Project Description:

Advanced Regression – Housing Price Prediction Assignment

Submitted by: Tandeep Sandhu

Project details

Problem Statement - Part II

Answering the Probelm Statement - Part II questions

- Build the model as per the questions in Probelm Statement - Part II
- This will be used to build the answer

Question 1

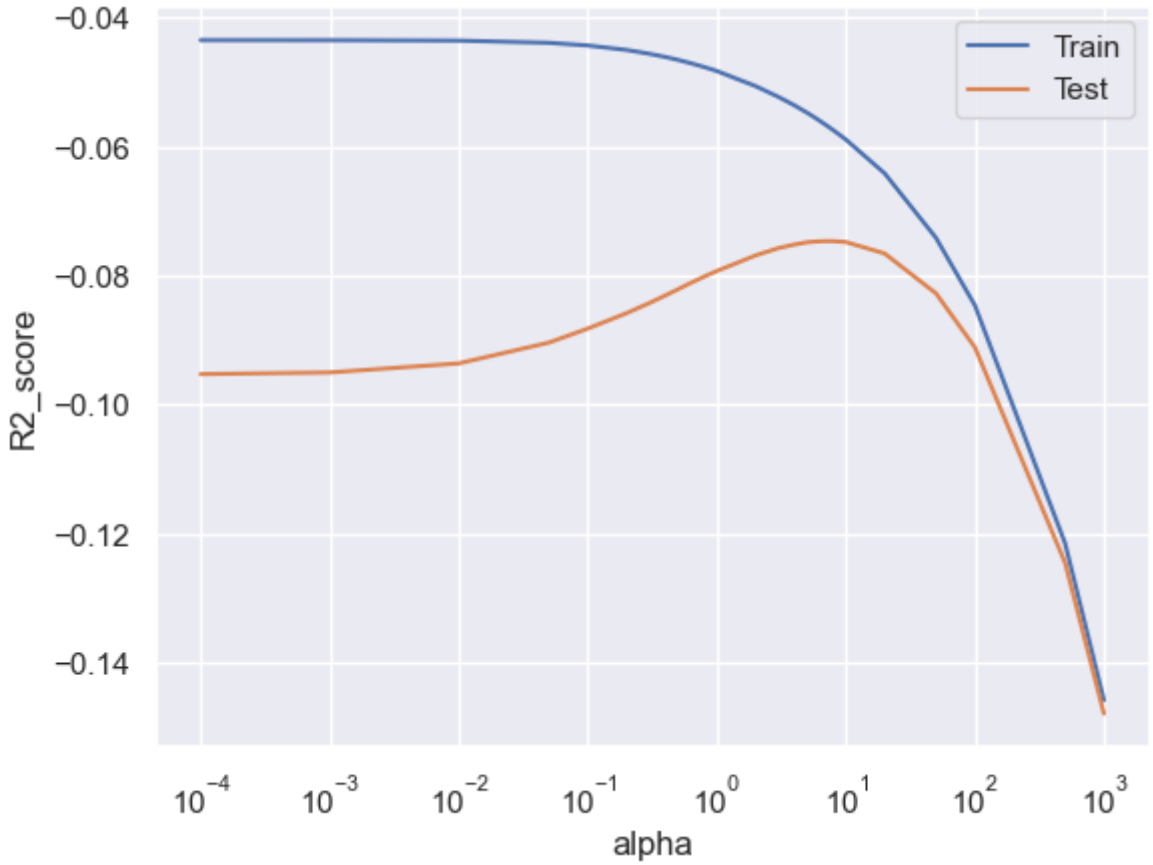
Question 1

- 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?

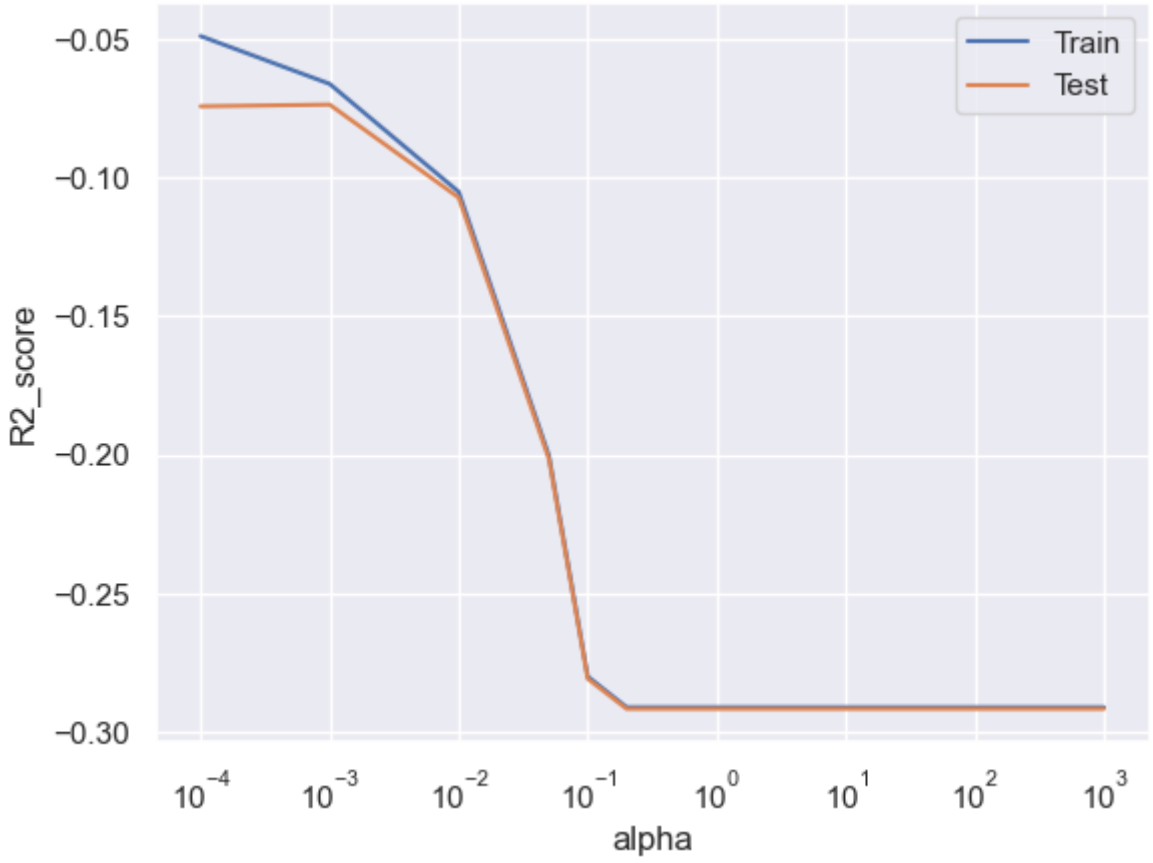
Answer 1

- As defined in the assignmemt in the previous sections:
  - The Ridge model has optimal value of alpha as 7
  - The Lasso model has an optimal value of alpha = 0.001
- The optimal values are also validated as per the graph shown below for the R2 score against the alpha values for both the models

In [125…
```
# Plot R2 score vs alpha values for Ridge Regression Model
```



In [126…
```
# Plot R2 score vs alpha values for Lasso Regression Model
```



- We will process the Ridge and Lasso models with dowble the alphas and see which predictors are better
  - Ridge alpha value = 2 * 7 = 14

- Lasso alpha value = 2 * 0.001 = 0.002

Important predictors when the value of alpha is doubled for Ridge and Lasso Model

- The most common predictors with double the alpha values is as below:
  - GrLivArea
  - TotalBsmtSF
  - OverallQual_9
  - OverallQual_8
  - GarageArea
  - YearRemodAdd
  - LotArea
  - Neighborhood_Crawfor
  - BsmtFinSF1

Question 2

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

- The model building has helped to define the optimal values
- The right model would need to be picked will be based on what we are trying to achieve
- The Lasso model would have a better usage here as the number of columns are very high.
- This would help to simplify and better understand the model

Question 3

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 3

- Now we will need to build the model after dropping the top five predictors in lasso Regression model with alpha equal to 0.001.
- The predictors which need to be removed are as below:
  - GrLivArea
  - TotalBsmtSF
  - OverallQual_9
  - OverallQual_8

    - YearBuilt
- Building the new Lasso model after removing the above columns

The top predictors after removing the top five features as per the earler Lasso model

- The value of optimal alpha as per the new dataset with removed columns is 0.0001
- The common predictors for the Ridge and Lasso are show below
    - Condition2_PosA
    - 1stFlrSF
    - 2ndFlrSF
    - SaleType_ConLD
    - BsmtFinSF1
- The model with five top predictors also has a lower R2 score on test compared to teh model with all predictors. The intial value was 0.89 and now it is 0.84

Question 4

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 4

- Definitions for generalizability and robustness
    - The classic approach towards the assessment of any machine learning model revolves around the evaluation of its generalizability i.e. its performance on unseen test scenarios.
    - Evaluating such models on an available non-overlapping test set is popular, yet significantly limited in its ability to explore the model's resilience to outliers and noisy data / labels (i.e. robustness).
    - The above definitions for generalizability and robustness are provided as per the paper - Paschali, M, Conjeti, S, Navarro, F & Navab, N 2018, 'Generalizability vs. Robustness: Adversarial Examples for Medical Imaging', *'arXiv:1804.00504v1'*, https://arxiv.org/pdf/1804.00504.pdf.
- The model is considered robust and generalized when it meets some of the criterias as outlined below:
    - The model is neither overfitting and neither underfitting based on the test or production data.
    - Any changes to test or the production data does not affect the performance of the model
    - The data is better able to adapt to the unseen data - test or production data
    - The model itself doen't have too many features so that it becomes overly complex and any small changes to the test data causes reduction in the score
    - The model is also not too simple as the features have been removed resulting in the model which can neither predict on train or test data

- The cross validation approach and regularization are some of the techniques to help build a robust and generalized model.
- If the model is not robust or generalized, then the model will be difficult to deploy in production environment