Problem Statement - Part II

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:

The Optimal Value of alphas for Ridge & Lasso is:

Ridge : 2.0

Lasso : 0.001

Metric------Ridge Regression ---------Lasso Regression

R2_Score (Train) ---------0.928670 ----------0.904682

R2_Score (Test) ---------0.869510 ----------0.874555

On Doubling alphas

Ridge : alpha = 4.0

- R-Squared score on Train set : 0.923855

- R-Squared Score on Test set : 0.872007

--> On doubling the alpha on ridge model, R2 score has reduced both on Train & Test set.

Lasso: alpha = 0.002

- R-Squared score on Train set : 0.904682

- R-Squared Score on Test set : 0.874555

--> On doubling the alpha on ridge model, R2 score is same on both Train & Test set.

The top 30 significant features on doubling alpha are :

GrLivArea,OverallQual_9,Fireplaces_2,OverallQual_8,GarageArea,CentralAir,BsmtExposure_Gd,GarageCars_3,MSZoning_RL,MSZoning_FV,

Fireplaces_1,Neighborhood_Crawfor,GarageCond_TA,SaleType_New,Neighborhood_NridgHt,OverallQual_7,TotRmsAbvGrd_10,

LotConfig_CulDSac,FullBath_3,BsmtFinType1_GLQ,Functional_Typ,HalfBath_1,Condition1_Norm,BsmtFullBath_1,KitchenAbvGr_1,

OverallCond_7,HouseStyle_1Story,ExterQual_Gd,Foundation_PConc,BedroomAbvGr_4

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?

- Lasso regression will eliminate the multicollinearity between predictors & the high dimensionality in the dataset.
- R2 Score of Lasso is slightly higher than the Ridge model on the housing dataset.
- So, we can choose Lasso over Ridge regression model as it is more robust

Question 3

After building the model, you realized 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 top 5 predictors from the first Lasso model are,

GrLivArea,
OverallQual_9,
OverallQual_8,
'Fireplaces_2',
'MSZoning_FV'

On dropping the top 5 predictors from train & test set, lets rebuild a new Lasso model, with alpha = 0.001, the scores on the newly built model are,

```
R-Squared score on Train set : 0.8908860846045172
R-Squared Score on Test set :  0.8507929144230305
Residual Sum of Squares in Train set : 17.537389338777892
Residual Sum of Squares in Test set : 10.753141446466026
mean Squared Error on Train : 0.017159872151446078
mean Squared Error on Test : 0.02455055124763933
```

R-Squared with top 5 features

Lasso Regression Model has the below Score on Train & Test

- R-Squared (Train) : 0.904682

- R-Squared (Test) : 0.874555

R-Squared without top 5 features

R-Squared score on Train set : 0.8908860846045172

R-Squared Score on Test set : 0.8507929144230305

- There is a slight drop in the R2-scores after dropping the top 5 significant features

Hence, there is a slight drop in the R2-scores after dropping the top 5 significant features

Question 4

How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

Answer:

1. **The model should be robust and generalizable with simple linear cost function & with less complexity**

2. **The model should be able to perform well on the unseen data or any missing features.**

3. **Proper scaling to be done on the dataset to handle predictors outliers & power or log transformation to be applied on the target to predict the unseen values of the target variable.**

4. **The model should be more generalizable so that the test accuracy is not more than the train accuracy.**