

Advanced Regression – Surprise Housing (Subjective Q & As)

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 Ridge Regression is observed to have an optimal lambda value of **5.0**, while the Lasso Regression optimal lambda value is found to be **0.001**.

Upon doubling the alpha value for both Ridge and Lasso, the subsequent changes in model performance metrics are as follows:

Ridge Regression:

- The R2 score for the training set experiences a slight decrease from 0.938 to 0.932.
- Similarly, the R2 score for the test set exhibits a minor decline, moving from 0.900 to 0.896.

Lasso Regression:

- The R2 score for the training set decreases from 0.926 to 0.908.
- Likewise, the R2 score for the test set witnesses a decline from 0.900 to 0.886.

Following the twofold increase in the alpha value for both Ridge and Lasso, the primary predictor variables identified are as follows:

Key Predictor Variables (Ridge):

1. Total_Sqr_Footage
2. GrLivArea
3. TotalBsmtSF
4. Total_Bathrooms
5. TotRmsAbvGrd
6. LotArea
7. OverallQual_Very Good
8. Neighborhood_Crawfor
9. GarageArea
10. OverallQual_Very Excellent

Key Predictor Variables (Lasso):

1. Total_Sqr_Footage
2. GrLivArea
3. BsmtUnfSF
4. GarageCars
5. OverallQual_Excellent
6. SaleType_New
7. CentralAir_Y
8. Total_Bathrooms
9. OverallQual_Very Good
10. OverallQual_Very Excellent

All these calculations are detailed in the attached Jupyter Notebook.

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:

When choosing a model, it's important to consider the specific situation.

If we have many variables and want to focus on picking the most important ones, we should go with **Lasso**. It naturally picks out the most influential variables, making the model more straightforward.

On the other hand, if we want to avoid having very large numbers as your answers, we should choose **Ridge Regression**. It works by balancing the impact of each variable, preventing any of them from dominating the results.

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:

If the lasso model's top 5 features are not accessible, it becomes necessary to eliminate them from both X_train and X_test datasets. Subsequently, a new lasso model is reconstructed using the remaining features. This procedure has been executed in the provided Jupyter notebook, resulting in the identification of the following new top 5 features:

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

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:

Robustness in a model implies that its performance remains relatively unaffected by variations in the input data.

A generalizable model demonstrates the ability to effectively adapt to new, previously unseen data originating from the same distribution used for model creation.

Ensuring a model is both robust and generalizable requires guarding against overfitting. Overfitting, characterized by high variance, results in significant model prediction changes with even slight variations in the data. While an overfit model may perfectly capture training data patterns, it struggles to generalize to unseen test data.

Put differently, a model's complexity should be moderated to achieve robustness and generalizability.

From an accuracy standpoint, an excessively complex model tends to exhibit high accuracy. However, for enhanced robustness and generalizability, a balance needs to be struck by introducing some level of bias, leading to a reduction in accuracy.

Achieving this balance between model accuracy and complexity is often addressed through regularization techniques such as Ridge Regression and Lasso. These methods help prevent overfitting,

ensuring the model is better equipped to generalize to new data while maintaining a reasonable level of accuracy.