Logistic Regression with Graduate Student

Md Mominul Islam

2/22/2022

Table of Contents

Project Question

In the file you will find a few hundred pieces of data. This is a fictitious student admissions dataset. The columns are defined as follows:

- ID: This is the ID
- Cohort: Semester of Enrollment Admission: 1= student applied and was admitted to SDSU. 0 = student applied and was not admitted.
- Enrolled: 1= admitted student subsequently enrolled in courses and was counted as part of the freshman 'cohort'. 0 = otherwise.
- HSGPA: This is the student's High School GPA.
- ACT: This is the student's ACT score for entrance examination.
- Graduated: 1 = graduated

I believe that we can predict with some level of accuracy whether or not a student will graduate. Create two models model using the data given that assigned a probability of graduation.

1. Logistic regression
2. A tree

General code syntax for a logistic regression is:
Mymodel=glm(Target ~ Var1 + Var2 + Var3, family=binomial, data=mydataset)

You should submit:

1. R code file….not an RMD file.
2. A powerpoint presentation of not more than 8 slides.
3. A Word document report.

Your analysis, model, and documentation should include.

1. Examination of the data, decisions made, and justification for those decisions.
2. Your model and any results you might have. Think in terms of answering the question 'did it work?'
3. If called on to present your model, I may ask you to calculate a predicted probability for a given student.

Project Summary

- Understand and compare logistic regression model and Tree based model.
- Using the fictitious SDSU Student data, we have compared tree based model and logistic regression model.
- Used Summary statistics and exploratory data analysis for categorical and continuous data in our SDSU student data set.
- In logistic regression response is a categorical variable with two levels. We code them with 1 or 0. In case of mathematical modeling, we model the probability of one level.
- A logistic regression model using glm() with Graduated as dependent (target) variable and HSGPA and ACT as independent (predictor) variables.
- A tree model using rpart() with Graduated as dependent (target) variable and HSGPA and ACT as independent (predictor) variables.
- Lastly, we compared our two models based on results that we got.

Data Cleaning

Our SDSU Student data consists of 350 rows and 7 columns.

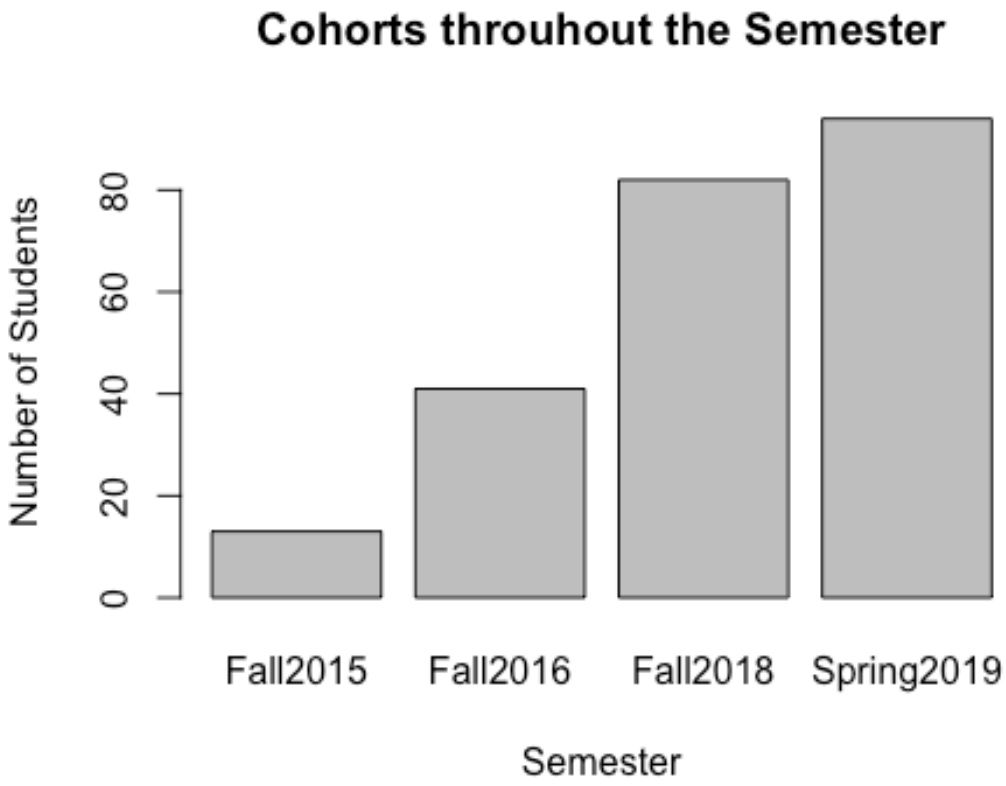
We are considering 'Graduated' as our dependent variable and other variables as independent variables. At first, we had a glimpse in our data set. We saw some missing values and null values in our dataset. Also, there were two typos in our HSGPA column.

Out of 350 observations, 112 missing values in our data set. We have cleaned the missing values.

Exploratory Analysis of Categorical Data

Cohort Column

Fall 2015 Session	Fall 2016 Session	Fall 2018 Session	Spring 2019
13	41	82	94



If we look at our cohort column, we can see that from 2015 to 2019, no. of enrollment increases.

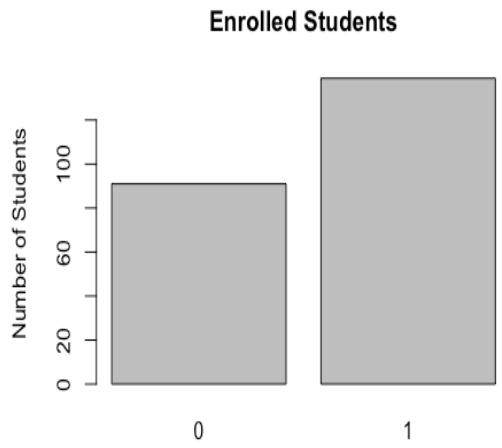
Enrollment and Graduated Column

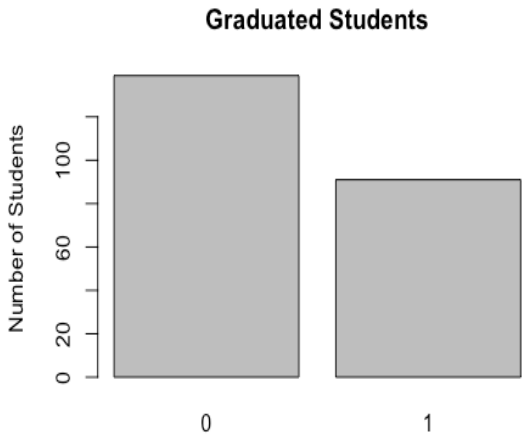
0 (Otherwise)	1 (Being Enrolled)
91	139

We looked at enrollment status of students who got admission in SDSU. Out of 230 admitted students, 139 are being enrolled.

We have got similar results but opposite for our Graduated column.

Not Graduated	Graduated
139	91





We have 230 students, out of which 91 of them graduated and 139 didn't.

After looking at the structure of our data set, we have seen that ID, Cohort, ACT are character variables and we converted ACT to numeric.

Variables	Type
ID	Character
Cohort	Character
Admission	Integer
Enrolled	Character
High School GPA	Number
ACT	Character
Graduated	Integer

Summary Statistics

Parameters	HSGPA	ACT
Minimum Value	0.0	14.00
First Quartile	3.235	22.00
Median	3.460	25.00
Mean	3.435	25.34
Third Quartile	3.770	29.00
Maximum Value	4.00	34.00

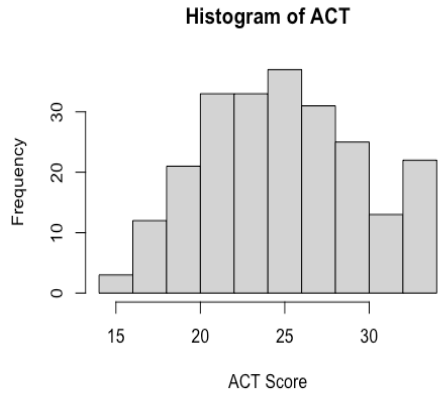
- From the output, we can infer that the average high school GPA of the students in South Dakota State University is 3.435, the average ACT score is 25.34.
- The middle most value of a variable in a data is its median value. From the output depicted in the table, we can infer that the median high school GPA of the students is 3.460, which is pretty close from the mean value, the median ACT is 25.00.
- Skewness is a measure of symmetry, or the lack of it, for a real-valued random variable about its mean. The skewness value can be positive, negative, or undefined. In a perfectly symmetrical distribution, the mean, median, and the mode will all have the same value.
- Here, for Act score, mean and median are almost equal so we can say that data is normally distributed

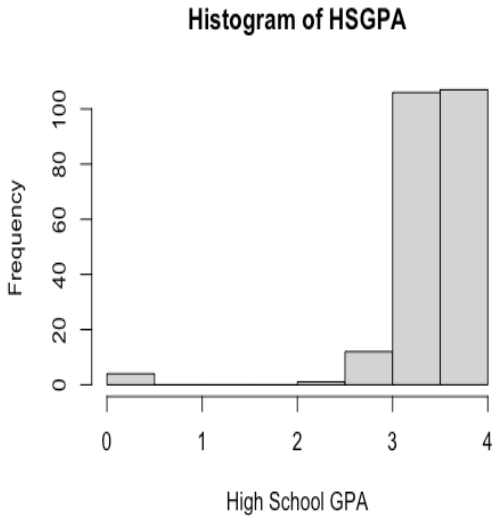
Two-way Contingency Table

	Enrolled	
Graduated	Otherwise	Enrolled
Not Graduated	91	48
Graduated	0	91

From the contingency table, we can say that out of all enrolled students, 91 of them graduated.

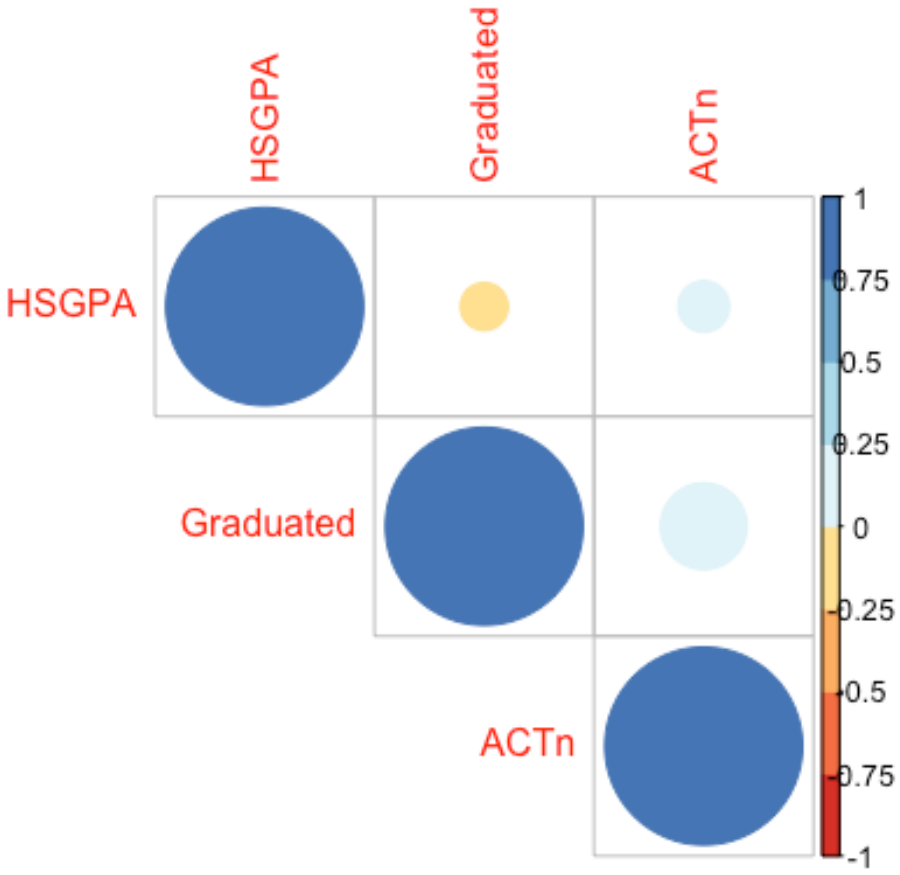
Histogram for Numeric Variable





From the ACT histogram, we can say that data is normally distributed and has uniform distribution. Whereas HSGPA is left skewed.

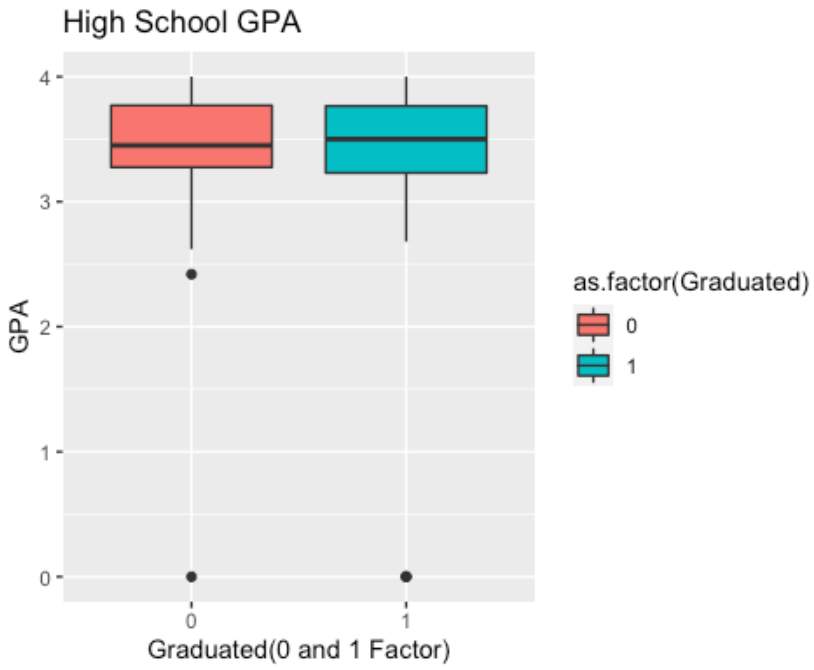
Correlation Plot

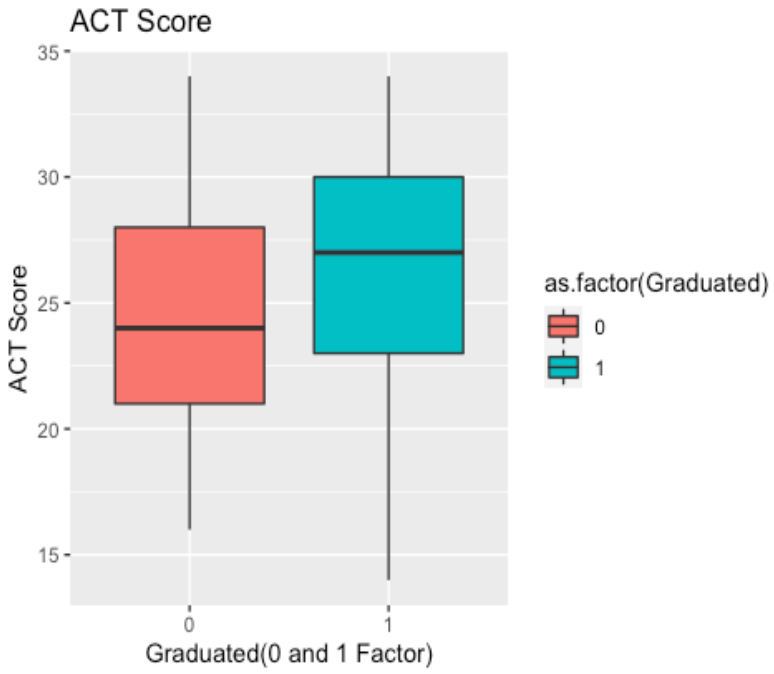


From the correlation plot depicted above, we can see weak correlation in between HSGPA and Graduated variables. On the contrary, we can see slightly positive correlation between ACT score and Graduated variables.

Boxplot of Numeric Variables

Next two figures show the boxplot of two numeric variables. For better visualization, we produced two separate plots for each group of the target





Interpretation of the box plot (alternatively box and whisker plot) rests in understanding that it provides a graphical representation of a five number summary.

The box encompasses 50% of the observations. So, we can say that 50% of the students have GPA in between 3.00~4.00. There are some extreme values in our data set. Also from our ACT boxplot, we can say that 50% of the students have ACT score in between 20~30.

Logistic Model

For logistic regression, our response is a probability, which has an interval of [0 1]. So we want such a function, that maps this interval to R.

Such a function is logit function and defined as follows:

$$logit(P) = \log\left(\frac{P}{1-P}\right)$$

Here, P means probability of being graduated.

	Estimate	Std. Error	t-value	Pr(>|t|)
Intercept	-1.76770	1.11539	-1.585	0.11301
HSGPA	-0.27025	0.25402	-1.064	0.28739
ACT	0.08906	0.03061	2.909	0.00362 **
Enrolled	20.30775	1093.46	0.019	0.98518

Initially to build our model, we have selected Graduated as Target variable, and HSGPA and ACT as predictor variables. After modeling, we found ACT score is significant in our data set with p-value less than the $\alpha = 0.05$.

Parameter Interpretation

- $b_{HSGPA} = -0.270$ and exp (-0.270) = 0.76. So, with unit increase of high school GPA, the odd of a student being graduated will increase by 76.33%.
- $b_{ACT} = -0.08906$ and exp (0.08906) = 1.093. So, with unit increase of ACT score, the odd of a student being graduated will increase by 109.3%

Confidence Interval

	2.5%	97.5%
(Intercept)	-3.96	0.455
HSGPA	-0.816	0.2110
ACT	0.030	0.150

Log Odd

The term log (P/1−P) is called log-odd. It is the natural log of the ratio of probability of something to happen (P) and the probability of that thing NOT to happen (1 − P).

Error rate

As it was mentioned, error rate is one way ti assess the performance of a logistic model. But we need to

classify the predictions as “Graduated” or “Not Graduated”. Let us use 0.5 as our cutoff. Any value less than 0.5 will be classified as not Graduated (0), and the others will be classified as Graduated (1)

Confusion/contingency matrix

	Predicted Class	
Original Class	0	1
0	123	16
1	71	20

From the contingency table, we can see that 87 were predicted wrong out of 230. So the error rate here is 87/230= 37.82%.

Validation of model: Training and Testing

For validation of the model, we almost always save a portion of the data to validate our model. We do not touch this portion while making the model. We took 70% of the data to train our model.

Contingency Table

	Predicted Class	
Original Class	0	1
0	34	8
1	21	6

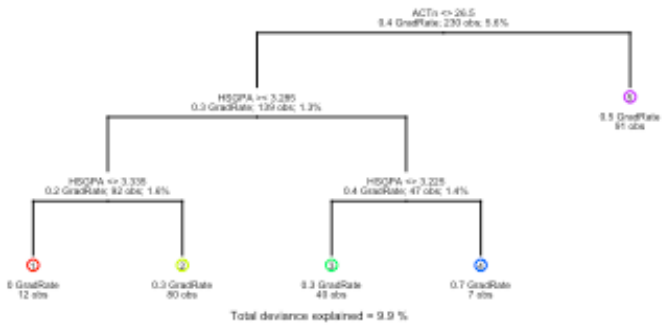
From the contingency table, we can see that 29 were predicted wrong out of 69. So the error rate here is 29/69= 42.02%.

Tree Based Model

High School GPA and ACT Score are dependent variable and Graduated as Target Variable in our tree regression model in R-Studio.

From our tree-based model, the ACT variable is contained in the top node, which is divided into two nodes initially.

- ACT Score greater than 26.5 will be likely to graduate with 50% chance.
- But with an ACT score less than 26.5, further divided into left node. We will now have HSGPA section.
- There are 92 observations with HSGPA less than 3.265 with 20% Graduation Rate.
- We ended up with a total of 5 nodes at the conclusion of the process. The tree diagram of the tree



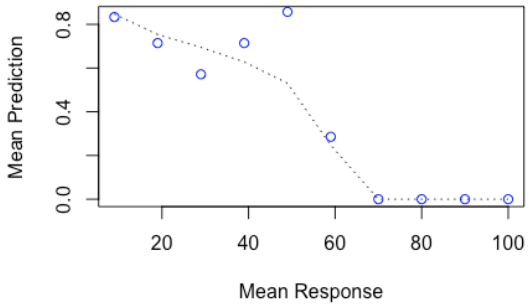
Comparison between Two models

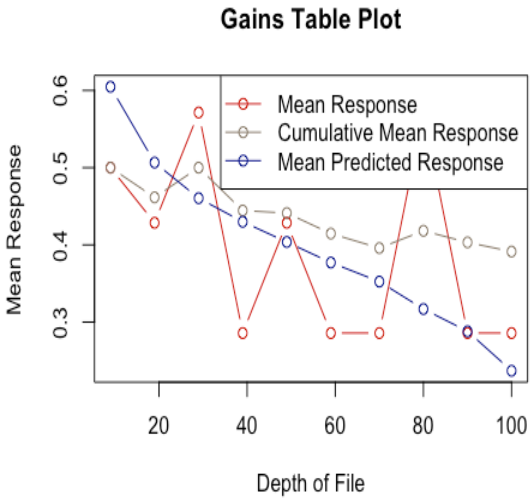
Gain and Lift charts are used to evaluate performance of classification model. They measure how much better one can expect to do with the predictive model comparing without a model. It's a very popular metrics in marketing analytics. It's not just restricted to marketing analysis. It can be used in other domains as well such as risk modeling, supply chain analytics etc. It also helps to find the best predictive model among multiple challenger models. In this tutorial, we will see how gain and lift metrics are calculated along with their interpretation.

For our logistic model, we have plotted mean Response vs Mean Prediction Curve. The prediction was higher with less mean response.

Gain at a given decile level is the ratio of cumulative number of targets (events) up to that decile to the total number of targets (events) in the entire data set.

Logistic Gain

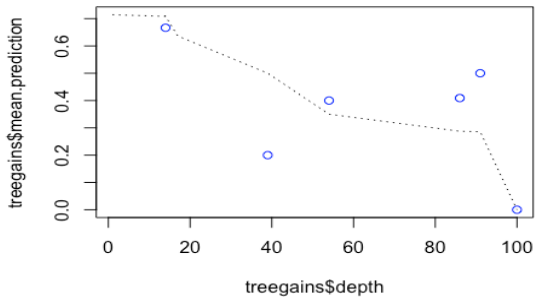


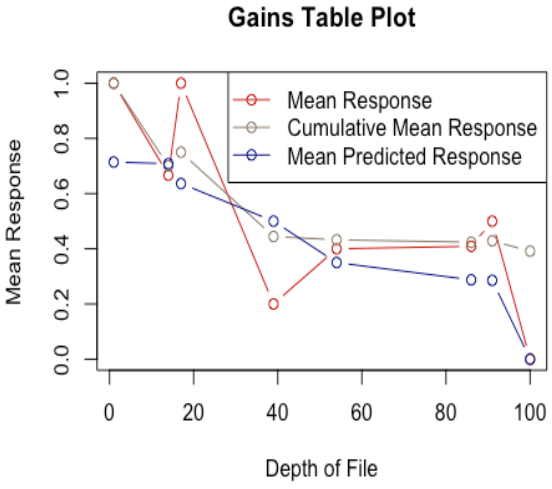


From our predicted curve, we can say that with increase of depth, mean response decreases.

For simplicity let's assume we have 1000 students. If we want to predict graduation with logistic regression to all our students, we might find that 35% (350 out of 1000) will be graduated at a depth of 80.

Tree Gain





For simplicity let's assume we have 1000 students. If we want to predict graduation with tree model to all our students, we might find that less than 35% (350 out of 1000) will be graduated at a depth of 80.

Conclusion

With our logistic model, the prediction curve form gains plot looks better than the tree model