ASSIGNMENT: ADVANCED REGRESSION SUBJECTIVE QUESTIONS

**Declaration: Following document has the response/answers for the questions shared as part Advanced Regression assignment.*

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?

Ans1: When we plot the curve between negative mean absolute error and alpha, we see that as the value of alpha increase from 0 the error term decrease and the train error is showing increasing trend when value of alpha increases. When the value of alpha is 2 the test error is minimum, so we decided to go with value of alpha equal to 2 for our ridge regression.

For lasso regression small value has been set as 0.01, when we increase the value of alpha the model will try to penalize more and try to make most of the coefficient value zero.

When we double the value of alpha for our ridge regression no, we will take the value of alpha equal to 10 the model will apply more penalty on the curve and try to make the model more generalized that is making model more simpler and no thinking to fit every data of the data set from the graph we can see that when alpha is 10 we get more error for both test and train.

Similarly, when we increase the value of alpha for lasso, we try to penalize more our model and more coefficient of the variable will reduced to zero, when we increase the value of our r2 square also decreases.

The most important variable after the changes has been implemented for ridge regression are as follows:

- MSZoning_FV
- MSZoning_RL
- Neighborhood_Crawfor
- MSZoning_RH
- MSZoning_RM
- SaleCondition_Partial
- Neighborhood_StoneBr
- GrLivArea
- SaleCondition_Normal
- Exterior1st_BrkFace

The most important variable after the changes has been implemented for lasso regression are as follows:

- GrLivArea
- OverallQual
- OverallCond
- TotalBsmtSF
- BsmtFinSF1
- GarageArea

- Fireplaces
- LotArea
- LotArea
- LotFrontage

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?

Ans2: Though the model performance by Ridge Regression was better in terms of R2 values of Train and Test, it is better to use Lasso, since it brings and assigns a zero value to insignificant features, enabling us to choose the predictive variables. It is always advisable to use simple yet robust model.

Q3: 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?

Ans3: Those 5 most important predictor variables that will be excluded are:

- GrLivArea
- OverallQual
- OverallCond
- TotalBsmtSF
- GarageArea

Q4: 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?

Ans4: A model needs to be made robust and generalizable so that they are not impacted by outliers in the training data. The model should also be generalizable so that the test accuracy is not lesser than the training score. The model should be accurate for datasets other than the ones which were used during training. Too much weightage should not be given to the outliers so that the accuracy predicted by the model is high.

To ensure that this is not the case, the outlier analysis needs to be done and only those which are relevant to the dataset need to be retained. Those outliers which does not make sense to keep must be removed from the dataset. This would help increase the accuracy of the predictions made by the model. Confidence intervals can be used (typically 3-5 standard deviations). This would help standardize the predictions made by the model. If the model is not robust, it cannot be trusted for predictive analysis. It is important to have balance in Bias and Variance to avoid overfitting and under-fitting of data.