**Upgrad\_AdvanceRegression\_Assignment**

***Assignment Part-I***

*A US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them on at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia. The data is provided in the CSV file below.*

*The company is looking at prospective properties to buy to enter the market. You are required to build a regression model using regularisation in order to predict the actual value of the prospective properties and decide whether to invest in them or not.*

***The company wants to know:***

*Which variables are significant in predicting the price of a house, and*

*How well those variables describe the price of a house.*

*Also, determine the optimal value of lambda for ridge and lasso regression.*

--------------------------------------------------------------------------------------------------------------------------------------

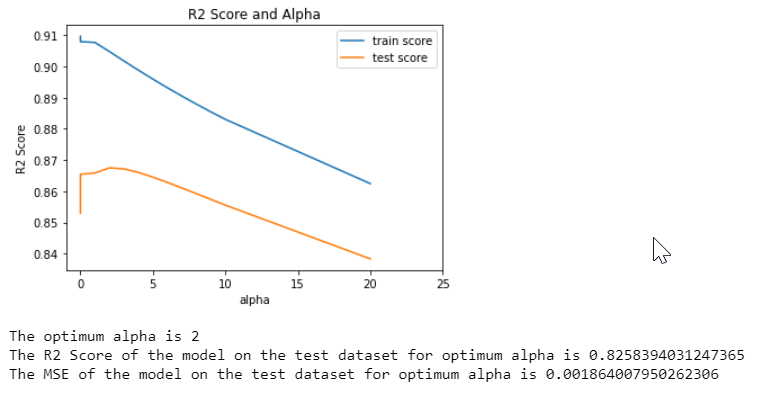
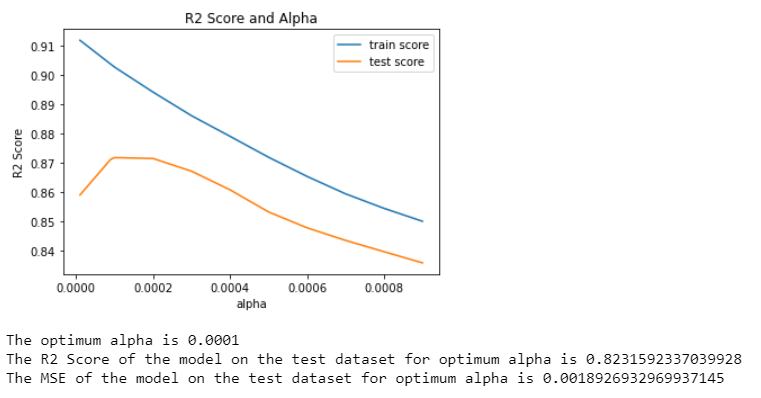
Ans:

We can achieve an R2 score of 0.82 approx. on both Ridge and Lasso Models. The following factors influence the house price the most as demonstrated by both the models: -

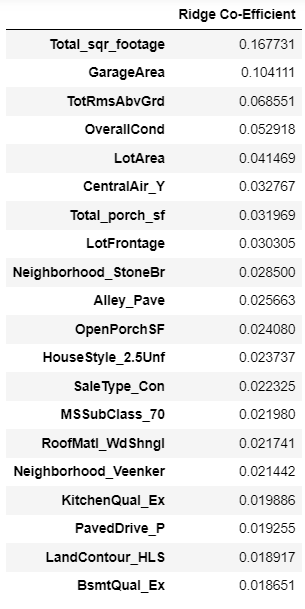
1. Total area in square foot
2. Total Garage Area
3. Total Rooms
4. Overall Condition
5. Lot Area
6. Centrally Air Conditioned
7. Total Porch Area (Open + Enclosed)
8. Kitchen Quality
9. Basement Quality

-----------------------------------------------------------------------------------------------------------------------------------------

Ridge: Lasso:

Below table shows the coefficient of variables for which impact the price of a house.

Optimum value of Alpha:

|  |  |  |
| --- | --- | --- |
|  | Ridge | Lasso |
| Optimum Value (Alpha) | 2 | 0.0001 |

---------------------------------------------------------------------------------------------------------------------------------------------------------

***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 to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?**

**Optimum value of Alpha:**

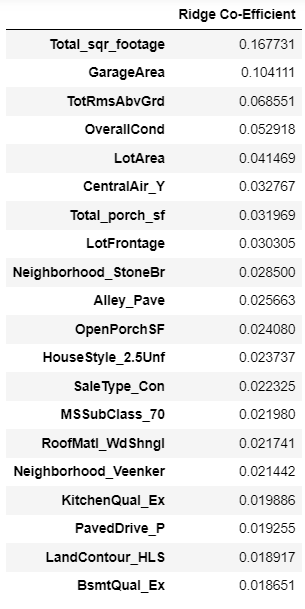
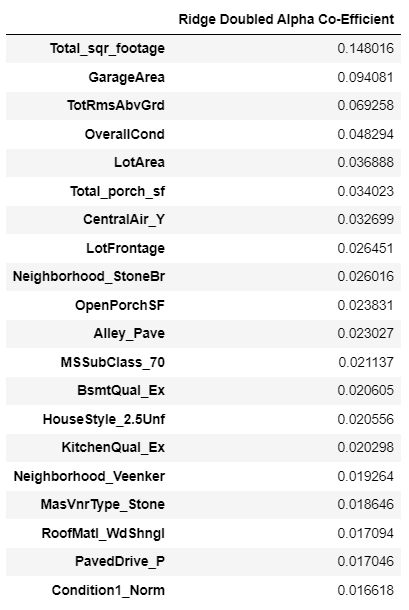
|  |  |  |
| --- | --- | --- |
|  | Ridge | Lasso |
| Optimum Value(Alpha) | 2 | 0.0001 |
| R^2 value | 0.825839403 | 0.823159234 |

**Ridge Doubled:**

The R2 Score of the model on the test dataset for doubled alpha is 0.8235972351125412

The MSE of the model on the test dataset for doubled alpha is 0.0018880054507046655

The most important predictor variables are as follows:



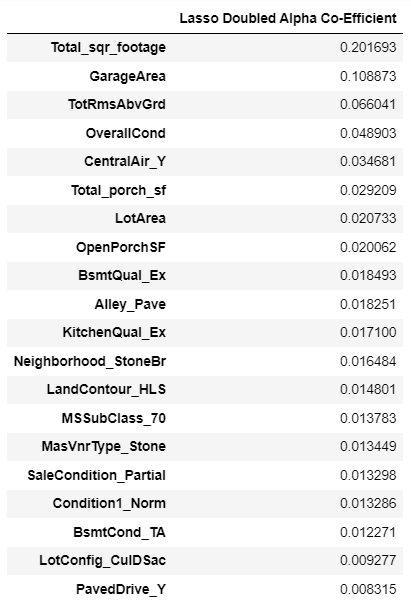
Remarks: Not much change is observed for Ridge model if we choose to double the alpha value. In general, if the alpha value changes, ridge regression shrinks coefficient slowly. So after the doubling the value if alpha, we observe the Coeff magnitude value changes slightly for the variables but overall structure of model remain as like as before.

**Lasso Doubled:**

The R2 Score of the model on the test dataset for doubled alpha is 0.8202721121912073

The MSE of the model on the test dataset for doubled alpha is 0.0019235936128502332

The most important predictor variables are as follows:



Remarks: In general, Lasso reg is sensitive to changes in alpha value, higher alpha means strong regularization, feature elimination, results simpler model. But in our case, alpha value is so small, so that much change is not observed while doubling the alpha value. Only CentralAir\_Y becomes the 5th imp variable for model (Lasso Doubled).

**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?**

|  |  |  |
| --- | --- | --- |
|  | Ridge | Lasso |
| Optimum Value(Alpha) | 2 | 0.0001 |
| R^2 value | 0.825839403 | 0.823159234 |
| MSE Value | 0.001864008 | 0.001892693 |

The purpose of Ridge and Lasso Reg is to prevent overfitting and generalize the performance of model by regularizing the Coeff value to balance between model complexity and performance.

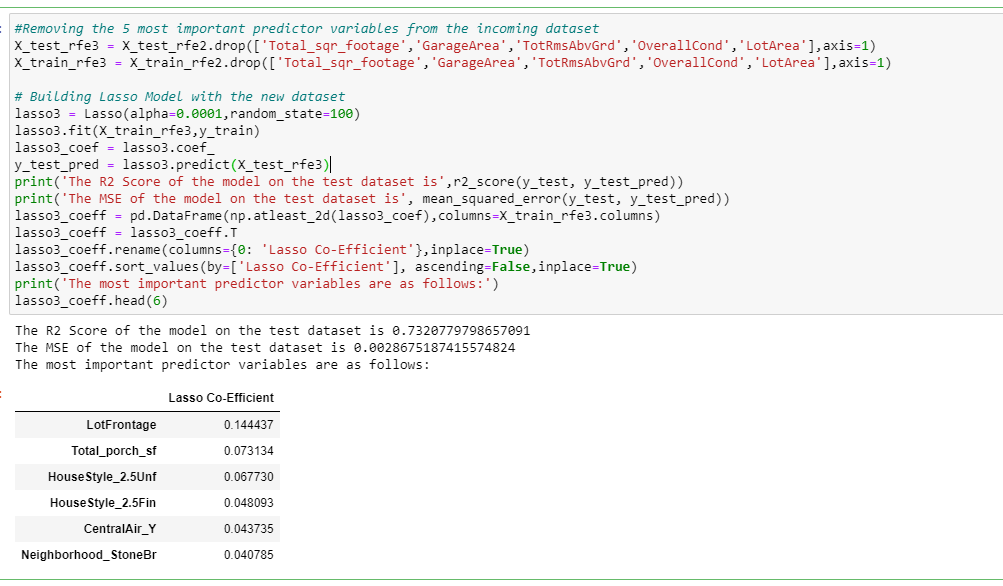
If we observe the MSE value, Ridge’s MSE value is slightly smaller than Lasso. It means the Ridge performs better than Lasso. But as we have lots of feature to predict the house price and we need to understand the impact of all these features and find out the most effective one, I will prefer for Lasso, as it helps in feature reduction.

**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?**

The R2 Score of the model on the test dataset is 0.7320779798657091

The MSE of the model on the test dataset is 0.0028675187415574824

****

Five most important predictor variable:

| **Lasso Co-Efficient** |
| --- |
| **LotFrontage** | **0.144437** |
| **Total\_porch\_sf** | **0.073134** |
| **HouseStyle\_2.5Unf** | **0.067730** |
| **HouseStyle\_2.5Fin** | **0.048093** |
| **CentralAir\_Y** | **0.043735** |

**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?**

**Ans**:

If two model provide similar performance, in the training and testing data, we should choose simple model, as simple model captures the essential patterns in the data without overfitting to noise. Simple model more robust to change in the training data. Simple model requires fewer training data.

So as per Occam’s razor, “A predictive model has to be as simple as possible, but no simpler”.

There are few techniques that helps to keep final model simpler.

Regularization techniques like Lasso and Ridge regularization used to prevent overfitting and improve generalization.

Avoiding overly complex models that might memorize the training data instead of learning general patterns. Simpler models often have better generalization.

**Making a model simpler, leads to Bias - Variance Trade off:** Bias -variance trade-off is a fundamental concept of machine learning which can deals with finding the right balance between model simplicity and complexity. High bias model underfit the data, it generates high training error, and high-test error. A high variance model, overfits the training data, capturing noise and fluctuations rather than underlying patterns. For that, it generates low train error but high-test error.

Our goal is to find out the optimal point, where the total error is minimized which leads to best possible generalize performance on unseen data. Accuracy of the model, can be maintained by keeping the balance between Bias and Variance as it minimizes the total error.

