Assignment - Advanced Regression

Housing Regression Model using Regularisation

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 for Question 1

1. Optimal value of alpha for ridge regression = 10
2. Optimal value of alpha for lasso regression = 0.0001

After doubling the alpha the top 5 predictors from Lasso Regression are still the same but their coefficient values have changed. RoofMatl is less significant now than OverallQual.

• GrLivArea	(**Above grade (ground) living area square fee**t)
• OverallQual:: Very Excellent	(Very Excellent & Excellent **Rate for the overall material and finish of the house**)
• RoofMatl:: WdShngl	(Wood Shingles for **Roof material**)
• Neighborhood:: NoRidge	(Northridge for **Physical locations within Ames city limits**)
• GarageCars	(**Size of garage in car capacity**)

* Steps performed to answer this is in the Python notebook

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 for Question 2

- I will apply Lasso Regression because it gives less difference when we compare R2, RSS and MSE scores of train and test sets. This means Lasso is more consistent in this case.
- Also, we had lot of variables here. Lasso helps in reducing the number of variables, as it also does feature selection.

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 for Question 3

If we remove the top 5 most important predictor variables in the lasso model the next top five most important predictor variables are:

• 1stFlrSF	(**First Floor square feet**)
• 2ndFlrSF	(**Second floor square feet**)
• GarageCars	(**Size of garage in car capacity**)
• LotArea	(**Lot size in square feet**)
• Exterior2nd:: ImStucc	(Imitation Stucco for **Exterior covering on house**)

* Steps performed to answer this is in the Python notebook

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 for Question 4

- To make sure that our model is robust and generalisable, we compare the difference between R2, RSS and MSE scores of train and test sets.
- If the difference is not very large it means our model is able to generalise and predict on unseen data.
- The accuracy is represented by R2. If R2 from both train and test set is high enough, and the difference is not too much, we can say our model is accurate enough.
- If the R2 score is not too high, it indicates that our model is not overfitted on the train data.