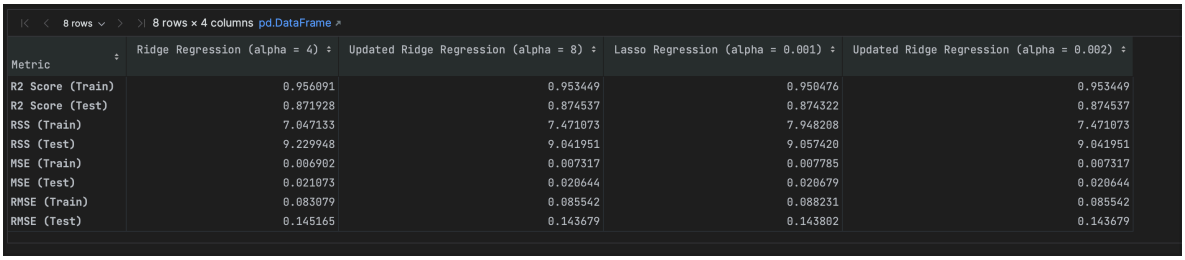
Housing Prices Case Study Assignment - Part 2

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 -
Optimal value of lambda for Ridge Regression = 4
Optimal value of lambda for Lasso = 0.001

If we double the values of alpha for both Ridge and Lasso i.e. to 8 and 0.002 respectively, we didn't saw any major change in the R2 scores for Train and Test data set -



8 rows × 4 columns pd.DataFrame

Metric	Ridge Regression (alpha = 4)	Updated Ridge Regression (alpha = 8)	Lasso Regression (alpha = 0.001)	Updated Ridge Regression (alpha = 0.002)
R2 Score (Train)	0.956091	0.953449	0.950476	0.953449
R2 Score (Test)	0.871928	0.874537	0.874322	0.874537
RSS (Train)	7.047133	7.471073	7.948208	7.471073
RSS (Test)	9.229948	9.041951	9.057420	9.041951
MSE (Train)	0.006902	0.007317	0.007785	0.007317
MSE (Test)	0.021073	0.020644	0.020679	0.020644
RMSE (Train)	0.083079	0.085542	0.088231	0.085542
RMSE (Test)	0.145165	0.143679	0.143802	0.143679

1. Ridge LinearRegression
- R2 Score for Train set remained pretty much same i.e. 0.95
- R2 Score for Test Set also remained same i.e. 0.87

2. Lasso LinearRegression
- R2 Score for Train set remained pretty much same i.e. 0.95
- R2 Score for Test Set also remained same i.e. 0.87

The most important predictor variables after the change is implemented will be -

1. RoofMatl_CompShg
2. GrLivArea
3. RoofMatl_Tar&Grv
4. OverallQual
5. RoofMatl_WdShngl
6. RoofMatl_WdShake
7. OverallCond
8. GarageCars
9. RoofMatl_Membran
10. RoofMatl_Roll

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 model we will choose to apply will depend on the use case.

- If we have too many variables and one of our primary goal is feature selection, then we will use **Lasso Linear Regression**.
- If we don't want to get too large coefficients and reduction of coefficient magnitude is one of our primary goals, then we will use **Ridge Linear Regression**.

Question 2 - 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-
We will have to drop the top 5 features in Lasso model and build the model again

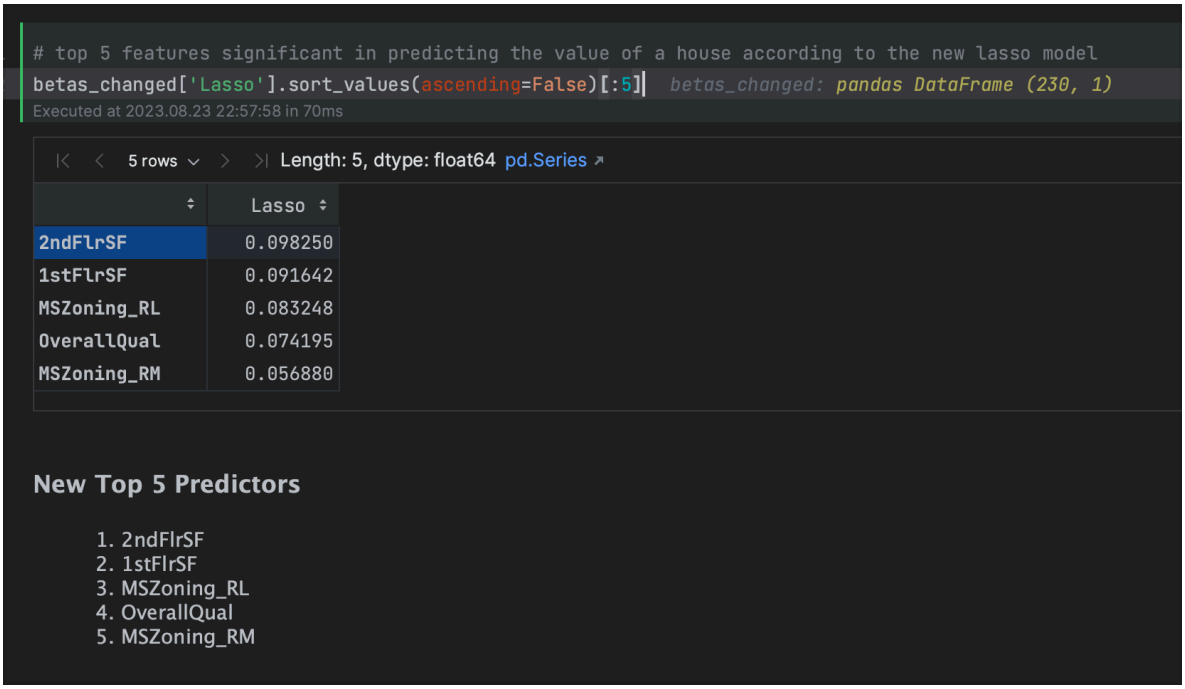
Top 5 features based on original optimal alpha values was -

Sl.No.	**Feature**	**Description**
1	RoofMatl_CompShg	Roof Material is "Standard (Composite) Shingle"
2	RoofMatl_Tar&Grv	Roof Material is "Gravel & Tar"
3	GrLivArea	Above grade (ground) living area square feet
4	RoofMatl_WdShngl	Roof Material is "Wood Shingles"
5	RoofMatl_WdShake	Roof Material is "Wood Shakes"

After we remove the above 5 columns and retrain the model using Lasso Regression, we get the following metrics -

```
R-Squared (Train) = 0.93
R-Squared (Test) = 0.87
RSS (Train) = 10.81
RSS (Test) = 9.23
MSE (Train) = 0.01
MSE (Test) = 0.02
RMSE (Train) = 0.10
RMSE (Test) = 0.15
```

And then, calculating the new top 5 predictors using betas -



New Top 5 Predictors

1. 2ndFlrSF
2. 1stFlrSF
3. MSZoning_RL
4. OverallQual
5. MSZoning_RM

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 -

1. A model is **robust** when any variation in the data does not affect its performance in a big way/by a huge margin.
2. A **generalisable** model is able to adapt to new, previously unseen data, drawn from the same distribution as the one used to create the model i.e. the variance between the train vs test data scores will not be too much.
3. A model should not overfit. If its not overfitting then it can be called robust and generalisable. This is because an overfitting model has very high variance and a smallest change in data affects the model prediction heavily. Such a model will identify all the patterns of a training data, but fail to pick up the patterns in unseen test data.
4. The model should not be too complex in order to be robust and generalisable.
5. From **Accuracy standpoint**, an overly complex model will have a very high accuracy. So, to make our model more robust and generalisable, we will have to decrease variance which will lead to some bias. Addition of bias means that accuracy will decrease.
6. In general, we have to find strike some balance between model accuracy and complexity. This can be achieved by Regularisation techniques like Ridge Regression and Lasso.