Name – Neelam Mahapatro

Batch – C46

Advanced Linear Regression Assignment

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 – From the experiment, the optimal value of alpha obtained for ridge regression is 3. The optimal value of alpha obtained 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	R square Test	MSE Train	MSE Test
Ridge	3	0.9364	0.9214	0.00118	0.00116
Ridge	6	0.9314	0.9202	0.00128	0.00118
Lasso	0.0001	0.9346	0.9265	0.00121	0.00108
Lasso	0.0002	0.9278	0.9270	0.0013	0.001079

   With doubling the alpha, both train and test score increased for Ridge regression. However, for Lasso regression, train score decreased and test score increased by 0.05 percent.

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 optimal value of lamda obtained for ridge and lasso regressions are 3.0 and 0.0001 respectively. I will choose Lasso regression over Ridge regression because of the following reasons –

a. Lasso gives better test performance then Ridge

b. Lasso has better feature selection as a lot of redundant features got eliminated during model building.

c. Out of 197 columns, lasso method was able to make 104 coefficients as 0.

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 – From the current model, the initial five most important predictor variables in lasso model are

a. 1stFlrSF

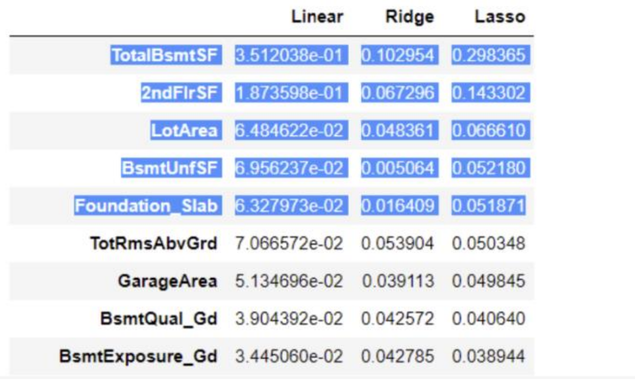
b. GrLivArea

c. BsmtFinSF1

d. Neighborhood

e. ExterQual

After eliminating these 5 features and building another model the five most important predictor variables obtained are



	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

a. TotalBsmtSF
b. 2ndFlrSF
c. LotArea
d. BsmtUnfSF
e. Foundation

The code snippet displaying above 5 features can be found in above section.

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 -

- A Model will be called robust when it will perform effectively very well on unseen, but similar test data .
- A Model can be called generalizable when it doesn't memorize the whole train data, but only those details which can be generalized to the dataset which is similar. Model should perform equally on test and train dataset.
- The steps and techniques which should be followed for creating a robust and generalized model are –

i. A clear and concise Exploratory Data Analysis (EDA) to understand the distribution of the data.
ii. Perform univariate and bivariate analysis.
iii. Removing unnecessary features and cleaning the dataset
iv. Check the prerequisites if output variable is linearly dependent on the input features. If linear relationship is not there, we should take approach for transforming data and developing non-linear regression model for the same
v. Find the numerical and categorical features and perform the transformation accordingly.
vi. Do a test train data split and build a Regression model having low bias and low variance.

vii. To meet the above requirements, we have to perform regularization by using lasso and ridge regression. We can do regularization such that the model learns generalized the feature but doesnot overfit on the train dataset.

viii. We have to check the R squared and mean squared values on train and test dataset.

ix. The accuracy of the model depends on how well the model has learnt the generalized properties of the dataset and how well it can predict on unseen/test dataset.

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