PROBLEM PART – II

NAME :ANINDITA DAS

CHAPETR: ADVANCE REGRESSION

DATE: 07-09-2022

GITHUB_LINK : https://github.com/Anindita2019/Advance_Regression

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?

Using an alpha value of **10**, the evaluation of the model, the train, and test data indicate better performance on the ridge model than on the linear regression model. **1.0** gave the best train(91%) and test(90%) results for ridge regression

For lasso regression, the alpha value is **1**. The output is the best cross-validated lambda, which comes out to be 0.001

In case of unnormalized features in ridge regression, the coefficients of other columns may also change depending upon the correlation between them. We can't tell about the coefficients of other columns. While in case of normalized features in ridge regression, the column whose values are doubled gets the coefficients halved while others remains unchanged.

The double lasso. method is **calibrated to not over-select potentially spurious covariates**, and simulations. demonstrate that using this method reduces error and increases statistical power. This method. can be used to identify which covariates have sufficient empirical support for inclusion in.

Generally **variable with highest correlation** is a good predictor. You can also compare coefficients to select the best predictor (Make sure you have normalized the data before you perform regression and you take absolute value of coefficients) You can also look change in R-squared value

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?

Selecting a good value for λ is critical. When λ=0, the penalty term has no effect, and ridge regression will produce the classical least square coefficients. However, as λ increases to infinite, the impact of the shrinkage penalty grows, and the ridge regression coefficients will get close zero. **Lasso regression** would be a better option it would help in feature elimination and the model will be more robust

1. The model we will choose to apply will depend on the use case.
2. If we have too many variables and one of our primary goal is feature selection, then we will use **Lasso**.
3. If we don't want to get too large coefficients and reduction of coefficient magnitude is one of our prime goals, then we will use **Ridge Regression**.

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?

Here, we will drop the top 5 features in Lasso model and build the model again.

Top 5 Lasso predictors were: 'GrLivArea', 'OverallQual', 'OverallCond', 'TotalBsmtSF', 'BsmtFinSF1'

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?

- A model is **robust** when any variation in the data does not affect its performance much.
- A **generalizable** model is able to adapt properly to new, previously unseen data, drawn from the same distribution as the one used to create the model.

- To make sure a model is robust and generalizable, we have to **take care it doesn't overfit**. This is because an overfitting model has very high variance and a smallest change in data affects the model prediction heavily. Such a model will identify all the patterns of a training data, but fail to pick up the patterns in unseen test data.
- In other words, the model should not be too complex in order to be robust and generalizable.

- If we look at it from the prespective of **Accuracy**, a too complex model will have a very high accuracy. So, to make our model more robust and generalizable, we will have to decrease variance which will lead to some bias. Addition of bias means that accuracy will decrease.
- In general, we have to find strike some balance between model accuracy and complexity. This can be achieved by Regularization techniques like Ridge Regression and Lasso.