Shrinivas Bhat

s.shrinivas.bhat@gmail.com

C44 Group

Subjective questions

Advanced Linear Regression Part-II

# Table of Contents

[Table of Contents 1](#_Toc123139702)

[Assignment-II Subjective Questions 2](#_Toc123139703)

[**Question 1** 2](#_Toc123139704)

[**Question 2** 4](#_Toc123139705)

[**Question 3** 6](#_Toc123139706)

[**Question 4** 7](#_Toc123139707)

[Appendix (Reference for Answers) 8](#_Toc123139708)

[Model Summary 8](#_Toc123139709)

[Across Models Inferences 11](#_Toc123139710)

Shrinivas Bhat

# Assignment-II Subjective Questions

## **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?

* Optimal alpha for Ridge = 7, Optimal value for Lasso = 0.0001 (Model-1: Target variable is log transformed and min-max scaled ***Refer: Model Summary***)
* With the alpha doubled there is degradation seen in R2-score of Ridge and Lasso
* With alpha doubling the RSS/RMSE of Ridge and Lasso regression degraded for training data but improved for test-data. This implies that increasing the regularization factor is improving the generalization by trading of residual errors in training data by improving it in test-data.

Table

Description automatically generated

Figure 1 Scores with alpha

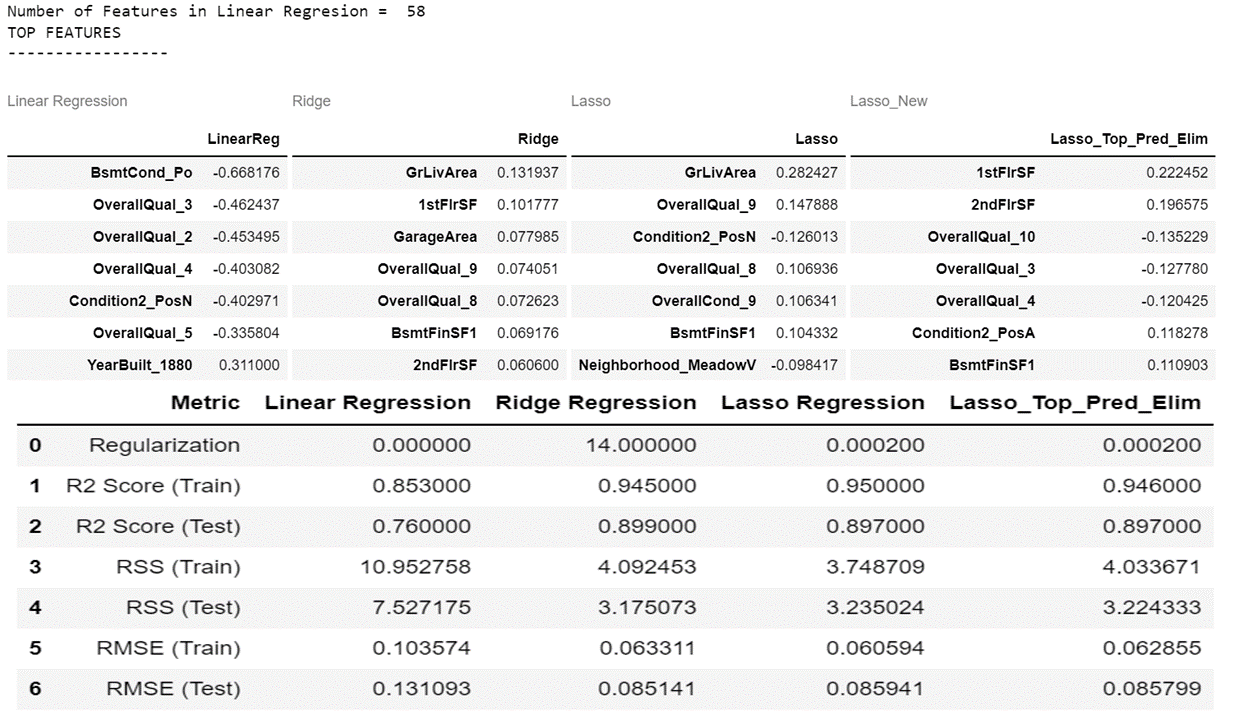


Figure 2 Scores with alpha\*2

* GrLivArea, Condition2\_PosN, Condition2\_PosA, OverallQual\_8,OverallQual\_9 are the top predictors with Lasso Regression with alpha
* GrLivArea, Condition2\_PosN, BsmtFinSF1, OverallQual\_8,OverallQual\_9 are the top variables with Lasso Regression with alpha\*2
* GrLivArea, GarageArea,1stFlrSF, OverallQual\_9, BsmntFinSF1 are the top-5 Predictors for Ridge Regression for alpha
* GrLivArea, GarageArea,1stFlrSF, OverallQual\_9, OverallQual\_8 are the top-5 Predictors for Ridge Regression for alpha\*2
* 3 to 4 predictors in the top-5 predictor variables don’t seem to change with the modification of regularization factors by a factor of 2, both in Lasso as well as ridge regression
* The Ridge and Lasso Columns in the following table indicate the top-most feature variables along with coefficients when the regularization value or the alpha value is doubled.

Table

Description automatically generated

Figure Ridge and Lasso Predictors and Coefficients with doubled alpha

## **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?

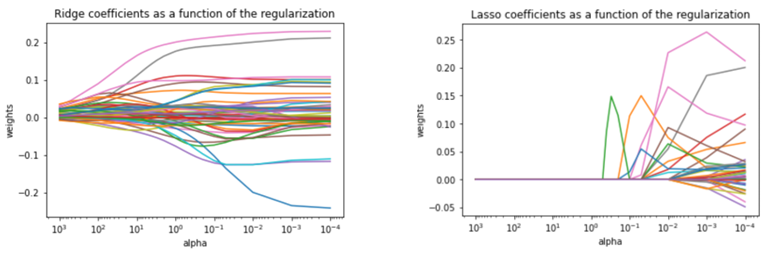
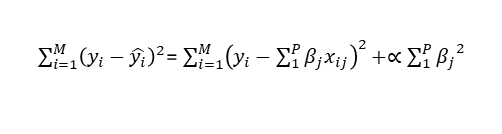


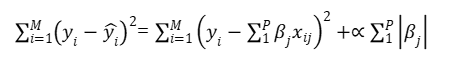
Figure 4 Coefficient Emergence with varying alpha

Shown below in Equation 1 and Equation 2 Ridge and Lasso Cost Functions minimized for determining , used in the regression for observations of the target Variable and Predictor Variables given by

Equation 1 Ridge Regression



Equation 2 Lasso Regression Equation



* In Ridge and Lasso the additional terms try to penalize the increase of coefficients by minimizing the joint term one of which is .The regularization term (alpha) regularizes the coefficients such that if the coefficients take large values the optimization function is penalized. By keeping the coefficient growth under check Ridge and Lasso Regression try to minimize the complexity of the model while preventing overfitting issues of linear regression.
* From the plots we can see how the coefficients start emerging for Ridge and Lasso with decreasing alphas. Decreasing the alpha will start the emergence of coefficients at orders of magnitude higher for Ridge than that of Lasso and hence Ridge optimal alphas are orders of magnitudes higher.
* The optimal value of the alpha can be chosen by the model in such a way that both mean-test-scores for testing data and training data is minimized, while maximizing their R2\_scores.
* The difference between the Lasso and Ridge is that with the alpha increase for Ridge the coefficients starts shrinking towards zero but doesn’t become zero where as in the Lasso the coefficients gets completely eliminated. Because of this the Lasso can lead to feature selection.
* The choice of the alpha and the model that we apply depends on the our business goal , in case we want to go for optimal feature selection's while making the model simpler then the Lasso is chosen, But if we want good accuracy with good scores(R2,Mean-test/train scores) include higher number of features without causing overfitting then we go for Ridge regression. However, the complexity will increase as the dynamic range of the coefficients becomes higher for Ridge and it demands higher bits to for processing and storage. Lasso will be useful when there are millions of features where ridge regression will lead to significant complexity.
* Ridge generally works well even in presence of highly correlated features as it will include all features where coefficients get decided based on the correlation between the features whereas in Lasso highly correlated features are selected and remaining are set to zero. Also change in Model parameters for Lasso changes the top predictor variable as compared to ridge regression. So, this also will be a factor for choosing the right model.

## **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?

1. The top-5 Predictor Variabes with Lasso Regression

- GrLivArea

- Condition2\_PosN

- Condition2\_PosA

- OverallQual\_8

- OverallQual\_9

***Refer: Model Summary***

1. The top-5 Predictor Variables obtained from Lasso regression after eliminating the above Variables

- YearBuilt\_1893

- 1stFlrSF

- 2ndFlrSF

- OverallQual\_10

- OverallQual\_3

## **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?

* More robust model is the model which does not change its accuracy when the training data changes. For a generalizable model it should not undergo changes when it sees a new data.
* Same model should fit properly to new, previously unseen data, assumed to be from the same distribution.
* To make the model more generalized and robust we must make the model simple.
* But when we try to make the model simple it suffers from bias as it does not perform well on the training data itself but when we try to make it complex it memorizes the data and suffers from overfitting, and it does not perform well on the test data and this variance of the model between training and test-data is termed as variance which is high for Complex models.
* Striking the right balance between the bias and variance where the variance of the model can be reduced by trading off against the training error\*(bias) will achieve an optimal model.
* So the accuracy of the model on the training data is traded off in order to make it more robust and generalizable model for unseen data.
* This is achieved using Regularization techniques like Lasso and Ridge regression. Regularization helps with managing the model complexity by essentially shrinking the model coefficients estimates to 0 (by adding a penalty term for increased coefficients) and discourages the model from becoming too complex and thus avoiding the risk of overfitting and essentially making it more robust and generalizable.
* Regularization also takes care of Multicollinearity by reducing the coefficients of collinear variables close to zero.
* Apart from this the data analysis and cleaning need to be performed on training data by treatment of outliers and missing data with appropriate imputations, removal of unreliable features, removal of collinear features…etc to ensure the generation of a robust model.
* Based on the distribution of the target data the data-transformation also can be performed in case distribution is significantly skewed.

## Appendix (Reference for Answers)

### Model Summary

Multiple Model options were tried with various configurations for Linear,Ridge and Lasso Regression

* Model-1 Min-Max Scaling for Target and Predictor Variables, Log Transformation for Target Variable
* Model-2 Min Max scaling for Target and Predictor Variables, Log Transformation for Target Variable, Doubling of Optimal Regularization parameters obtained for Lasso and Ridge
* Model-3 Min-Max Scaling applied for Predictor Variables only, Log transformation of Target Variable
* Model-4 Min-Max Scaling applied for Predictor Variables, No Transformation of Target Variable
* Model-5 Min-Max Scaling applied for Predictor Variables+ Target Variable, No Transformation of Target Variable

Lasso Regression was also tried with elimination of the top-5 predictors for each Model.In some of the models the log transformation on the target variable was performed as the target variable showed some skewness in its distribution.

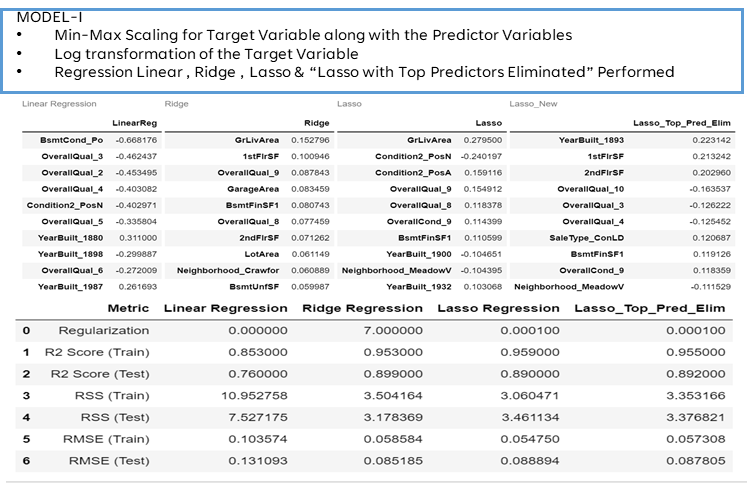


Figure 5 Results for Model-I (Performance Scores, Top-Predictors and Coefficients)

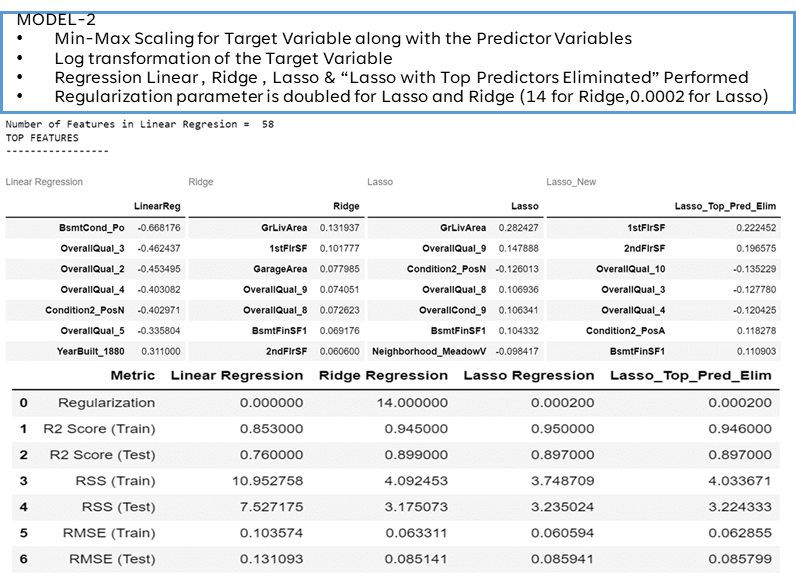


Figure 6 Results for Model-2 (Performance Scores, Top-Predictors and Coefficients)

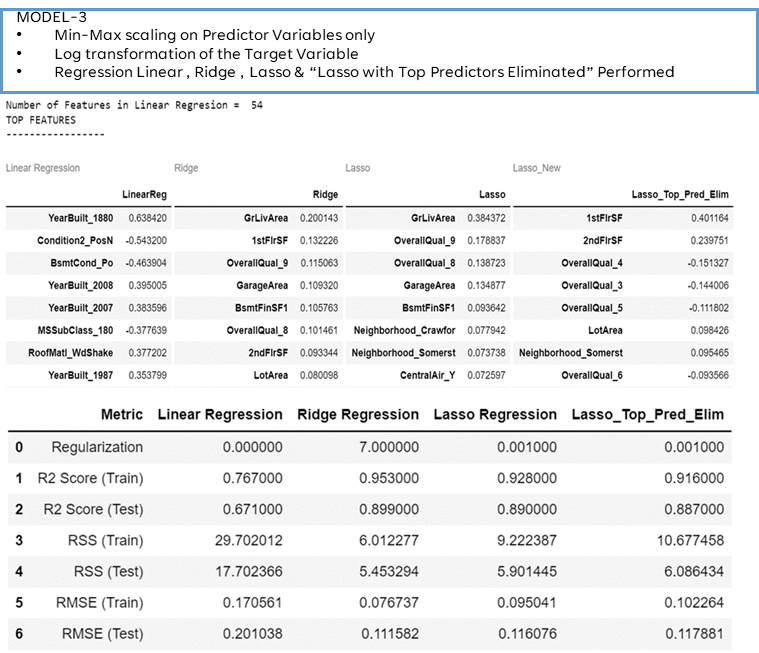


Figure 7 Results for Model-3 (Performance Scores, Top-Predictors and Coefficients)

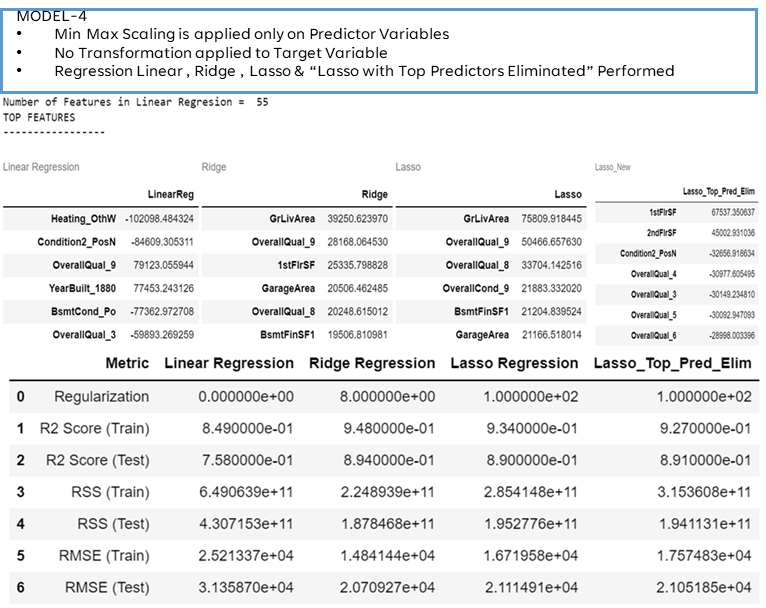


Figure 8 Results for Model-4 (Performance Scores, Top-Predictors and Coefficients)

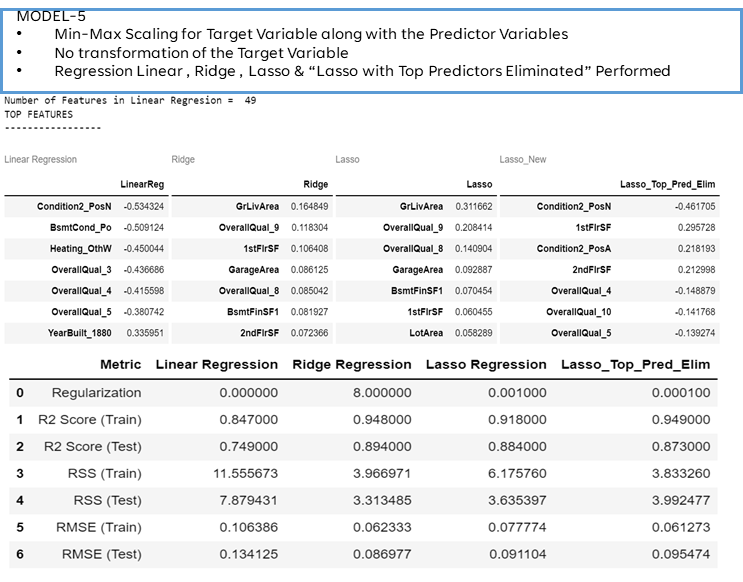


Figure 9 Results for Model-5 (Performance Scores, Top-Predictors and Coefficients)

### Across Models Inferences

* 3-4 Predictors among the Top-5 Predictors found to be mostly same across all the Model Variations for Lasso Regression. (GrLivArea, GarageArea, OverallQual\_8, OverallQual\_9)
* 3-4 Predictors among the Top-5 Predictors found to be mostly same across all the Model Variations for Ridge Regression (GrLivArea,1stFlrSF, OverallQual\_8, OverallQual\_9)
* 3 Predictors among the Top-5 Predictors are mostly same between Ridge and Lasso Regression for most of the Models.
* Top predictor variables with and without Min-Max scaling of the Target variable is same. (Model-4 and Model-5)