Predictive Analytics for Real Estate Investment

IITB, UPGRAD :
SOUHARDYABISWAS02@GMAIL.COM

QUESTION 1

1.1. WHAT IS THE OPTIMAL VALUE OF ALPHA FOR RIDGE AND LASSO REGRESSION?

The optimal value of alpha for lasso regression is 0.001, and for ridge regression, it is 2.0.

1.2. WHAT WILL BE THE CHANGES IN THE MODEL IF YOU CHOOSE DOUBLE THE VALUE OF ALPHA FOR BOTH RIDGE AND LASSO?

Increasing the value of alpha has different effects on Ridge and Lasso regression models:

For Ridge Regression: Doubling the value of alpha in Ridge regression increases the penalty for large coefficients (L2 regularization). This stronger regularization effect leads to more shrinkage of coefficient values towards zero, reducing the complexity of the model. Consequently, the model becomes more parsimonious, with smaller coefficients for predictor variables.

For Lasso Regression: Doubling the value of alpha in Lasso regression increases the penalty for non-zero coefficients (L1 regularization). This promotes sparsity in the coefficient matrix, forcing more coefficients to exactly zero. As a result, Lasso regression tends to perform feature selection, with many coefficients being zero. Consequently, doubling alpha in Lasso regression increases the level of feature sparsity, leading to a simpler model with fewer features.

Overall, both Ridge and Lasso regression models become more regularized and simpler with higher values of alpha, potentially preventing overfitting.

1.3. WHAT WILL BE THE MOST IMPORTANT PREDICTOR VARIABLES AFTER THE CHANGE IS IMPLEMENTED?

After implementing the change (doubling the value of alpha), the most important predictor variables will likely be those that survive the increased regularization and remain significant in predicting the target variable.

For Ridge Regression:

The most important predictor variables will still be those with larger coefficients, albeit reduced in magnitude due to increased regularization. These variables will have a significant impact on the predicted outcome, even after the regularization penalty.

For Lasso Regression:

The most important predictor variables will be those with non-zero coefficients. Doubling the value of alpha in Lasso regression increases the penalty for non-zero coefficients, leading to more coefficients being pushed towards zero. However, variables that are highly correlated with the target variable and have strong predictive power will likely survive the regularization and retain non-zero coefficients.

In both cases, the most important predictor variables will be those that contribute the most to explaining the variability in the target variable while still surviving the regularization penalty imposed by the increased alpha value. It's essential to analyze the coefficients of the predictors after the change to identify the most important variables for predicting the outcome accurately.

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?

After evaluating both Ridge and Lasso regression models and determining the optimal values of lambda for regularization, we can compare their performance metrics, such as R2 scores, to make a decision on which model to choose.

If we find that the R2 scores are similar for both Lasso and Ridge regression, it indicates that both models are performing comparably in terms of explaining the variance in the target variable. However, since Lasso regression penalizes more aggressively by forcing some coefficients to zero, it effectively performs feature selection and helps reduce the complexity of the model.

Therefore, if the R2 scores are similar and there is a need for feature selection or a preference for a simpler model, we will choose Lasso Regression as the final model. This choice ensures that we prioritize model simplicity and interpretability while maintaining comparable predictive performance to Ridge Regression.

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?

After removing the five most important predictor variables in the Ridge model, the new top predictor variables are:

```
Neighborhood_StoneBr LandContour_HLS
SaleCondition_Partial Neighborhood_Crawfor
LandContour_Lvl
```

```
The most important predictor variables after removing
the five most important predictor variables in the
Ridge model are: ['Neighborhood_StoneBr',
'LandContour_HLS', 'SaleCondition_Partial',
'Neighborhood_Crawfor', 'LandContour_Lvl',
'Neighborhood_NoRidge', 'Neighborhood_NridgHt',
'OverallQual', 'Condition1_RRAn', 'Functional_Typ']
```

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?

The model should be as simple as possible. Even though its accuracy might fall significantly, as a trade-off, it will be more robust and generalizable. In layman’s terms, the simpler the model the more the bias but less variance and more generalizable. This concept is explained in the bias-variance trade-off, where simplicity correlates with higher bias but lower variance, ultimately yielding greater generalizability. The practical implication of this trade-off on accuracy manifests in the model's consistent performance across both training and test datasets. A robust and generalizable model demonstrates minimal deviation in accuracy between these datasets, signifying its capacity to maintain reliable performance levels regardless of data variation.