House Pricing Assignment, Part-II

Submitted by Debadutta Pahadsing
Date 08-Nov-2022

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?

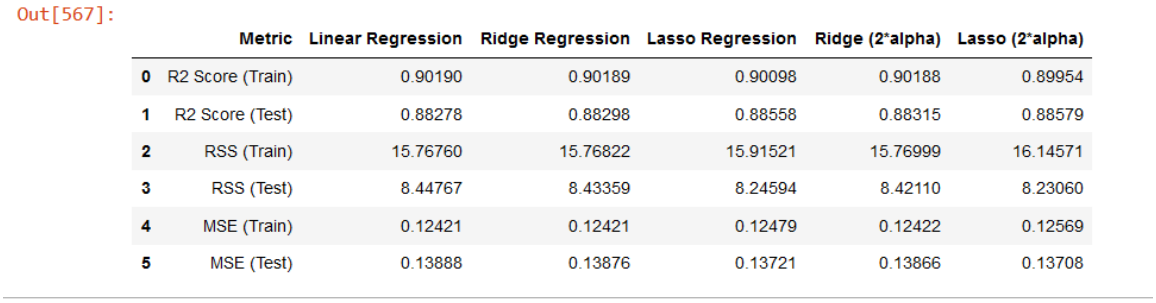
Answer:

Optimal value of alpha for ridge and lasso regression found as **2.0** and **0.001** respectively after hyperparameter tuning using Grid Search.

The above requirement was simulated and submitted as part of the notebook (in the last section **Part-II of Assignment**).

In our case, doubling the value of alpha resulted in the following (screenshot of the notebook inserted below for reference):

- The Test **R2 Score increased** for both Ridge & Lasso Regression
- The Test **RSS decreased** for both Ridge & Lasso Regression
- The Test **RSME** also **decreased** for both Ridge & Lasso Regression



Out[567]:

	Metric	Linear Regression	Ridge Regression	Lasso Regression	Ridge (2*alpha)	Lasso (2*alpha)
0	R2 Score (Train)	0.90190	0.90189	0.90098	0.90188	0.89954
1	R2 Score (Test)	0.88278	0.88298	0.88558	0.88315	0.88579
2	RSS (Train)	15.76760	15.76822	15.91521	15.76999	16.14571
3	RSS (Test)	8.44767	8.43359	8.24594	8.42110	8.23060
4	MSE (Train)	0.12421	0.12421	0.12479	0.12422	0.12569
5	MSE (Test)	0.13888	0.13876	0.13721	0.13866	0.13708

Top Ten Strong Predictors for Ridge Regressor are:

- GrLivArea
- OverallQual
- OverallCond
- GarageCars
- GarageQual
- Neighborhood_Somerst
- Neighborhood_NridgHt
- BsmtFullBath
- Foundation_PConc
- FireplaceQu

Top Ten Strong Predictors for Lasso Regressor are:

- GrLivArea
- OverallQual
- OverallCond
- GarageCars

- Neighborhood_Somerst
- BsmtFullBath
- FireplaceQu
- Neighborhood_NridgHt
- BsmtQual
- MSZoning_RL

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:

Interpretation 1 of the question:
Optimal value of alpha for ridge and lasso regression was determined during the assignment as 2.0 and 0.001 respectively. Whether we choose Ridge Regressor or Lasso Regressor?

Since the Lasso Regressor offers better performance on the Test data, we will choose and apply Lasso Regressor.

	Ridge Regressor	Lasso Regressor	Observations	Remarks
R2 Score on Test Data	0.88298	0.88558	Higher for Lasso	Higher the Better
RSS on Test Data	8.43359	8.24594	Lower for Lasso	Lower the Better
RSME on Test Data	0.13876	0.13721	Lower for Lasso	Lower the Better

Interpretation 2 of the question:

a. Optimal value of alpha for ridge and lasso regression was determined during the assignment as 2.0 and 0.001 respectively.
b. In response to Question 1, we observed better Model performance on Test Data after doubling the value of alpha.

Between a. and b. above, which values of alpha will we choose?

Between a. & b. above, we will choose the new values of alpha (doubled value) as they are resulting in better model performance.

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:

Five most important predictor variables in the original lasso model are:

- GrLivArea

- OverallQual
- OverallCond
- GarageCars
- Neighborhood_Somerst

Now that the five most important predictors are not in the incoming data, new model needs to be created.

The following steps were followed for the same:

1. The Clean Data (saved earlier) was read
2. Five most important predictor variables as found in the original lasso model were dropped from the dataset
3. Dataset was split into Training & Test Set
4. Features were standardized (scaled)
5. Feature Selection done using RFE (selected 45 top features, same numbers as before)
6. Residual Analysis was carried out and found OK
7. Optimal value of alpha was determined by Grid Search
8. New Lasso Model was created and fit using this Optimal alpha value and above selected features
9. Model Performance was checked, accuracy (R2 score) dropped from 0.88558 to 0.85576
10. Residual Analysis was performed and found OK
11. Top 5 important predictors were checked out

The above requirement was simulated and submitted as part of the notebook (in the last section **Part-II of Assignment**).

Top 5 important predictors now are:

- 1stFlrSF
- 2ndFlrSF
- FireplaceQu
- BsmtQual
- Neighborhood_NridgHt

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?

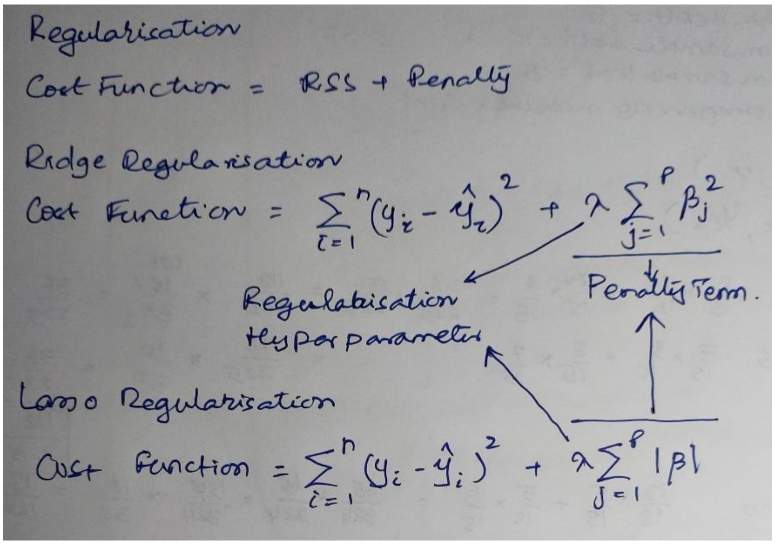
Answer:

A model is considered robust if the dependent variable is consistent even if one or more of the features are changed in the unseen data due to unforeseen circumstances. This means the model should have low variance so that is not sensitive to changes in the data. Model robustness also comes from model stability i.e., delivering same performance every time. The key to this is Bias-Variance Trade-Off, where a reasonable sacrifice in Bias is allowed for reduction in the Variance. This is achieved through cross-validation to ensure that training is repeated with different test-folds.

A generalized model is one which does not overfit the training data. In case of overfit models, the performance on the training data is excellent but when the model performance is checked on the unseen data, it is poor. This happens when the model is overly complex enough to memorize the training data, rather than generalizing the same. In order to ensure Generalization in the model, and to prevent overfitting, regularization techniques are used. Some of the most widely used regularization techniques are Ridge and Lasso Regularization.

Regularization adds a penalty term to the cost function using a regularization factor (hyperparameter) alpha (also referred as lambda in some books). Regularization reduces the coefficients of the non-important features to near zero in case of Ridge and to zero in case of Lasso. Therefore, Lasso regularization also results in feature selection.

The cost function for Ridge and Lasso regularization is depicted below for reference:



Regularisation

Cost Function = RSS + Penalty

Ridge Regularisation

$$\text{Cost Function} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \underbrace{\sum_{j=1}^{p} \beta_j^2}_{\text{Penalty Term}}$$

Regularisation Hyperparameter

Lasso Regularisation

$$\text{Cost Function} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \overline{\sum_{j=1}^{p} |\beta|}$$

One important aspect one has to remember that Ridge and Lasso regularization require the features to be standardized (scaled).

As explained above, a robust and generalized model is more likely perform better, i.e. deliver higher accuracy on the unseen data.