Advanced Regression Assignment – PART 2

Q-1

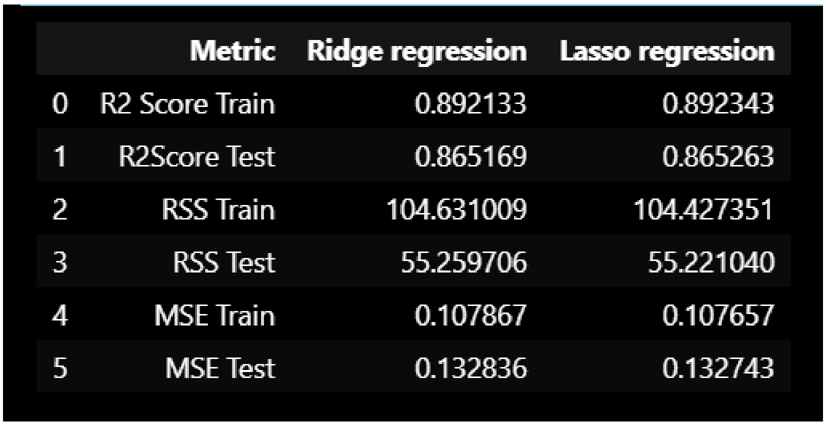
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Ans:

Optimal value of alpha:

Ridge – 0.6

Lasso – 0.0001

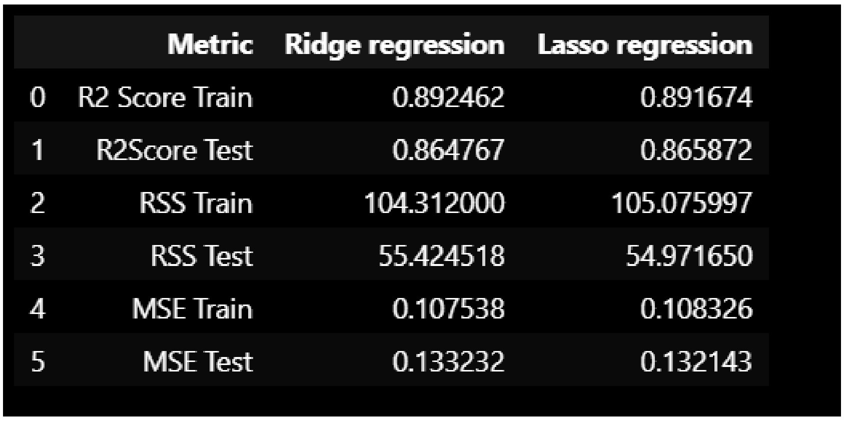


	Metric	Ridge regression	Lasso regression
0	R2 Score Train	0.892133	0.892343
1	R2Score Test	0.865169	0.865263
2	RSS Train	104.631009	104.427351
3	RSS Test	55.259706	55.221040
4	MSE Train	0.107867	0.107657
5	MSE Test	0.132836	0.132743

Double the values

Ridge – 0.12

Lasso – 0.0002



	Metric	Ridge regression	Lasso regression
0	R2 Score Train	0.892462	0.891674
1	R2Score Test	0.864767	0.865872
2	RSS Train	104.312000	105.075997
3	RSS Test	55.424518	54.971650
4	MSE Train	0.107538	0.108326
5	MSE Test	0.133232	0.132143

After doubling the value,

R2 score doesn't show any noticeable difference

Whereas, RSS and MSE values shows noticeable difference for Ridge and Lasso.

The most important feature after double the value of alpha is

- *MSZoning_FV*
- *MSZoning_RL*
- *MSZoning_RH*
- *Neighborhood_Veenker*
- *MSZoning_RM*
- *SaleType_Oth*
- *GrLivArea*
- *OverallQual*
- *HouseStyle_2.5Fin*
- *RoofStyle_Shed*

Q – 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?

Ans:

The values I got during the assignment for ridge and lasso are 0.6 and 0.0001.

Based on alpha/lamda values I have got, Ridge regression does not have zero value for any co efficient,

Whereas Lasso nearly zeroed for one or two co efficient, So Lasso is better option and it also helps in eliminating some of the features. So I would choose Lasso.

Q-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?

Ans:

After removing the five most important features we have got new top 5 features as "MSSubClass", "BuiltOrremodelAge", "OverallCond", "OverallQual", "ExterCond".

Code is included in the jupyter notebook.

Q – 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:

The model is robust when,

- *Test scores does not differ much from the training score*
- *The model should not be impacted by the outliers.*
- *Outlier treatment is most important to get the robust model. We can detect outliers in the dataset using box plots, Z score etc. Treating the outliers will not effect mean, median etc. So that we can impute correct values to missing values., the outlier analysis needs to be done and only those which are relevant.*
- *This would standardize the predictions made by model. If the model is not robust, it cannot be trusted for predictive analysis*
- *The predicted variables should be significant.*
- *Model significance can be determined by the P-Values, R2 and adjusted R2.*

Implications Of Accuracy Of a Model:

1. *Gain more data as much as you can.*
   a. *Having more data allows data to train itself more accurately, instead of depending on weak correlation and assumptions, it is good to have more data.*
2. *Fix missing values and outliers:*
   a. *If the data has missing values and outliers can lead to inaccurate model. Outliers can affect the mean, median that we are imputing to continuous variables*
   b. *You can get the outlier values using a boxplot, treating the outliers in the data will make our model more accurate*
3. *Featuring Engineering or newly derived columns/standardize the values:*
   a. *We can extract the new data from existing data, e.g. from DOB we can get the age of the person, after extracting new data required we can drop the existing features.*
   b. *Scaling the values, e.g. one value is in grams and other values are in kilograms then it is necessary to convert each value in common unit.*

4. *Feature Selection:*
    1. *It is purely based on domain knowledge, so that we can select important features that have good impact on the target variable.*
    2. *Data visualisation also helps in selecting the features.*
    3. *Statistical parameters like P-Values, VIF can give us significant variables.*
5. *Applying the right algorithm:*
    1. *Choosing the right ML algorithm is very important to build an accurate model.*
6. *Cross validation:*
    1. *Some times more accuracy will cause overfitting, then we can use cross validation technique, i.e. leave a sample on which you do not train the model & test the model on this sample before got to the final model.*