ADVANCED REGRESSION – SUBJECTIVE QUESTIONS

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?

- Optimal value of lambda for Ridge Regression = 0.4

- Optimal value of lambda for Lasso = 0.0001

For Ridge Regression:

- We can see the the R2 score for Train has changed from 92.37 to 91.67.

- R2 score of Test changed from 88.4 to 88.22

For Lasso Regression:

- We can see the the R2 score for Train has changed from 92.47 to 91.62.

- R2 score of Test changed from 88.65 to 88.3.

Top 5 predictor variables:

1. GrLivArea
2. TotalSF
3. TotalBsmtSF
4. OverallQual_9
5. OverallQual_8

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?

R2 scores of Lasso regression is slightly higher than that of Ridge regression. Additionally, our main intention is to find out the major features or variables effecting the sales price. Lasso regression helps in this by performing feature selection process.

Hence Lasso will be the choice considered.

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?

Top 5 features after excluding the GrLivArea, OverallQual_9, OverallQual_8, LotArea, TotalBsmtArea are:

I. TotalSF
II. Neighborhood_Crawfor
III. SaleCondition_Partial
IV. BsmtExposure_Gd
V. OverallCond_9

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?

Ensuring that a model is robust and generalisable involves multiple steps during the model development process.

Some of which are as below:

1. Suffucient and Relevant Data: In order to build a good model, we need to have enough data. This ensures that the we have adequte data for training and testing. Relevant data is also necessary to build an appropriate solution for the problem statement.
2. Data Preprocessing: This is one of the most important steps in the model building that affects the model accuracy. This involves imputing the null values, changing the datatypes, removing the outliers, etc.
3. Feature Engineering and Selection: Choosing features that are relevant to the problem and have predictive power in another step. Additionally, we can consider techniques such as regularization to prevent overfitting and improve generalization.
4. Model Evaluation Metrics: Choosing appropriate evaluation metrics that reflect the model performance on unseen data is also significant. Metrics such as accuracy, precision, recall, etc are commonly used for classification tasks, while metrics such as R-squared, root mean squared error (RMSE), and mean absolute error (MAE) are used for regression problems.
5. Hyperparameter Tuning: Optimize the model's hyperparameters using techniques such as grid search or randomized search to find the best combination of hyperparameters that maximize performance on unseen data.

A model that is robust and generalisable is more likely to perform well on unseen data, leading to more accurate predictions in real-world scenarios. While a model that is overfitted to the training data or lacks generalisation may perform well on the training data but poorly on unseen data, resulting in lower accuracy and reliability in practical applications.A robust and generalisable model is essential for making accurate predictions in real-world scenarios and minimising the risk of performance degradation when deployed in production**.**