Problem Statement - Part II

Q. 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:

Lasso Regression:

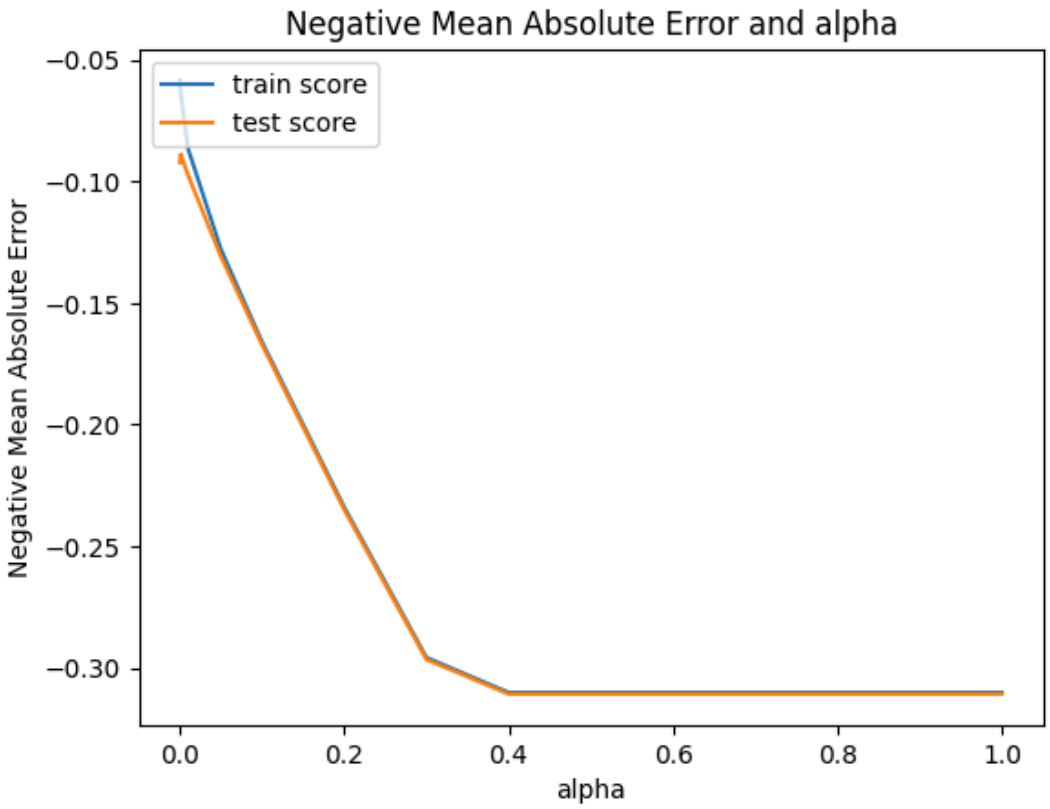

Figure 1: Negative Mean Absolute Error vs alpha (for Lasso Regression)

After applying Lasso regression to the train and test sets and plotting negative mean absolute error against alpha (our regularization term), we observe that as alpha values increase, the negative mean absolute error decreases and stabilizes around 0.4 for both the train and test scores. An alpha value of 0.4 heavily penalizes the variables in the model, driving their values towards 0. Consequently, I opted for a smaller alpha, specifically 0.04, which resulted in the best test and train r-squared scores. This choice also yielded a reasonable number of predictor variables, totaling 17, as a result of the filtration that lasso incorporates into the model.

Ridge Regression:

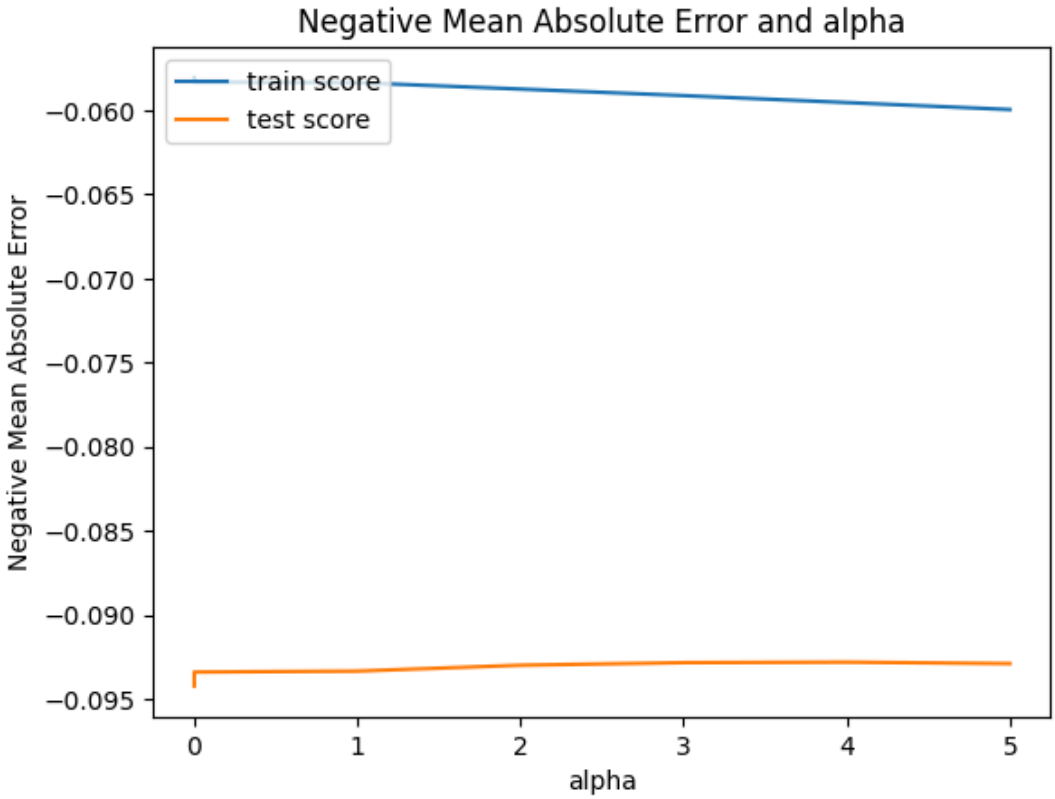

Figure 2: Negative Mean Absolute Error vs alpha (for Ridge Regression)

We note a decrease in the negative mean absolute error for the train score as alpha increases, while the negative mean absolute error of the test score shows a slight increase with higher alpha values. Both the train and test scores exhibit a stable negative mean absolute error when alpha is set to 4.0. Consequently, 4.0 emerges as the optimal alpha value, demonstrating favorable test and train r-squared scores.

When we double the value of alpha for our Ridge regression, setting it to 8, the model applies a more significant penalty to the curve, aiming to enhance generalization and simplicity. The graph indicates increased errors for both the test and train sets when alpha is 8. Similarly, elevating the alpha value for Lasso results in a higher penalty on the model, leading more coefficients of the variables to be reduced to zero. As we increase the alpha, the R-squared value also decreases. Thus, the statement correctly highlights the tendency of higher alpha values to promote model simplicity and regularization, contributing to higher errors and reduced coefficients.

If we implement the change for Lasso i.e. double the alpha from 0.04 to 0.08, the most important predictor variables are:
OverallQual,
GrLivArea,
GarageArea,

TotalBsmtSF,
1stFlrSF,
GarageType_Attchd,
FireplaceQu_none
And If we implement the change for Ridge i.e. double the alpha from 4 to 8, the most important predictor variables are:
RoofMatl_CompShg
RoofMatl_Tar&Grv
RoofMatl_WdShngl
MSZoning_RL
Followed by 218 more variables.

Q 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 the optimal values of lambda for Ridge and Lasso regression, I would choose to apply Lasso regression. The decision is driven by the observation that Lasso not only yields better R-squared scores for both the train and test sets but also provides a valuable feature selection mechanism. Lasso's regularization approach penalizes insignificant variables by reducing their coefficient values to zero. This feature selection property is particularly beneficial as it filters out non-essential predictor variables from the dataset, promoting a more parsimonious and interpretable model. Therefore, the dual advantages of improved predictive performance and feature selection make Lasso regression a preferable choice over Ridge in this context.

Q. 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:
If the five most important predictor variables in the lasso model i.e. 1. OverallQual, 2.GrLivArea, 3. GarageArea, 4. TotalBsmtSF, and 5. CentralAir_Y are not available in the data, the five most important predictor variables after creating another model without the aforementioned variables will be not be the next top 5 predictor variable(6. GarageType_Attchd,7. BsmtFinType1_GLQ, 8. 1stFlrSF, 9. BsmtFullBath and 10. Foundation_PConc) but the lasso outputs different top 5 predictor variable as follows:
1stFlrSF,
2ndFlrSF,
Foundation_PConc,
CentralAir_Y,
BsmtFinType1_GLQ

Note: the r-squared scores for both train and test set also changes(lowers).

Q.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:

Going by the principle of Occam's razor, a model should be simple but robust. Occam's razor does not advocate indiscriminate simplification until further simplification becomes impossible. Instead, it proposes that when confronted with a choice between a complex and a simple model, and assuming all other factors are approximately equal, it is advisable to opt for the simpler model.

To ensure that a model is robust and generalizable, we can employ some of the following strategies but not limited to:

1. Cross-Validation: Use techniques like k-fold cross-validation to evaluate the model's performance on multiple subsets of the data.
2. Feature Engineering: Choose relevant features and avoid overfitting by removing irrelevant or redundant variables.
3. Regularization: Apply regularization techniques (e.g., Lasso or Ridge Regression) to prevent overfitting by penalizing large coefficients which encourages the model to be more generalizable to new data.
4. Hyperparameter Tuning: Optimize hyperparameters through techniques like grid search or random search.

The implication of aptoping the above strategies such as regularization might reduce the accuracy of the model as they introduce a penalty to prevent overfitting by simplifying the model. However, this simplification can lead to increased bias, underfitting, and the exclusion of important features, causing a decrease in overall model accuracy on unseen data. Although it may not attain the highest accuracy on the training set, its effectiveness is anticipated to be reliable when applied to new and unseen data. The objective is to achieve a balance where the model captures fundamental patterns without memorizing noise specific to the training data, ultimately resulting in enhanced performance across diverse situations.

Submitted by:
Imliyangla Longkumer