# Problem Statement - Part II – Question & Answers

Submitted by: Tandeep Sandhu

## 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 to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

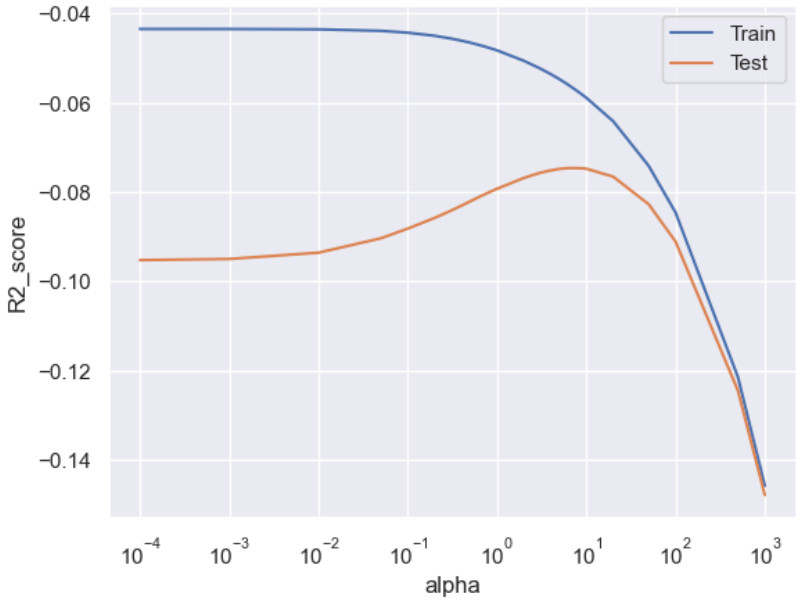
*Answer 1*

In accordance with the specifications outlined in the sections of Jupyter notebook document (Housing_v06.ipynb):

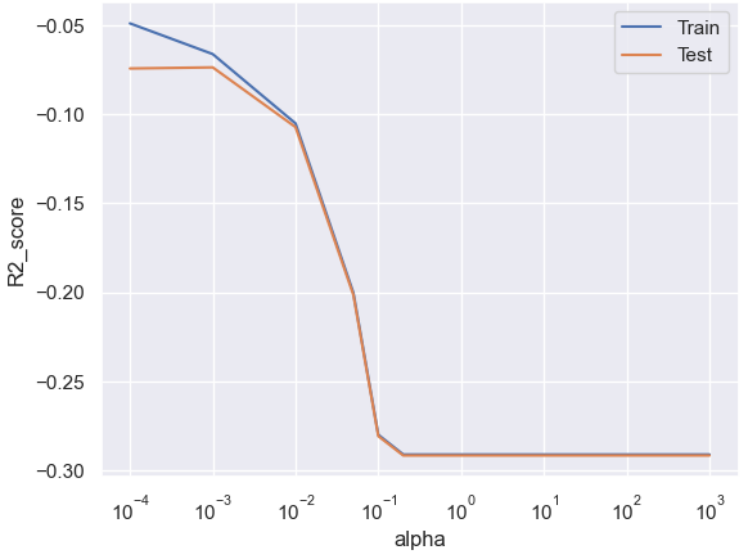
- The Ridge model demonstrates an optimal alpha value of 7.
- The Lasso model exhibits an optimal alpha value of 0.001.

- Moreover, these optimal values are corroborated through validation, as illustrated in the graph below, depicting the R2 score against the alpha values for both models.

*Plot of R2 score vs Alpha for the Ridge Regression Model*



*Plot of R2 score vs Alpha for the Lasso Regression Model*




## 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 process of constructing the model has facilitated the identification of optimal values.
- The selection of the appropriate model depends on the specific objectives we aim to accomplish.
- In this context, the Lasso model proves to be more suitable due to the significantly high number of columns. Utilizing Lasso would contribute to simplifying and enhancing our understanding of 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*

- Our next step involves constructing the model by excluding the top five predictors in the Lasso Regression model with an alpha value of 0.001.
- The specified predictors that are slated for removal are listed below:
    - GrLivArea
    - TotalBsmtSF
    - OverallQual_9
    - OverallQual_8
    - YearBuilt

- Creating the updated Lasso model post the elimination of the aforementioned columns results in a shift in the top predictors.
- The revised list of top predictors, following the exclusion of the initial five features based on the earlier Lasso model, is as follows:
- The shared predictors for both the Ridge and Lasso models are presented below:
    - Condition2_PosA
    - 1stFlrSF
    - 2ndFlrSF
    - SaleType_ConLD
    - BsmtFinSF1

- The model featuring the top five predictors exhibits a lower R2 score on the test set when contrasted with the model encompassing all predictors. The initial R2 score, which stood at 0.89, has now decreased to 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*

As outlined by the authors Xu, H & Mannor, S (2010), the concepts of generalizability and robustness assume distinct meanings:

- The traditional method of assessing machine learning models centers on evaluating their generalizability, which refers to their performance on unseen test scenarios.
- While assessing these models on a non-overlapping test set is a common practice, it has notable limitations in thoroughly exploring the model's resilience to outliers and its ability to handle noisy data or labels, indicating its robustness.

The definitions for generalizability and robustness mentioned above are derived from the work of Xu, H & Mannor, S in their 2010 paper titled "Robustness and Generalization," which can be accessed at *arXiv:1005.2243* via https://doi.org/10.48550/arXiv.1005.2243.

A model is deemed robust and generalized when it satisfies certain criteria, outlined below:

- The model avoids overfitting or underfitting when evaluated on test or production data.
- Changes to the test or production data do not adversely affect the model's performance.
- The model demonstrates adaptability to unseen data, whether in test or production scenarios.
- It maintains a balanced complexity by not having an excessive number of features that could lead to over-complexity, causing a decline in performance with slight changes in the test data.
- Simultaneously, the model isn't excessively simplistic, avoiding issues where crucial features are removed, rendering it ineffective in predicting both train and test data.
- Techniques such as cross-validation and regularization are employed to facilitate the development of a robust and generalized model.
- A lack of robustness and generalization makes deploying the model in a production environment challenging.