SUBJECTIVE QUESTIONS

IITB, UPGRAD :

VAMSHI.KRISHNA.PRIME@GMAIL.COM

QUESTION 1

1.1. WHAT IS THE OPTIMAL VALUE OF ALPHA FOR RIDGE AND LASSO REGRESSION?

The optimal value of alpha for lasso and ridge are 0.001 and 4.0 respectively.

1.2. WHAT WILL BE THE CHANGES IN THE MODEL IF YOU CHOOSE DOUBLE THE VALUE OF ALPHA FOR BOTH RIDGE AND LASSO?

If we increase the alpha(hyper parameter value) the accuracy of the model starts dropping gradually. It might increase a bit till the optimal hyper parameter value but the accuracy will dececrease with the increase in alpha and model will become more parse as the strength of regularization applied to the coefficients is increased.

For Ridge Regression: As Ridge regression penalizes the sum of squared coefficients (L2 regularization). Doubling the value of alpha in Ridge regression will increase the penalty for large coefficients, leading to a stronger regularization effect. Which in turn will result in more shrinkage of coefficient values towards zero, reducing the complexity of the model and potentially preventing overfitting.

For Lasso Regression: As Lasso regression penalizes the sum of absolute values of coefficients (L1 regularization). Doubling the value of alpha in Lasso regression will increase the penalty for non-zero coefficients, promoting sparsity in the coefficient matrix. As a result, Lasso regression tends to force more coefficients to exactly zero, effectively performing feature selection. Hence, doubling alpha in Lasso regression will increase the level of feature sparsity and may lead to a simpler model with fewer features.

Overall, increasing the value of alpha for both Ridge and Lasso regression models will strengthen the regularization effect, leading to simpler models with potentially better generalization performance, especially in cases where overfitting is a concern. However, it's essential to choose the optimal value of alpha through techniques like cross-validation to balance between bias and variance in the model.

1.3. WHAT WILL BE THE MOST IMPORTANT PREDICTOR VARIABLES AFTER THE CHANGE IS IMPLEMENTED?

Lasso Regression:
The most important predictor variables after the change are: ['BsmtFinType1', 'RoofStyle_Mansard', 'GarageCond', 'BldgType_2fmCon', 'LandSlope', 'RoofStyle_Hip', 'GarageFinish', 'OverallCond', 'Condition1_RRAn', 'BldgType_TwnhsE']

Ridge Regression:
The most important predictor variables after the change are: ['RoofStyle_Mansard', 'LandSlope', 'GarageCond', 'BldgType_2fmCon', 'BsmtFinType1', 'BsmtCond', 'Condition1_PosN', 'GarageFinish', 'RoofStyle_Hip', 'ExterCond']

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?

After creating model in both Ridge and Lasso regressions, we can see that the R2 scores are similar for both Lasso and Ridge regression. However, as Lasso regression penalize more on the dataset, in-turn helping in more feature elimination, we will consider Lasso Regression model as the final model.

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?

GarageCond, RoofStyle_Mansard, BsmtFinType1, BldgType_2fmCon, LandSlope are the five most important predictor variables in the lasso model. Hence we need to exclde these 5 variables from our model.

The most important predictor variables after removing the five most important predictor variables in the final lasso model are : ['OpenPorchSF', 'RoofMatl_Tar&Grv', 'BsmtFinType2', 'BldgType_Duplex', 'OverallQual', 'RoofStyle_Gable', 'RoofMatl_Roll', 'Condition1_PosN', 'GarageArea', 'MasVnrArea']

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?

The model should be as simple as possible. Even though its accuracy might fall significantly, as a trade off, it will be more robust and generalisable. In layman's terms, the simpler the model the more the bias but less variance and more generalizable. This concept is explained in the Bias-Variance trade-off, where simplicity correlates with higher bias but lower variance, ultimately yielding greater generalizability.

The practical implication of this trade-off on accuracy manifests in the model's consistent performance across both training and test datasets. A robust and generalizable model demonstrates minimal deviation in accuracy between these datasets, signifying its capacity to maintain reliable performance levels regardless of data variation.

NOTE: Please refer the Jupyter Notebook for the detailed step by step approach on determining above solutions at the end of the analysis.