Advanced Regression - Subjective Questions

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 – 0.1

Lasso – 100

After doubling the alpha values, there’s no significant impact on R2 score for Lasso. However, for Ridge, the R2 decreases significantly by 10%.

Predictors after doubling alpha for Ridge –

- OverallQual
- TotRmsAbvGrd
- Fireplaces
- GarageCars
- ExterQual_TA

Lasso –

- GrLivArea
- OverallQual
- Neighborhood_NoRidge
- Neighborhood_StoneBr
- 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:

I would choose Ridge as it has higher R2 score for both, train and test. Also, the gap between train and test R2 scores for Ridge is negligible. Although, Lasso provides a simpler model but Ridge provides a better metric here. So, it depends on the business requirement. If there’s a requirement for simpler model, we can go with Lasso as it has comparable results to Ridge; if there’s a requirement for better performance, then we can go with 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:

Top 5 predictors before –

- (194100.248, 'GrLivArea'),
- (84633.856, 'OverallQual'),
- (50790.41, 'Neighborhood_NoRidge'),
- (42681.419, 'Neighborhood_StoneBr'),
- (35883.676, 'GarageCars')

Top 5 predictors after –

- (187791.503, 'TotalBsmtSF'),
- (119802.959, '1stFlrSF'),
- (-41361.433, 'ExterCond_TA'),
- (39387.481, 'YearRemodAdd'),
- (-38585.914, 'Foundation_PConc')

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:

Generalization plays a crucial role, emphasizing the necessity for test accuracy to surpass the training score. However, it is essential to strike a balance, avoiding an excessively high disparity. While the model should generalize effectively during training, an alarmingly high training score coupled with a lower testing score indicates overfitting—suggesting that the model has memorized the data. Ideally, the results should not exhibit substantial differences.

The robustness of a model extends beyond merely achieving high test scores; it hinges on the condition that training scores outperform test scores. Both scores must meet the criteria set by the specific business case and the model's expectations. Additionally, attention should be directed towards the values derived from both training and test phases, ensuring the model's competence with unseen data. This underscores the importance of retaining some outliers in the data to enhance predictive capabilities.

While there might not be a flawless model, various steps are available to guarantee the model's suitability for the specific context and the unique aspects of the business case. This aligns with Occam's razor, emphasizing the selection of a model that is not unnecessarily complex for the given purpose.