Assignment -Advance Regression

~ Kiran Sumit Dalmiya

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 and lasso regression

Ridge Alpha 4

lasso Alpha 2

Now If we choose to double the value of alpha for both ridge and lasso regression:

Ridge Alpha 8

lasso Alpha 4

So in Ridge regression :

Alpha =4

R2score(train)= 0.873194882513791

R2score(test) = 0.8620092899244391

Alpha =8

R2score(train) = 0.8632600332904024

R2score(test)= 0.8564643650812065

R2score in ridge has decreased on training data as well as testing data

So in Lasso regression :

Alpha =2

R2score(train)= 0.8820008198275171

R2score(test) = 0.8535711495616316

Alpha =4

R2score(train) = 0.881980926349407

R2score(test)= 0.8541521933646786

R2score in lasso has decreased slightly on training data and increased slightly on testing data

Predictors are same but the coefficent of these predictor has changed

LotArea - Lot size in square feet

OverallQual - Rates the overall material and finish of the house

YearBuilt - Original construction date

BsmtFinSF1 - Type 1 finished square feet

TotalBsmtSF - Total square feet of basement area

GrLivArea - Above grade (ground) living area square feet

TotRmsAbvGrd - Total rooms above grade (does not include bathrooms)

Street_Pave - Pave road access to property

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 :

The r2_score of ridge(0.86) is slightly higher than lasso(0.85) for the test dataset so we will choose ridge regression to solve this problem

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:

After building the model the five most important variable predictor variables are:
LotArea - Lot size in square feet
OverallQual - Rates the overall material and finish of the house

YearBuilt - Original construction date

BsmtFinSF1 - Type 1 finished square feet

TotalBsmtSF - Total square feet of basement area

Now as given these variable are not available in incoming data .

So we have to remove this top 5 variable and create a model again

After doing it in jupyter notebook the top 5 predictor variables now are :

five most important predictor variables

1stFlrSF --- First Floor square feet

GrLivArea --- Above grade (ground) living area square feet

Street_Pave --- Pave road access to property

RoofMatl_Metal --- Roof material_Metal

RoofStyle_Shed --- Type of roof(Shed)

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:

We want to create a model that works well not only on the data it was trained on but also on new, unseen data. If the model performs significantly better on the training data compared to the test data, it might not generalize well. The objective is to develop a model with robust generalization capabilities, ensuring that its performance on unseen data is not significantly inferior to its performance on the training dataset. Discrepancies between training and test accuracy should be minimized, indicating a robust and reliable predictive model. Additionally, we aim for the model to be robust, meaning it shouldn't be overly affected by outliers in the data. Outliers are extreme values that can skew predictions. We need to analyse and keep only the relevant outliers that make sense for our dataset. Removing outliers that don't contribute meaningfully to the data helps improve the model's accuracy.

In summary, a good model should generalize well to new, unseen data, perform similarly on both the training and test datasets, be robust to outliers, focusing on the relevant ones and disregarding outliers that don't make sense for the context.