Advanced Regression Assignment

(Problem Statement – Part II) By – Manish Kumar Singh

Q1. 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 1. The optimal values of alpha are:

a) For ridge regression, it is 20 and
b) For lasso regression, it is 0.002

After choosing double the alpha value, the R2-squared values for train & test sets have decreased slightly in both ridge and lasso regression.
Whereas, the RMSE values have increased slightly in both ridge and lasso regression. The same can be seen in the tabular form below.

Metric	Ridge(α)	Ridge(2*α)	Lasso(α)	Lasso(2*α)
R2-Train	88.14%	87.11%	88.75%	86.38%
R2-Test	86.84%	86.29%	86.81%	86.52%
RMSE-Train	0.11851	0.12885	0.11244	0.13620
RSME-Test	0.13302	0.13852	0.1333	0.13625

The top most five important predictor variables after the changed is implemented are:
a) Neighborhood_NridgHt
b) Neighborhood_OldTown
c) BsmtFullBath
d) OverallCond
e) Neighborhood_Timber

Q2. 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 2: After determining the optimal value of lambda for both ridge and lasso regression, we got almost the similar values for metrics like R2 score & RMSE but since lasso penalizes more on the data and it also helps in feature selection by eliminating variables which are not significant. We will choose lasso regression model as the final model.

Q3. 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 predictors variables. Which are the five most important predictor variables now?

Answer 3: After releasing that the five most important predictor variables are not available & re-building the new model excluding that five variables, now we have the following five most important predictor variables are:

a) Exterior1s_AsphShn
b) Neighborhood_SWISU
c) BsmtHalfBath
d) RoofMatl_Membran
e) Exterior2nd_MetalSd

Q4. How can you make sure that a model is robust and generalisable? What is the implications of the same for the accuracy of the model and why?

Answer 4: Since we have finalize our model as lasso regression model which is having less number of variables as compare to ridge regression model. Also, we saw that our model performs very well on the training as well as testing data sets, we also observed that the metrics (R2 & RMSE) for both training and testing almost similar, we can say our model is robust and generalize.

The model generated should be such that it performs very well on the unseen data other than trained data on which has it learnt. The model should not give much weightage to the outliers present in the data which are not significant otherwise it will not perform well on the unseen data. However, outlier treatment has to be performed to evaluate the importance of outliers and the redundant outliers present in the data should be removed. In this way the model predictive performance or the accuracy of the model can be increased.