Advanced Regression subjective Question 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:

Optimal Value of alpha for Ridge Regression: 4.0

Optimal Value of alpha for Lasso Regression: 0.001

After doubling alpha values for Ridge and Lasso there is no much change in model performance in both regression models.

Ridge Regression (alpha=8):

- R-squared Value on training data: 92.0%
- R-squared Value on test data: 88.3%
- Important Predictor Variables (top 5) with Beta values after doubling alpha:
    - SaleCondition_Partial (0.1407)
    - SaleCondition_Others (0.1257)
    - SaleCondition_Normal (0.1221)
    - GarageFinish_Unf (0.1191)
    - GarageFinish_RFn (0.0978)

Lasso Regression (alpha= 0.002):

- R-squared Value on training data: 89.4%
- R-squared Value on test data: 87.3%.
- Important Predictor Variables (top 5) with Beta values after doubling alpha:
    - SaleCondition_Partial (0. 1995)
    - SaleCondition_Others (0. 1782)
    - SaleCondition_Normal (0. 1302)
    - GarageFinish_Unf (0. 128)
    - GarageFinish_RFn (0. 1109)

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:

As per the results Ridge and Lasso have very similar metrics.

Ridge: Alpha = 4, Train R-Squared=92%, Test R-Squared=88.0%

Lasso: Alpha=0.001, Train R-Squared=91%, Test R-Squared=87%

Both models look fine and resulted in good R-Squared scores.

Ridge has slightly higher values of R-Squared value (1% more) for both train, test set.

As we have more of features, Lasso helps to do feature elimination by making coefficients of less significant features to zero. Which makes model simple and robust so we proceed with Lasso.

Final model will be Lasso with Alpha = 0.001

Train R-Squared=91%, Test R-Squared=87%

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:

Below are five most important predictor variables before excluding

- Top 5 predictor variables:
  - SaleCondition_Partial
  - SaleCondition_Others
  - SaleCondition_Normal
  - GarageFinish_Unf
  - GarageFinish_RFn

New model is created by excluding the five most important predictor variables and below new set of most important predictor variables are:

  - GarageFinish_No Garage
  - GarageType_No Garage
  - GarageType_Detchd
  - GarageType_CarPort
  - GarageType_BuiltIn

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 model should be simple in general, to that it can be generalisable, and robust.

- **Generalisable model**, can be able to adopt properly for unseen dataset.
- **Robust Model**, ensure performance / results are not affecting much, thought there is variation in the data.
- To make sure model is **robust, generalisable**, we need to ensure:
    - Model doesn't cause overfitting. As an overfitting model has high variance. Overfitting is an undesirable behaviour as it gives accurate predictions for training data but not for new data.
    - **Bias, Variance Trade off:**
        - Bias (Correctness of model) refers to how much error does the model make in test data. Variance(Consistency of model) refers to how sensible is the model to input data.
        - Simple model might have higher Bias, but lower variance. So, this leads to underfitting. Complex models might have lower Bias, and higher variance, causing overfitting. So, it is important to have balance between both to avoid overfitting, underfitting of data.
- **Accuracy of Model:**
    - Complex models generally will have high accuracy. For making model to be more generalisable, robust, need to reduce the variance, which incurs to some additional bias. This helps in having an optimal model, this mean accuracy will be reduced. o
    - The balance between accuracy, and complexity can be achieved with help of **regularization techniques** such as Ridge Regression, Lasso Regression.