Subham Sahoo
C37

subjective question for Advance Regression Assignment

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?

Ans : The optimal value of alpha for ridge regression is 3. The optimal value of alpha for lasso regression is 0.0001.

Changes in R squared (train+test), RSS(train+test), MSE(train+test) for both ridge and lasso when alpha value is doubled are shown below.

Model	alpha	R square Train	RSS Test	RSS Train	RSS Test	MSE Train	MSE Test
Ridge	3	0.9364	0.9214	0.9731	0.4086	0.00118	0.00116
Ridge	6	0.9314	0.9202	1.0489	0.4149	0.00128	0.00118
Lasso	0.0001	0.9346	0.9265	1.00002	0.3822	0.00121	0.00108
Lasso	0.0002	0.9278	0.9270	1.104	0.3798	0.0013	0.001079

After the change, it was found that for Ridge regression, both train score and test score also decreased slightly.

But for lasso regression, train score decreased while test score increased slightly.

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?

I will choose lasso regression for below reasons.

1. Lasso gives slightly better test performance that Ridge regression
2. Lasso does feature selection as a result of which a lot of redundant features get eliminated during model building.
3. In our model, there are total 197 columns in input data among which lasso regression was able to remove 104 columns by making their co-efficient as 0.

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 initial five most important predictor variables in lasso model are

1. **1stFlrSF**
2. **GrLivArea**
3. **BsmtFinSF1**
4. **Neighborhood**
5. **ExterQual**

After creating another model without these 5 features, the five most important predictor variables came out to be

1. **TotalBsmtSF**
2. **2ndFlrSF**
3. **LotArea**
4. **BsmtUnfSF**
5. **Foundation**

Pasted the snapshot the code editor after building and displaying absolute of 5 highest beta values.

```
In [249]: final_betas1
```

Out[249]:

	Linear	Ridge	Lasso
TotalBsmtSF	3.512038e-01	0.102954	0.298365
2ndFlrSF	1.873598e-01	0.067296	0.143302
LotArea	6.484622e-02	0.048361	0.066610
BsmtUnfSF	6.956237e-02	0.005064	0.052180
Foundation_Slab	6.327973e-02	0.016409	0.051871
TotRmsAbvGrd	7.066572e-02	0.053904	0.050348
GarageArea	5.134696e-02	0.039113	0.049845
BsmtQual_Gd	3.904392e-02	0.042572	0.040640
BsmtExposure_Gd	3.445060e-02	0.042785	0.038944

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?

Model will be called robust when the trained model will perform effectively on unseen, but similar data (test data).

Model will be called generalizable when the model doesn't memorize the whole train data, but only those details which can be generalized to any similar dataset (test data). In this case, model will be performing almost equally in train as well as test set.

We have to follow below techniques to make sure model is robust and generalisation.

a. Do EDA to understand data distribution
b. Check the prerequisites like output should have some linear relationship with some input variables.
c. In case of direct linear relationship is absent, we can do data transformation, polynomial regression or non-linear regression
d. After building regression model, we have to make sure model is having low bias and low variance
e. To meet low bias and low variance, we have to do regularisation.
f. We can do regularisation such that model learns generalized feature, but doesn't over-fit on train data.
g. To do that we can go either for Ridge Regression or Lasso Regression
h. We have to look into R squared value for train data as well as test data to check test accuracy is not much less compared to train data.

The accuracy of a model depends on how well the model has learnt the generalized properties of the data so that it can apply that on unseen data to predict.

Robustness and generalization capability of a model increases the overall accuracy of a model on unseen data.