Problem Statement - Part II

(Submitted by: Shen Shaji)

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 values of alpha are

Lasso: 0.01

Ridge: 2

The r2 score for train and test data changes when we double the alpha values and the model tends to have more bias and less overfit/variance when the alpha value increases. Because the model is penalized for higher coefficients.

Lasso regression:

Old r2 score train: 0.885

Old r2 score test: 0.889

Old RMSE : 0.12573331036519156

The most important variables (old alpha):

1. **GrLivArea**
2. **OverallQual**
3. **OverallCond**
4. **TotalBsmtSF**
5. **BsmtFinSF1**
6. **GarageArea**

New Train R2: 0.871

New Train R2: 0.881

New RMSE : 0.130

The most important variables (new alpha):

1. **OverallQual**
2. **GrLivArea**

3. **TotalBsmtSF**
4. **OverallCond**
5. **BsmtFinSF1**

Ridge regression:

Old r2 score train r2: 0.9365668597616261

Old r2 score test r2: 0.9078617486572108

Old RMSE : 0.11479430768918945

The most important variables (old alpha):

1. **MSZoning_FV**
2. **MSZoning_RL**
3. **Neighborhood_Crawfor**
4. **MSZoning_RH**
5. **MSZoning_RM**

New Train R2: 0.934

New Test R2: 0.908

New RMSE : 0.114

The most important variables (new alpha):

1. **Neighborhood_Crawfor**
2. **MSZoning_FV**
3. **MSZoning_RL**
4. **Neighborhood_StoneBr**
5. **SaleCondition_Partial**

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

Though the model performance by Ridge Regression was better in terms of R2 values of Train and Test, it is better to use Lasso, since it brings and assigns a zero value to insignificant features, enabling us to choose the predictive variables. The number of predictor variables in the Ridge regression is much more than that of lasso regression and the r2 score is only slightly different with almost similar RMSE scores. It is always advisable to use simple yet robust model.

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 most five important variables before(in ascending order):

1. **GrLivArea**
2. **OverallQual**
3. **OverallCond**
4. **TotalBsmtSF**
5. **BsmtFinSF**

After dropping them the new most important variables are (in ascending order):

1. **1stFlrSF**
2. **2ndFlrSF**
3. **GarageArea**
4. **Fireplaces**
5. **BsmtFullBath**

New model metrics after dropping the 5 columns:

Train R2: 0.8278068685543116

Train R2: 0.8302688539612822

RMSE : 0.155804910852560541'

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

The model should as be simple and robust as possible. Even though this will decrease the accuracy of the model, it will perform good in the case of un-seen data. This should be done by taking care of the Bias-Variance trade off. When the model becomes too simple it has high bias and when the model memorizes the training data, it overfits and has high variance. The implications of accuracy and other scoring metrics in this scenario is that, these metrics should be comparable same for both training and testing data. Even though one among them would be low in case of underfitting or overfitting, it should be almost similar in case of a general model that can do well with un-seen data.

Bias: High bias means model is unable to learn details in the data. Model performs poor on training and testing data but well with un-seen data.

Variance: high variance is when model tries to over learn every pattern from the training data. High variance means it performs very poor on testing data as it was unseen data for the model. It is important to have balance in Bias and Variance to avoid overfitting and under-fitting of data.