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

After processing the dataset with various data cleaning mechanisms, the optimal value of alpha achieved for the Ridge model is 6.0. This alpha value resulted in the current Ridge model with the following metrics:

- R2 score (training data): 0.85
- R2 score (test data): 0.84
- RMSE (training data): 0.37
- RMSE (test data): 0.39

In the current Ridge model, the top predictor variable is Neighborhood_NoRidge with a coefficient of 0.56.

When the optimal value of alpha is doubled and the Ridge model is re-built and evaluated, the following metrics are observed:

- R2 score (training data): 0.85
- R2 score (test data): 0.84
- RMSE (training data): 0.38
- RMSE (test data): 0.40

Despite doubling the optimal alpha value, there is only a slight change in the RMSE metrics, indicating robustness to changes in regularisation strength.

After processing the dataset with various data cleaning mechanisms, the optimal value of alpha for the Lasso model is found to be 0.001. This alpha value resulted in the Lasso model with the following performance metrics:

- R2 score (training data): 0.85
- R2 score (test data): 0.84
- RMSE (training data): 0.38
- RMSE (test data): 0.39

In the initial Lasso model, the top predictor variable is Neighborhood_NoRidge with a coefficient of 0.67.
When the optimal value of alpha is doubled and the Lasso model is re-built and evaluated, the following metrics are observed:

- R2 score (training data): 0.85
- R2 score (test data): 0.84
- RMSE (training data): 0.38
- RMSE (test data): 0.40

Despite doubling the optimal alpha value, there is a no change in the R2 score for the test data and a slight increase in the RMSE for test data. This indicates that the model's performance is relatively stable with respect to changes in the regularisation strength.

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 contrasting the Ridge and Lasso models, the Lasso model demonstrates a slightly higher R2 score for the training set compared to the Ridge model. The R2 score signifies the proportion of variance in the dependent variable that is explainable by the independent variables. Since both models have identified the optimal lambda value, opting for the Lasso model seems preferable due to its marginally better fit to the data, as indicated by the R2 score and RMSE values. Moreover, the Lasso model offers simplicity in terms of complexity, thanks to its feature selection capability, which eliminates non-significant features.

Here's a summary of the comparison:

- Ridge model R2 score (training): 0.856012234589521
- Lasso model R2 score (training): 0.8561282143975366

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

Top 5 most important predictor variables in the initial Lasso model:
Neighborhood_NoRidge: 0.6789563652984398
Neighborhood_NridgHt: 0.6290938522176681
2ndFlrSF: 0.38455487760300516
BldgType_Twnhs: -0.3539277561187091
Neighborhood_Somerst: 0.32083001129642535
Dropped top 5 predictors and rebuild model
Top 5 most important predictor variables in the new Lasso model:
HouseStyle_2Story: 0.3100357843875774
1stFlrSF: 0.2970814060440534
BldgType_TwnhsE: -0.2628779803113354
LandContour_Low: 0.2330225213946334 8
OverallQual: 0.2307903034892456 2

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 that a model is robust and generalizable involves several key steps and considerations:

1. Cross-Validation: Utilize techniques such as k-fold cross-validation to assess the model's performance on multiple subsets of the data. This helps to evaluate the model's stability and generalization across different data partitions.
2. Train-Test Split: Divide the dataset into separate training and testing sets. Train the model on the training set and evaluate its performance on the unseen testing set. This helps to simulate real-world scenarios where the model encounters new data.
3. Validation Set: Apart from the training and testing sets, use a validation set to fine-tune model hyperparameters and assess its performance during training. This helps to prevent overfitting to the training data.
4. Regularization: Employ regularization techniques such as Ridge and Lasso regression to prevent overfitting and improve the model's generalization ability.
5. Feature Engineering: Carefully select and engineer features that are relevant and informative for the

prediction task. Avoid overfitting by not including irrelevant or redundant features.

6. Bias-Variance Tradeoff: Strive to strike a balance between bias and variance in the model. A model with high bias may underfit the data, while a model with high variance may overfit the data. Aim for a model that generalises well to unseen data while capturing the underlying patterns in the data.
7. Evaluate Performance Metrics: Assess the model's performance using appropriate evaluation metrics such as R-squared, RMSE, accuracy, precision, recall, etc. These metrics provide insights into how well the model is performing and its ability to generalise to new data.

The implications of having a robust and generalizable model for its accuracy are significant. A robust and generalizable model is more likely to perform well on unseen data, making it more reliable and trustworthy for real-world applications. It reduces the risk of overfitting to the training data and provides more accurate predictions for new instances. Additionally, a generalizable model ensures that the insights and conclusions drawn from the model are applicable across different datasets and scenarios, enhancing its utility and value. Overall, prioritising robustness and generalisation leads to more accurate and reliable machine learning models.