STA363_Lab_6

Zishan Shao

2022-10-11

Question 1:

Is this a regression or classification task? Based on this, what metric will we likely use to assess predictive accuracy?

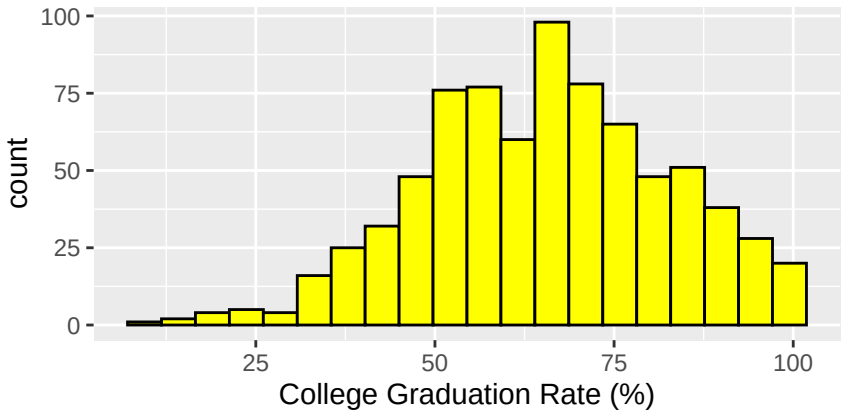
ANS: The graduation rate is a continuous response variable. Therefore, we are dealing with a regression task to make prediction based on features given. We are expected to use the root mean squared error (RMSE) or mean squared error (MSE), which we try to minimize the metric.

Question 2:

Make a visualization to explore the distribution of graduation rate. What is the smallest value of graduation rate? The largest?

ANS:

Figure 1:




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    10.0    53.0    65.0    65.4    78.0   100.0
```

Based on the distribution of the Y, we could see that the smallest graduation rate is 10.0% and the maximum graduation rate is 100.0%.

Question 3:

Compute the MSE (training) and RMSE (training) for these data.

ANS:

```
## [1] "The MSE: 291.512300"
```

```
## [1] "The RMSE: 17.073731"
```

Based on the outputs, MSE should be around 291.5, while the RMSE should be about 17.07.

Question 4:

Train an LSLR (OLS) regression model using the whole data set and the single feature X = the student faculty ratio. Find the MSE (training) and RMSE (training). How much has our training predictive accuracy improved from the values you got in Question 3?

ANS:

```
##
## Call:
## lm(formula = Grad.Rate ~ S.F.Ratio, data = collegedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.377 -11.038   0.351  11.588  47.938
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  84.1616     2.1581  38.998   <2e-16 ***
## S.F.Ratio    -1.3319     0.1475  -9.032   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.26 on 774 degrees of freedom
## Multiple R-squared:  0.09535,    Adjusted R-squared:  0.09418
## F-statistic: 81.58 on 1 and 774 DF,  p-value: < 2.2e-16
```

The RMSE in this case is 16.26 and the MSE is about 264.39. Compared with the MSE in Q3, which is 291.5, it has been improved by (291.5 - 264.39) = 27.1124. For RMSE, it has been improved by about 0.814.

Question 5:

Which column in the data set cannot be used as a feature (aside from the response variable)? Explain why this column cannot be used.

ANS: The first column, X, could not be used as a feature because it is the name of the school, which could not provide any explanation of the value of the graduation rate.

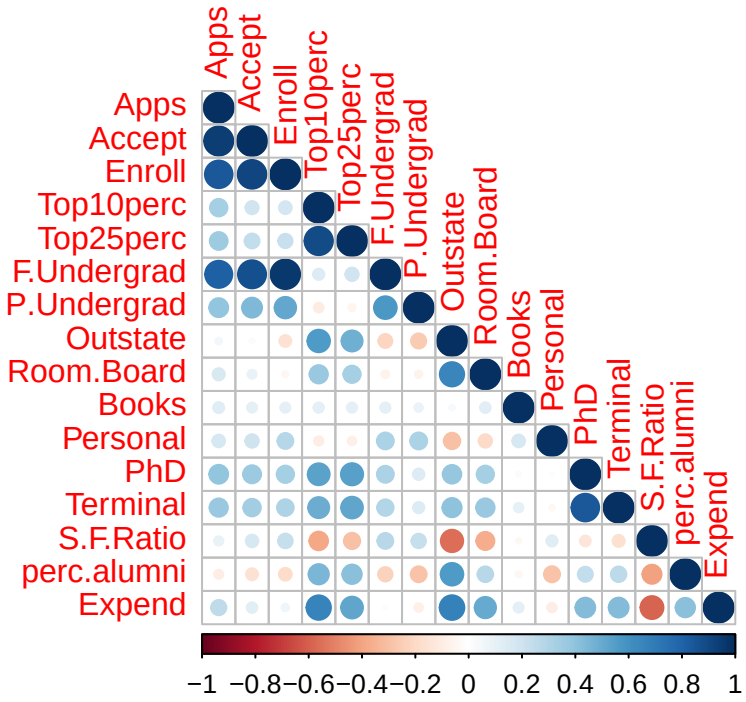
Question 6:

Build a correlation plot (you can choose the styling you like best!) to explore the correlations in the features in this data set.

ANS:

```
## corrplot 0.92 loaded
```

Figure 2: correlation plot



The correlation plot shows that some of the features are strongly related with each other.

Question 7:

Based on the plot in Question 6, why would you suggest we start with Ridge Regression instead of LSLR?

ANS: From the correlation plot, there are some variables that are strongly related. For instance, the accepted applications and the enrolled students number, which has correlation more than 0.8. Due to this densely dependency of features, OLS may not perform well and we should use ridge regression, which is a penalized regression model.

Question 8:

Using the code above, create the design matrix. State the dimensions of this matrix.

ANS:

```
## Loading required package: Matrix

## Loaded glmnet 4.1-4

## [1] 776  18
```

The dimension of the matrix is [776, 18].

Question 9:

Suppose we let lambda = 2. Adapting the code above, what is Beta_hat ridge?

ANS:

Table 1: Coefficients with Tuning Parameter = 2

	Ridge
(Intercept)	33.8140140
PrivateYes	3.5800642
Apps	0.0005856
Accept	0.0002645
Enroll	0.0002597
Top10perc	0.0930465
Top25perc	0.1109139
F.Undergrad	-0.0001430
P.Undergrad	-0.0013784
Outstate	0.0007933
Room.Board	0.0017864
Books	-0.0024418
Personal	-0.0016920
PhD	0.0707884
Terminal	-0.0298726
S.F.Ratio	0.0604338
perc.alumni	0.2533070
Expend	-0.0002480

Question 10:

What is the sum of the coefficients (excluding the intercept) when lambda=2? Why do we exclude the intercept? Well, the intercept is not actually penalized in ridge.

ANS:

```
## [1] "The sum of the coefficients: 4.136468"
```

Question 11:

Suppose we let lambda = 200. Adapting the code above, what is beta Ridge?

ANS:

Table 2: Coefficients with Tuning Parameter = 200

	Ridge
(Intercept)	56.3360365
PrivateYes	0.8082141
Apps	0.0000433
Accept	0.0000333
Enroll	-0.0000179
Top10perc	0.0292154
Top25perc	0.0255111
F.Undergrad	-0.0000166
P.Undergrad	-0.0001935
Outstate	0.0001500
Room.Board	0.0004040
Books	-0.0000980
Personal	-0.0004365
PhD	0.0194660
Terminal	0.0193521
S.F.Ratio	-0.0723084
perc.alumni	0.0432018
Expend	0.0000700

Question 12:

What is the sum of the coefficients (excluding the intercept) when lambda=200?

ANS:

```
## [1] "The sum of the coefficients (lambda = 200): 0.872590"
```

Question 13:

What do you notice happens to the sum of the coefficients as lambda increases?

ANS: the sum of the coefficients decreased significantly from about 4.14 to 0.87. The lambda serve as a penalty term that essentially constains the size of the coefficients.

Question 14:

Briefly explain the process of choosing the tuning parameter in ridge regression. Use words, not code!

ANS: Top choose the tuning parameter, we need to use a loop to go through a range of possible values for the tuning parameter and choose the one of the value that the estimated $\hat{\beta}$ will minimize the RSS + shrinkage penalty ($RSS + \lambda\hat{\beta}^T\hat{\beta}$).

Question 15:

Why do we need to set a random seed here?

ANS: This is because we used the cross validation technique in this case (cv stands for cross validation). Therefore, we need to set a seed so that the experiment was reproducable.

Question 16:

Why do we want to make sure to include

$$\lambda$$

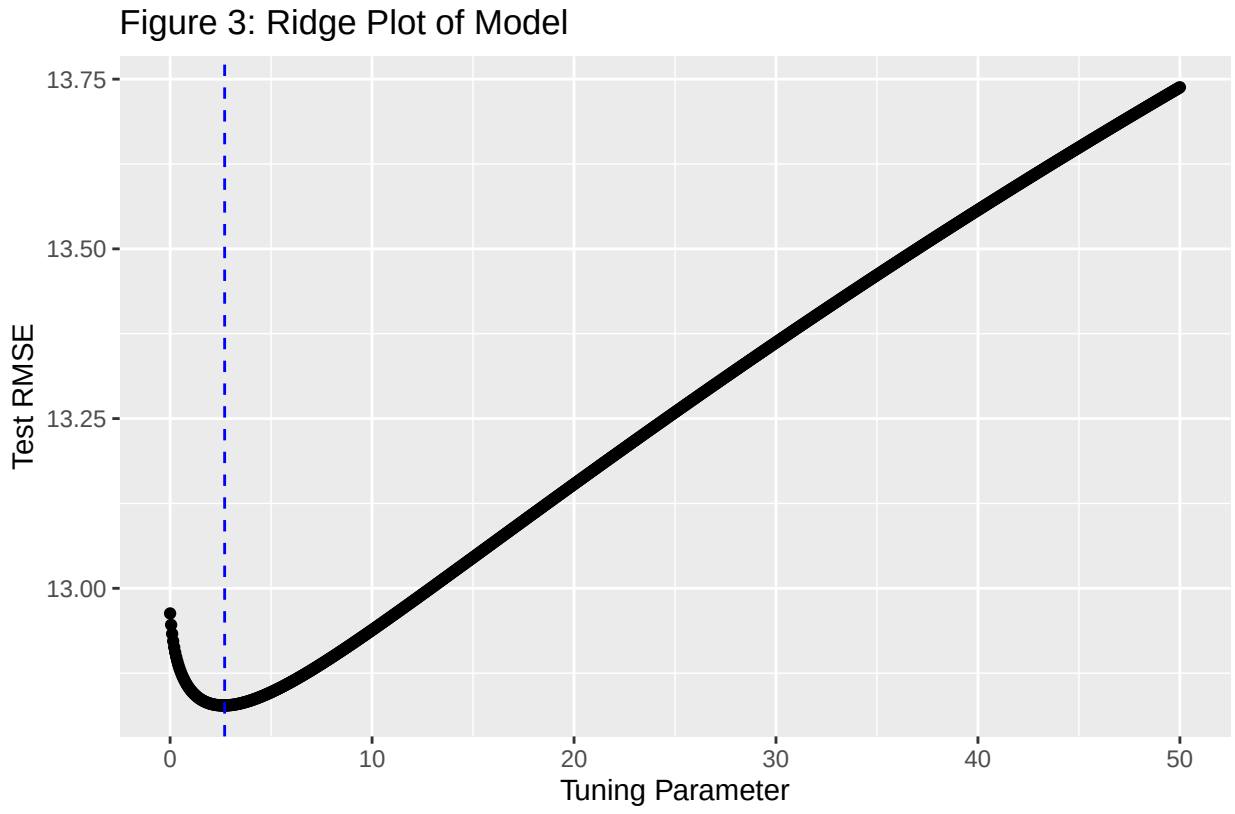
= 0 in our list of possible tuning parameters?

ANS: The $\lambda = 0$ is the case of ordinary linear regression. We add the $\lambda = 0$ because we want to see if it is actually necessary to apply a penalized model. If the OLS works well, this means the features are not strongly correlated so we do not need a penalized model.

Question 17:

Plot your results using the code above, and show your result. Approximately what choice of lambda gives the best value of our test metric?

ANS:




```
## [1] "Minimum of Lambda: 2.70"
```

Question 18:

Why might it be helpful to look at the plot instead of just using ridge.mod's lambda.min all the time?

ANS: This is because we can see the trend of RMSE in changing value of the λ. The minimum value could, in some cases, be the lowest only because of the range limit (it could continue decrease later, which can be determined by the plot but not the minimum value). Therefore, the plot is also necessary along with the minimum value.

Question 19:

Using your choice of lambda, what is the estimated test RMSE?

ANS:

```
## [1] "Estimated Test RMSE: 12.83"
```

The minimum test MSE is 164.5418 for $\lambda = 2.7$, so the RMSE is around 12.83.

Question 20:

How much has our RMSE improved from the values you got in Question 3? This allows us to see how much better our model does than using no feature information.

ANS: The RMSE in Q3 is around 17.07, compared with 12.83 in Q19, it has been improved by 4.24.

Question 21:

Using the code above as a template, train your ridge regression model. Then, train the LSLR model using the same code template and call the result lslr.final. Once we have trained the model, we can make a data frame holding the coefficients for both LSLR and ridge using the following code. Show the resultant data frame as the answer to this question. Make sure you have formatted it using knitr::kable().

ANS:

Table 3: Final Coefficients

	LSLR	Ridge
(Intercept)	32.8167769	34.1566828
PrivateYes	3.5008372	3.5723435
Apps	0.0013787	0.0005231
Accept	-0.0009829	0.0002962
Enroll	0.0022849	0.0002151
Top10perc	0.0508966	0.0937669
Top25perc	0.1351723	0.1081846
F.Undergrad	-0.0004258	-0.0001233
P.Undergrad	-0.0014938	-0.0013396
Outstate	0.0010399	0.0007551
Room.Board	0.0018326	0.0017507
Books	-0.0023240	-0.0024053

	LSLR	Ridge
Personal	-0.0015421	-0.0017193
PhD	0.1098040	0.0656695
Terminal	-0.0775325	-0.0209880
S.F.Ratio	0.0712191	0.0508069
perc.alumni	0.2754379	0.2451920
Expend	-0.0004725	-0.0002064

Question 22:

State the estimated test RMSE for both of the two trained models (LSLR or ridge regression). Which has the best predictive accuracy? By how much (in percent difference)?

ANS:

```
## [1] "Percent of Difference: 1.06 "
```

The RMSE of ridge regression is 12.83, while the RMSE of LSLR is 12.96. The ridge regression has the best performance by 1% difference.