Assignment - Surprise Housing

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 Value of Alpha for Ridge and Lasso Regression

- **Ridge Regression:** Optimal alpha (lambda) value is 10.
- **Lasso Regression:** Optimal alpha (lambda) value is 0.001.

Effect of Doubling the Alpha Value

Doubling the alpha value for both Ridge and Lasso regression will have the following effects:

Ridge Regression:

- **Increased Regularization:** The coefficients will be further reduced, increasing the model's bias but reducing its variance. This can lead to a simpler model, potentially resulting in underfitting.
- **Smaller Coefficients:** All coefficients will decrease in magnitude, but none will be zero, as Ridge regression does not eliminate features.

Lasso Regression:

- **Increased Regularization and Feature Selection:** More coefficients will be reduced to zero, effectively eliminating more features from the model. This increases bias but reduces variance, enhancing model interpretability.
- **Feature Exclusion:** Less important features are more likely to have their coefficients set to zero, simplifying the model further.

Most Important Predictor Variables After Doubling Alpha

- **Ridge Regression:** The most important predictor variables will still be those with the largest coefficients, although their absolute values will decrease due to stronger regularization.
- **Lasso Regression:** The most important predictor variables will be those that retain non-zero coefficients. Due to increased regularization, some previously significant variables may now have zero coefficients. The remaining non-zero coefficients will indicate the most significant predictors.

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

Optimal Value of Alpha for Ridge and Lasso Regression:

- **Ridge Regression:** Optimal lambda value is 10.
- **Lasso Regression:** Optimal lambda value is 0.001.

Model Performance:

- **Ridge Regression:**
  - Training Score: 90.9
  - Testing Score: 87.4
- **Lasso Regression:**
  - Training Score: 89.8
  - Testing Score: 86.4

Choice and Justification:

Given that both models yield good performance scores, I would choose to apply **Lasso Regression**. The reasons for this choice are:

1. **Feature Selection:** Lasso regression results in some coefficients being exactly zero, effectively performing feature selection. This simplifies the model by removing less important features, making it easier to interpret and potentially improving the generalizability of the model.

2. **Model Interpretability:** By reducing the number of features, Lasso regression makes it clearer which variables are the most significant predictors. This can provide more straightforward insights for decision-making.
3. **Performance:** Although both models have similar performance, the slight difference in scores (Ridge: 87.4, Lasso: 86.4) is negligible. The benefits of feature selection and interpretability with Lasso regression outweigh this minor difference.

Therefore, Lasso regression is preferred for its ability to produce a more parsimonious and interpretable model without significantly compromising 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

Upon re-running the model after removing the top 5 significant variables, the next five significant predictor variables identified are:

Lasso Regression:

1. **GarageType_BuiltIn**: 0.089
2. **GarageType_Detchd**: 0.094
3. **GarageType_No Garage**: 0.101
4. **GarageType_Others**: 0.12
5. **GarageFinish_No Garage**: 0.195

Ridge Regression:

1. **GarageType_BuiltIn**: 0.089
2. **GarageType_Detchd**: 0.093
3. **GarageType_No Garage**: 0.096

4. **GarageType_Others**: 0.103
5. **GarageFinish_No Garage**: 0.14

These variables have emerged as the most significant predictors in the model after excluding the initial top five important variables. Both Ridge and Lasso regression show a consistent set of significant variables, emphasizing different aspects of the garage features in predicting house prices.

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

Changes to Improve Model Robustness and Generalizability:

1. **Use a Model That's Resistant to Outliers:**
    - **Tree-Based Models:** Tree-based models like Random Forests and Gradient Boosting are generally less affected by outliers compared to linear models.
    - **Non-Parametric Tests:** If performing statistical tests, use non-parametric tests which do not assume a specific distribution of data.
2. **Use a More Robust Error Metric:**
    - **Mean Absolute Error (MAE):** Switch from mean squared error to mean absolute error to reduce the influence of outliers.
    - **Huber Loss:** Use Huber loss, which is less sensitive to outliers than mean squared error but more sensitive than mean absolute error, offering a balance.

Changes to Improve Data Quality:

1. **Winsorize Your Data:**
    - **Capping Data:** Artificially cap extreme values to reduce the impact of outliers. This involves setting limits on the values in the dataset, effectively reducing the influence of extreme values.
2. **Transform Your Data:**

- **Log Transformation:** Apply log transformation to skewed data to make it more normally distributed. This is especially useful for data with a pronounced right tail.

3. **Remove Outliers:**

   - **Identify and Remove Outliers:** If there are very few outliers, and they are likely anomalies, removing them can help improve model performance and accuracy.

Implications for Model Accuracy:

1. **Increased Bias but Reduced Variance:**

   - Implementing these changes may increase the model's bias but reduce its variance. This means the model will be less sensitive to fluctuations in the training data, leading to better generalization on unseen data.

2. **Improved Generalizability:**

   - By making the model more robust to outliers and using more robust error metrics, the model is likely to perform better on new, unseen data, which is crucial for real-world applications.

3. **Balanced Trade-Off:**

   - There is always a trade-off between bias and variance. By reducing variance and improving robustness, you may slightly compromise on accuracy on the training data, but this typically results in better performance on test data and in real-world scenarios.

In summary, ensuring a model is robust and generalizable involves using models and metrics that are less sensitive to outliers, as well as preprocessing the data to reduce the impact of extreme values. These practices help create a model that performs consistently well on new data, making it more reliable for practical applications.