House Price Predictions: Advanced Regression Assignment Part -II - Subjective Questions and Answers

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:

The optimal lambda values obtained for Ridge and Lasso are as follows:
Ridge: 2.0
Lasso: 0.0001

Doubling the alpha value for both Ridge and Lasso results in the following changes in the model:

- For Ridge, doubling the alpha value leads to a slight increase in mean squared error, while the R2 values for both train and test remain largely unchanged.

- In the case of Lasso, doubling the alpha value results in a slight increase in mean squared error. The R2 value for train decreases slightly, but there is a significant decrease in the R2 value for test, indicating a worsening of the model's predictive performance.

Moreover, the model is further penalized, causing more coefficients of variables to shrink towards zero.

After implementing these changes, the most important predictor variables are:

- **For Ridge:**
    1. Total_sqr_footage
    2. OverallQual
    3. GrLivArea
    4. Neighborhood_StoneBr
    5. OverallCond
    6. TotalBsmtSF
    7. LotArea
    8. YearBuilt
    9. Neighborhood_Crawfor
    10. Fireplaces

- **For Lasso:**
  1. Total_sqr_footage
  2. OverallQual
  3. YearBuilt
  4. GrLivArea
  5. Neighborhood_Crawfor
  6. Neighborhood_StoneBr
  7. OverallCond
  8. Neighborhood_NridgHt
  9. LotArea
  10. GarageCars

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:
The optimal lambda values obtained for Ridge and Lasso are as follows:
Ridge - 2.0
Lasso - 0.0001

Regarding the R2 values:
For Ridge, the R2 values are: Train = 0.928, Test = 0.902, with a difference of 0.026.
For Lasso, the R2 values are: Train = 0.927, Test = 0.913, with a difference of 0.014.

The Mean Squared Error for Ridge and Lasso is:
Ridge - 0.00297
Lasso - 0.00280

Observations:

- The Mean Squared Error of Lasso is slightly lower than that of Ridge.

- The difference in R2 values between train and test is smaller in Lasso compared to Ridge.

- Additionally, Lasso facilitates feature reduction, as it shrinks the coefficient value of certain features toward 0. This feature reduction enhances model interpretation by emphasizing the magnitude of coefficients. **Consequently, Lasso offers a superior advantage over Ridge.**

Question 3

After building the model, you realized 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:
After excluding the top 5 most significant predictor variables in the Lasso model, re-evaluated the model and identified the following top five predictor variables:

1. TotalBsmtSF
2. OverallCond
3. TotRmsAbvGrd
4. LotArea
5. Total_Bathrooms

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:
The aim of the model is to be **as uncomplicated as possible**, as simpler models are more generic, they tend to be more resilient. This principle aligns with the Bias-Variance trade-off concept. Simplified models exhibit higher bias but lower variance, making them more generalizable. Conversely, complex models exhibit high variance and low bias.

Issues such as underfitting and overfitting may arise in models. Therefore, **achieving a balance between Bias and Variance** is crucial to mitigate these problems. **Regularization** serves as a solution to manage model complexity by shrinking coefficients towards zero, preventing the model from becoming overly complex and reducing the risk of overfitting.

Regularization techniques aim to maintain the model's optimal **simplicity by penalizing it if it becomes overly complex**. This approach helps in achieving the Bias-Variance trade-off, where bias is increased to an optimal level to minimize Total Error.

This optimal model complexity, also known as the Optimum Model Complexity, strikes a balance where the model is sufficiently simple to be generalizable yet complex enough to be robust. Simplifying the model involves making trade-offs between bias and variance.