Advanced Regression Assignment-II

Subjective Questions:

Q1: What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

A1. Optimal value of alpha for ridge and lasso regression are:

- Optimal value of lambda for Ridge Regression: 10
- Optimal value of lambda for Lasso Regression: 0.001

If we double the alpha values:

- Ridge: Increasing alpha will decrease coefficients, potentially making some less influential.
- Lasso: As alpha increases, more coefficients may become zero, emphasizing the most important predictors.

For all the models, the training score has decreased slightly, but the testing score has increased significantly.

The most important predictor variables post-change are those with significant non-zero coefficients. i.e., 'GrLivArea', 'OverallQual', 'TotalBsmtSF', 'OverallCond', are the most important predictor variables

Regularization with a higher alpha helps to emphasise the most influential features while potentially avoiding overfitting.

Q.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?

A.2. The optimal values of lambda for Ridge and Lasso are determined as follows:

- Optimal value of lambda for Ridge: 10
- Optimal value of lambda for Lasso: 0.001

Model performance scores:

- Ridge: Train = 90.9, Test = 87.4
- Lasso: Train = 89.8, Test = 86.4

The decision to proceed with Lasso over Ridge is based on the good scores obtained for both models. Lasso gives model parameters in which less significant feature coefficients are absolutely zero, resulting in a kind of feature selection.

As a result, Lasso Regression has been chosen for this specific circumstance.

This selection is determined by the model's interpretability and sparsity.

Q.3. After building the model, you realized 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?

A.3. The initial top 5 predictors were determined as OverallQual_9, GrLivArea, OverallQual_8, Neighborhood_Crawfor, and Exterior1st_BrkFace. These features were dropped from the training and test datasets.

Later, a new Lasso regression model was built using cross-validation to find the optimum alpha value, resulting in alpha = 0.001. The model's performance metrics, such as R-squared, RSS, MSE, and RMSE, were evaluated.

After regularization, the coefficients of the independent variables were examined, and the top 5 predictors were identified as 2ndFlrSF, Functional_Typ, 1stFlrSF, MSSubClass_70, and Neighborhood_Somerst.

These predictors are considered to be the most important in predicting home values.

Q.4. How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

A.4.

- Ensuring a model's robustness and generalizability is crucial for reliable predictions. Robustness, or resistance to data variations, ensures stable performance, while generalizability allows proper adaptation to new, unseen data.
- Avoiding overfitting is key, as it arises from excessive model complexity, leading to high variance and a lack of adaptability to new patterns in test data.
- The model should strike a balance between complexity and accuracy, as overly complex models, while achieving high accuracy on training data, may struggle to generalize. This trade-off involves reducing variance to enhance generalizability, even if it introduces some bias.
- Regularization techniques, such as Ridge Regression and Lasso, offer systematic approaches to manage this complexity and strike the right balance.