For Question 1 and Question 3 work please refer Jupyter Notebook

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:

- **Ridge Regression**:
  - Optimal value of (\lambda): `10`
- **Lasso Regression**:

Optimal value of (\lambda): `0.001`

Performance Metrics Update

Changes in Ridge Regression metrics:

- **R2 Score**:
  - **Train Set**:
    - Previous: `0.94`
    - Updated: `0.93`
  - **Test Set**:
    - Remained consistent at: `0.93`

Changes in Lasso Regression metrics:

- **R2 Score**:
  - **Train Set**:
    - Previous: `0.92`
    - Updated: `0.91`
  - **Test Set**:
    - Previous: `0.93`
    - Updated: `0.91`

Notes: The R2 score, or the coefficient of determination, is a statistical measure that indicates the proportion of the variance for the dependent variable that is explained by independent variables in a regression model. A decrease in the R2 score might suggest that the model's predictive capability has reduced slightly. However, small changes may not always indicate significant degradation, and it's essential to consider other evaluation metrics and domain-specific context.

Important Predictor Variables

After doubling the alpha (or (\lambda)) values in our regularization techniques, the most impactful predictor variables in our model are:

1. **`GrLivArea`**

   - **Description**: Represents the above grade (ground) living area in square feet.
   - **Relevance**: The total living area often has a strong correlation with the price of a house.

2. **`OverallQual_8`**

   - **Description**: Represents a rating of the overall material and finish of the house, with a score of 8.
   - **Relevance**: Houses with higher material quality and finish tend to fetch higher prices.

3. **`OverallQual_9`**

   - **Description**: Similar to `OverallQual_8`, but with a score of 9.
   - **Relevance**: Higher scores generally indicate better overall quality, which can increase the house value.

4. **`Functional_Typ`**

   - **Description**: Represents the home's functionality, where "Typ" indicates typical functionality.
   - **Relevance**: Homes with better functionality generally appeal more to potential buyers.

5. **`Neighborhood_Crawfor`**

   - **Description**: Indicates that the house is located in the Crawford neighborhood.
   - **Relevance**: Location plays a vital role in house pricing; certain neighborhoods might have higher desirability.

6. **`Exterior1st_BrkFace`**

   - **Description**: Represents that the exterior covering of the house is brick face.
   - **Relevance**: Certain exterior materials can enhance the aesthetic and durability of the home, influencing its price.

7. **`TotalBsmtSF`**

   - **Description**: Total square feet of the basement area.
   - **Relevance**: A larger basement can provide more functional space, potentially increasing the house's value.

8. **`CentralAir_Y`**

   - **Description**: Indicates that the house has central air conditioning.
   - **Relevance**: Central air conditioning can significantly increase the comfort level of a home, making it more appealing to potential buyers.

Conclusion: These predictor variables are critical when estimating house prices. By adjusting the regularization parameter (alpha or (\lambda)), we have identified that these predictors play a substantial role in the model's performance.

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:

he decision to apply either Ridge or Lasso regression hinges upon the specific requirements and goals of our modeling process:

1. Lasso Regression:

- **Ideal For**:
  - Feature selection when faced with a high number of predictors.
  - Simplifying models to make them more interpretable.

Detailed Benefits:

- **Feature Reduction**: Lasso can reduce some coefficients to zero, effectively excluding irrelevant features.
- **Model Simplicity**: Models with fewer features are often easier to interpret and understand.

When to Choose:

- If the primary objective is to pinpoint and select only the most relevant features from a vast set, Lasso is the way to go.

2. Ridge Regression:

- **Ideal For**:
  - Managing multicollinearity in data.
  - Controlling the magnitude of coefficients without entirely excluding any.

Detailed Benefits:

- **Coefficient Shrinkage**: Ridge Regression shrinks the magnitude of coefficients, thereby ensuring they don't become excessively large.
- **Handling Correlated Features**: Ridge performs stably even when predictors are highly correlated.

When to Choose:

- If the model's main concern is to prevent coefficients from growing too large, which could lead to overfitting, Ridge Regression is a suitable choice.

Conclusion:

The choice between Ridge and Lasso Regression must align with the specific needs of the model and the overarching objectives of the analysis. Both techniques have their strengths, and understanding the nuances of each ensures a more informed modeling decision.

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

Based on our analysis, we have decided to refine our Lasso model. The approach we'll take is outlined below:

Action Plan:

- **Feature Omission**: We will remove the top 5 predictors identified by the Lasso model.
- **Model Rebuilding**: After removing these predictors, the model will be rebuilt to assess its performance without the influence of these variables.

Top 5 Lasso Predictors to be Dropped:

1. `OverallQual_9`
    - **Description**: Represents a rating of the overall material and finish of the house with a score of 9.
2. `GrLivArea`
    - **Description**: Indicates the above grade (ground) living area in square feet.
3. `OverallQual_8`
    - **Description**: Similar to `OverallQual_9`, but with a score of 8.
4. `Neighborhood_Crawfor`
    - **Description**: Denotes that the property is located in the Crawford neighborhood.
5. `Exterior1st_BrkFace`
    - **Description**: Signifies that the exterior covering on the house is brick face.

Next Steps:

Once the features are dropped and the model is rebuilt, we will evaluate the model's performance to determine if this approach leads to better or more interpretable results.

Updated Model Analysis

After omitting the initial top 5 predictors from the Lasso model, we've re-evaluated the model and identified a new set of dominant predictors:

New Top 5 Lasso Predictors:

1. `2ndFlrSF`
    - **Description**: Refers to the square footage of the second floor.
2. `Functional_Typ`
    - **Description**: Represents the home's functionality, where "Typ" indicates typical functionality.
3. `1stFlrSF`
    - **Description**: Corresponds to the square footage of the first floor.
4. `MSSubClass_70`
    - **Description**: Indicates the building class of the property, specifically a type represented by the number 70.
5. `Neighborhood_Somerst`
    - **Description**: Denotes that the property is situated in the Somerst neighborhood.

Conclusion:

These newly emerged predictors will now play a pivotal role in our Lasso model's predictive performance. Continuous monitoring and evaluation are essential to ensure model accuracy and reliability.

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?
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 Robustness**:
    - **Definition**: A model is considered robust when variations in the data do not significantly affect its performance.
- **Model Generalizability**:
    - **Definition**: A model is generalizable if it can accurately adapt and make predictions on new, previously unseen data, drawn from the same underlying distribution as the training data.
- **Avoiding Overfitting**:
    - A primary concern for model robustness and generalizability is overfitting.
    - An overfitting model will recognize all patterns in the training data but might struggle with previously unseen test data due to its high variance.
    - Overly complex models are susceptible to this, which means the model might not be robust or generalizable.
- **Accuracy vs. Complexity**:
    - A highly complex model might boast high accuracy on training data.
    - However, to ensure robustness and generalizability, some decrease in model accuracy might be necessary to reduce variance, even if it introduces some bias.
- **Balancing Act**:
    - The goal is to strike a balance between model accuracy and complexity.

- Regularization techniques like **Ridge Regression** and **Lasso** are effective methods to achieve this balance.

Conclusion:

In essence, the journey towards a robust and generalizable model often involves navigating the trade-off between accuracy and complexity, ensuring the model is versatile enough to perform well on unseen data without being overly tailored to the training data.