

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?

Optimal value for Ridge and Lasso is as follows:

Ridge regression: Alpha = 6

Lasso regression: Alpha = 0.001

Ridge regression:

Train R2 Score - 0.8956364397476351

Test R2 Score - 0.8543453911689542

RMSE - 0.03932933964590328

Variable	Coefficient
OverallQual	0.078
2ndFlrSF	0.070
GrLivArea	0.065
RoofMatl_WdShngl	0.064

Lasso regression:

Train R2 Score - 0.9107462494496534

Test R2 Score - 0.8359580153509536

RMSE - 0.04173803956113555

Variable	Coefficient
GrLivArea	0.317
RoofMatl_WdShngl	0.184
OverallQual	0.121
Neighborhood_NoRidge	0.065
GarageCars	0.058

After doubling the alpha for both Ridge and Lasso, the model r2 score (both test and train) decreases. Most important predictor variables after the change are:

Ridge regression:

Variable	Coefficient
OverallQual	0.064
Neighborhood_NoRidge	0.056

Variable	Coefficient
2ndFlrSF	0.054
GrLivArea	0.051

Lasso Regression:

Variable	Coefficient
GrLivArea	0.266
OverallQual	0.137
RoofMatl_WdShngl	0.130
Neighborhood_NoRidge	0.068

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?

We will choose Lasso regression, since it has almost same R2 score for optimal alpha value, but also provides the feature selection. Features which are not important in predicting target variables are pushed to zero.

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?

After dropping the top 5 most important predictor variables, following are the predictor variables:

Variable	Coefficient
1stFlrSF	0.319
2ndFlrSF	0.191
intercept	0.076
MasVnrArea	0.076
GarageArea	0.047

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?