Module 3 | Assignment: GLM and Logistic Regression

Trang Tran

CPS, Northeastern University

ALY6015 | Intermediate Analytics

Professor Steve Morin

May 03, 2023

Introduction

We are looking at a college dataset from ISLR library that contains a large number of US Colleges from the 1995 issue of US News and World Report. The dataset has 18 variables with 777 observations. The variables dictionary attached as below:

Figure 1: Data Dictionary

`Private`
A factor with levels `No` and `Yes` indicating private or public university
`Apps`
Number of applications received
`Accept`
Number of applications accepted
`Enroll`
Number of new students enrolled
`Top10perc`
Pct. new students from top 10% of H.S. class
`Top25perc`
Pct. new students from top 25% of H.S. class
`F.Undergrad`
Number of fulltime undergraduates
`P.Undergrad`
Number of parttime undergraduates
`Outstate`
Out-of-state tuition
`Room.Board`
Room and board costs
`Books`
Estimated book costs
`Personal`
Estimated personal spending
`PhD`
Pct. of faculty with Ph.D.'s
`Terminal`
Pct. of faculty with terminal degree
`S.F.Ratio`
Student/faculty ratio
`perc.alumni`
Pct. alumni who donate
`Expend`
Instructional expenditure per student
`Grad.Rate`
Graduation rate

Analysis

Firstly, we create train and test sets with the ratio of 80/20 from the College dataset by using createDataPartition() function that helps to preserve the ratio or balance of the factor classes [Private]. The train and test sets have 622 and 155 observations respectively.

Secondly, we perform an EDA on the train set by using the skim() function. The descriptive table and plots are shown as follow:

Figure 2: Descriptive Statistic of the train set

— Variable type: factor

	skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
1	Private	0	1	FALSE	2	Yes: 452, No: 170

— Variable type: numeric

	skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
1	Apps	0	1	3110.	3995.	141	838	1617	3754.	48094	
2	Accept	0	1	2100.	2574.	118	665	1234.	2451.	26330	
3	Enroll	0	1	806.	972.	46	266	456	890.	6392	
4	Top10perc	0	1	27.8	17.9	1	15	24	36	96	
5	Top25perc	0	1	56.1	20.0	12	41	55	69.8	100	
6	F.Undergrad	0	1	3833.	5072.	199	1048.	1790	3993	31643	
7	P.Undergrad	0	1	888.	1597.	1	95.8	346	1082.	21836	
8	Outstate	0	1	10579.	4109.	2340	7436.	10206.	13106.	21700	
9	Room.Board	0	1	4376.	1110.	1780	3600	4201	5059	8124	
10	Books	0	1	553.	174.	96	475	514.	600	2340	
11	Personal	0	1	1332.	685.	250	850	1200	1655	6800	
12	PhD	0	1	73.2	16.4	8	63	76	86	100	
13	Terminal	0	1	80.2	14.7	24	72	83	92	100	
14	S.F.Ratio	0	1	14.0	3.85	2.9	11.4	13.6	16.5	28.8	
15	perc.alumni	0	1	22.9	12.4	0	13	21	31	64	
16	Expend	0	1	9799.	5335.	3480	6830	8538.	10890.	56233	
17	Grad.Rate	0	1	66.2	17.2	10	54	66	78	118	

Figure 3: Bar plot of 'Private'

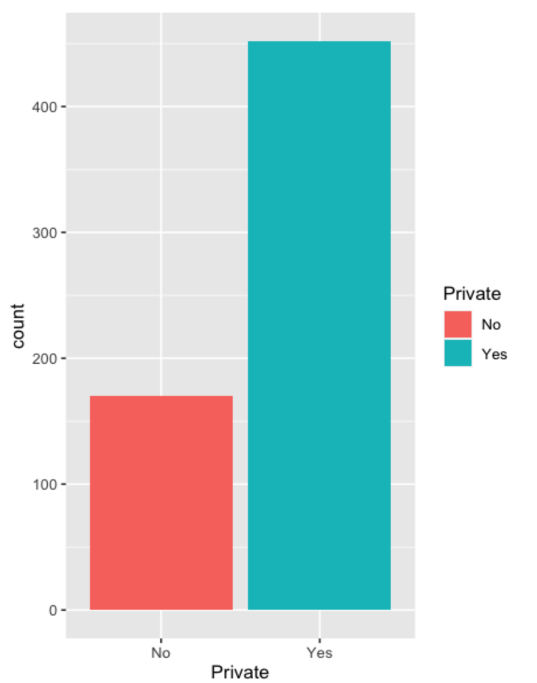
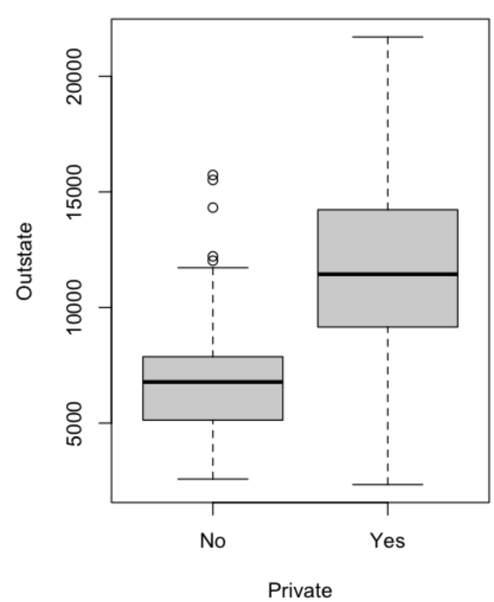


Figure 3: Boxplot of 'Outstate' by 'Private'



Private schools outnumber non-private schools by more than two times. Additionally, there are several outliers in the 'Outstate' variable (out-of-state tuition) for non-private schools that exceed $12,000.

Next, we try to fit a logistic regression model to the training set using four predictors: ‘Top10perc’, ‘Outstate’, ‘PhD’, and ‘S.F.Ratio’. By looking at p-values, the model (Figure 4) reports that only Outstate, PhD, and S.F.Ratio are statistically significant predictors of Private, while Top10perc is not significant. The coefficients for these significant predictors tell us that:

- For a one-unit increase in Outstate (Out-of-state tuition), the log-odds of a school being private increases by 0.000725.
- For a one-unit increase in PhD (percentage of faculty with PhDs), the log-odds of a school being private decreases by 0.123040.
- For a one-unit increase in S.F.Ratio (student/faculty ratio), the log-odds of a school being private decreases by 0.222718.
- The intercept of the model is 6.965594, which represents the log-odds of a school being private when all other predictors are equal to 0.

Figure 4: Fit a logistic regression model to the train set

```
Call:
glm(formula = Private ~ Top10perc + Outstate + PhD + S.F.Ratio,
    family = binomial(link = "logit"), data = caret_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2523  -0.2529   0.1084   0.3465   3.2592

Coefficients:
               Estimate  Std. Error z value             Pr(>|z|)
(Intercept)  6.96559449  1.32186033   5.270 0.000000136766848383 ***
Top10perc   -0.00106474  0.01178633  -0.090                0.928
Outstate     0.00072484  0.00007203  10.064 < 0.0000000000000002 ***
PhD         -0.12304030  0.01514098  -8.126 0.000000000000000443 ***
S.F.Ratio   -0.22271831  0.04636829  -4.803 0.000001561138508559 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 729.64  on 621  degrees of freedom
Residual deviance: 322.72  on 617  degrees of freedom
AIC: 332.72

Number of Fisher Scoring iterations: 7
```

 # Create a confusion matrix and report the results of model 1 for both sets

Figure 5: Confusion matrix for the train set

```
          Reference
Prediction  No Yes
       No  133  29
       Yes  37 423

               Accuracy : 0.8939
                 95% CI : (0.867, 0.917)
    No Information Rate : 0.7267
    P-Value [Acc > NIR] : <0.0000000000000002

                  Kappa : 0.7289

 Mcnemar's Test P-Value : 0.3889

            Sensitivity : 0.9358
            Specificity : 0.7824
         Pos Pred Value : 0.9196
         Neg Pred Value : 0.8210
             Prevalence : 0.7267
         Detection Rate : 0.6801
   Detection Prevalence : 0.7395
      Balanced Accuracy : 0.8591

       'Positive' Class : Yes
```

Figure 7: Confusion matrix for the test set

```
          Reference
Prediction  No Yes
       No   37   8
       Yes   5 105

               Accuracy : 0.9161
                 95% CI : (0.8608, 0.9546)
    No Information Rate : 0.729
    P-Value [Acc > NIR] : 0.000000005148

                  Kappa : 0.7924

 Mcnemar's Test P-Value : 0.5791

            Sensitivity : 0.9292
            Specificity : 0.8810
         Pos Pred Value : 0.9545
         Neg Pred Value : 0.8222
             Prevalence : 0.7290
         Detection Rate : 0.6774
   Detection Prevalence : 0.7097
      Balanced Accuracy : 0.9051

       'Positive' Class : Yes
```

Figure 5 shows the confusion matrix for the train set. The sensitivity of the model is 0.9358, indicating that it correctly classified 93.58% of the private schools in the test set, while the specificity is 0.7824, indicating that it correctly classified 78.24% of the non-private schools.

The positive predictive value (PPV) is 0.9196, meaning that out of all the instances predicted as private schools by the model, 91.96% are actually private schools. The negative predictive value (NPV) is 0.8210, meaning that out of all the instances predicted as non-private schools by the model, 82.10% are actually non-private schools. In terms of misclassifications, I think false negatives (predicting a non-private school as private) could be more damaging for this analysis as it could result in a non-private school missing out on potential funding or resources that they would have received if they were correctly classified as non-private or public schools.

Figure 6 presents the confusion matrix for the test set. The model correctly classified 913% of the private schools as private (sensitivity) and 88% of the non-private schools as non-private (specificity). The overall accuracy of the model on the test set was 92%, meaning that it correctly classified 92% of all schools in the test set. The positive predictive value (PPV) of the model was 95%, which means that when the model predicted a school to be private, it was correct 95% of the time. The negative predictive value (NPV) was 82%, which means that when the model predicted a school to be non-private, it was correct 82% of the time.

 # Report and interpret metrics for Accuracy, Precision, Recall, and Specificity.

```
> conf_matrix_train$overall["Accuracy"]
 Accuracy
0.8938907
> conf_matrix_train$byClass["Precision"]
Precision
0.9195652
> conf_matrix_train$byClass["Recall"]
   Recall
0.9358407
> conf_matrix_train$byClass["Specificity"]
Specificity
  0.7823529
```

```
> conf_matrix_test$overall["Accuracy"]
Accuracy
0.916129
> conf_matrix_test$byClass["Precision"]
Precision
0.9545455
> conf_matrix_test$byClass["Recall"]
   Recall
0.9292035
> conf_matrix_test$byClass["Specificity"]
Specificity
  0.8809524
```

For the training set, the metrics are:

- Accuracy is 0.8939 or 89.39%. It means that 89.39% of the time, the model correctly classified the schools as private or non-private.
- The precision of this model for private schools is 0.9196, meaning that among all schools predicted as private, 91.96% of them are actually private schools.
- The recall for private schools is 0.9358, meaning that among all actual private schools, 93.58% of them were correctly identified as private schools by the model.
- The specificity of this model for non-private schools is 0.7824, meaning that among all schools predicted as non-private, 78.24% of them are actually non-private schools.

For the test set, the metrics are:

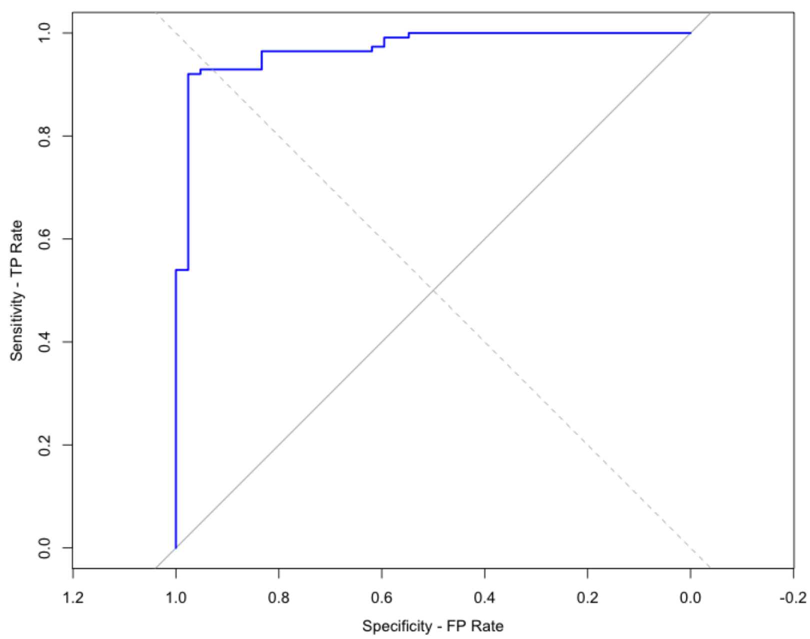
- Accuracy: the model correctly predicted 91.61% of the cases in the test set.

- The precision for the positive class is 95.45%, which means that out of all the predictions made for the positive class, 95.45% were correct.
- The recall for the positive class is 92.92%, which means that out of all the actual positive cases, the model correctly identified 92.92%.
- The specificity for the negative class is 88.10%, which means that out of all the actual negative cases, the model correctly identified 88.10%.

#Plot and interpret the ROC curve for the test set

Figure 8 shows a good look of a ROC curve that represents the trade-off between TPR and FPR for different threshold values.

Figure 8: ROC plot



#Calculate and interpret the AUC for the test set

The closer the AUC is to 1, the better the model is at distinguishing between positive and negative classes. In this case, the AUC for the test set is 0.9701, which indicates a very good performance of the model.

Conclusion

In this analysis, we used logistic regression to predict whether a university is private or not based on several predictors such as the percentage of students in the top 10% of their high school class, out-of-state tuition, the percentage of faculty with a Ph.D., and the student-to-faculty ratio. The model achieved an accuracy of 89.39% on the training set and 91.61% on the test set, indicating good predictive performance.

The confusion matrices and associated metrics showed that the model had high precision, recall, and specificity for both the training and test sets, indicating a low false positive rate and a low false negative rate. This means that the model can correctly identify most private and non-private universities.

The ROC curve for the test set showed that the model had good discrimination power, with an AUC of 0.9701. This indicates that the model is able to distinguish between private and non-private universities with a high degree of accuracy.

Overall, the logistic regression model performed well in predicting the private/public status of universities based on the given predictors, with high accuracy and discrimination power. However, further validation and testing may be needed to confirm the generalizability of the model to other datasets.

References

1. ISLR. Dataset. Retrieved May 03, 2023. https://rdrr.io/cran/ISLR/man/College.html#heading-0