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?

Answer:

As per the calculations done in the python Notebook-

Optimal value of alpha for ridge regression- 10

Optimal value of alpha for lasso regression- 0.001

If we double the Alpha values for Ridge Regression, then:

- **For Alpha = 10**

Prediction Score on Train Data: 0.9092068605070026

Prediction Score on Test Data: 0.8744204967072813

And top 5 variable coefficients are:

```
('SaleCondition_Partial', 0.143)
('SaleCondition_Others', 0.105)
('SaleCondition_Normal', 0.099)
('GarageFinish_Unf', 0.094)
('GarageFinish_RFn', 0.092)
```

- **For Alpha = 20**

Prediction Score on Train Data: 0.9011410632113473

Prediction Score on Test Data: 0.8670533292350965

And top 5 variable coefficients are:

```
('SaleCondition_Partial', 0.108)
('SaleCondition_Others', 0.089)
('SaleCondition_Normal', 0.079)
('GarageFinish_Unf', 0.077)
('GarageFinish_RFn', 0.075)
```

If we double the Alpha values for Lasso Regression, then:

- **For Alpha = 0.001**

Prediction Score on Train Data: 0.898288939025357

Prediction Score on Test Data: 0.8646575331441891

And top 5 variable coefficients are:

```
('SaleCondition_Partial', 0.198)
('SaleCondition_Others', 0.12)
('SaleCondition_Normal', 0.098)
('GarageFinish_Unf', 0.084)
('GarageFinish_RFn', 0.079)
```

- **For Alpha = 0.002**

Prediction Score on Train Data: 0.8804814856818384

Prediction Score on Test Data: 0.8493873663439893

And top 5 variable coefficients are:

```
('SaleCondition_Partial', 0.156)
('SaleCondition_Others', 0.104)
('SaleCondition_Normal', 0.07)
('GarageFinish_Unf', 0.066)
('GarageFinish_RFn', 0.059)
```

- **When we double the Alpha values for Ridge Regression & Lasso Regression, R2 score on training & testing data has been decreased.**
- **Predictors are same but the coefficient of these predictor has changed in both Ridge & Lasso Regression.**

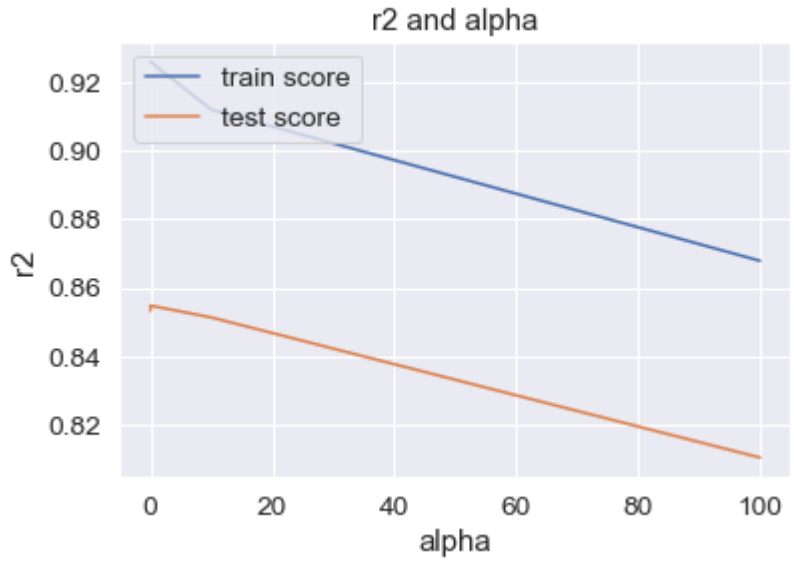
Question 2

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:

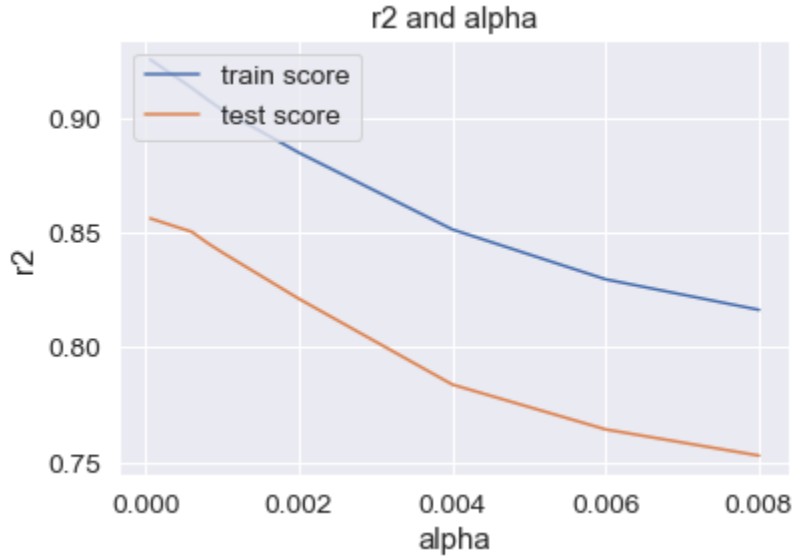
I would decide that on the basis of plots and choose a value of alpha where I have good training as well as the test score.

Ridge regression plot:



Based on the plot, I choose 10 as the value of lambda for Ridge Regression, since it has the best train as well as the test score.

Lasso Regression Plot:



Based on the plot, I choose 0.001 as the value of lambda for Lasso Regression, since it has the best train as well as the test score.

Lasso regression would be a better option it would help in feature elimination and the model will be more robust.

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:

1) **Before-The Model excluding the five most important predictor variables**-

Prediction Score on Train Data: 0.898288939025357

Prediction Score on Test Data: 0.8646575331441891

The five most important predictor variables

('SaleCondition_Partial', 0.198)

-('SaleCondition_Others', 0.12)

-('SaleCondition_Normal', 0.098)

-('GarageFinish_Unf', 0.084)

-('GarageFinish_RFn', 0.079)

2) After-The Model excluding the five most important predictor variables-

- R2 score of training and testing data has decreased.

Prediction Score on Train Data: 0.897442956866729

Prediction Score on Test Data: 0.8637289839318595

The new five most important predictor variables

('GarageFinish_No Garage', 0.203)

('GarageType_Others', 0.122)

('GarageType_No Garage', 0.097)

('GarageType_Detchd', 0.084)

('GarageType_BuiltIn', 0.08)

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:

A model is considered to be robust if the model is stable, i.e. does not change drastically upon changing the training set. The model is considered generalisable if it does not overfits the training data, and works well with new data. The model should be generalized so that the test accuracy is not lesser than the training score. The model should be accurate for datasets other than the ones which were used during training.

Its implication in terms of accuracy is that a robust and generalisable model will perform equally well on both training and test data i.e. the accuracy does not change much for training and test data.