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:

Optimum value of alpha for Ridge: 500

Optimum value of alpha for Lasso: 0.01

Changes in model metrics if value of alpha is doubled:

	Ridge (alpha = 500)	Ridge (alpha = 1000)	Lasso (alpha = 0.01)	Lasso (alpha = 0.02)
R2 Score (Train)	0.87	0.85	0.88	0.86
R2 Score (Test)	0.85	0.84	0.84	0.83
RSS (Train)	0.85	0.84	114.1	134.4
RSS (Test)	65.77	71.05	70.4	73.56
MSE (Train)	0.12	0.14	0.11	0.13
MSE (Test)	0.15	0.16	0.16	0.16

The values shows that models with optimum alpha have better performance matrices.

Most important predictor variables (top 10) can be identified based on the size of coefficients:

Ridge Regression, alpha = 500:

Feature	Coef_val
OverallQual	0.104721
TotalBuiltUpArea	0.096827
1stFlrSF	0.081233
Neighborhood_NoRidge	0.077805
2ndFlrSF	0.066682
RoofMatl_WdShngl	0.066168
Neighborhood_NridgHt	0.065672
BsmtExposure	0.063889
BsmtQual	0.062004
KitchenQual	0.061950

Ridge Regression, alpha = 1000:

Feature	Coef_val
OverallQual	0.088530
TotalBuiltUpArea	0.086396
1stFlrSF	0.070657
Neighborhood_NoRidge	0.067042
KitchenQual	0.058209
GarageCars	0.055376
TotRmsAbvGrd	0.055260
Neighborhood_NridgHt	0.054138
BsmtQual	0.054061
BsmtExposure	0.051745

Lasso Regression, alpha = 0.01:

Total non-zero coefficients: 91

Feature	Coef_val
2ndFlrSF	0.247941
1stFlrSF	0.191377
OverallQual	0.184735
Neighborhood_NridgHt	0.099755
Neighborhood_NoRidge	0.089190
BsmtExposure	0.087815
RoofMatl_WdShngl	0.084671
BsmtQual	0.077648
TotalBuiltUpArea	0.072646
KitchenQual	0.065230

Lasso Regression, alpha = 0.02:

Total non-zero coefficients: 58

Feature	Coef_val
OverallQual	0.199119
TotalBuiltUpArea	0.175209
2ndFlrSF	0.132537
1stFlrSF	0.110531
Neighborhood_NoRidge	0.090492
Neighborhood_NridgHt	0.078864
BsmtQual	0.078839
KitchenQual	0.076901
BsmtExposure	0.073672
RoofMatl_WdShngl	0.062733

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:

The top 10 predictor variables are same for both Ridge and Lasso regression.

Feature	Coef_val
OverallQual	0.104721
TotalBuiltUpArea	0.096827
1stFlrSF	0.081233
Neighborhood_NoRidge	0.077805
2ndFlrSF	0.066682
RoofMatl_WdShngl	0.066168
Neighborhood_NridgHt	0.065672
BsmtExposure	0.063889
BsmtQual	0.062004
KitchenQual	0.061950

Ridge (alpha=500)

Feature	Coef_val
2ndFlrSF	0.247941
1stFlrSF	0.191377
OverallQual	0.184735
Neighborhood_NridgHt	0.099755
Neighborhood_NoRidge	0.089190
BsmtExposure	0.087815
RoofMatl_WdShngl	0.084671
BsmtQual	0.077648
TotalBuiltUpArea	0.072646
KitchenQual	0.065230

Lasso (alpha=0.01)

Thus feature elimination in Lasso is not having any impact on top predictor variables.

Also, Ridge model is having better performance matrices compared to that of Lasso.

Hence we can go with Ridge model for prediction.

At the same time, variables with non-zero coefficients in Lasso model can be used as the most significant set of predictor variables.

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?

Answer:

Top 5 predictor variables of Lasso model (alpha = 0.01) are:

Feature	Coef_val
2ndFlrSF	0.247941
1stFlrSF	0.191377
OverallQual	0.184735
Neighborhood_NridgHt	0.099755
Neighborhood_NoRidge	0.089190

To accommodate incoming data without these variables, we will drop these columns from training data and rebuild the model with remaining data.

The new model have optimum alpha as: 0.01

Giving top 5 predictor variables as:

Feature	Coef_val
LotArea	0.034283
YearRemodAdd	0.049444
MasVnrArea	0.081042
ExterQual	0.100533
BsmtQual	0.117038

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?

Answer:

To make a model robust and generalizable, it has to be ensured that the model doesn't overfit.

Linear regression models generally overfit if:

- The coefficient values are too large
- The number of feature variables are too large

making the model highly complex. High complexity leads to unstable models (high variance).

This can also be identified by analyzing bias – variance trade-off.

A robust and generalized model will show similar performance for both training and test data keeping the accuracy almost same for both.