**Logo

Description automatically generatedIntermediate Analytics**

**Module 4 RIDGE and LASSO Regression**

**ALY6015, Spring 2022**

**Week-4**

**Professor: Roy Wada**

**Submitted by: Abhinav Jain**

**Date: 05/08/2022**

Introduction

**Background**

In this paper, we used the college dataset from the ISLR library to analyze and forecast the results of the college, whether it is private or public. To get the findings, used R Language, and then execute both (Ridge and LASSO regression) models to forecast the outcome of the College dataset.

**Implementations**

Importing the college dataset into R via the ISLR module will allow insights to run Ridge and LASSO based on a model of while dependent on the lambda value. Using the regression approach, we will determine the lamda.min and lampbda.1se. Furthermore, utilizing the dplyr library description () function to analyse the dataset in R will aid in understanding statistics and cleaning the dataset as needed. We will make various graphs and obtain a basic understanding of the data to analyze the dataset.

Divide the dataset into test and train data while executing the regression model. Partition the data using the feature selection technique and fit the model using the model.matrix function with one predictor using the Ridge and lasso model. A stepwise selection will aid in analyzing the model's stability in predicting the variable, as well as identifying misclassification of false positives and false negatives. The RMSE (root mean square error) will, however, reveal the discrepancy between the observed and predicted variables.

**Task 1**: In this task Imported the dataset which has a dimension of 777 rows and 18 columns

Table 1: Descriptive Analysis of College Dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | mean | sd | min | max |
| Private | 0.727156 | 0.445708 | 0 | 1 |
| Apps | 3001.638 | 3870.201 | 81 | 48094 |
| Accept | 2018.804 | 2451.114 | 72 | 26330 |
| Enroll | 779.973 | 929.1762 | 35 | 6392 |
| Top10perc | 27.55856 | 17.64036 | 1 | 96 |
| Top25perc | 55.79665 | 19.80478 | 9 | 100 |
| F.Undergrad | 3699.907 | 4850.421 | 139 | 31643 |
| P.Undergrad | 855.2986 | 1522.432 | 1 | 21836 |
| Outstate | 10440.67 | 4023.016 | 2340 | 21700 |
| Room.Board | 4357.526 | 1096.696 | 1780 | 8124 |
| Books | 549.381 | 165.1054 | 96 | 2340 |
| Personal | 1340.642 | 677.0715 | 250 | 6800 |
| PhD | 72.66023 | 16.32815 | 8 | 103 |
| Terminal | 79.7027 | 14.72236 | 24 | 100 |
| S.F.Ratio | 14.0897 | 3.958349 | 2.5 | 39.8 |
| perc.alumni | 22.74389 | 12.3918 | 0 | 64 |
| Expend | 9660.171 | 5221.768 | 3186 | 56233 |
| Grad.Rate | 65.46332 | 17.17771 | 10 | 118 |

Plots1: Created Scatter plot which show the private university and with blue dots whereas public university as red dots with respect to the application accepted in the private university.

**Chart1: Scatterplot Applications accepted in Private University**

Chart, scatter chart

Description automatically generated

Plot 2: In this below graph purple bar show the number of application received in private university in the given data

Chart2: Number of Applications in Private University

Chart, bar chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart

Description automatically generated

**Chart 4: Boxplot**

Task 2: In this task, we split the dataset in two using an 80/20 ratio to gain the best prediction accuracy. The training set has 621 observations and 18 variables, while the test set has 156 observations and 18 variables. The training and test sets are described in the table below.

Table 2: Descriptive Analysis of Train Dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | mean | sd | min | max |
| Private | 0.723027 | 0.447864 | 0 | 1 |
| Apps | 2846.911 | 3371.385 | 81 | 21804 |
| Accept | 1912.808 | 2158.324 | 72 | 18744 |
| Enroll | 745.5733 | 861.6725 | 51 | 6392 |
| Top10perc | 27.35105 | 17.83488 | 1 | 96 |
| Top25perc | 55.36715 | 19.96789 | 9 | 100 |
| F.Undergrad | 3541.176 | 4515.978 | 139 | 31643 |
| P.Undergrad | 854.4992 | 1541.677 | 1 | 21836 |
| Outstate | 10380.54 | 4020.853 | 2580 | 21700 |
| Room.Board | 4343.641 | 1098.987 | 1780 | 8124 |
| Books | 554.7279 | 171.3886 | 110 | 2340 |
| Personal | 1347.615 | 678.5384 | 250 | 6800 |
| PhD | 72.44928 | 16.44785 | 8 | 103 |
| Terminal | 79.5153 | 14.80764 | 24 | 100 |
| S.F.Ratio | 14.2066 | 4.006399 | 2.5 | 39.8 |
| perc.alumni | 22.63285 | 12.37911 | 0 | 60 |
| Expend | 9566.366 | 5211.562 | 3186 | 56233 |
| Grad.Rate | 65.47021 | 17.31505 | 10 | 118 |

Table 3: Descriptive analysis for Test Dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | mean | sd | min | max |
| Private | 0.723027 | 0.447864 | 0 | 1 |
| Apps | 2846.911 | 3371.385 | 81 | 21804 |
| Accept | 1912.808 | 2158.324 | 72 | 18744 |
| Enroll | 745.5733 | 861.6725 | 51 | 6392 |
| Top10perc | 27.35105 | 17.83488 | 1 | 96 |
| Top25perc | 55.36715 | 19.96789 | 9 | 100 |
| F.Undergrad | 3541.176 | 4515.978 | 139 | 31643 |
| P.Undergrad | 854.4992 | 1541.677 | 1 | 21836 |
| Outstate | 10380.54 | 4020.853 | 2580 | 21700 |
| Room.Board | 4343.641 | 1098.987 | 1780 | 8124 |
| Books | 554.7279 | 171.3886 | 110 | 2340 |
| Personal | 1347.615 | 678.5384 | 250 | 6800 |
| PhD | 72.44928 | 16.44785 | 8 | 103 |
| Terminal | 79.5153 | 14.80764 | 24 | 100 |
| S.F.Ratio | 14.2066 | 4.006399 | 2.5 | 39.8 |
| perc.alumni | 22.63285 | 12.37911 | 0 | 60 |
| Expend | 9566.366 | 5211.562 | 3186 | 56233 |
| Grad.Rate | 65.47021 | 17.31505 | 10 | 118 |

**Model1 : Ridge Regression**

Finding the estimated coefficient of multiple regression models when linearly independent variables are associated in the ridge regression model. After, implementing the model.matrix function and cross validation in the dataset to fine the lambda value. Here split the dataset in x and y while using the predictive variable F.Undergrad to make the training and test set for generating the lambda minimum and lambda 1se which are as follows:

lambda.min = 3.240402

lambda.1se = 6.403549

Lambda.min denotes the model's least mean cross-validation error of 3.2404, whereas lambda.1se = 6.40 denotes the model's most regularized lambda within one standard error of the minimum. Here we are finding the amount of penalty of lambda by cross-validation. We will search for the minimum lambda.

Graph 1 (Penalty type where alpha = 0 is Ridge)

Chart, histogram

Description automatically generated

We can see the lambda is affecting estimated coefficient where lambds is at its height on the other hand coefficient is approaching to zero. We can see in the below matrix also of lambda.min and lambda.1se.

**Graph 2: Ridge Path**

Chart

Description automatically generated

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| #Model with minimum lambda   |  |  |  |  | | --- | --- | --- | --- | |  | Df | %Dev | Lambda | | 1 | 17 | 94.43 | 40.67 |   Table4: Estimated coefficient of value of variables lambda.1se  18 x 1 sparse Matrix of class "dgCMatrix" | |
|  | s0 |
| (Intercept) | -147.66 |
| PrivateYes | -549.31 |
| Apps | 0.05 |
| Accept | -0.01 |
| Enroll | 4.43 |
| Top10perc | -1.82 |
| Top25perc | 12.00 |
| P.Undergrad | 0.23 |
| Outstate | -0.04 |
| Room.Board | 0.01 |
| Books | 0.02 |
| Personal | 0.17 |
| PhD | -8.42 |
| Terminal | 14.20 |
| S.F.Ratio | 16.26 |
| perc.alumni | -1.89 |
| Expend | -0.01 |
| Grad.Rate | -8.45 |

#Model with 1se lambda

|  |  |  |  |
| --- | --- | --- | --- |
|  | Df | %Dev | Lambda |
| 1 | 17 | 93.01 | 501.4 |

Table5: Estimated coefficient of value of variables lambda.min

18 x 1 sparse Matrix of class "dgCMatrix"

|  |  |
| --- | --- |
|  | s0 |
| (Intercept) | 5.33E+01 |
| PrivateYes | -8.28E+02 |
| Apps | 9.26E-02 |
| Accept | 3.06E-01 |
| Enroll | 2.96E+00 |
| Top10perc | 6.04E+00 |
| Top25perc | 8.96E+00 |
| P.Undergrad | 3.09E-01 |
| Outstate | -5.74E-02 |
| Room.Board | -7.71E-02 |
| Books | 1.97E-01 |
| Personal | 3.09E-01 |
| PhD | -2.77E+00 |
| Terminal | 1.24E+01 |
| S.F.Ratio | 2.51E+01 |
| perc.alumni | 5.61E-01 |
| Expend | -8.62E-03 |
| Grad.Rate | -7.92E+00 |

**Root mean square error** allow us to measure the difference between the observed value and the predicted value of the regression analysis after calculating the root mean square we understand the gap is 11.35

RMSE = 11.35708

After fitting the model and finding the predictive value of the ridge model, the difference between the predictive variable is small which means the regression model fit indicates a good gap between the predictive and observed value. Whereas, if the gap will be larger the regression model suggests the model is not capable of integrating the dataset. Here we can define the min value to give a better prediction value to determine the variable.

With the train set, the value of the predictor variable is 1143.39, whereas with the test set the value of the predictor variable is 1154.747 which signifies the ridge regression model with the value of alpha = 0 satisfies the requirement with the dataset.

Train set value of the predictor variable = 1143.39

Test set value of the predictor variable =1154.747

=========================================================================================================================================================================================================

**# LASSO Regression**

Finding the estimated coefficient of multiple regression models when linearly independent variables are associated in the ridge regression model. After, implementing the model.matrix function and cross validation in the dataset to fine the lambda value. Here split the dataset in x and y while using the predictive variable F.Undergrad to make the training and test set for generating the lambda minimum and lambda 1se which are as follows:

lambda.min = 3.240402

lambda.1se = 6.403549

Lambda.min denotes the model's least mean cross-validation error of 3.2404, whereas lambda.1se = 6.40 denotes the model's most regularized lambda within one standard error of the minimum. Here we are finding the amount of penalty of lambda by cross-validation. We will search for the minimum lambda.

Graph 3:(Penalty type where alpha = 1 is Lasso)

Graphical user interface

Description automatically generated with medium confidence

Graph4: Lasso Path

Chart

Description automatically generated

After fitting the lasso regression model from the training and test dataset the coefficient of the variable is different while analysing the lambda.min and lambda.1se

Table 6: Estimated coefficient with lambda.min

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Df | | %Dev | | Lambda |
| 1 | 11 | | 94.37 | | 40.67 |
|  | | s0 | |
| (Intercept) | | 2.19E+02 | |
| PrivateYes | | -5.62E+02 | |
| Apps | | 3.40E-03 | |
| Accept | | . | |
| Enroll | | 4.64E+00 | |
| Top10perc | | . | |
| Top25perc | | 5.11E+00 | |
| P.Undergrad | | 2.05E-01 | |
| Outstate | | -3.22E-02 | |
| Room.Board | | . | |
| Books | | . | |
| Personal | | 1.21E-01 | |
| PhD | | . | |
| Terminal | | 2.73E+00 | |
| S.F.Ratio | | 9.76E+00 | |
| perc.alumni | | -6.62E-01 | |
| Expend | | . | |
| Grad.Rate | | -5.03E+00 | |

Table 7: Estimated coeffcient with lambda.1se

|  |  |  |  |
| --- | --- | --- | --- |
|  | Df | %Dev | Lambda |
| 1 | 3 | 92.55 | 501.4 |

|  |  |
| --- | --- |
|  | s0 |
| (Intercept) | 429.9325037 |
| PrivateYes | -297.629527 |
| Apps | . |
| Accept | . |
| Enroll | 4.45224988 |
| Top10perc | . |
| Top25perc | . |
| P.Undergrad | 0.05593763 |
| Outstate | . |
| Room.Board | . |
| Books | . |
| Personal | . |
| PhD | . |
| Terminal | . |
| S.F.Ratio | . |
| perc.alumni | . |
| Expend | . |
| Grad.Rate | . |

After fitting the model and finding the predictive value of the ridge model, the difference between the predictive variable is small which means the regression model fit indicates a good gap between the predictive and observed value. Whereas, if the gap will be larger the regression model suggests the model is not capable of integrating the dataset. Here we can define the min value to give a better prediction value to determine the variable.

With the train set, the value of the predictor variable is 1150.394, whereas with the test et the value of the predictor variable is 1149.647 which signifies the ridge regression model with the value of alpha = 0 satisfies the requirement with the dataset.

Train set value of the predictor variable = 1150.394

Test set value of the predictor variable =1149.647

Root mean square error allow us to measure the difference between the observed value and the predicted value of the regression analysis after calculating the root mean square we understand the gap is 0.7471506

RMSE = 0.7471506

**# Performance of the models**

Lasso and Ridge are trying to identify the relationship between the observed variable and the predictor variables, whereas both models focus on the penalty term (lambda) on the coefficients of predictors. We have seen while doing the prediction that lasso shrinks the coefficient all the way to zero resulting while doing the feature selection. On the other hand, if we look at the Ridge model coefficients close to zero, it will not set to zero if no feature selection applies while applying the model.

Even, after fitting a linear regression model we can see that the R squared and the adjusted R square is 94% by using the same predictor variable

**#Fit a Model**

Table

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

Below graph show that few variable turn to zero like application, acceptance of aaplication, Top 10 percent, books, PhD, Expend and Graduation Rate which show these variable will not help in predicting the variable if in formation will be provided in the training and test dataset.

A picture containing calendar

Description automatically generated

**Conclusion and Comparison:**

Finally, we can see from the previous research that the lasso and ridge both perform differently when gathering correlated data. Ridge shrinks the coefficient with linked variables, while lasso provides a good way to getting the variable close to zero. It stated the model's regularization route when it will be zero. In contrast, lasso removes the co-linear predictors from the fit, making it fascinating to analyze the lasso model's practicality. To comprehend the analysis and estimate the efficiency, look at the correlation between the variables. After adopting both models, we can see that the lasso has more collinearity after deleting the predictive variables, however the ridge has the coefficient with the same variable accountability. Both Lasso and Ridge are looking for a relationship between the observed variable and the predictor variables, with both models focusing on the penalty term (lambda) on predictor coefficients. While performing the prediction, we saw that lasso reduces the coefficient to zero, resulting in feature selection. If we look at the Ridge model coefficients close to zero, however, it will not be zero if no feature selection is used when applying the model.

**References:**

[1]Machine Learning for Biostatistics

Teixeira-Pinto

<https://bookdown.org/tpinto_home/Regularisation/lasso-regression.html>

[2]Ridge Regression in R (Step-by-Step)

Zach

<https://www.statology.org/ridge-regression-in-r/>

[3]Deepika Singh

Singh

<https://www.pluralsight.com/guides/linear-lasso-and-ridge-regression-with-r>

[4]Penalized Regression Essentials: Ridge, Lasso & Elastic Net

Kassambara & Chandrakumaran

<http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/153-penalized-regression-essentials-ridge-lasso-elastic-net/>

[5]Lasso vs Ridge vs Elastic Net: ML

<https://www.geeksforgeeks.org/lasso-vs-ridge-vs-elastic-net-ml/>

[6]R-bloggers

R-Ladies Cologne joins the family. R-Ladies Cologne is a new R-Ladies chapter which was founded in September 2021. It joins R-Ladies' mission to promote and support gender equality in the field. Together with Luciana and Gabe, we have hosted three events so far.

[https://www.r-bloggers.com](https://www.r-bloggers.com/)