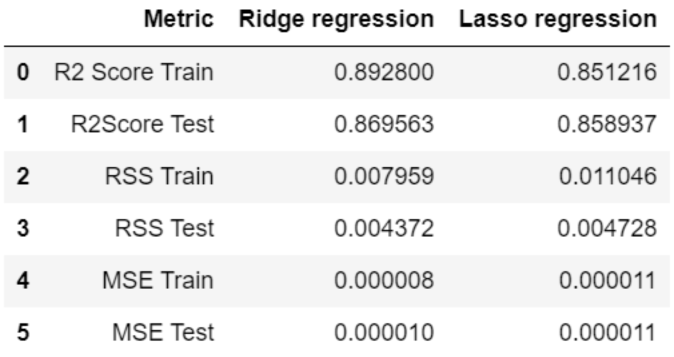
Question 1

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?

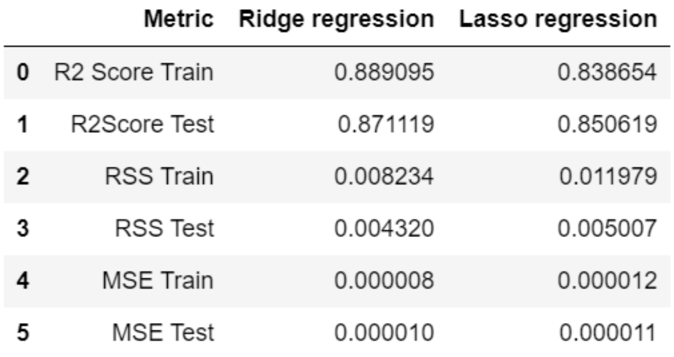
Answer:

For Ridge Regression, optimal value of alpha is 5 and for Lasso Regression optimal value is 0.0001.



	Metric	Ridge regression	Lasso regression
0	R2 Score Train	0.892800	0.851216
1	R2Score Test	0.869563	0.858937
2	RSS Train	0.007959	0.011046
3	RSS Test	0.004372	0.004728
4	MSE Train	0.000008	0.000011
5	MSE Test	0.000010	0.000011

When we double it then Ridge optimal value of alpha becomes 10 and Lasso optimal value becomes 0.0002

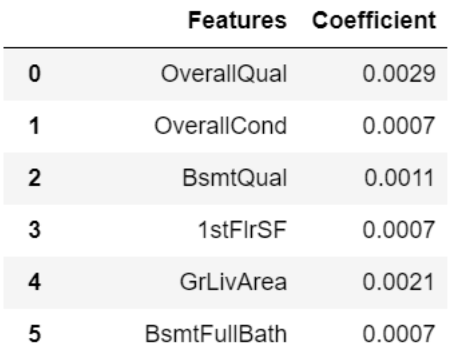


	Metric	Ridge regression	Lasso regression
0	R2 Score Train	0.889095	0.838654
1	R2Score Test	0.871119	0.850619
2	RSS Train	0.008234	0.011979
3	RSS Test	0.004320	0.005007
4	MSE Train	0.000008	0.000012
5	MSE Test	0.000010	0.000011

The difference in value of both has increased by approximately 1 point but has similar values

There is an increase in R2 score and RSS for Ridge regression

Difference between R2 score test and train has decreased for Ridge and increased for Lasso,



	Features	Coefficient
0	OverallQual	0.0029
1	OverallCond	0.0007
2	BsmtQual	0.0011
3	1stFlrSF	0.0007
4	GrLivArea	0.0021
5	BsmtFullBath	0.0007

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:

After working on models, I got 5 as my optimal value for the model for Ridge and .0001 for Lasso , Lasso can take coefficients to 0 which also helps in feature selection. But on the other hand, Ridge doesn’t change or takes any of the variable coefficient values to 0.

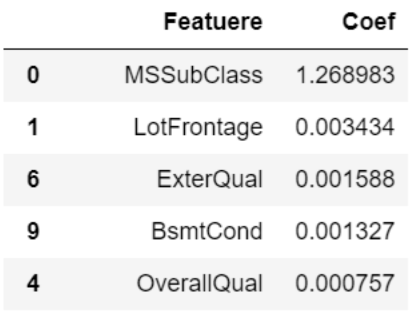
So technically Lasso has an upper hand over Ridge which brings me no confusion on using Lasso as my final regression model to apply

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?

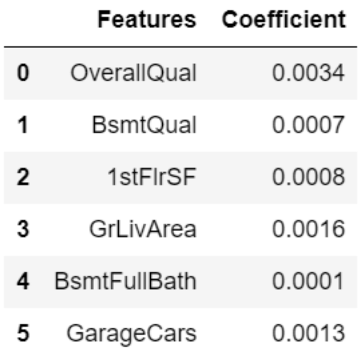
Answer:

The five most important variable were excluded from the earlier model.



	Featuere	Coef
0	MSSubClass	1.268983
1	LotFrontage	0.003434
6	ExterQual	0.001588
9	BsmtCond	0.001327
4	OverallQual	0.000757

A new model was created. After recreation, I got below 5 most important predictors



	Features	Coefficient
0	OverallQual	0.0034
1	BsmtQual	0.0007
2	1stFlrSF	0.0008
3	GrLivArea	0.0016
4	BsmtFullBath	0.0001
5	GarageCars	0.0013

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:

The model is robust and generalizable when test score is lower than training score and it is approximately max to max -5% to it. Predicted Variables should be also significant and it depends on R-squared and Adjusted R-Squared values. Also, Complex model is less robust than simple model so always try to get a simple model.

Implications on accuracy of the model

a. Standardizing values is really important to bring accurate models as without standardizing variable values would be in different tangent and thus accuracy will affect
b. P value and VIF help us to get accurate model as they provide us the required robustness and stability in the model
c. More data gives you more training data for analysis and checking accuracy with less assumptions

d. Overfitting and underfitting will affect accuracy so need cross validation set within training set. Its eats up the training set minuscule but helps to get the right model