Problem Statement - Part II

Assignment Part-II

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 of alpha for ridge : {'alpha': 1.0}

optimal value of alpha for lasso: alpha = 0.0001

We will observe the changes in R2 on changing the alpha. Here are the observations:

Before changing Alpha to double

==========

R2 Score (Train)

- Ridge Regression: Train R2 Score: 0.8402983680070933
- Lasso Regression: Train R2 Score: 0.8401992901314188

R2 Score(Test)

- Ridge Regression: Test R2 Score: 0.8397113771466141
- Lasso Regression : Test R2 Score: 0.840888929980365

After changing Alpha to double

==========

R2 Score (Train)

- Ridge Regression: Train R2 Score: 0.838767

- Lasso Regression: Train R2 Score: 0.840816

R2 Score(Test)

- Ridge Regression: Test R2 Score: 0.838767
- Lasso Regression : Test R2 Score: 0.838242

Observations:

==========

- R2 score drops for both Ridge and Lasso regression on train and test set on doubling the values of alpha.
- Lasso penalizes and removes one of the features known as “Exterior1st_Stone”. According to data dictionary this depicts Exterior covering stone on the house

Out[761]:

	Feature	Ridge	Lasso
0	LotFrontage	0.056692	0.046399
1	BsmtFinSF1	0.087933	0.083275
2	TotalBsmtSF	0.167703	0.184122
3	2ndFlrSF	0.100434	0.104895
4	FullBath	0.050002	0.040758
5	Fireplaces	0.049094	0.045706
6	GarageCars	0.092509	0.091183
7	LotShape_IR2	0.034917	0.029178
8	Neighborhood_Crawfor	0.034170	0.030882
9	Neighborhood_Veenker	0.032168	0.012037
10	OverallQual_Excellent	0.164144	0.170478
11	OverallQual_Good	0.038590	0.039173
12	OverallQual_Very Excellent	0.203227	0.221557
13	OverallQual_Very Good	0.096258	0.097541
14	Exterior1st_Stone	0.025564	0.000000

The most important predictor variables after the change is implemented are shown by top 5 predictors. For this Sorted Lasso column coefficients in descending order:

Out[769]:

	Feature	Ridge	Lasso
0	OverallQual_Very Excellent	0.203227	0.221557
1	TotalBsmtSF	0.167703	0.184122
2	OverallQual_Excellent	0.164144	0.170478
3	2ndFlrSF	0.100434	0.104895
4	OverallQual_Very Good	0.096258	0.097541

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?

As we see in final_metric result that Train R2 score for Linear, Ridge and Lasso Regression are almost same but Test R2 score is high for Lasso compared to Linear and Ridge. Therefore we will consider Lasso as our final model. Lasso also helps in feature selection apart from reducing overfitting.

Out[753]:

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.840945	0.840298	0.840199
1	R2 Score (Test)	0.840300	0.839711	0.840889
2	RSS (Train)	1.915096	1.922877	1.924070
3	RSS (Test)	0.776483	0.779344	0.773618
4	MSE (Train)	0.043609	0.043698	0.043711
5	MSE (Test)	0.042396	0.042474	0.042318

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?

A large drop in R2 scores are there after dropping 5 features and training the model on remaining 9 features.

```
final_metric
```

Out[789]:

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.637938	0.629586	0.623372
1	R2 Score (Test)	0.627044	0.620981	0.625494
2	RSS (Train)	4.359388	4.459950	4.534766
3	RSS (Test)	1.813361	1.842841	1.820894
4	MSE (Train)	0.065796	0.066550	0.067106
5	MSE (Test)	0.064789	0.065313	0.064923

Five most important predictor variables now :

Out[798]:

	Feature	Linear	Ridge	Lasso
0	GarageCars	0.203447	0.181556	0.203914
1	FullBath	0.173786	0.149038	0.154403
2	BsmtFinSF1	0.136060	0.120391	0.116916
3	Fireplaces	0.103369	0.102878	0.101504
4	LotFrontage	0.109452	0.085686	0.054472

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?

A model is robust and generalisable only when it can see low bias and low variance on the unseen data. As unseen data of production environment is not there while model building.

Therefore, while model building, we divide the data set in training and test data set. Its
accuracy on test set should be as good as seen on training set. R2 score is the most important parameter for model evaluation

As we seen in below table R2 behaviour is almost same in Train and Test data set.

Out[753]:

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.840945	0.840298	0.840199
1	R2 Score (Test)	0.840300	0.839711	0.840889
2	RSS (Train)	1.915096	1.922877	1.924070
3	RSS (Test)	0.776483	0.779344	0.773618
4	MSE (Train)	0.043609	0.043698	0.043711
5	MSE (Test)	0.042396	0.042474	0.042318

While model building, we need to take care of EDA and data cleaning part.
We must handle the outliers and make imputation of missing values.
Apart from that removing the highly skewed features is also an important step.

We must care the 4 assumptions of regression like handling multicollinearity, Error term normally distributed,Residuals should not follow any pattern and Homoscedasticity.

Homoscedasticity in a model means that the error is constant along the values of the dependent variable. The best way for checking homoscedasticity is to make a scatterplot with the residuals against the dependent variable

In case of Regression problem, we can select relevant features using Lasso and RFE rather than picking up all independent
features.