Problem Statement - Part II

Assignment Part-II

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?

Ans – The optimal value of alpha for ridge regression is 1 and for lasso regression is 0.001. If we double the value then it will increase the regularization strength, if we see specifically in ridge, it will cause the coefficients to more shrinkage towards zero. Similarly for lasso regression if we double the value of alpha then it will increase the penalty on the absolute value of coefficient's, which will result in more coefficients being pushed to zero.

Now the most important Predictor variables are shown below:

Snippet from my code:

```
Most important predictor variables after doubling alpha for Ridge regression:
   Predictor_Variable  Coefficient  Absolute_Coefficient
41            TotalSF     0.221440              0.221440
3         OverallQual     0.163105              0.163105
2             LotArea     0.100177              0.100177
4         OverallCond     0.085492              0.085492
42          TotalBath     0.070376              0.070376

Most important predictor variables after doubling alpha for Lasso regression:
   Predictor_Variable  Coefficient
3         OverallQual     0.167032
6        YearRemodAdd     0.034981
8           ExterQual     0.002416
10           BsmtQual     0.005121
```

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?

Ans - Ridge regression has a lower MSE compared to Lasso, indicating that Ridge regression performs better in terms of minimizing prediction errors. Ridge regression has a higher R-squared score compared to Lasso, saying that it explains more variance in the target variable same goes with adjusted R-squared as well. Adding to it, Ridge regression has more interpretability, which can be beneficial in real estate scenarios where stakeholders may want to understand the impact of various features.

Coming to coefficients many coefficients are zero in Lasso regression model, but for our aspect since it is house price prediction those features can play a vital role. Considering the importance of domain knowledge some features may be important but can be ignored by Lasso due to its method of calculating. So, finally after considering all the analysis, Ridge regression will be chosen.

Therefore, for our problem statement like house pricing where understanding the relationship between predictors and the target variable is crucial, Ridge regression offers a good balance between interpretability and predictive performance.

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?

Ans – Now to see which are top most 5 predictor variables now as follows:

Snippet from my code:

```
Top five predictor variables after excluding the original top five:
   Predictor_Variable  Coefficient
20       TotRmsAbvGrd     0.189995
24         GarageArea     0.125655
7           ExterQual     0.082710
9            BsmtQual     0.077361
22        FireplaceQu     0.053887
```

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?

Ans - For a model to be robust and generalizable, there are some steps to be performed. First, splitting the data into train and test datasets for training and testing. Here the quality of the dataset is important as a limited dataset can lead to bad performance. Now performing feature engineering to select relevant features for the model and remove any unnecessary features.

Then use techniques like regularization, which includes Ridge (L2 regularization) and Lasso (L1 regularization), which can prevent overfitting by penalizing the large coefficients, while tuning the hyperparameters using methods like random search or grid search. Lastly, evaluating the model based on a few metrics like R-squared score, F1 score, MSE, P-value, etc.

Lastly, for checking the robustness and generalizability of the model, model diagnosis can be performed where techniques like prediction plot and residual analysis are done to check all the factors.

Finally, by doing all this leads to more accurate predictions on unseen data, reduced overfitting, improved stability, and better adaptation to real-world scenarios.