ADVANCE REGRESSION

PROBLEM STATEMENT – 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?

Answer - The optimal value of alpha for Ridge regression is found to be 1, while for Lasso regression, it is 10. Upon doubling the value of alpha for both Ridge and Lasso regression, we observed changes in the models' performance and the importance of predictor variables:

1**. Ridge Regression:**

- Original Alpha: 1

- Doubled Alpha: 3

- R2 Score (Train/Test):

  - Original: 0.8843 / 0.8696

  - Doubled: 0.8797 / 0.8710

- Changes:

  - The R2 score decreased slightly on the training data but increased on the testing data, indicating potential improvement in generalization

  - LotArea, OverallQual, OverallCond, YearBuilt, BsmtFinSF1, TotalBsmtSF, GrLivArea, TotRmsAbvGrd, Street_Pave, RoofMatl_Metal are important predictor variables.

2. **Lasso Regression:**

- Original Alpha: 10

- Doubled Alpha: 20

- R2 Score (Train/Test):

  - Original: 0.8859 / 0.8647

  - Doubled: 0.8854 / 0.8670

- Changes:

  - The R2 score decreased slightly on the training data and increased marginally on the testing data.

  - This suggests that the model's generalization performance remained relatively stable.

- LotArea, OverallQual, OverallCond, YearBuilt, BsmtFinSF1, TotalBsmtSF, GrLivArea, TotRmsAbvGrd, Street_Pave, RoofMatl_Metal are important predictor variables.

In summary, doubling the value of alpha for both Ridge and Lasso regression leads to adjustments in the models' performance and the relative importance of predictor variables. While the exact changes may vary, the models generally become more parsimonious, potentially improving their generalization ability and interpretability.

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

I would choose Lasso regression for its ability to perform feature selection by driving certain coefficients to exactly zero, effectively removing irrelevant predictors from the model. This feature is particularly beneficial when dealing with datasets containing a large number of predictors, as it helps in simplifying the model and improving its interpretability. Additionally, Lasso regression can handle multicollinearity more effectively than Ridge regression by selecting one variable from a group of highly correlated predictors. Therefore, Lasso regression not only offers regularization but also aids in feature selection, making it a preferred choice in many practical scenarios.

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

After excluding the five unavailable predictors – LotArea, OverallQual, YearBuilt, BsmtFinSF1, and TotalBsmtSF – from the dataset, the top five important predictors in the retrained Lasso model (Lasso21) are as follows:

1. 1stFlrSF: First Floor square feet

2. GrLivArea: Above grade (ground) living area square feet

3. Street_Pave: Pave road access to property

4. RoofMatl_Metal: Roof material_Metal

5. RoofStyle_Shed: Type of roof (Shed)

The Lasso model's coefficients indicate that these predictors have the most significant impact on the target variable after the exclusion. Notably, the model's performance slightly decreased, as reflected in the decreased R2 scores on both training and testing data, indicating the importance of the excluded predictors in the original model. (refer python notebook for the detailed code solution)

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

To ensure model robustness and generalizability, implement cross-validation, feature selection, hyperparameter tuning, validation on diverse datasets, and ensemble methods. These practices help assess the model's performance across different data subsets, select relevant features, optimize hyperparameters, validate on diverse datasets, and combine multiple models for improved performance. While prioritizing robustness may slightly reduce accuracy on the training set due to regularization, it enhances performance on unseen data, making the model more reliable in real-world scenarios. This balance ensures stable and consistent performance across various datasets, ultimately enhancing the model's predictive power and utility.
