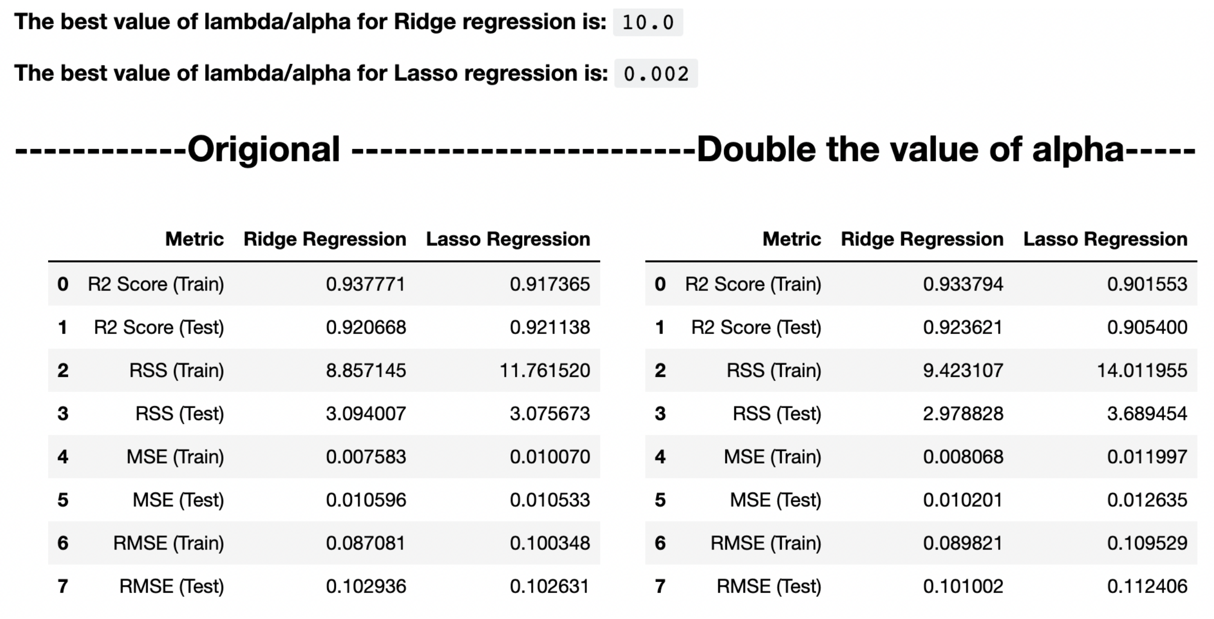
Question 1

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

The optimal alpha value for **Ridge Regression *is* 5.0**, *and **for*** **Lasso is 0.001**.

If we choose to double the alpha values following will be the numbers:



The best value of lambda/alpha for Ridge regression is: `10.0`

The best value of lambda/alpha for Lasso regression is: `0.002`

------------Origional ------------------------Double the value of alpha-----

	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.937771	0.917365
1	R2 Score (Test)	0.920668	0.921138
2	RSS (Train)	8.857145	11.761520
3	RSS (Test)	3.094007	3.075673
4	MSE (Train)	0.007583	0.010070
5	MSE (Test)	0.010596	0.010533
6	RMSE (Train)	0.087081	0.100348
7	RMSE (Test)	0.102936	0.102631

	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.933794	0.901553
1	R2 Score (Test)	0.923621	0.905400
2	RSS (Train)	9.423107	14.011955
3	RSS (Test)	2.978828	3.689454
4	MSE (Train)	0.008068	0.011997
5	MSE (Test)	0.010201	0.012635
6	RMSE (Train)	0.089821	0.109529
7	RMSE (Test)	0.101002	0.112406

Changes in Ridge Regression

- R2 score for Train data changed from 0.938 to 0.934
- R2 score for Test data changed from 0.921 to 0.924

Changes in Lasso Regression

- R2 score for Train data changed from 0.917 to 0.924
- R2 score for Test data changed from 0.921 to 0.905

The most important predictor variables after the changes are implemented are:

- Neighborhood_Crawfor
- GrLivArea
- OverallQual_Excellent

- OverallQual_Very Good
- OverallCond_Excellent
- Functional_Typ

Note: Please see the attached Jupyter notebook for the logic used to calculate above

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

I would choose **Lasso** regression. It has simplified the model significantly by removing over **72%** of features, yet it is still robust and comparable to the Ridge regression.

Note: the number 72% was calculated using this piece of code.

```
nprcnt=round(betas[round(betas['Lasso'],3)==0].shape[0] /
betas.shape[0],2)

printmd(f'**Insights:** </br> From above we can see `Lasso` regression has
removed around  **{nprcnt}%** variables')
```

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

Here are the top five(5) features after we removed the prior top 5 variables

- 2ndFlrSF
- MSSubClass_2-STORY 1945 & OLDER
- 1stFlrSF
- Exterior1st_BrkFace
- SaleCondition_Alloca

Note: Please see the attached Jupyter notebook for the logic used to calculate the above

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

First, define the terms:

Robust: A model is considered robust when data variation does not significantly affect its performance.

Generalizable: A model can adapt appropriately to new, previously unseen data drawn from the same distribution as the one used to create the model

Accuracy: How close calculated values are to their actuals.

Let's break the question into two parts:

1) How can you make sure that a model is robust and generalisable?

- To ensure a model is robust and generalizable, we need to make it `simple` and `not overfit` as it may fail on unseen data. The overfit can be sensed by looking at the R2 scores of the test and train data. The model has learned the train data if the score is very high on train data but much lower on test data.

II) What are the implications of the same for the accuracy of the model and why?

- A complex model will have very high accuracy, but with a complex model, the robustness and generalization of the model will be compromised. Also, such a model's maintenance will be far higher than a simpler model, so we need to balance model accuracy and complexity. This can be achieved by Regularization techniques like Ridge Regression and Lasso.