Surprise Housing: Multivariate Linear Regression

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 1

The optimal value of alpha for ridge regression is 10. While it is 0.0006 for lasso regression. By doubling the value of alpha for ridge regression, R2 score drops from 0.908 (alpha =10) to 0.900 for alpha = 20 on the training set
For lasso regression, doubling the value of value of alpha results in R2 score dropping from 0.907 to 0.893 in the training set

The most important predictor after the alpha is doubled are:

a. Ridge:

- LotFrontage: 0.112
- LotArea: 0.091
- MasVnrArea: 0.082
- BsmtFinSF1: 0.079
- BsmtFinSF2: 0.073

b. Lasso:

- LotFrontage: 0.206
- LotArea: 0.123
- MasVnrArea: 0.102
- BsmtFinSF1: 0.089
- BsmtFinSF2: 0.086

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 2

Optimal value of lambda:

- Ridge Regression: 10
- Lasso Regression: 0.0006

Model Performance

- Ridge Regression: Training: 0.908, Testing: 0.876
- Lasso Regression: Training: 0.907 Testing: 0.876

Although Ridge regression had comparable model performance with Lasso in this instance, however, Lasso is preferred as it reduced number of features through feature selection by reducing some of the coefficients to zero.

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 3

The next 5 important variables and their coefficients after excluding the five most important predictor variables in:

1. Lasso regression model:
    - BsmtUnfSF: 0.098
    - TotalBsmtSF: 0.077
    - 1stFlrSF: 0.06
    - 2ndFlrSF: 0.059
    - LowQualFinSF: 0.058

2. Ridge regression model:
    - BsmtUnfSF: 0.071
    - TotalBsmtSF: 0.065
    - 1stFlrSF: 0.058
    - 2ndFlrSF: 0.055
    - LowQualFinSF: 0.055

Question 4

How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

Answer 4

- Retrain the model on more data if available
- Further reduce the number of features
- Remove outliers
- Try alternative imputation approaches to treat missing values (k-means)
- Feature engineering

More robust and generalized model may see a drop in accuracy, however it will likely not overfit and may perform better on the test data