

Advance Regression:

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 value of alpha for Ridge regression is 0.9

The optimal value of alpha for Lasso regression is 0.001

When I double the value of alpha in lasso regression and ridge regression, there is no significant change in the r^2 score of train and test data.

There is no change in the predictor variable after the changes in alpha values.

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

Both Lasso and Ridge have almost same r^2 scores. We can consider Lasso as our final model since it has feature elimination

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

The five most important predictor variables are

- 1. MiscVal: \$Value of miscellaneous feature**
- 2. BsmtUnfSF: Unfinished square feet of basement area**

- 3. **BsmtFinType2: Rating of basement finished area (if multiple types)**
- 4. **CentralAir: Central air conditioning**
- 5. **LotArea: Lot size in square feet**

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

To make model robust and generalisable we need to handel the outliers. To handle the outliers we can take following actions based on the situation:

- 1. Drop the outliers:** If there are very few records with extermne values, we can drop those records.
- 2. Winsorizing:** In this we set the extreme values to some specific values like we can set value of 95th percentile to the upper 5% of values.
- 3. Log-Scale Transformation:** In this method we change the value y to $\log(y)$. We do this mostly when response variable follows exponential distribution.
- 4. Binning:** in this method we create groups of continuous variables. We do this to indentify the hidden patterns in the continuous variables.