.000000
Fireplaces_1	0.089580	0.103548	0.086361	0.103548	0.083584
MoSold_7	0.042150	0.032412	0.028530	0.032412	0.015844

What will be the most important predictor variables after the change is implemented?

GrLivArea: Above grade (ground) living area square feet is most important predictor variable before and after.

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?

I will choose value which model selected during first time, as we doubled the value r2 score dipped little bit so I am going with first parameter only.

1. optimal value of alpha for Ridge = 100
2. optimal value of alpha for Lasso = 0.01

Question 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?

For first Lasso model below are top 5 predictors

```
GrLivArea             0.429572
OverallQual_Good      0.202596
Age                   -0.196963
BsmtFinType1_Unf      -0.119960
BsmtFinType1_Other    -0.107985
```

After removing top 5 variable and creating new Lasso model these are to5 predictors

```
GarageArea           0.330825
Fireplaces_1         0.213147
MasVnrArea           0.183652
BsmtExposure_No      -0.138614
HalfBath_1           -0.110978
```

Question 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?

Model should not be very simple and very complex means if we create model with very less variables the model could become under fit and would perform poor both on Train and actual data.

If model is more complex it could learn all train data and perform very well (Over fit) on Train data but produce high variance in actual/test data.

To make model robust and generalizable we have balance both cases and implement model which fits in between above case.