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 for Ridge and Lasso Regression

- The optimal alpha for Ridge and Lasso regression is typically determined through a process known as cross-validation.

Effect of Doubling Alpha

- If you double the alpha value, the model becomes more regularized and potentially simpler.
- The model relies less on the input features, which could lead to an increase in bias and a decrease in the model's R2 score.

Important Predictor Variables After Change

- The most important predictor variables after the change would depend on the specific dataset and the relationships between the variables.
- It's recommended to re-evaluate the model after changing the alpha value.

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 :

- **Occam's Razor Principle:** This principal advocates for simplicity in models. The complexity of a model is usually determined by the number of features or independent variables and the magnitude of the beta coefficients. Techniques like Ridge and Lasso help in reducing this complexity by shrinking the beta coefficients towards zero.
- **Comparison of Lasso and Ridge:** Both Lasso and Ridge models yield similar r2 scores and Mean Absolute Error (MAE) on the test dataset. However, the Lasso model is simpler as it has eliminated a significant number of features, while the Ridge model retains all original features. Despite this, the Lasso model maintains similar r2 score and MAE.
- **Performance Metrics:** The performance of both the Ridge and Lasso models on the test dataset is characterized by specific values of r2 score, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and MAE.

RIDGE:

r2 score on training dataset: 0.9464358813869536

MSE on training dataset: 0.008784000624663424

RMSE on training dataset: 0.09372299944337795

MAE on training dataset: 0.06288727925962699

LASSO:

r2 score on training dataset: 0.9460236043786002

MSE on training dataset: 0.008851610091461008

RMSE on training dataset: 0.09408299576151372

MAE on training dataset: 0.0625552592635115

- **Final Model Selection:** Given the similar performance of these two models on the test dataset, the principle of Occam's Razor suggests choosing the simpler model. Therefore, the Lasso model is selected as the final model due to its simplicity and comparable performance.

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 :

The most significant predictor variables can be identified as follows:

Initial Top 5 Features in Lasso Model:

- GrLivArea: 0.339067
- OverallQual: 0.311471
- GarageCars: 0.188410
- OverallCond: 0.156641
- Neighborhood_StoneBr: 0.132668

Given that Neighborhood_StoneBr is a dummy variable, we decided to drop the entire Neighborhood feature. After dropping GrLivArea, OverallQual, OverallCond, GarageArea, and Neighborhood features, we rebuilt the Lasso model with the remaining features.

Top 5 Predictor Variables After Rebuilding the Model:

- 1stFlrSF: 0.340473
- 2ndFlrSF: 0.303063
- GarageCars: 0.223494
- Exterior1st_BrkFace: 0.131593
- YearRemodAdd: 0.13097

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 :

Model Complexity and Generalization

A model should strike a balance in complexity - it should be complex enough to learn the patterns in the training dataset, but not so complex that it learns the noise or memorizes every data point.

Underfitting and Overfitting

An underfitting model, characterized by high bias and low variance, fails to capture the data pattern in the training dataset, resulting in poor performance on both the training and testing datasets. On the other hand, an overfitting model, characterized by low bias and high variance, performs well on the training dataset but poorly on the testing dataset or unseen data.

Identifying Overfitting

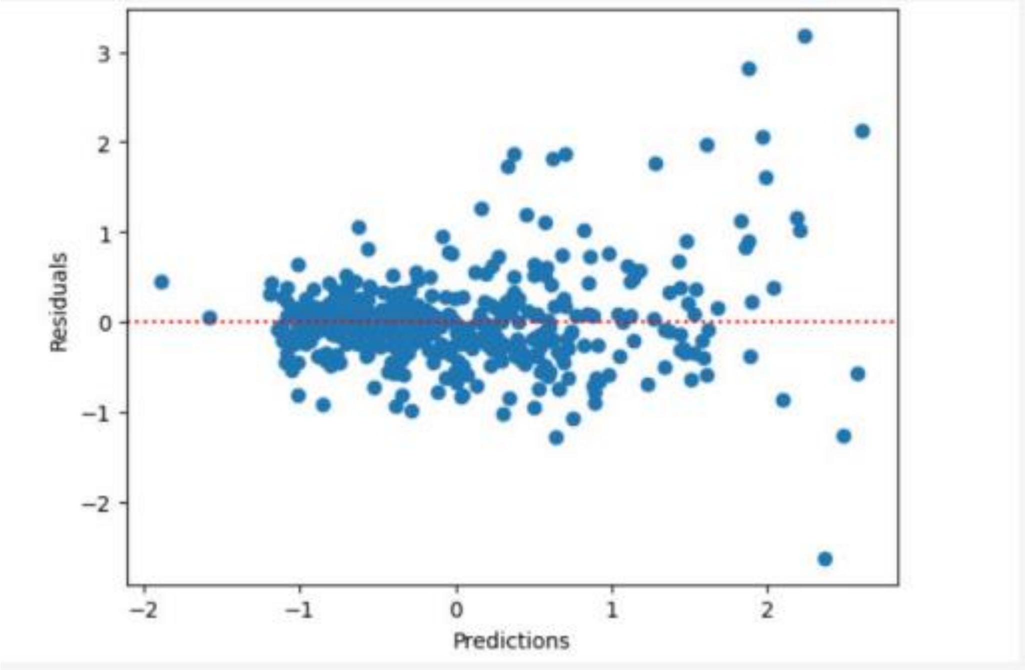
Overfitting can be identified by comparing model performance on training and testing datasets. If there's a significant difference in performance metrics (like r2 score, model accuracy, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Confusion Matrix, etc.) on these datasets, it's likely a case of overfitting.

Creating a Robust Model

A robust model should have low bias and low variance, avoiding both underfitting and overfitting. This can be achieved by striking a balance between bias and variance. One way to mitigate overfitting and create a robust, generalizable model is to reduce model complexity.

Case Study: CV and Lasso

In the given case study, we used Cross-Validation (CV) and Lasso to identify the optimal parameters using an appropriate value of alpha. The Scatter plot shows that the residuals are evenly distributed across the X-axis with no apparent trend, and the Distribution plot (Distplot) shows a normal (0) distribution.


Residuals
Predictions
3
2
1
0
−1
−2
−2
−1
0
1
2