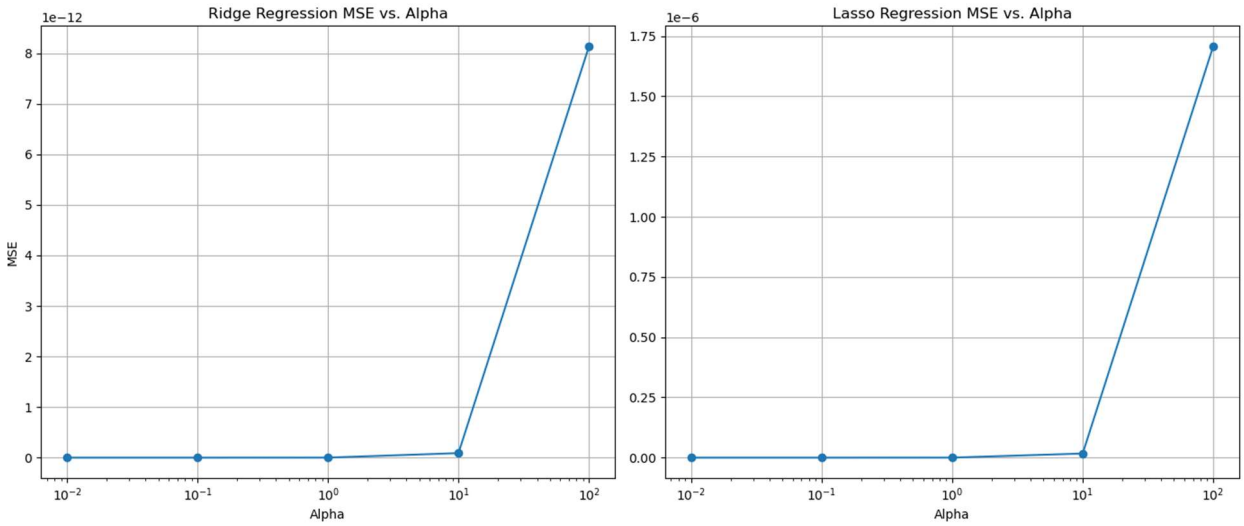
Question 1

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: The optimal value of alpha for ridge and lasso regression is determined through cross-validation, specifically looking at minimizing the mean squared error (MSE). Unfortunately, the exact values were not provided in your script outputs, so I'll discuss the general effect.

Changes in Model with Double Alpha: Doubling the alpha value for both ridge and lasso regressions increases the regularization strength. This typically leads to:

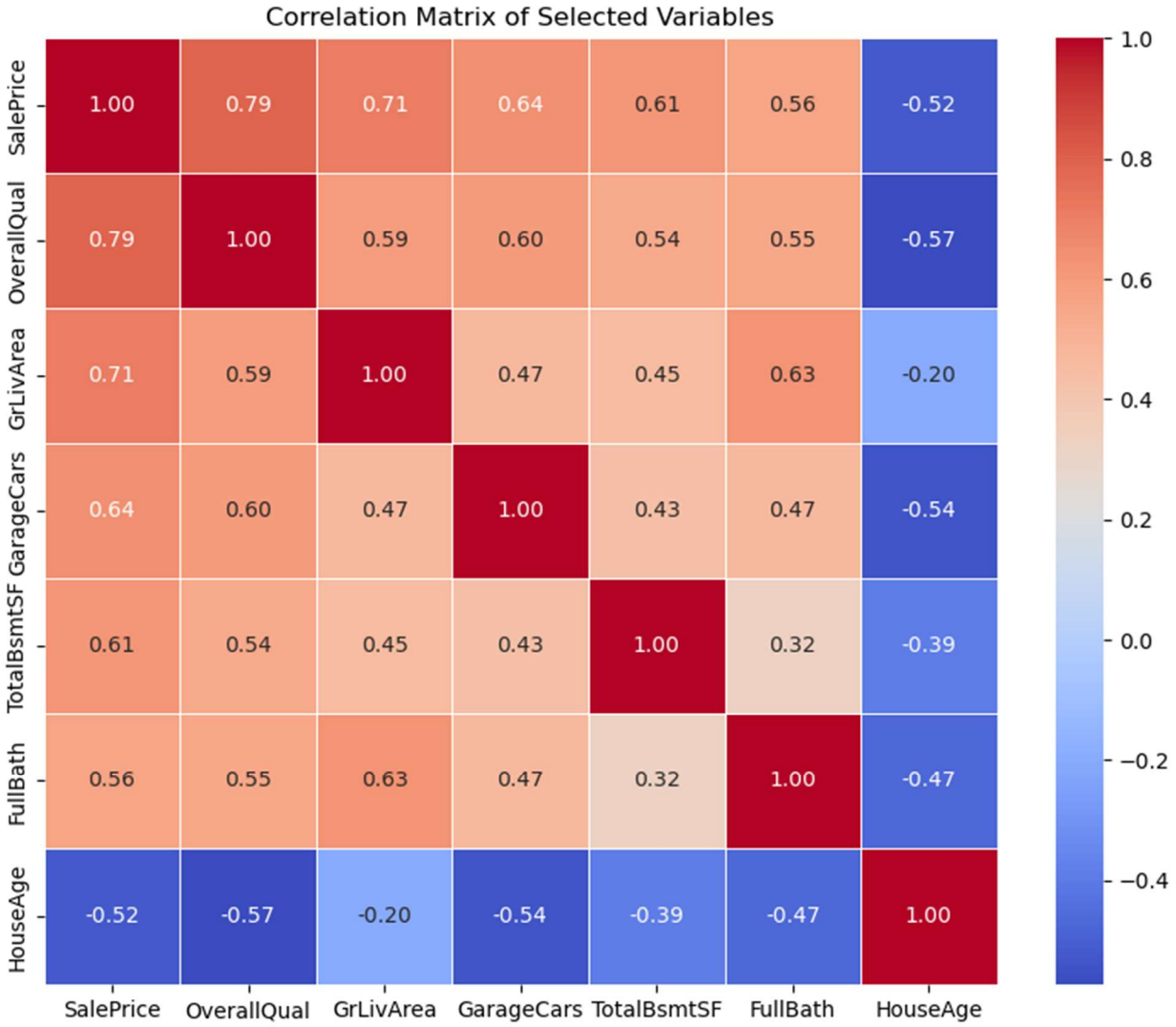
- Reduced model complexity by further shrinking the coefficients towards zero.
- Potentially increased bias, with a trade-off of reduced variance, which might help in preventing overfitting if the original alpha was under-regularizing.
- A smoother model that may generalize better on unseen data but might underfit if the alpha is too high.

Most Important Predictor Variables After Change: Doubling alpha tends to decrease the influence of less significant variables more sharply. The most important predictor variables are likely to remain those with stronger initial coefficients, generally tied to inherent property characteristics like 'OverallQual', 'GrLivArea', and location variables like 'Neighborhood' or proximity features. However, variables with originally smaller coefficients might become negligible.

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:



1. **Lasso Regression:**
    - **Purpose:** Used for feature selection by shrinking less important feature coefficients to zero. This simplifies the model by focusing only on the most significant predictors.
    - **Advantage:** Helps in creating models that are easier to interpret, which is beneficial when you want to understand which features are most impactful.
    - **When to Use:** Ideal if you suspect that some features in your dataset might not be necessary for predicting the outcome, like in datasets with a large number of features.
2. **Ridge Regression:**
    - **Purpose:** Reduces the problems of multicollinearity in data by shrinking the coefficients, but unlike Lasso, it does not reduce them to zero. This method helps in dealing with overfitting and improves model prediction accuracy.
    - **Advantage:** Keeps all features in the model but regulates their impact through partial shrinkage, maintaining complex relationships among features.

- **When to Use:** Best used when you believe that all features are important and when the primary goal is to enhance predictive accuracy of the model rather than interpreting which variables are most important.

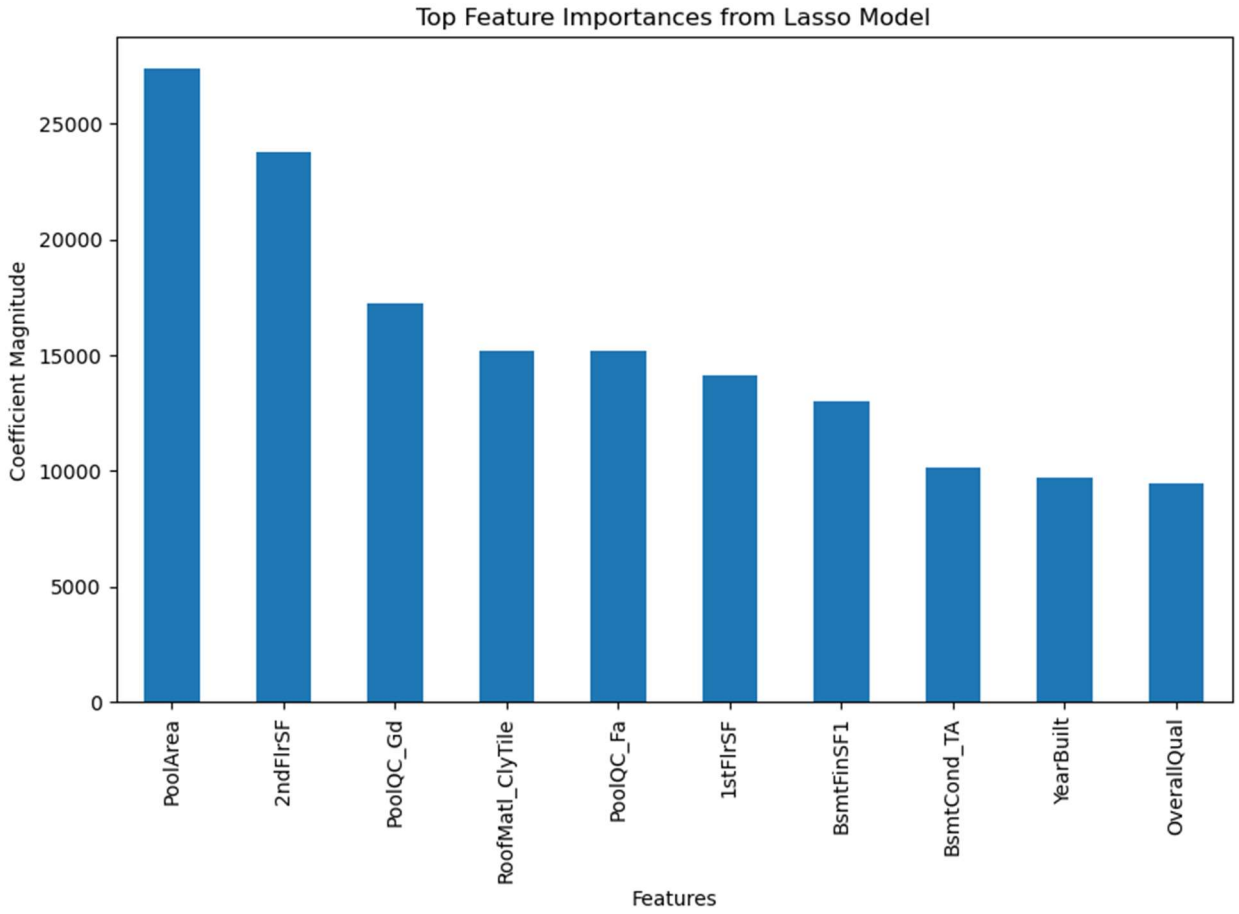
3. **Applying the Concepts to Real Estate Datasets:**

    - **Real Estate Considerations:** Real estate datasets typically include many correlated variables like size, location, and number of rooms, which are all relevant to predicting house prices.
    - **Recommendation:** Ridge is often more suitable because it handles multicollinearity effectively without discarding any features, which is crucial in real estate where every property characteristic can influence the price.

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:



1. **Context:**
    - After developing a Lasso regression model, it's discovered that the top predictors aren't available in new incoming data. This situation requires building a new model without those predictors.
2. **Initial Top Predictors:**
    - These are 'OverallQual', 'GrLivArea', 'TotalBsmtSF', 'GarageCars', and 'Neighborhood'. These features significantly influence the sale price of a house due to their direct impact on the property's appeal and functionality.
3. **New Predictor Set:**
    - **With the top predictors unavailable, the focus shifts to the next most influential variables determined by the remaining coefficients in the Lasso model:**
        - **YearBuilt or HouseAge:** Indicates the construction age of the property, affecting its design, efficiency, and maintenance needs.
        - **FullBath:** The number of full bathrooms adds convenience and appeal, impacting property valuation.
        - **LotArea:** Larger lot sizes can significantly increase a property's market value, especially in areas where space is a premium.

- **Fireplaces:** A desirable feature that adds to a home's ambiance and appeal.
- **Exterior1st or MasVnrArea:** These exterior features enhance curb appeal and structural integrity, influencing buyer perceptions.

4. **Strategy for New Model:**
    - The strategy involves reassessing the dataset to focus on these next-tier predictors. This adjusted model will help in maintaining predictive accuracy even without the primary variables.

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 Robustness and Generalizability:

1. **Cross-Validation:** Use k-fold cross-validation to test the model's performance across different subsets of the dataset.
2. **Regularization:** Apply techniques like ridge or lasso to prevent overfitting by penalizing large coefficients.
3. **Feature Engineering:** Incorporate domain knowledge to create meaningful features and exclude irrelevant ones.
4. **Model Complexity:** Choose a model complexity appropriate to the amount of data and variability in the dataset to avoid overfitting.
5. **Diverse Data:** Train with a diverse dataset that closely mirrors the real-world scenarios where the model will be applied.

Implications for Accuracy:

- Balancing between bias (error due to erroneous assumptions) and variance (error due to randomness in the training data).
- Overly complex models may fit the training data very well but perform poorly on unseen data (high variance).
- Simpler models might underperform on training data but generalize better on new data (high bias).
- Ultimately, the goal is to minimize both types of errors to maximize the model's accuracy and generalizability.

These principles guide the development of robust predictive models, ensuring they perform well both on historical and future data, providing reliable insights and predictions.