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:

For Ridge Model:

For the given data set and after processed the data set with various data cleaning mechanisms, the optimal value of alpha achieved for Ridge model is 8.0 and this alpha value resulted into current ridge model with below metrics

R2 score (training data) : 0.85

R2 score (test data) : 0.84

RMSE (training data) : 0.39

RMSE (test data) : 0.40

With the current ridge model, the top predictor variables in the ridge model:

Neighborhood_NoRidge: 0.52

When the optimal value of alpha is doubled and the ridge model is re-built and evaluated, we observed below metrics

R2 score (training data) : 0.84

R2 score (test data) : 0.84

RMSE (training data) : 0.39

RMSE (test data) : 0.41

For Lasso Model:

For the given data set and after processed the data set with various data cleaning mechanisms, the optimal value of Alpha for Lasso is 0.001 and this resulted into lasso model with below metrics

R2 score (training data) : 0.85

R2 score (test data) : 0.84

RMSE (training data) : 0.38

RMSE (test data) : 0.40

With the current lasso model, the top predictor variables in the initial Lasso model:

Neighborhood_NoRidge: 0.67

When the optimal value of alpha is doubled and the lasso model is re-built and evaluated, we observed below metrics

R2 score (training data) : 0.85

R2 score (test data) : 0.83

RMSE (train data) : 0.39

RMSE (test data) : 0.41

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:

When comparing the Ridge and Lasso models, the R2 score for the training set is slightly higher for the Lasso model compared to the Ridge model. The R2 score represents the proportion of the variance in the dependent variable that is predictable from the independent variables. Given that both models have determined the optimal value of lambda, I would choose the Lasso model because it provides a slightly better fit to the data based on the R2 score and RMSE values. Also, the lasso model is quite simpler in terms of complexity as Lasso also does elimination of non-significant features.

Here's a summary of the comparison:

Ridge model R2 score (training): 0.850975850819784

Lasso model R2 score (training): 0.8524719647682271

Ridge model RMSE (train) : 0.38603646094665206

Lasso model RMSE (train) : 0.3840937844222072

Question 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?

Answer:

Top 5 most important predictor variables in the initial Lasso model:

Neighborhood_NoRidge: 0.6690544166467738

Neighborhood_NridgHt: 0.6105164809085014

2ndFlrSF: 0.37353198589266806

BldgType_Twnhs: -0.3339088014999202

Neighborhood_Somerst: 0.2953973727254121

Now, dropped off the above high significant predictor variables and re-built the lasso model. These are the top 5 high coefficient values.

Top 5 most important predictor variables in the new Lasso model:

OverallQual: 0.353154203253468

1stFlrSF: 0.1730974471335822

GarageArea: 0.08961515649491918

KitchenQual: 0.08116533812372087

GarageFinish: 0.026472489622987802

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:

Below are the important aspects of considerations to build a robust and generalizable model.

1) Data Preprocessing and Feature Engineering:

Handle missing values. Common methods include mean/median imputation or removing non-significant columns. Drop the highly correlated data points. Feature importance scores can help identify the most impactful ones.

Transformation: Consider transforming features (e.g., log transformation) to improve linearity or normality of residuals. Recursive feature elimination (RFE) to identify the most important features. Removing irrelevant features helps in simplifying the model, reducing overfitting, and improving generalization.

Feature Scaling: Scale the features if they are on different scales. Standardization or normalization can help the model converge faster and can make the coefficients more comparable.

2) Cross-Validation:

Implement cross-validation techniques such as k-fold cross-validation, which further helps in evaluating the model's performance and robustness by training and testing the model on different subsets of the data.

3) Regularization:

Apply regularization techniques like Ridge Regression (L2 regularization) or Lasso Regression (L1 regularization) to prevent overfitting by penalizing large coefficients. This helps generalizing the model.

4) Check for Multicollinearity:

Detect and handle multicollinearity among the predictor variables. Multicollinearity can lead to unstable estimates of regression coefficients and affect the model's generalizability. Techniques like Variance Inflation Factor (VIF) can be used to identify multicollinearity.

5) Residual Analysis:

Analyze the residuals to check if the assumptions of linear regression are met. Residual plots can help identify if there is any pattern left in the residuals, which might suggest a need for a more complex model.

6) Evaluate Performance Metrics:

Use appropriate performance metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), or R-squared to evaluate the model's performance on both the training and testing datasets.

Impact on Accuracy:

By ensuring robustness and generalizability, we are making the model less susceptible to biases and errors in the training data. It is a good tradeoff between bias and variance which optimizes the accuracy. This leads to several benefits for accuracy:

Robustness techniques prevent overfitting by making the model less sensitive to specific data points. A generalizable model performs well on new data, reflecting a truer representation of the underlying relationship. This translates to more accurate predictions on unseen data points.

In essence, robustness and generalizability lead to a model that is more reliable and trustworthy in its predictions. It captures the actual relationship between variables instead of just memorizing specific features of the training data.