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:

Ridge Regression

Optimal Alpha for Ridge Regression was defined at 10. After doubling it to 20 below are the model metrics

R-Squared (Train) = 0.93
R-Squared (Test) = 0.93
RSS (Train) = 0.07
RSS (Test) = 0.02
MSE (Train) = 0.00
MSE (Test) = 0.00
RMSE (Train) = 0.01
RMSE (Test) = 0.01

Below are the new top features with their coefficients, post revising Alpha -

GrLivArea	1.006681
OverallQual_8	1.005682
Neighborhood_Crawfor	1.005363
Functional_Typ	1.005166
OverallQual_9	1.005150
Exterior1st_BrkFace	1.004774
OverallCond_9	1.004537
CentralAir_Y	1.004148
TotalBsmtSF	1.003989
OverallCond_7	1.003513
SaleCondition_Alloca	1.003239
MSSubClass_70	1.003237
BsmtCond_Gd	1.003232
OverallCond_8	1.002948
Neighborhood_StoneBr	1.002936

Lasso Regression

Optimal Alpha for Lasso Regression was defined at 0.001. After doubling it to 0.002 below are the model accuracy metrics -

R-Squared (Train) = 0.83

R-Squared (Test) = 0.86
RSS (Train) = 0.17
RSS (Test) = 0.04
MSE (Train) = 0.00
MSE (Test) = 0.00
RMSE (Train) = 0.01
RMSE (Test) = 0.01

Below are the new top features with their coefficients, post revising Alpha -
rLivArea 1.010396
YearRemodAdd 1.004679
TotalBsmtSF 1.004469
Fireplaces 1.003015
GarageCars 1.002705
GarageArea 1.001670
BsmtFinSF1 1.001396
LotArea 1.001093
GarageYrBlt_1922 1.000000
GarageYrBlt_1923 1.000000
GarageYrBlt_1924 1.000000
LotFrontage 1.000000
GarageYrBlt_1925 1.000000
GarageYrBlt_1921 1.000000
GarageYrBlt_1927 1.000000
Name: Lasso, dtype: float64

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:

Ridge Regression has a higher R-sq for Train as well as test set at 92% compared to Lasso at 87%.

However, as Lasso Regression reduces the features significantly, to less than 15, it's less complex and hence chances of performing better on unseen data is higher. For this reason we will be using Lasso Regression 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:

Post removing the top 5 variables, we have the below accuracy metrics -
R-Squared (Train) = 0.83
R-Squared (Test) = 0.85
RSS (Train) = 0.17
RSS (Test) = 0.04
MSE (Train) = 0.00
MSE (Test) = 0.00
RMSE (Train) = 0.01
RMSE (Test) = 0.01

Below are the new top 5 predictors -
1stFlrSF 0.008581
2ndFlrSF 0.006938
GarageArea 0.004283
BsmtFinSF1 0.003227
LotArea 0.002661

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:
For a model to be robust and generalized, we need to ensure that it's not very complex and in other words it does not overfit. If the model is dependent on a lot of variables, it tends to memorize the training dataset and inherently will not perform well on test or unseen data. While the model accuracy metrics like R-sq would be high on the training set, they can be significantly off on the test data set.

To improve this we will need to strike a trade off between bias and variance, where variance needs to be reduced at the cost of bias. This can be done by using Lasso/Ridge regression techniques which penalize the model for complexity by reducing weightages of features and hence performing feature selection.