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:

Optimal value of alpha for:

Ridge: 3.0

Lasso: 0.01

When we double the value of alpha for both Ridge & Lasso, the model will be more penalized.

	Ridge_alpha	Ridge_alpha_double	Lasso_alpha	Lasso_alpha_double
R2 Score Train	0.936906	0.931628	0.919407	0.905082
R2 Score Test	0.924754	0.924618	0.923860	0.913476
RMSE Train	0.092306	0.096089	0.104324	0.113216
RMSE Test	0.103044	0.110496	0.103654	0.103137

After doubling the alpha values in the Ridge and Lasso, the prediction accuracy remains the same for Ridge i.e., 0.93 for Train data & 0.92 for Test data, whereas for Lasso there is a slight reduction in Train data by a difference of 0.014 & in Test data by 0.01.

There is a small change in the co-efficient values as well. The new model is created and demonstrated in the Jupiter notebook. Below are the changes in the co-efficients & the important predictor variables in descending order after the changes.

For Ridge :

	Ridge_Coeff
OverallQual	0.275794
TotalAreaSF	0.240176
OverallCond	0.230851
GrLivArea	0.198578
1stFlrSF	0.179150
BsmtFinSF1	0.135322
TotalBsmtSF	0.120741
Neighborhood_Crawfor	0.120463
MSZoning_FV	0.116619
LotArea	0.112661
MSZoning_RL	0.109183
MSZoning_RH	0.099670
Functional_Typ	0.094657
2ndFlrSF	0.093082
Neighborhood_StoneBr	0.091420
MSZoning_RM	0.089799
CentralAir_Y	0.087112
FullBath	0.086636
Exterior1st_BrkFace	0.080350
GarageArea	0.078695

	Ridge Doubled Alpha Co-Efficient
OverallQual	0.233427
TotalAreaSF	0.220051
OverallCond	0.189286
GrLivArea	0.178974
1stFlrSF	0.166509
BsmtFinSF1	0.122843
TotalBsmtSF	0.113155
Neighborhood_Crawfor	0.111835
LotArea	0.108515
CentralAir_Y	0.089842
FullBath	0.084805
2ndFlrSF	0.082179
Functional_Typ	0.081594
GarageArea	0.079505
Neighborhood_StoneBr	0.079411
MSZoning_RL	0.077317
MSZoning_FV	0.076998
TotRmsAbvGrd	0.074395
Exterior1st_BrkFace	0.067073
HalfBath	0.063237

For Lasso:

	Lasso_Coeff
TotalAreaSF	0.789795
OverallQual	0.442333
OverallCond	0.270512
LotArea	0.120258
BsmtFinSF1	0.105805
GarageArea	0.104385
Neighborhood_Crawfor	0.100787
CentralAir_Y	0.088062
SaleCondition_Partial	0.080142
MSZoning_FV	0.061119
Functional_Typ	0.056518
MSZoning_RL	0.054348
Condition1_Norm	0.051063
BsmtExposure_Gd	0.046622
SaleCondition_Normal	0.042682
GarageCars	0.038316
Foundation_PConc	0.038088
PavedDrive_Y	0.037920
WoodDeckSF	0.036442
TotalPorchSF	0.031098

	Lasso Doubled Alpha Co-Efficient
TotalAreaSF	0.760670
OverallQual	0.467099
OverallCond	0.185629
GarageArea	0.138628
LotArea	0.115091
BsmtFinSF1	0.102525
CentralAir_Y	0.092878
Neighborhood_Crawfor	0.070669
SaleCondition_Partial	0.061688
MSZoning_RL	0.048050
Functional_Typ	0.044637
Foundation_PConc	0.043577
PavedDrive_Y	0.042524
Condition1_Norm	0.040154
BsmtExposure_Gd	0.035367
MSZoning_FV	0.029000
BsmtFinType1_GLQ	0.028788
SaleCondition_Normal	0.026300
GarageCond_TA	0.025353
Neighborhood_NridgHt	0.019875

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:

	Optimum Alpha
Ridge	3.000
Lasso	0.001

	Ridge	Lasso
R2 Score Train	0.936906	0.919407
R2 Score Test	0.924754	0.923860
RMSE Train	0.092306	0.104324
RMSE Test	0.103044	0.103654

We can see the RMSE values are almost the same for both the models.

Since Lasso helps in feature selection, 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:

Top 5 features before: 'TotalAreaSF', 'OverallQual', 'OverallCond', 'LotArea', 'BsmtFinSF1'

We dropped the top 5 most important predictor variables in the lasso model and created a model which gives us the below five most important predictor variables:

1. GrLivArea
2. TotalBsmtSF
3. MSZoning_FV
4. MSZoning_RL
5. MSZoning_RH

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:

A model is said to be robust when its performance doesn't change significantly when the training data undergoes small changes. And said to be generic when it performs better on unseen data.

For the model to be robust & generalisable, we must make sure the model doesn't overfit. Because an overfitted model becomes highly specific to the data on which it is trained that it fails to generalize the unseen data points in a larger domain. If the accuracy of the model on training data is high it is likely to have overfitting in the model.

Simpler models failing predicting complex real world problems. Complex models have high variance. We have to make sure the model is not too simple & not too complex.

To prevent the model from becoming too complex we use regularisation techniques like Ridge & Lasso Regression.