Subjective Assignment – Advanced Regression

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:

Optimal Alpha Values:

In the final model, the determined optimal value of alpha for Ridge regression was 15, while for Lasso regression, it stood at 100.

Doubling Alpha Impact:

Upon doubling these alpha values for both Ridge and Lasso regressions, the model's overall performance remained unchanged. This indicates that increasing the regularization strength by doubling alpha didn't significantly alter the predictive power of the models.

Changes in Important Predictor Variables after implementing the changes:

Ridge Regression:

The important predictor variables that emerged were Neighborhood_StoneBr, GarageArea, Neighborhood_NridgHt, TotalBsmtSF, GrLivArea, KitchenQual, Neighborhood_Names, Neighborhood_Edwards, BldgType_TwnhsE, and GarageFinish. These variables retained their significance even with the doubled alpha.

Lasso Regression:

The crucial predictor variables included YearBuilt, YearRemodAdd, GarageArea, ScreenPorch, EnclosedPorch, GrLivArea, WoodDeckSF, TotalBsmtSF, and 2ndFlrSF. Similar to Ridge, these variables remained important despite the increased penalty imposed by doubling alpha.

The stability of important predictor variables in both Ridge and Lasso regressions post-doubling alpha suggests that these specific features consistently contribute significantly to the models, regardless of higher regularization.

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:

Upon evaluating both Ridge and Lasso regression models, it was noted that they exhibited quite similar performance characteristics. While assessing their performance on the test dataset, it was observed that the Ridge regression model marginally outperformed Lasso by approximately 1.21%.

Despite the slightly better performance of the Ridge model on the test dataset,my decision was to choose he Lasso regression model in the final analysis.The crucial factor influencing this decision was Lasso's inherent capability for feature selection and elimination. Given the dataset's extensive nature with over more than 200 columns, Lasso's ability to zero out or discard less impactful features provides a substantial advantage.

Handling High Dimensionality: Dealing with a dataset encompassing numerous columns, Lasso's feature elimination property becomes particularly advantageous. It aids in identifying and emphasizing the most pertinent predictor variables while discarding less influential ones. This feature selection process simplifies the model and can enhance its interpretability.

Selected Model: The final chosen model for application is the Lasso regression.

Model Performance: The Lasso regression model achieved an R-squared score of 80 on the training dataset and 71 on the test dataset, respectively.

The decision to proceed with the Lasso regression model despite the slight test dataset performance difference in favor of Ridge was primarily driven by Lasso's ability to facilitate feature elimination. In dealing with a high-dimensional dataset, where feature reduction and identification of the most impactful predictors are crucial, Lasso's feature selection capability was deemed advantageous, leading to its selection as the final model for application.

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:

Initially Important Predictor Variables in Lasso Model:

The five most crucial predictor variables identified in our Lasso model were:

GrLivArea: Represents the above-grade living area in square feet, suggesting that larger living spaces contribute significantly to higher property sale values.

YearBuilt: Signifies the original construction date of the property.

YearRemodAdd: Reflects the remodeling date, indicating property modifications or additions, if any.

TotalBsmtSF: Represents the total square footage of the basement area, indicating a positive influence on sale prices with larger basement spaces.

GarageArea: Indicates the size of the garage in square feet, with larger garage spaces contributing positively to property values.

Considering the unavailability of the initially identified top five important predictor variables, the revised set of five most crucial predictors is as follows WoodDeckSF,2ndFlrSF,EnclosedPorch, TotalBsmtSF and GarageArea.

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:

Ensuring Robustness and Generalisability of a Model:

Avoiding Overfitting: Overfitting occurs when a model learns from noise or specific patterns in the training data to such an extent that it performs exceptionally well on the training set but poorly on new, unseen data. By preventing overfitting, a model can generalize better to unseen data.

Model Simplicity: Keeping the model as simple as possible while still capturing the underlying patterns in the data contributes to its robustness. A simpler model with fewer complex relationships is more likely to generalize well to new data.

Implications on Model Accuracy:

Overfitting a model can lead to higher accuracy on the training dataset as it memorizes the noise or intricacies specific to that dataset. However, this doesn't guarantee its performance on new data. A highly overfitted model might exhibit significantly reduced accuracy when presented with unseen data.

A robust and generalizable model might not achieve the highest possible accuracy on the training dataset. However, it's more likely to maintain a reasonably good level of accuracy on both the training and testing datasets. This balance ensures that the model is not excessively tailored to the training data and can be applied effectively to new, real-world scenarios.