GLM prediction: Private/Public college

Heejae Roh
Northeastern University
ALY 6015: Intermediate Analytics
Paromita Guha
January 29, 2023

INTRODUCTION

Generalized linear models (GLM) enable analysis including categorical data. In the real world, we have a lot of binary data like churn or not. In this project, I will practice the overall process of modeling GLM, how to interpret the results of GLM, and how to find out the most fitted model.

ABSTRACTION

Data itself

Dataset is about statistics for US colleges from the 1995 issue of US news and World Report. There is one categorical variable which has information whether the college is private or not. There are 17 numeric variables which contain information about admissions, estimated cost of college life, student/faculty ratio of each college.

EDA

Histogram of Private yes or no shows that there is 565 Yes and 212 No in target variable. Boxplot which is made by formular that is 'Absolute[Mean(Yes) – Mean(No)] / Mean(Yes + No)' shows the each variable's mean difference in Yes & No category. With boxplots based on barplot the biggest mean difference ratio is in F.Undergrad.

PREREQUISITE for MODELING

I made a model with every variable and arrange it by ascending order of p-value. After that I made a correlation matrix to check the multicollinearity between variables.

FINDING BEST FITTED MODEL

With the prerequisite for modeling, I used backward selection. I choose the factors that are significant in the backward modeling. In this process, I check the multicollinearity with correlation matrix. I add and delete the variables comparing AIC of each model.

MODEL on my own

The model I made has AIC 192.55 with train dataset. The first model which contains every variable was 197.48. The second model is best model I found.

CONFUSION MATRIX & ROC

After making model, I check the confusion matrix and interpret it, and then I applied it to the test dataset. I also check the confusion matrix in test dataset. The Accuracy of model with train dataset is 94% and Sensitivity & Specificity is 96% & 89%. With the test dataset, Accuracy & Sensitivity & Specificity is 90%, 96%, and 76% each. After that I draw a ROC curve and calculate the AUC. AUC is 0.9829 & 0.9606 in train & test dataset.

FALSE POSITIVE, FALSE NEGATIVE

I think about the which is worse between False Positive and False Negative. Generally, the false negative is worse than False Positive, but we have to think about the conditions.

CONCLUSION

I learn that I have to be careful using accuracy in confusion matrix when the number of positive & negative data differs. I realize I must check the sensitivity and specificity respectively.

PART 0. INTRODUCTION

Generalized linear models enable analysis including categorical data. Previous linear models could only predict continuous outcomes, but with GLM, outcomes including binary, or count can be predicted. While doing GLM model, the formular use the log value which enable symmetric analysis. Therefore, it is necessary to exponentiate and analyze the exponentiated log value. In this module, I will practice the overall process of doing GLM and understand GLM.

PART 1. ANALYSIS

Data Analysis by data explanation

1. Dataset is about statistics for US Colleges from the the 1995 issue of US News and World Report (R-DATA, n.d.)
2. Data has 777 observations and 18 variables. There is one categorical variable which has information about whether the college is private or not. There are 17 numeric variables about information about admissions, estimated cost of college life, student/faculty ratio, and etc.
3. In this analysis, I will find out the model which can predict private or not with other numeric variables

PART 1-1. SUMMARY of EDA

Headtail of Data 1

	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.undergrad
Abilene Christian University	Yes	1660	1232	721	23	52	2885
Adelphi University	Yes	2186	1924	512	16	29	2683
Adrian College	Yes	1428	1097	336	22	50	1036
…					…		
Xavier University of Louisiana	Yes	2097	1915	695	34	61	2793
Yale University	Yes	10705	2453	1317	95	99	5217
York College of Pennsylvania	Yes	2989	1855	691	28	63	2988

Headtail of Data 2

	P.undergrad	Outstate	Room.Board	Books	Personal	Phd
Abilene Christian University	537	7440	3300	450	2200	70
Adelphi University	1227	12280	6450	750	1500	29
Adrian College	99	11250	3750	400	1165	53
…						
Xavier University of Louisiana	166	6900	4200	617	781	67
Yale University	83	19840	6510	630	2115	96
York College of Pennsylvania	1726	4990	3560	500	1250	75

Headtail of Data 3

	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Abilene Christian University	78	18.1	12	7041	60
Adelphi University	30	12.2	16	10527	56
Adrian College	66	12.9	30	8735	54
…					
Xavier University of Louisiana	75	14.4	20	8323	49
Yale University	96	5.8	49	40386	99
York College of Pennsylvania	75	18.1	28	4509	99

Detailed explanation about variables

1. Private indicates private or public university
2. Apps: number of applications and Accept: number of applications accepted, Enroll is number of new students enrolled
   - With these variables I want to make proportion of accepted and enrolled rate
3. Outstate: out of state tuition, it could be indicate estimated or average of tuition
4. Room.Board is room and board costs, Books is estimated book costs
5. Terminal is pct. Of faculty with terminal degree (PhDs or doctorates) and S.F.Ratio: Student/faculty ratio are about education quality
6. Perc.alumni: Pct. Alumni who donate and Expend: Instructional expenditure per student, these are about financial conditions and support
7. Grad.Rate: Graduation rate, I think it could explain the loyalty about school, because if you don't like or trust your school curriculum, it's very likely that you won't graduate (R-DATA, n.d.)

Summary of Data 1

Private			Apps	Accept	Enroll	Top10perc	Top25perc	F.undergrad	P.undergrad	Outstate
No	212	Min.	81	72	35	1.00	9.0	139	1.0	2340
Yes	565	1st Qu.	776	604	242	15.00	41.0	992	95.0	7320
		median	1558	1110	434	23.00	54.0	1707	353.0	9990
		mean	3002	2019	780	27.56	55.8	3700	855.3	10441
		3rd Qu.	3624	2424	902	35.00	69.0	4005	967.0	12925
		Max.	48094	26330	6392	96.00	100.0	31643	21836.0	21700

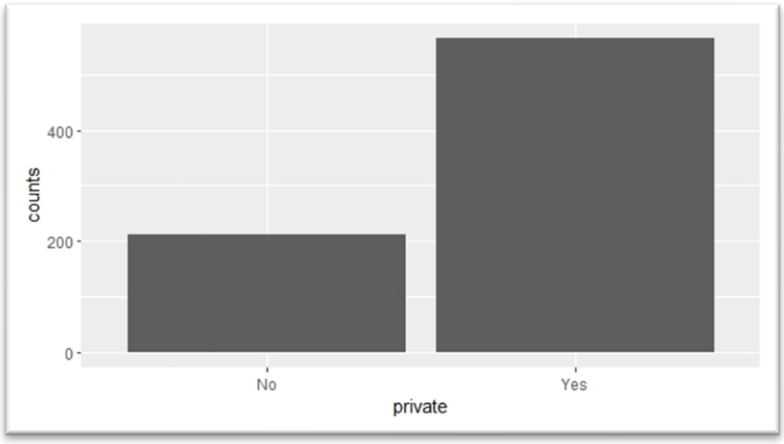
Summary of Data 2

	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Min.	1780	96.0	250	8.00	24.0	2.50	0.00	3186	10.00
1st Qu.	3597	470.0	850	62.00	71.0	11.50	13.00	6751	53.00
median	4200	500.0	1200	75.00	82.0	13.60	21.00	8377	65.00
mean	4358	549.4	1341	72.66	79.7	14.09	22.74	9660	65.46
3rd Qu.	5050	600.0	1700	85.00	92.0	16.50	31.00	10830	78.00
Max.	8124	2340.0	6800	103.00	100.0	39.80	64.00	56233	118.00

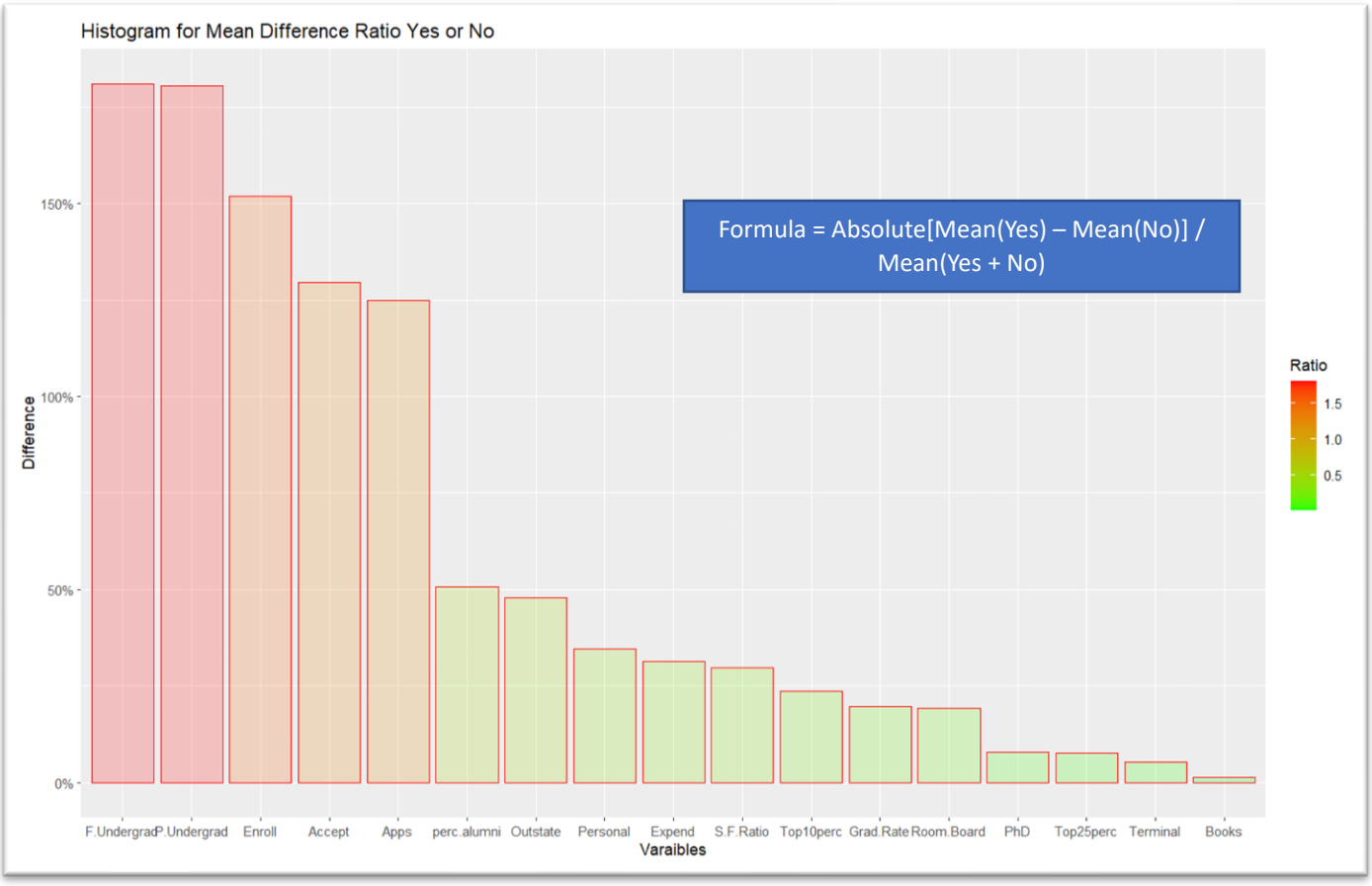
Data Cleaning

1. Checking summary(dataset)
2. There is no <NA> value in this dataset

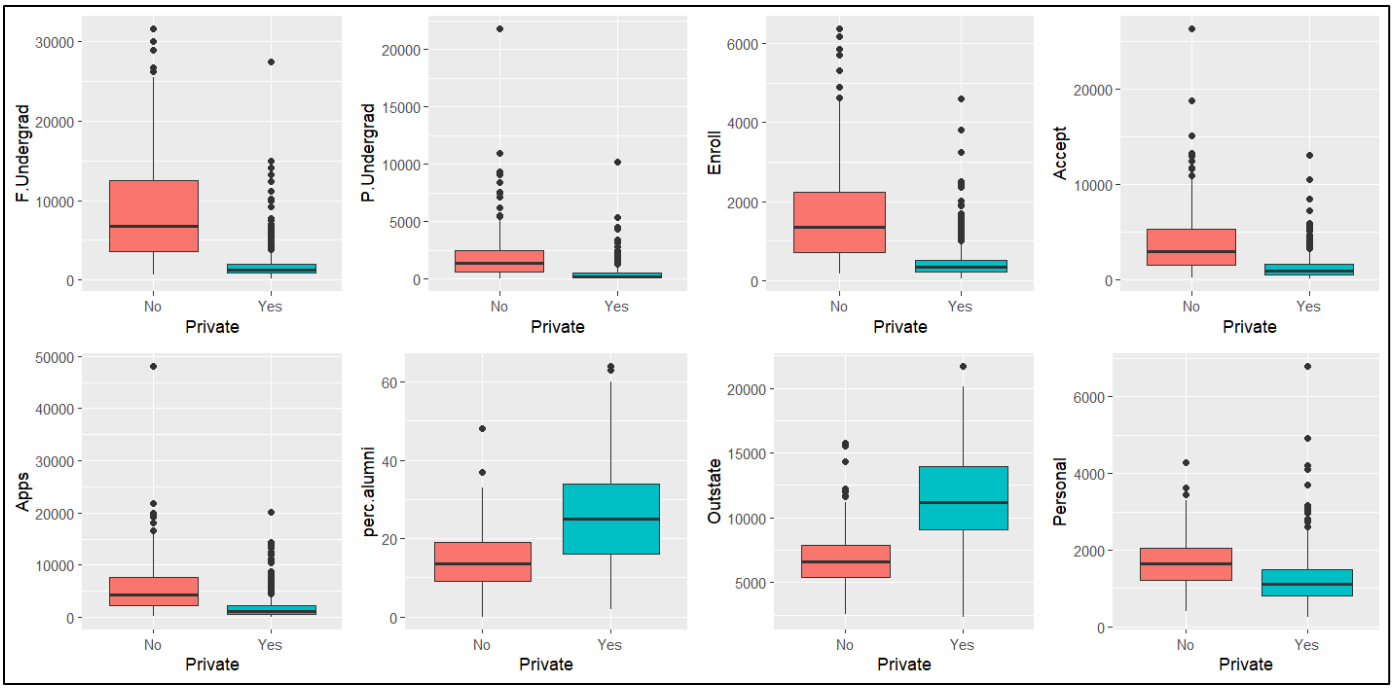
Histogram of Private yes or no



Barplot of Variables by Private yes or no



Box plots of Variables by Private yes or no



Explanation:

1. Private Yes: 565, No: 212. Difference is 353 and Yes is 2.67 times bigger than No
2. Barplot shows the difference ratio of Yes or No as can see in the formular. This explain which variable shows highest difference between Yes and No. Based on this I will find out the best model
3. Boxplots are made based on Barplot which is descending order of difference ratio. F.Undergrad > P.Undergrad > Enroll > Accept (First row) > Apps > perc.alumni > Outstate > Personal (Second row)

Trends and interesting stuffs about dataset

1. F.Undegrad & P.Undergrad have biggest differences, actually I can guess that the student number of Private and Public college is different. I have to check that if we check the number of student, we can predict Private or not
2. Enroll & Accept & Apps are 3rd, 4th, and 5th biggest difference variables. It's also highly related to the number of student. I have to think about overfitting in these variables
3. perc.alumni & Outstate & Personal is all about the money. I can guess that by checking the financial factors, It might be possible to predict Private or not

PART 1-2. PREREQUISITE for MODELING

Preparing dataset

1. Creating partition with the p = 0.80, train dataset has 622 observations and test dataset has 155 observations. I used backward selection which starts with the model which contains all variable
2. Splitting helps to avoid overfitting and to improve the training dataset accuracy. Finally, we need a model that can perform well on unknown data; therefore, we utilize test data to test the trained model's performance at the end (finnstats, 2021)

Call: glm(Private ~ ., data = train, family = binomial(link="logit"))

Deviance Residuals

Min	1Q	Median	3Q	Max
-3.9102	-0.0108	0.0385	0.1340	2.9381

Coefficients:

	Estimate	Std. Error	z value	Pr(>|t|)
(Intercept)	-0.1010071	2.628919	-0.045	0.96440
Outstate	0.0006827	0.0001445	4.723	2.33e-06 ***
F.Undergrad	-0.0008378	0.0002750	-3.046	0.00232 **
perc.alumni	0.0586747	0.0274391	2.138	0.03249 *
Books	0.0035014	0.0016765	2.089	0.03674 *
Apps	-0.0006244	0.0003116	-2.004	0.04511 *
Grad.Rate	0.0286088	0.0155768	1.837	0.06627 .
Terminal	-0.0581717	0.0337417	-1.724	0.08470 .
P.Undergrad	0.0003143	0.0002197	1.431	0.15254
Enroll	0.0017105	0.0012004	1.425	0.15418
PhD	-0.0451865	0.0334054	-1.353	0.17616
Personal	-0.0003974	0.0003281	-1.211	0.22576
S.F.Ratio	-0.0870070	0.0748791	-1.162	0.24525
Expend	0.0001652	0.0001565	1.056	0.29095
Accept	0.0006231	0.0006025	1.034	0.30102
Top10perc	0.0169635	0.0344779	0.492	0.62271
Top25perc	0.0105871	0.0248007	0.427	0.66946
Room.Board	0.0000495	0.0003132	0.158	0.87440

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

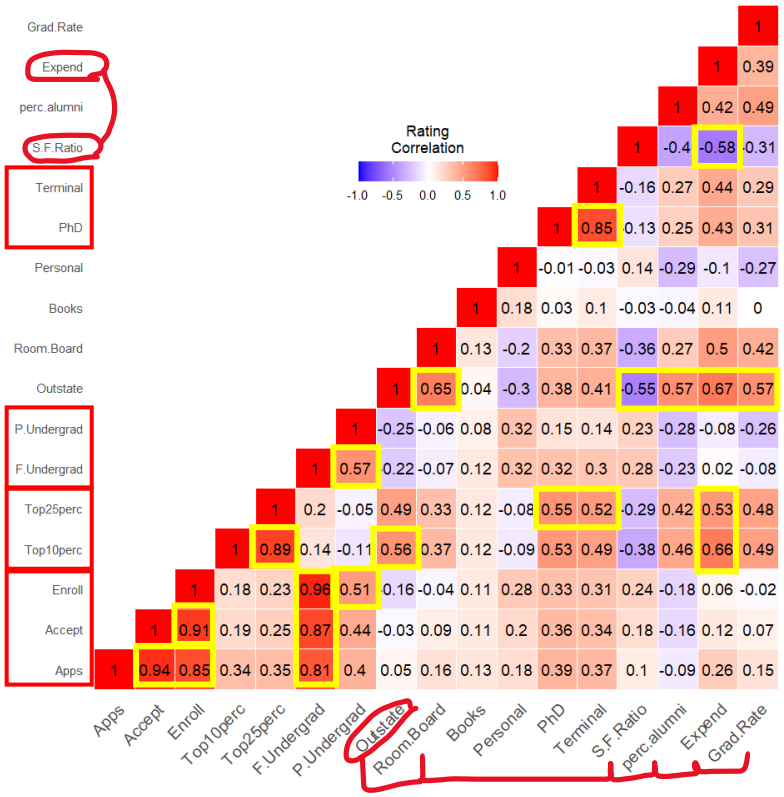
Null deviance: 729.64 on 621 degrees of freedom
Residual deviance: 161.48 on 604 degrees of freedom
AIC: 197.48

Number of Fisher Scoring iterations: 8

Explanation:

1. Coefficients table shows descending order of significant codes (ascending order by P-value)
2. Outstate, F.Undergrad, perc.alumni, and Apps is one of the biggest difference variables in EDA boxplot. On the other hand, Grad.rate, Terminal, and books was not significant in difference ration of Yes and No
3. I will make model based on the Call contains every variable, but before start I will check multicollinearity by correlation matrix. Because Multicollinearity is a common problem when estimating linear or generalized linear models (Paul, 2012)

Correlation Matrix of numeric variables



Interpretation:

1. (From left above) Expend is highly related to S.F.Ratio in negative 0.58
2. Terminal and PhD have high correlation 0.85
3. P.Undergrad and F.Undergrad is highly correlated as 0.57
4. Top25perc and Top10perc is also highly correlated as 0.89
5. Enroll & Accept & Apps is also highly correlated each other
6. Outstate which is about financial status is correlated Room.Board, S.F.Ratio, Expend, and Grad.Rate
7. If possible, I will choose one of correlated variables to make model

PART 1-3. FINDING BEST FITTED MODEL

Model	AIC
glm: Private~ Every variable	197.48
glm: Private~ Outstate + F.Undergrad + perc.alumni + Books + Apps + Grad.Rate + Terminal	192.55
glm: Private~ Outstate + F.Undergrad + perc.alumni + Books + Apps + Grad.Rate + Terminal + Personal	194.22
glm: Private~ Outstate + F.Undergrad + perc.alumni + Books + Apps + ~~Grad.Rate + Terminal~~	217.22

Explanation:

1. Second model is using the significant variables only from '***' to '.'
2. The AIC of second model is 192.55 which is lower than first model which use every variable and AIC is 197.48
3. In the third model the AIC getting higher after add the variable 'Personal' which is not significant
4. In the fourth model the AIC getting steeply higher after delete the significant variables which are 'Grad.Rate' and 'Terminal'
5. I will decide second model which is 'Private~ Outstate + F.Undergrad + perc.alumni + Books + Apps + Grad.Rate + Terminal' as fitted model. Do test and interpret the result of analysis with this model

Coef (model2)

Section	(Intercept)	Outstate	F.Undergrad	perc.alumni
coef	-2.1060490156	0.0008376268	-0.0005021876	0.0651541929
Section	Books	Apps	Grad.Rate	Terminal
coef	0.0035569073	-0.0001018381	0.0214126472	-0.0794765150

exp (Coef (model2))

Section	(Intercept)	Outstate	F.Undergrad	perc.alumni
coef	0.1217179	1.0008380	0.9994979	1.0673236
Section	Books	Apps	Grad.Rate	Terminal
coef	1.0035632	0.9998982	1.0216435	0.9235997

Interpretation:

1. Coef (model2) contains the log-odds, to convert log-odds to odds, exponentiate function is used (exp)
2. This means that if I maintain all other variable, 1.067 (perc.alumni) rises according to the change of 1 unit of the perc.alumni variable.
3. Odds of Private increase by a factor of 1.067 for each 1 unit increate in perc.alumni, 1.003 for each 1 unit increate in books, 1.021 for each 1 unit increate in Grad.Rat, 1.0008 for each 1 unit increate in Outstate
4. Odds of Private increase by a factor of 0.9999 for each 1 unit increate in Apps, 0.9994 for each 1 unit increate in F.Undergrad, 0.9236 for each 1 unit increate in Terminal

Calculate probs with

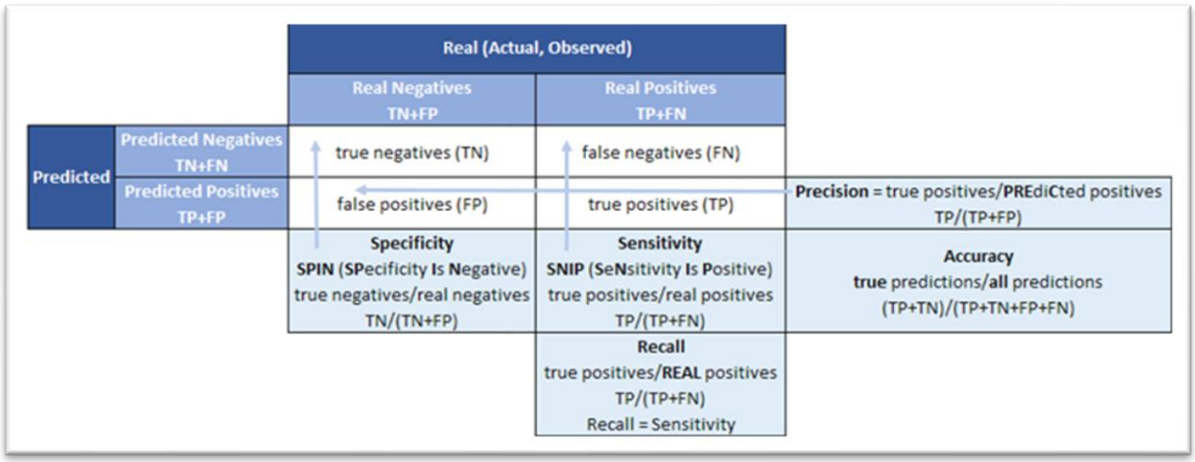
	Outstate	F.Undergrad	perc.alumni	Books	Apps	Grad.Rate	Terminal	Private	probs
Min	2340.00	139.00	0.00	96.00	81.00	10.00	24.00	Yes	0.1714
Mean	10440.67	3699.90	22.74	549.38	3001.64	65.46	79.70	Yes	0.9516
Max	21700.00	31643.00	64.00	2340.00	48094.00	118.00	100.00	Yes	0.9133

PART 1-4. CONFUSION MATRIX AND ROC with TRAIN.DATA

Confusion Matrix and Statistics with train.set

		Actual Values	
		No	Yes
Predicted Values	No	152	17
	Yes	18	435

Accuracy	0.9437	95% CI	(0.9226, 0.9605)
No Information Rate	0.7267	P-Value [Acc > NIR]	<2e-16
Kappa	0.8581	Mcnemar's Test P-Value	1
Sensitivity	0.9624	Specificity	0.8941
Pos Pred Value	0.9603	Neg Pred Value	0.8994
Prevalence	0.7267	Detection Rate	0.6994
Detection Prevalence	0.7283	Balanced Accuracy	0.9283
'Positive' Class	Yes		



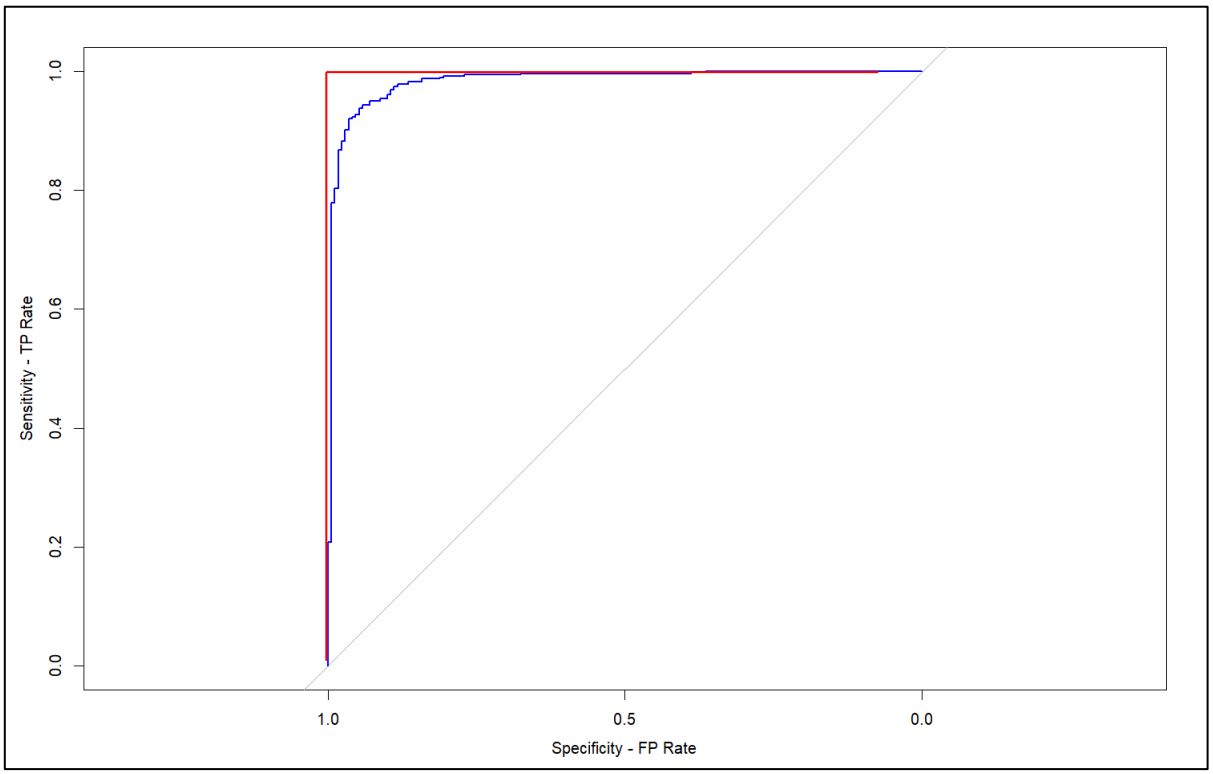
Interpretation:

1. Accuracy of this model is 94.37% (TP+TN)/(TP+TN+FP+FN). Remember, accuracy is a very useful metric when all the classes are equally important. But this might not be the case if we are predicting if a patient has cancer (Vipul, 2023)
2. Sensitivity (=Recall) of this model is 96.24% (TP/(TP+FN). 96% of the college which is Private were correctly predicted by the Logistic Regression model
3. Specificity is 89.41% (TN/(TN+FP). 89% of the college which is Public were

correctly predicted by the Logistic Regression model

4. The model is more accurate in predicting True Positive than True Negative, I think it's because there are 565 Yes vs 212 No
5. Accuracy alone doesn't tell the full story when you're working with a class-imbalanced data set, like this one, where there is a significant disparity between the number of positive and negative labels (MachineLearning, n.d.)

ROC curve and AUC with train.set



Interpretation:

1. AUC (Area under the curve) is 0.9829. X axis is False Positive Rate which is specificity and Y axis is True Positive Rate which is sensitivity
2. If the specificity and sensitivity is close to 100%, the graph seems like red line. It means that the Logistic Regression model predict the results perfectly
3. An ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. The following figure shows a typical ROC curve To compute the points in an ROC curve, we could evaluate a logistic regression model many times with different classification thresholds, but this would be inefficient. Fortunately, there's an efficient, sorting-based algorithm that can provide this information for us, called AUC (MachineLearning, n.d.)

PART 1-4. CONFUSION MATRIX AND ROC with TEST.DATA

Confusion Matrix and Statistics with test.set

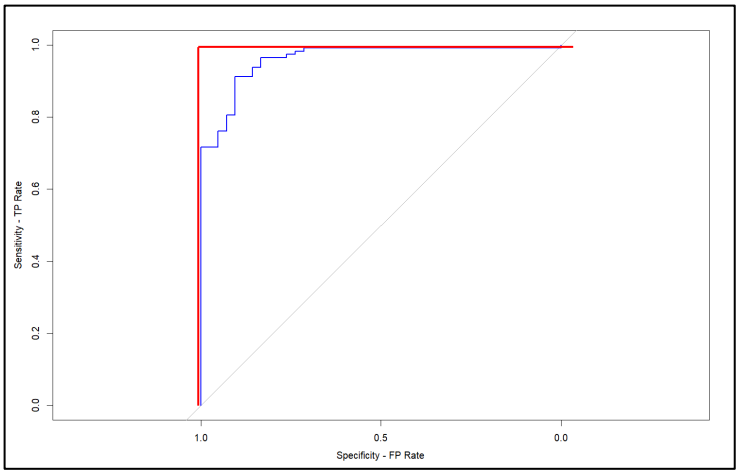
		Actual Values	
		No	Yes
Predicted Values	No	32	4
	Yes	10	109

Accuracy	0.9097	95% CI	(0.8531, 0.9497)
No Information Rate	0.729	P-Value [Acc > NIR]	1.993e-08
Kappa	0.7606	Mcnemar's Test P-Value	0.1814
Sensitivity	0.9646	Specificity	0.7619
Pos Pred Value	0.9160	Neg Pred Value	0.8889
Prevalence	0.7290	Detection Rate	0.7032
Detection Prevalence	0.7677	Balanced Accuracy	0.8633
'Positive' Class	Yes		

Interpretation:

1. Accuracy of this model is 90.97% (TP+TN)/(TP+TN+FP+FN). It is smaller than the accuracy of train dataset
2. Sensitivity (=Recall) of this model is 96.46% (TP/(TP+FN). 96% of the college which is Private were correctly predicted by the Logistic Regression model in test dataset
3. Specificity is 76.19% (TN/(TN+FP). 76% of the college which is Public were correctly predicted by the Logistic Regression model. It is smaller than the accuracy of train dataset whose specificity was 89.41%. This is also because in trainset there is 170 No, but in testset there is only 42 No

ROC curve and AUC with test.set



Interpretation:

1. AUC (Area under the curve) is 0.9606 which is smaller than the test dataset's 0.9829
2. The train dataset's curve is closer to the x=1.0 & y=1.0 (red line) than the test dataset's. In other words, AUC of train dataset is bigger than the test dataset. With the train dataset, the curve is more likely to be able to separate the classes

ANSWER FOR THE QUESTION

1. Number 4 Deliverable, which misclassifications are more damaging for the analysis, False Positives or False Negative?
   Depending on what the question is, I can tell which is more damaging between False Positive or False Negative. In this problem, a false positive is to conclude that it is private even though it is public. A false negative is to conclude that it is public even though it is private. If the private college is under pandemic and has to be closed, a false negative can be more fatal. Because the private college should be closed, it won't. In the reverse case, the result is same but vise-versa.
   Generally, since false-negative results pose greater risks, most testing applications are set up to minimise the occurrence of false-negative results. This means that false-positive results are more likely to occur and are therefore more often found as a topic of discussion (Mina, 2020). It's because if you have a disease, but the test result is negative (False negative), it could be more dangerous than false positive which enable you to take a test.

CONCLUSION

In this project, I focused on testing with Generalized linear models (GLM). To understand GLM, I have to know the log logic which make symmetric graphs with possibility. With GLM, I can handle the binary questions which are many in our real world. I learn GLM with college dataset. There was huge difference of positive and negative data. I learn that I have to be careful using accuracy when the number of positive & negative data differs. I realize that I must look at sensitivity and specificity respectively.

I want to make more accurate model by making additional variables with original dataset, but failed (in Appendix & R-code appendix). Before I check the AIC, I think that it's reasonable to using additional variables like Accept.Rate, Enroll.rate, and Fulltime student ratio. However, I think the additional variable is made by original dataset, which is simpler than additional variables. In other words, additional variable is not a brand-new variable.

APPENDIX

Making new columns

1. Accept.rate <- Accept/ Apps * 100
2. Enroll.rate <- Enroll/ Accept * 100
3. Full.ratio <- F.Undergrad/ P.Undergrad
4. Spending <- Room.Board + Books + Personal

Making model (train2 is new dataset with new columns)

mod1 <- glm(Private ~ ., data=train2, family = binomial(link="logit"))
AIC: 201.95

(Original) mod2 <- glm(Private~ Outstate + F.Undergrad + perc.alumni + Books + Apps + Grad.Rate + Terminal, data = train2, family = binomial(link="logit"))
AIC: 192.55

(new) mod3 <- glm(Private~ Outstate + **Full.ratio** + perc.alumni + Books + Apps + Grad.Rate + Terminal, data = train2, family = binomial(link="logit"))
AIC: 212.12

(new) mod4 <- glm(Private~ Outstate + F.Undergrad + perc.alumni + **Spending** + Apps + Grad.Rate + Terminal, data = train2, family = binomial(link="logit"))
AIC: 197.69

(new) mod5 <- glm(Private~ Outstate + F.Undergrad + perc.alumni + Books + Apps + Grad.Rate + Terminal + **Full.ratio**, data = train2, family = binomial(link="logit"))
AIC: 194.4

Interpretation:

1. The AIC of original model is 192.55, but the AIC of other models which contains dummy variables which is made by original variables is bigger than original model
2. I think, because of collinearity and simplicity, making model with additional variable which is made by original variable does not affect on making better model. But I will try to make better model with every inch of the way.

REFERENCE

Kabacoff, Robert I. (2015). R in Action 2nd Edition. Manning Publisher.

R-DATA. (n.d.). R dataset / package ISLR / college. Retrieved from https://r-data.pmagunia.com/dataset/r-dataset-package-islr-college

Pragati. (2023, January 3). Train test validation split: how to & best practices [2023]. V7. Retrieved from https://www.v7labs.com/blog/train-validation-test-set

APA. (n.d.). Title page setup. Retrieved from https://apastyle.apa.org/style-grammar-guidelines/paper-format/title-page

STHDA. (n.d.). ggplot2 Quick correlation matrix heatmap - R software and data visualization. Retrieved from http://www.sthda.com/english/wiki/ggplot2-quick-correlation-matrix-heatmap-r-software-and-data-visualization

PERDUE. (n.d.). General format. Retrieved from https://owl.purdue.edu/owl/research_and_citation/apa_style/apa_formatting_and_style_guide/general_format.html

RobJan. (2018, August 9). Histogram with count R. stackoverflow. Retrieved from https://stackoverflow.com/questions/51762720/histogram-with-count-r

ZACH. (2020, October 9). How to Calculate the Mean by Group in R (With Examples). STATOLOGY. Retrieved from https://www.statology.org/r-mean-by-group/

Naveen. (2022, June 28). How to Select Rows in R with Examples. SparkBy{Examples}. Retrieved from https://sparkbyexamples.com/r-programming/select-rows-in-r/

Deepanshu Bhalla. (2016). R : KEEP / DROP COLUMNS FROM DATA FRAME. LISTEN DATA. Retrieved from https://www.listendata.com/2015/06/r-keep-drop-columns-from-data-frame.html

Jitender_1998. (2020, May 10). Calculate the absolute value in R programming – abs() method. geeksforgeeks. Retrieved from https://www.geeksforgeeks.org/calculate-the-absolute-value-in-r-programming-abs-method/

Quick-R. (n.d.). Sorting data. Retrieved from https://www.statmethods.net/management/sorting.html#:~:text=To%20sort%20a%20data%20frame,sign%20to%20indicate%20DESCENDING%20order.
MrFlick. (2017, April 5). R ggplot histogram bars in descending order. stackoverflow.

Retrieved from https://stackoverflow.com/questions/43216628/r-ggplot-histogram-bars-in-descending-order

finnstats. (2021, December 14). How to split data into train and test in R. R-bloggers. Retrieved from https://www.r-bloggers.com/2021/12/how-to-split-data-into-train-and-test-in-r/#:~:text=Splitting%20helps%20to%20avoid%20overfitting,model's%20performance%20at%20the%20end

Paul Allison. (2012, September 10). When can you safely ignore multicollinearity? STATISTICAL HORIZONS. Retrieved from https://statisticalhorizons.com/multicollinearity/

ZACH. (2021, March 23). How to use the predict function with glm in R (With Examples). STATOLOGY. Retrieved from https://www.statology.org/r-glm-predict/

Tripartio. (2021, March 22). Classification matrix. StackExchange. Retrieved from https://stats.stackexchange.com/questions/122225/what-is-the-best-way-to-remember-the-difference-between-sensitivity-specificity

Vipul Jain. (2023, January 17). Idiot's guide to precision, recall, and confusion matrix. KDnuggets. Retrieved from https://www.kdnuggets.com/2020/01/guide-precision-recall-confusion-matrix.html

MechineLearning. (n.d.). Classification: ROC Curve and AUC. Retrieved from https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

MechineLearning. (n.d.). Classification: Accuracy. Retrieved from https://developers.google.com/machine-learning/crash-course/classification/accuracy

Miha Mozina. (2020, November 9). Which is worse: false-positive or false-negative?. Retrieved from https://www.linkedin.com/pulse/which-worse-false-positive-false-negative-miha-mozina-phd/

R-Codes

```
install.packages("ISLR")
library(ISLR)
library(caret)
library(ggplot2)
library(gridExtra)
library(pROC)
library(psych)
library(dplyr)
library(scales)
library(reshape2)
install.packages("rsample")
library(rsample)
library(rpart)
library(recipes)

#Load Data
C1 <- College
C2 <- College
str(C2)
#Checking Data
headTail(C2,5)
# Na.omit
C2 <- na.omit(C2)

######
#EDA
#####

#histogram
table(C1$Private)
private <- c("No", "Yes")
counts <- c(212,565)
pri <- data.frame(x=private, counts=counts)
pri
plt <- ggplot(pri) + geom_bar(aes(x=private, y=counts, labels=counts), stat="identity")
print(plt)

attach(C2)
#scatter
qplot(x=Terminal, y=Outstate, color=Private, shape=Private,
geom='point')+scale_shape(solid=FALSE)
qplot(x=Top10perc, y=Expend, color=Private, shape=Private,
geom='point')+scale_shape(solid=FALSE)
qplot(x=S.F.Ratio, y=Grad.Rate, color=Private, shape=Private,
```

```
geom='point')+scale_shape(solid=FALSE)

# mean differece comparison by variables in Private vs Public
str(C1)

# lots of plots stories ~
# boxplot of each with categorical variable
table(C1$Private)
summary(C1)
mean(C1$Apps)
describe(C1)

str(C1)
C1$all <- c("all")
str(C1)
C1.mean0 <- aggregate(C1,list(C1$all), FUN=mean)
C1.mean0

C1.mean <- aggregate(C1, list(C1$Private), FUN=mean)
C1.mean
C1.mean2 <- abs(C1.mean[1,] - C1.mean[2,])
C1.mean2

C1.mean2 <-  subset(C1.mean2, select=-c(Group.1,Private,all))
C1.mean2
C1.mean0 <-  subset(C1.mean0, select=-c(Group.1,Private,all))
C1.mean0
new <- C1.mean2/C1.mean0
new

#histogram of mean difference
barplot(new)
new1 <- as.data.frame(colnames(new))
new2 <- as.data.frame(as.numeric(new[1:17]))

new3 <- cbind(new1,new2)
colnames(new3) <- c("variables", "Mean.Differ.Ratio.by.Private")

new4 <- new3[order(-new3$Mean.Differ.Ratio.by.Private),]
new4$variables
ggplot(new4) +
  geom_bar(aes(reorder(variables, -Mean.Differ.Ratio.by.Private), Mean.Differ.Ratio.by.Private,
fill=Mean.Differ.Ratio.by.Private),
           col="red", alpha = .2, stat="identity") +
  scale_fill_gradient("Ratio", low = "green", high = "red") +
  scale_y_continuous(labels = percent_format()) +
```

```
  labs(title="Histogram for Mean Difference Ratio Yes or No") +
  labs(x="Varaibles", y="Difference")

# Box plot fixed
b1 <- qplot(x=Private, y=F.Undergrad, fill=Private, geom='boxplot') + guides(fill=FALSE)
b2 <- qplot(x=Private, y=P.Undergrad, fill=Private, geom='boxplot') + guides(fill=FALSE)
b3 <- qplot(x=Private, y=Enroll, fill=Private, geom='boxplot') + guides(fill=FALSE)
b4 <- qplot(x=Private, y=Accept, fill=Private, geom='boxplot') + guides(fill=FALSE)
b5 <- qplot(x=Private, y=Apps, fill=Private, geom='boxplot') + guides(fill=FALSE)
b6 <- qplot(x=Private, y=perc.alumni, fill=Private, geom='boxplot') + guides(fill=FALSE)
b7 <- qplot(x=Private, y=Outstate, fill=Private, geom='boxplot') + guides(fill=FALSE)
b8 <- qplot(x=Private, y=Personal, fill=Private, geom='boxplot') + guides(fill=FALSE)
grid.arrange(b1, b2, b3, b4, b5, b6, b7, b8, nrow = 2)

###################
# Split dataset into train and test sets with original dataset
###################

set.seed(123)
trainIndex <- createDataPartition(Private, p = 0.80, list = FALSE)
train <- C2[trainIndex,]
test <- C2[-trainIndex,]

headTail(train,5)
headTail(test,5)
str(train)
str(test)

########################
#fit a logistic regression model

#AIC 197.48
model1 <- glm(Private ~., data=train, family = binomial(link="logit"))
summary(model1)

str(C2)
#4. cor()
#Correlation heatmap Continuous
cordata3 <- subset(C2, select = -c(Private))
cordata3
str(cordata3)

aR <- round(cor(cordata3), 2)

cor(cordata3)
aR <- round(cor(cordata3), 2)
```

```
get_upper_tri<-function(aR){
  aR[lower.tri(aR)] <- NA
  return(aR)}

upper_tri <- get_upper_tri(aR)
upper_tri
melted_cormat <- melt(upper_tri, na.rm = TRUE)
melted_cormat

reorder_cormat <- function(aR){
  # Use correlation between variables as distance
  dd <- as.dist((1-cormat)/2)
  hc <- hclust(dd)
  aR <-aR[hc$order, hc$order]}

aR <- reorder_cormat(aR)
aR
upper_tri <- get_upper_tri(aR)

melted_cormat <- melt(upper_tri, na.rm = TRUE)

#NEW
ggheatmap <- ggplot(melted_cormat, aes(Var2, Var1, fill = value))+
  geom_tile(color = "white")+
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name="Rating\nCorrelation") +
  theme_minimal()+ # minimal theme
  theme(axis.text.x = element_text(angle = 45, vjust = 1,
                                   size = 12, hjust = 1))+
  coord_fixed()

ggheatmap +
  geom_text(aes(Var2, Var1, label = value), color = "black", size = 4) +
  theme(
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank(),
    axis.ticks = element_blank(),
    legend.justification = c(1, 0),
    legend.position = c(0.6, 0.7),
    legend.direction = "horizontal")+
  guides(fill = guide_colorbar(barwidth = 7, barheight = 1,
```

```
                    title.position = "top", title.hjust = 0.5))

# AIC 192.55
model2 <- glm(Private~ Outstate + F.Undergrad + perc.alumni + Books + Apps + Grad.Rate +
Terminal, data = train, family = binomial(link="logit"))
summary(model2)

# AIC 194.22
model3 <- glm(Private~ Outstate + F.Undergrad + perc.alumni + Books + Apps + Grad.Rate +
Terminal + Personal, data = train, family = binomial(link="logit"))
summary(model3)

# AIC 217.22
model4 <- glm(Private~ Outstate + F.Undergrad + perc.alumni + Books + Apps, data = train,
family = binomial(link="logit"))
summary(model4)

# Display regression coefficients (log-odds) with model2
coef(model2)

# Display regression coefficients (odds) with model2
exp(coef(model2))

# View min, mean and max values for pedigree
summary(C2$perc.alumni)

# Create dataset to see how probability changes for different values of balance
# Create dataset to see how probability changes for different values of balance
testdata <- data.frame(Outstate = c(min(C1$Outstate), mean(C1$Outstate), max(C1$Outstate)),
F.Undergrad = c(min(C1$F.Undergrad), mean(C1$F.Undergrad), max(C1$F.Undergrad)),
perc.alumni = c(min(C1$perc.alumni), mean(C1$perc.alumni), max(C1$perc.alumni)), Books  =
c(min(C1$Books), mean(C1$Books), max(C1$Books)), Apps = c(min(C1$Apps),
mean(C1$Apps), max(C1$Apps)), Grad.Rate = c(min(C1$Grad.Rate), mean(C1$Grad.Rate),
max(C1$Grad.Rate)),  Terminal = c(min(C1$Terminal), mean(C1$Terminal), max(C1$Terminal)),
Private="Yes")
testdata$probs <- predict(model2, testdata, type='response')
testdata

######################
# Train set predictions
######################
# Make predictions on the test data using lambda.min
probabilities.train <- predict(model2, newdata = train, type='response')
probabilities.train
predicted.classes.min <- as.factor(ifelse(probabilities.train >= 0.5, "Yes", "No"))
```

```
#Model accuracy
confusionMatrix(predicted.classes.min, train$Private, positive = 'Yes')

# ROC of train.set
ROC1 <-  roc(train$Private, probabilities.train)
ROC1
plot(ROC1, col= "blue", ylab = "Sensitivity - TP Rate", xlab = "Specificity - FP Rate")

# Calculate the area under the ROC curve
auc <- auc(ROC1)
auc

##################
# Test set predictions
##################
probabilities.test <-  predict(model2, newdata = test, type = 'response')
predicted.classes.min <-  as.factor(ifelse(probabilities.test >= 0.5, "Yes", "No"))

#Model accuracy
confusionMatrix(predicted.classes.min, test$Private, positive = 'Yes')

# Plot the Receiver operator characteristic curve
ROC1 <-  roc(test$Private, probabilities.test)
ROC1
plot(ROC1, col= "blue", ylab = "Sensitivity - TP Rate", xlab = "Specificity - FP Rate")

# Calculate the area under the ROC curve
auc <- auc(ROC1)

# For demonstration only - convert logodds to odds to probabilities
# Predict the log odds
pred_logodds <- predict(model2, test)
head(pred_logodds)

# Exponentiate the log odds
pred_odds <- exp(pred_logodds)
head(pred_odds)

# Calculate probabilities from the odds
probs <- pred_odds/(1+pred_odds)
head(probs)

# Predicted probability
pred_probs <- predict(model2, test, type = 'response')
head(pred_probs)
```

```
### APPENDIX additional analysis
# Making new columns
C1$Accept.rate <- C2$Accept/C2$Apps * 100
C1$Accept.rate

C1$Enroll.rate <- C2$Enroll/C2$Accept * 100
C1$Enroll.rate

C1$Full.ratio <- C2$F.Undergrad/C2$P.Undergrad
C1$Full.ratio

C1$Spending <- C2$Room.Board + C2$Books + C2$Personal

#
###################
# Split dataset into train and test sets with original dataset
###################
set.seed(123)
trainIndex <- createDataPartition(Private, p = 0.80, list = FALSE)
train2 <- C1[trainIndex,]
test2 <- C1[-trainIndex,]

########################
#fit a logistic regression model
########################

str(C1)
C1 <- subset(C1, select = -c(all))

#AIC 197.48
mod1 <- glm(Private ~ ., data=train2, family = binomial(link="logit"))
summary(mod1)

mod2 <- glm(Private~ Outstate + F.Undergrad + perc.alumni + Books + Apps + Grad.Rate +
Terminal, data = train2, family = binomial(link="logit"))
summary(mod2)

mod3 <- glm(Private~ Outstate + Full.ratio + perc.alumni + Books + Apps + Grad.Rate +
Terminal, data = train2, family = binomial(link="logit"))
summary(mod3)

mod4 <- glm(Private~ Outstate + F.Undergrad + perc.alumni + Spending + Apps + Grad.Rate +
Terminal, data = train2, family = binomial(link="logit"))
summary(mod4)

mod5 <- glm(Private~ Outstate + F.Undergrad + perc.alumni + Books + Apps + Grad.Rate +
```

```
Terminal + Full.ratio, data = train2, family = binomial(link="logit"))
summary(mod5)
```