
# Advanced Regression

Assignment Part II - Subjective Questions

Sanjeev S

10/30/22

Master of Science in Machine Learning & Artificial Intelligence

# Table of Contents

# List of Figures

# List of Tables

# 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?

Code for these 3 questions is included in the python notebook.

## 1.1 What is the optimal value of alpha for ridge and lasso regression?

Optimal value of alpha for Ridge Regression is 50

Optimal value of alpha for Lasso is 0.0001

## 1.2 What will be the changes in the model if you choose double the value of alpha for both ridge and lasso?

New value of alpha for Ridge Regression is 100
New value of alpha for Lasso is 0.002

From Ridge - R2 score against Alpha values indicates how the model complexity varies with Alpha.

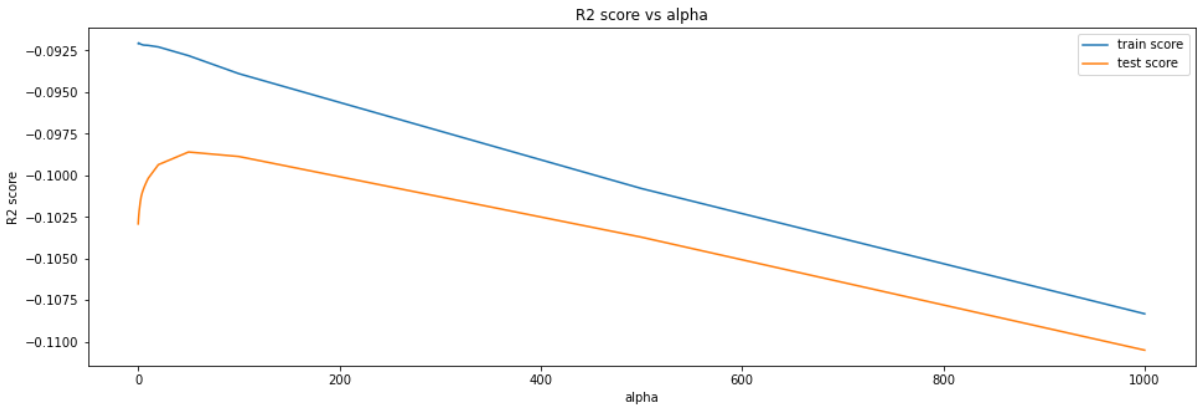


*Figure 1: Ridge - R2 Score vs Alpha*

From Lasso - R2 score against Alpha values indicates how the model coefficients reaches zero with va rying alpha values.

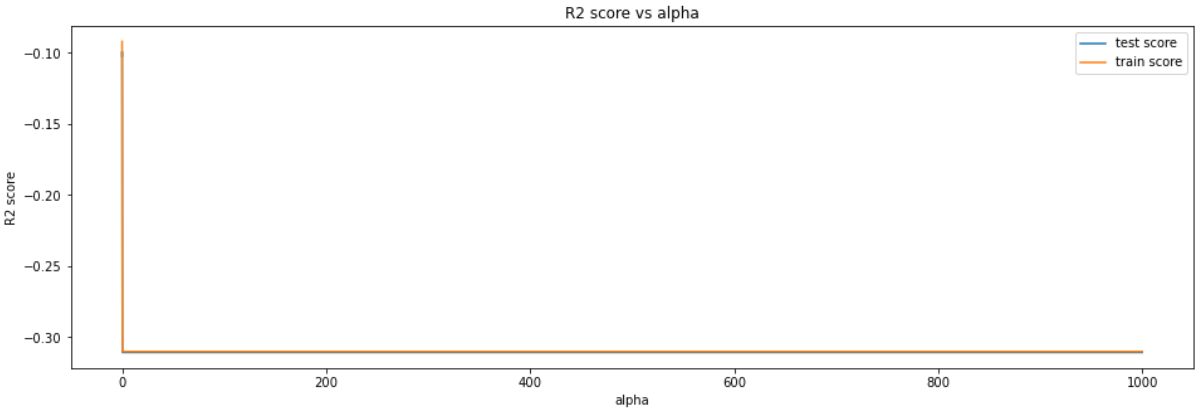


*Figure 2: Lasso - R2 Score vs Alpha*

Changes in Ridge Regression metrics:

- R2 score of train set decreased from 0.8825 to 0.8790
- R2 score of test set decreased from 0.8826 to 0.8798

Changes in Lasso metrics:

- R2 score of train set decreased from 0.8824 to 0.8789
- R2 score of test set decreased from 0.8821 to 0.8784

## 1.3 What will be the most important predictor variables after the change is implemented?

Top ten features of Ridge Regression after the change were implemented:

*Table 1: Ridge - New top predictors*

| Predictors | Coefficients (Beta) |
|---|---|
| OverallQual | 0.0837 |
| GrLivArea | 0.0613 |
| FireplaceQu | 0.0377 |
| OverallCond | 0.0361 |
| GarageCars | 0.036 |
| Condition1_Norm | 0.035 |
| FullBath | 0.0288 |
| GarageType_Attchd | 0.0263 |
| MSZoning_RL | 0.0255 |
| KitchenQual | 0.0254 |

Top ten features of Lasso after the change were implemented:

*Table 2: Lasso - New top predictors*

| **Predictors** | **Coefficients (Beta)** |
|---|---|
| OverallQual | 0.1031 |
| GrLivArea | 0.0891 |
| GarageCars | 0.048 |
| Condition1_Norm | 0.0472 |
| HouseStyle_1Story | 0.0416 |
| OverallCond | 0.0387 |
| FireplaceQu | 0.0337 |
| GarageType_Attchd | 0.0336 |
| KitchenQual | 0.0236 |
| BsmtFinType1 | 0.0232 |

## 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?

- It depends on what kind of business problem we are dealing with and use cases. Both Ridge and Lasso regression, allows some bias to get a significant decrease in variance, thereby pushing the model coefficients towards 0.
- The primary difference between Lasso and Ridge regression is their penalty term. The penalty term here is the sum of the absolute values of all the coefficients present in the model. As with Ridge regression, Lasso regression shrinks the coefficient estimates towards 0.

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

*Figure 3: Ridge Regression Equation*

$$\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

*Figure 4: Lasso Regression Equation*

- If we have a high dimensionality and high correlation in the dataset, then we would want to prefer Lasso regularisation since it penalises less important features more and makes them zero which gives you the benefit of model feature selection and would make robust predictions.
- Ridge regularisation handles the model complexity by focusing more on the important features which contribute more to the overall error than the less important features.

Generally, Lasso should perform better in situations where only a few among all the predictors that are used to build our model have a significant influence on the response variable. So, feature selection, which removes the unrelated variables, should help. But Ridge should do better when all the variables have almost the same influence on the response variable.

It is not the case that one of the techniques always performs better than the other – the choice would depend upon the data that is used for modelling.

## 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?

Code for this question is included in the python notebook.

The 5 most important predictor variables that will be excluded are:

*Table 3: Lasso - Excluded top 5 Predictors*

| Predictors | Coefficients (Beta) |
|---|---|
| OverallQual | 0.1031 |
| GrLivArea | 0.0891 |
| GarageCars | 0.048 |
| Condition1_Norm | 0.0472 |
| HouseStyle_1Story | 0.0416 |

The 5 most important predictor variables in the newly designed model are:

*Table 4: Lasso - New top 5 Predictors*

| Predictors | Coefficients (Beta) |
|---|---|
| FireplaceQu | 0.0738 |
| GarageArea | 0.056 |
| FullBath | 0.0546 |
| TotRmsAbvGrd | 0.0522 |
| Condition2_Norm | 0.0485 |

# 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?

## 4.1 How can you make sure that a model is robust and generalisable?

- A model should not be impacted by outliers in the training data. In simple terms, model performance should remain same even when variation happens in data set.
- A model should be able to predict when new, previously unseen data, drawn from the same distribution as the one used to create the model is provided.
- A model should not overfit on training data, which can affect accuracy in predicting unseen data
- A model should not underfit as well, which fails to identify any relationship between target and predictor variables

**Bias:** Bias is error in model, when the model is weak to learn from the data. High bias means model is unable to learn details in the data. Model performs poor on training and testing data.

**Variance**: Variance is error in model, when model tries to over learn from the data. High variance means model performs exceptionally well on training data as it has very well trained on this of data but 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.

In general, model should not be too complex. It should try to bring the variance to constant level by lightly adjusting bias.

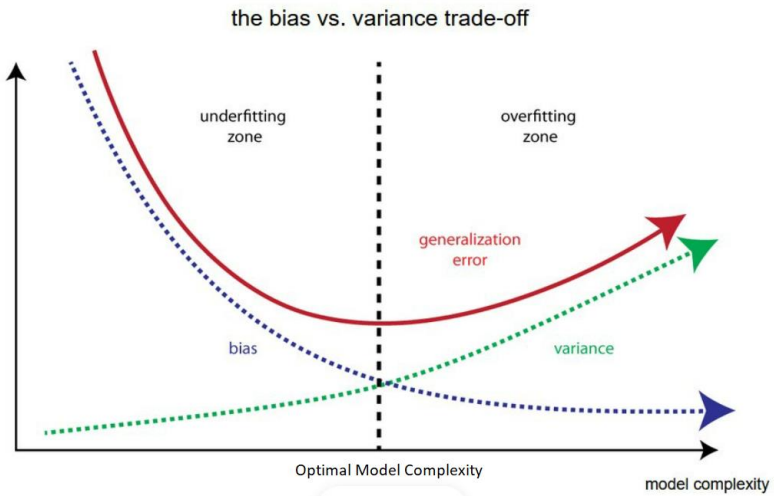


*Figure 5: Bias Variance Trade Off*

## 4.2 What are the implications of the same for the accuracy of the model and why?

- We decrease variance which will lead to some bias, to make a robust and generalized model. Adding bias will decrease the accuracy of the model.
- Regularization helps in finding an optimal solution between accuracy and robustness to build a better model.