Advanced 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?

Answer.

The optimal value of alpha for Ridge regression is: **0.001**

The optimal value of alpha for Lasso regression is: **0.0002**

If we double the value of alpha for the ridge regression, the absolute value of the coefficients will see a slight increase. For lasso regression, a higher alpha value means that a greater number of variables will be dropped. The selected variables dropped from **66** to **59**. In this case the most important predictor variables may also change.

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?

Answer.

The choice between lasso or ridge depends on our business requirements. If you have lots of variables and want to reduce the complexity then use lasso. If we know that the variables are important and you need in your model then use ridge. But it is advisable to use lasso if you have very large dataset. There are also some practical considerations. The ridge is a bit easier to implement and faster to compute, which may matter depending on the type of data you have.

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?

Answer.

The five most important features for lasso regression were:

['GrLivArea', 'OverallQual', '1stFlrSF', 'OverallCond', 'Neighborhood']

After dropping these five variables, the most important new variables are:

['TotalBsmtSF', 'TotRmsAbvGrd', 'LotArea', 'PoolQC', 'GarageCars']

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?

Answer.

Robust means the model works for a broad range of inputs. If the model gets really good results at training time but won't generalize to out-of-sample data (i.e. it isn't robust) then we call it overfitting and it's considered a bad thing exactly because it's not accurate. We can use regularization to make model more robust and generalizable. The accuracy generally takes a hit when we use regularization but at least we know that the model will give improved performance on the test data or never-seen-before data.