Module 3- GLM and Logistic Regression

ALY6015-Intermediate Analytics

NORTHEASTERN UNIVERSITY

DEVIKA PATIL
MAJOR-PROJECT MANAGEMENT
DATE OF SUBMISSION: 03-12-2024
Prof. Zhi He

Contents

Introduction

This analysis explores a dataset containing information about various colleges in the United States. The College dataset contains information for 777 universities and colleges across 18 different attributes. These attributes include details like the type of institution (private/public), tuition fees, student enrollment, and graduation rates.

The goal of this analysis is to understand how certain college characteristics might influence whether a college is private or public. I used stepAIC() with the direction set to "both" to find the optimal set of features for my logistic regression model. This approach automatically adds or removes variables based on their impact on the Akaike Information Criterion (AIC) value.
A logistic regression model is built using the following features from the dataset:

- Applications received ("Apps")
- Top 25th percentile SAT score ("Top25perc")
- Out-of-state tuition ("Outstate")
- Percentage of alumni who donate ("perc.alumni")
- Room and board cost ("Room.Board")
- Books and supplies cost ("Books")
- Per capita student spending ("Personal")
- Whether the college offers a Ph.D. program ("PhD")
- Expendture(“Expend”)

The goal of this report is to analyze a dataset containing information about colleges and universities to build predictive models for determining whether a college is private or public. We will utilize two machine learning algorithms, logistic regression and support vector machines (SVM), to train and evaluate the models. The report will provide a comprehensive analysis of the dataset, model training, evaluation metrics, and interpretation of results.

Dataset- This provides a snapshot of the initial 24 rows and 18 columns from the complete dataset.

	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Abilene Christian University	Yes	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	12	7041	60
Adelphi University	Yes	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	16	10527	56
Adrian College	Yes	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	30	8735	54
Agnes Scott College	Yes	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	37	19016	59
Alaska Pacific University	Yes	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	2	10922	15
Albertson College	Yes	587	479	158	38	62	678	41	13500	3335	500	675	67	73	9.4	11	9727	55
Albertus Magnus College	Yes	353	340	103	17	45	416	230	13290	5720	500	1500	90	93	11.5	26	8861	63
Albion College	Yes	1899	1720	489	37	68	1594	32	13868	4826	450	850	89	100	13.7	37	11487	73
Albright College	Yes	1038	839	227	30	63	973	306	15595	4400	300	500	79	84	11.3	23	11644	80
Alderson-Broaddus College	Yes	582	498	172	21	44	799	78	10468	3380	660	1800	40	41	11.5	15	8991	52
Alfred University	Yes	1732	1425	472	37	75	1830	110	16548	5406	500	600	82	88	11.3	31	10932	73
Allegheny College	Yes	2652	1900	484	44	77	1707	44	17080	4440	400	600	73	91	9.9	41	11711	76
llentown Coll. of St. Francis de Sales	Yes	1179	780	290	38	64	1130	638	9690	4785	600	1000	60	84	13.3	21	7940	74
Alma College	Yes	1267	1080	385	44	73	1306	28	12572	4552	400	400	79	87	15.3	32	9305	68
Alverno College	Yes	494	313	157	23	46	1317	1235	8352	3640	650	2449	36	69	11.1	26	8127	55
American International College	Yes	1420	1093	220	9	22	1018	287	8700	4780	450	1400	78	84	14.7	19	7355	69
Amherst College	Yes	4302	992	418	83	96	1593	5	19760	5300	660	1598	93	98	8.4	63	21424	100
Anderson University	Yes	1216	908	423	19	40	1819	281	10100	3520	550	1100	48	61	12.1	14	7994	59
Andrews University	Yes	1130	704	322	14	23	1586	326	9996	3090	900	1320	62	66	11.5	18	10908	46
Angelo State University	No	3540	2001	1016	24	54	4190	1512	5130	3592	500	2000	60	62	23.1	5	4010	34
Antioch University	Yes	713	661	252	25	44	712	23	15476	3336	400	1100	69	82	11.3	35	42926	48
Appalachian State University	No	7313	4664	1910	20	63	9940	1035	6806	2540	96	2000	83	96	18.3	14	5854	70
Aquinas College	Yes	619	516	219	20	51	1251	767	11208	4124	350	1615	55	65	12.7	25	6584	65
rizona State University Main campus	No	12809	10308	3761	24	49	22593	7585	7434	4850	700	2100	88	93	18.9	5	4602	48

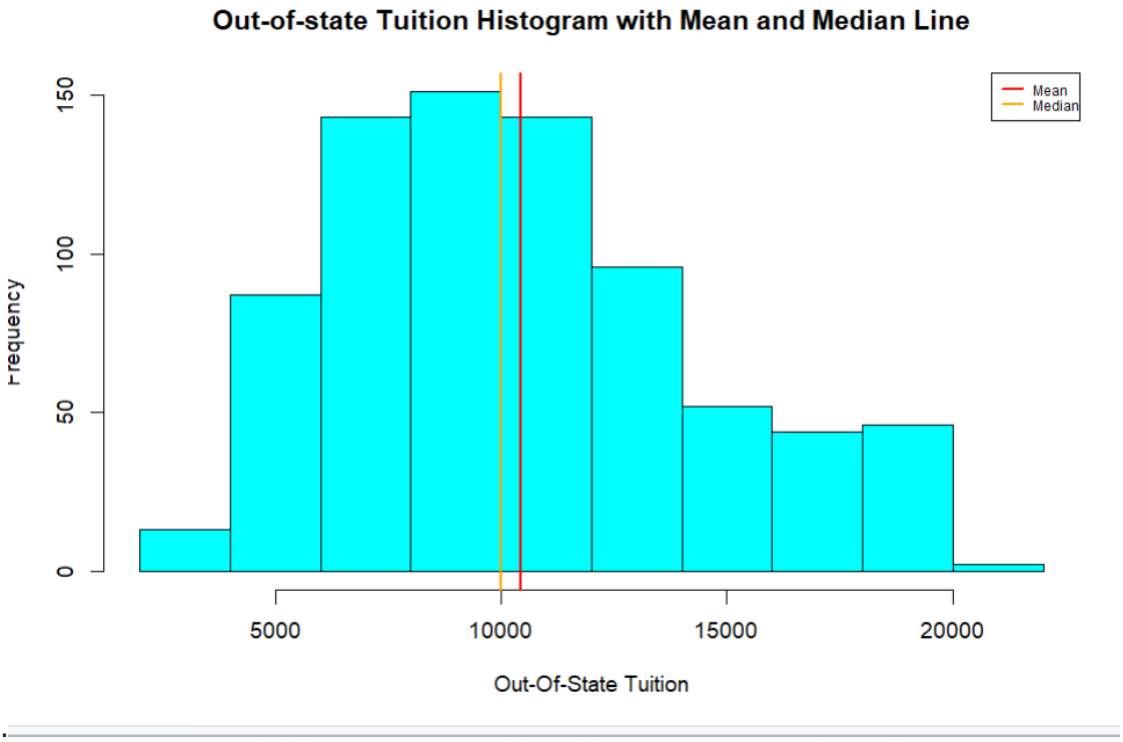
Analysis

Q.1. This summary of the dataset provides key statistical measures including the minimum, first quartile, median, mean, third quartile, and maximum values, offering insights into the distribution and range of the data.

```
> summary(college_df_lm)
 Private        Apps           Accept          Enroll       Top10perc       Top25perc      F.Undergrad
 No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00   Min.   :  9.0   Min.   :  139
 Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00   1st Qu.: 41.0   1st Qu.:  992
           Median : 1558   Median : 1110   Median : 434   Median :23.00   Median : 54.0   Median : 1707
           Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56   Mean   : 55.8   Mean   : 3700
           3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00   3rd Qu.: 69.0   3rd Qu.: 4005
           Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00   Max.   :100.0   Max.   :31643
  P.Undergrad         Outstate       Room.Board       Books           Personal         PhD            Terminal
 Min.   :    1.0   Min.   : 2340   Min.   :1780   Min.   :  96.0   Min.   : 250   Min.   :  8.00   Min.   : 24.0
 1st Qu.:   95.0   1st Qu.: 7320   1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00   1st Qu.: 71.0
 Median :  353.0   Median : 9990   Median :4200   Median : 500.0   Median :1200   Median : 75.00   Median : 82.0
 Mean   :  855.3   Mean   :10441   Mean   :4358   Mean   : 549.4   Mean   :1341   Mean   : 72.66   Mean   : 79.7
 3rd Qu.:  967.0   3rd Qu.:12925   3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00   3rd Qu.: 92.0
 Max.   :21836.0   Max.   :21700   Max.   :8124   Max.   :2340.0   Max.   :6800   Max.   :103.00   Max.   :100.0
   S.F.Ratio      perc.alumni        Expend        Grad.Rate
 Min.   : 2.50   Min.   : 0.00   Min.   : 3186   Min.   : 10.00
 1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751   1st Qu.: 53.00
 Median :13.60   Median :21.00   Median : 8377   Median : 65.00
 Mean   :14.09   Mean   :22.74   Mean   : 9660   Mean   : 65.46
 3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830   3rd Qu.: 78.00
 Max.   :39.80   Max.   :64.00   Max.   :56233   Max.   :118.00
>
```

Charts

1.1 Out-of-state Tuition Histogram with Mean and Median Line



Histogram with a mean and median line for out-of-state tuition in the United States. Here's what we can interpret from the graph:

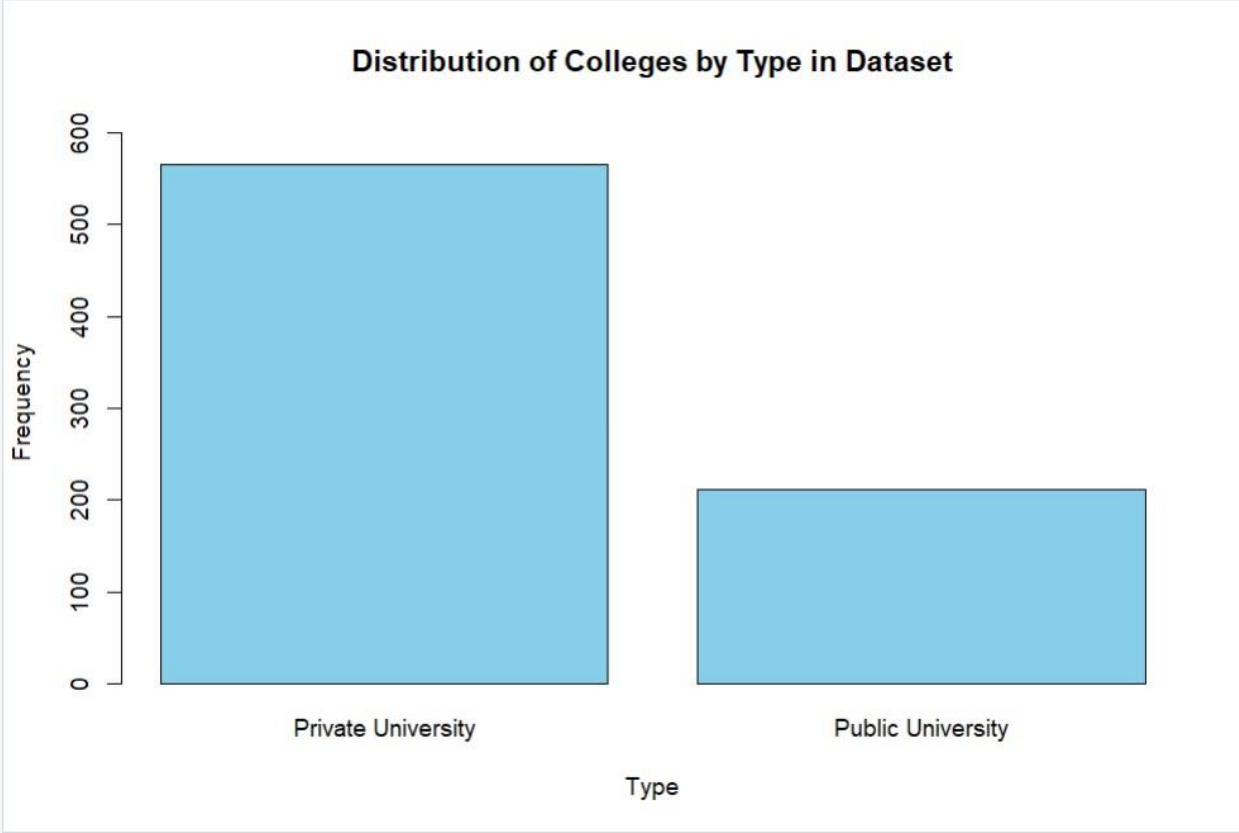
- **Distribution of out-of-state tuition:** The histogram shows the frequency of colleges across different tuition ranges. The x-axis represents the out-of-state tuition cost, and the y-axis represents the number of colleges in each cost range. Because it's a histogram, the exact tuition values for each college aren't shown, but we can see the distribution of costs.
- **Center of the distribution:** The vertical line in the center of the distribution represents the median out-of-state tuition. The median is the value that separates the lower half of the colleges from the higher half in terms of out-of-state tuition costs. In this graph, the median cost is around $18,000.
- **Average cost:** The dotted line on the graph represents the mean out-of-state tuition. The mean is the average tuition cost after considering all the colleges in the data set. It is possible for the mean and median to be different. In this case, the mean appears to

be slightly higher than the median, which suggests there may be a few colleges with very high out-of-state tuition costs that skew the average upwards.

- **Spread of the data:** The shape of the histogram shows the spread of the data. A wider histogram indicates a larger range of tuition costs among colleges. This histogram appears wider on the right side, suggesting that there may be more colleges with higher out-of-state tuition costs than there are colleges with lower costs.

Overall, the histogram provides a quick visual representation of the distribution of out-of-state tuition costs in the dataset. It shows that costs can vary significantly and that there are more colleges with higher tuition costs than lower costs in this particular dataset.

1.2. Distribution of Colleges by Type in Dataset

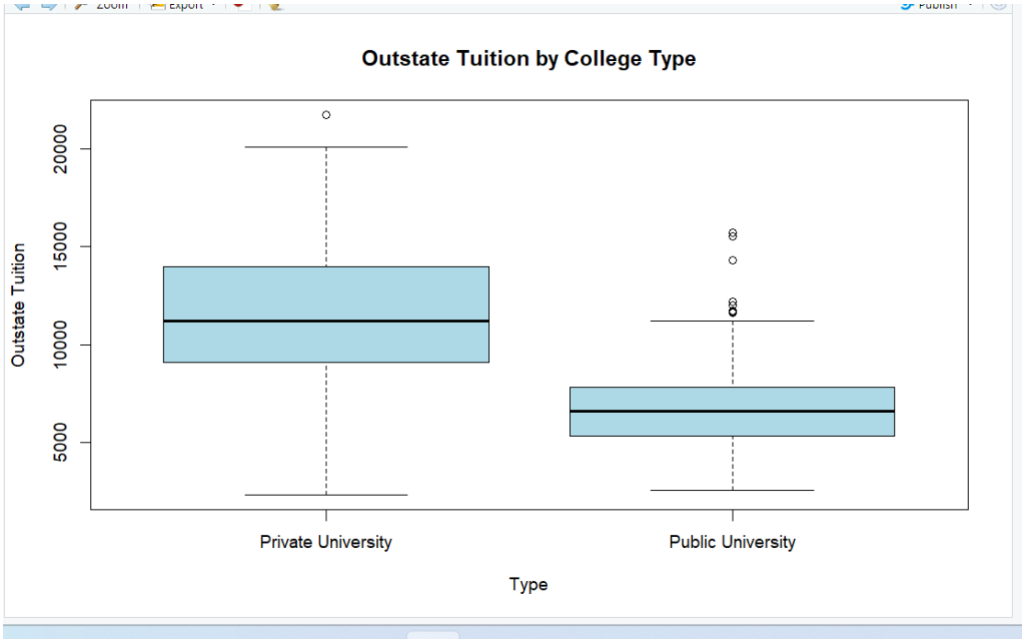


- There are two main categories: **Private Universities** and **Public Universities**.
- The frequency (number of colleges) for each type is represented on the y-axis.
- The college type with the higher frequency will have a taller bar on the chart.

Based on the provided counts (565 private universities and 212 public universities), the chart would likely show a higher bar for private universities, indicating there are more private universities in the dataset compared to public universities.

1.3.Boxplot (Outstate Tuition by College Type)

3..



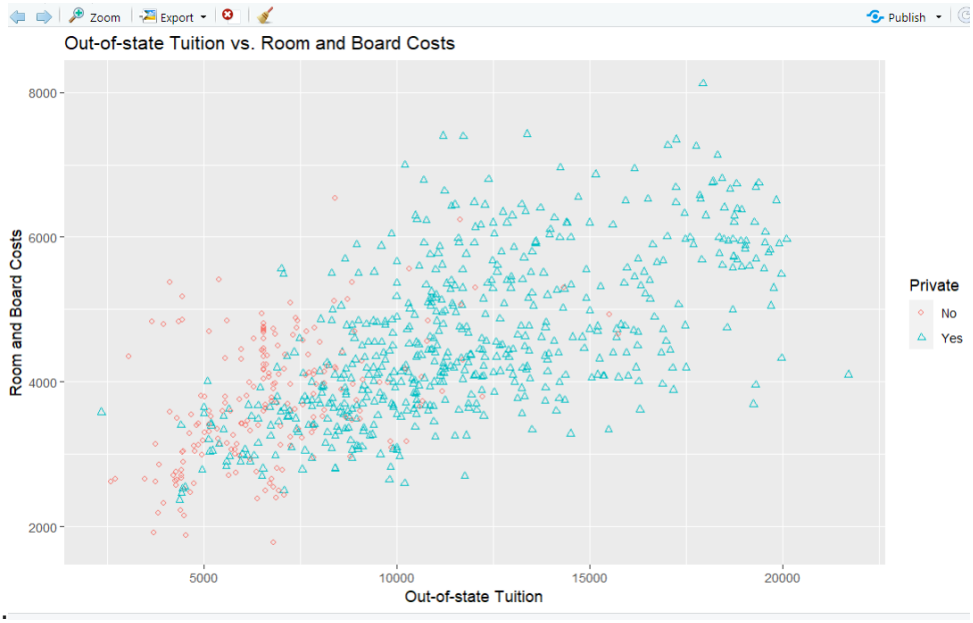
Box plot showing out-of-state tuition by college type in the United States. The box plot divides the data into quartiles. The box in the center represents the middle 50% of the data, with the line in the middle representing the median. The whiskers extend to the most extreme values within 1.5 times the interquartile range (IQR) from the quartile lines. Data points beyond the whiskers are considered outliers.

In this graph, we can see that out-of-state tuition is higher at private universities than at public universities. The median out-of-state tuition for private universities is around $20,000, while the median for public universities is around $15,000. There is also a wider range of out-of-state tuition at private universities than at public universities, as shown by the longer whiskers on the private university box. This means that some private universities have much higher out-of-state tuition than public universities.

Here are some additional insights that can be drawn from the graph:

- There are some outliers for both private and public universities. These outliers represent colleges with much higher out-of-state tuition than most other colleges in their category.
- The overall cost of attending a college out-of-state can vary depending on a number of factors, including tuition, fees, room and board, and other expenses.

1.4. Scatterplot (Out-of-state Tuition vs. Room and Board Costs)



Scatter plot of out-of-state tuition costs compared to room and board costs for colleges in the United States. Each data point represents a single college. The x-axis shows the out-of-state tuition and the y-axis shows the room and board costs. There is a positive correlation between the two variables, which means that colleges with higher out-of-state tuition also tend to have higher room and board costs. However, there is a lot of variability in the data. Some colleges with high out-of-state tuition also have high room and board costs, while others have lower room and board costs. There are also some colleges with lower out-of-state tuition that have high room and board costs.

Here are some additional insights that can be drawn from the graph:

- The most expensive colleges can cost over $20,000 per year for out-of-state tuition and over $15,000 per year for room and board.
- There are a few colleges that have out-of-state tuition costs below $10,000 per year.
- There is a wider range of costs for room and board than for out-of-state tuition.

It's important to consider both out-of-state tuition and room and board costs when making a decision about which college to attend. You can use this graph to get a sense of the range of costs for different colleges, but it's always best to check the specific costs for the colleges you're interested in.

Here are some additional factors to consider when making a decision about which college to attend:

- The academic programs offered by the college
- The size and location of the college
- The financial aid you are eligible for

Q.2. Utilized the **sample()** function to randomly partition the data into training and testing datasets.

```
> cat("Train Data Dimensions:\nNo. Rows =", dim(trainSetData)[1],"\nNo. Columns=",dim(trainSetData)[2],"\n")
Train Data Dimensions:
No. Rows = 582
No. Columns= 18
> cat("Test Data Dimensions:\nNo. Rows =", dim(testSetData)[1],"\nNo. Columns=",dim(testSetData)[2],"\n")
Test Data Dimensions:
No. Rows = 195
No. Columns= 18
>
```

2.1. Train Data

```
> print("Train Data")
[1] "Train Data"
> head(trainSetData)
                              Private Apps Accept Enroll Top10perc Top25perc F.Undergrad P.Undergrad
Gustavus Adolphus College         Yes 1709   1385    634        36        72        2281          50
Fayetteville State University      No 1455   1064    452         1        16        2632         617
Marietta College                  Yes 1611    960    342        27        60        1089         210
Fordham University                Yes 4200   2874    942        30        55        4740        1646
Concordia University CA           Yes  688    497    144        30        75         641         101
North Carolina Wesleyan College   Yes  812    689    195         7        24         646          84
                              Outstate Room.Board Books Personal PhD Terminal S.F.Ratio perc.alumni Expend
Gustavus Adolphus College        14125       3600   400      700  79       89      12.5          58   9907
Fayetteville State University     6806       2550   350      766  75       75      15.1          10   6972
Marietta College                 13850       3920   470      810  80       97      13.2          30  10223
Fordham University               14235       6965   600     1735  86       97      14.4          14  10864
Concordia University CA          10800       4440   570     1515  55       60      13.1          13   8415
North Carolina Wesleyan College   8242       4230   600     1295  77       77      12.7          11  10090
                              Grad.Rate
Gustavus Adolphus College            80
Fayetteville State University        24
Marietta College                     96
Fordham University                   80
Concordia University CA              55
North Carolina Wesleyan College      52
>
```

2.2 Test Data

```
> print("Test Data")
[1] "Test Data"
> head(testSetData)
                                Private  Apps Accept Enroll Top10perc Top25perc F.Undergrad P.Undergrad
Abilene Christian University        Yes  1660   1232    721        23        52        2885         537
Albion College                      Yes  1899   1720    489        37        68        1594          32
Andrews University                  Yes  1130    704    322        14        23        1586         326
Arizona State University Main campus  No 12809  10308   3761        24        49       22593        7585
Arkansas Tech University            No  1734   1729    951        12        52        3602         939
Augustana College IL                Yes  1879   1658    497        36        69        1950          38
                                Outstate Room.Board Books Personal PhD Terminal S.F.Ratio perc.alumni
Abilene Christian University        7440       3300   450     2200  70       78      18.1          12
Albion College                     13868       4826   450      850  89      100      13.7          37
Andrews University                  9996       3090   900     1320  62       66      11.5          18
Arizona State University Main campus 7434       4850   700     2100  88       93      18.9           5
Arkansas Tech University            3460       2650   450     1000  57       60      19.6           5
Augustana College IL               13353       4173   540      821  78       83      12.7          40
                                Expend Grad.Rate
Abilene Christian University      7041        60
Albion College                   11487        73
Andrews University               10908        46
Arizona State University Main campus  4602        48
Arkansas Tech University          4739        48
Augustana College IL              9220        71
>
```

Q.3.

```
> summary(logistic_r_model)

Call:
glm(formula = Private ~ Apps + Top25perc + Outstate + Room.Board +
    Books + Personal + PhD + Terminal + perc.alumni + Expend,
    family = binomial(link = "logit"), data = trainSetData)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.677e+00  1.446e+00  -1.160   0.2461
Apps        -6.034e-04  8.653e-05  -6.973 3.09e-12 ***
Top25perc    2.370e-02  1.211e-02   1.957   0.0504 .
Outstate     7.846e-04  1.276e-04   6.151 7.71e-10 ***
Room.Board   4.213e-04  2.735e-04   1.540   0.1236
Books        2.175e-03  1.569e-03   1.386   0.1657
Personal    -6.100e-04  2.814e-04  -2.168   0.0302 *
PhD         -5.832e-02  3.360e-02  -1.736   0.0827 .
Terminal    -5.089e-02  3.191e-02  -1.595   0.1108
perc.alumni  5.594e-02  2.287e-02   2.446   0.0145 *
Expend       2.302e-04  1.236e-04   1.863   0.0625 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 692.15  on 581  degrees of freedom
Residual deviance: 188.61  on 571  degrees of freedom
AIC: 210.61

Number of Fisher Scoring iterations: 8
```

In our analysis, we identified several predictors that demonstrate significant effects on the response variable (Private). Firstly, the predictor 'Apps' exhibits a highly significant effect, as evidenced by a remarkably low p-value ($p < 0.001$). Similarly, 'Outstate' emerges as another significant predictor, with a p-value indicating strong significance ($p < 0.001$). Furthermore, 'Personal' shows significance at the 0.05 level, as denoted by the asterisk (*) alongside its p-value ($p = 0.0302$), while 'perc.alumni' also proves to be significant at this level with a p-value of 0.0145. On the other hand, several predictors were found to be statistically insignificant. Notably, 'Top25perc' yields a p-value slightly exceeding 0.05 ($p = 0.0504$), suggesting marginal significance. Additionally, 'Room.Board', 'Books', 'PhD', 'Terminal', and 'Expend' all display p-values greater than 0.05, indicating a lack of statistical significance at the 0.05 threshold.

These are the coefficients which are not in scientific notations.

```
> coef(logistic_r_model) # log odds (Display Regression Coefficients)
  (Intercept)          Apps     Top25perc      Outstate    Room.Board         Books      Personal
-1.6770938713 -0.0006034095  0.0236954387  0.0007845756  0.0004212575  0.0021747496 -0.0006100387
          PhD      Terminal   perc.alumni        Expend
-0.0583205003 -0.0508901017  0.0559370829  0.0002301922
> exp(coef(logistic_r_model)) # odds (Display Regression Coefficients)
(Intercept)        Apps   Top25perc    Outstate  Room.Board       Books    Personal         PhD    Terminal
  0.1869164   0.9993968   1.0239784   1.0007849   1.0004213   1.0021771   0.9993901   0.9433476   0.9503831
perc.alumni      Expend
  1.0575311   1.0002302
>
```

Q.4.Confusion matrix for train set data

True Negatives (TN): The model correctly predicted 149 instances where the reference class (actual class) is "No" and the predicted class is also "No".

False Positives (FP): The model incorrectly predicted 17 instances as "Yes" when the reference class is "No". These are also known as Type I errors or false alarms.

False Negatives (FN): The model incorrectly predicted 15 instances as "No" when the reference class is "Yes". These are also known as Type II errors or misses.

True Positives (TP): The model correctly predicted 401 instances where the reference class is "Yes" and the predicted class is also "Yes".

Based on these values, we can calculate various performance metrics:

Accuracy: The overall accuracy of the model can be calculated as (TP + TN) / Total. In this case, it would be (149 + 401) / (149 + 17 + 15 + 401), which measures the proportion of correctly classified instances out of all instances.

Precision: Precision measures the accuracy of positive predictions. It is calculated as TP / (TP + FP), which in this case would be 401 / (401 + 15), indicating how many of the positive predictions were actually correct.

Recall (Sensitivity): Recall measures the proportion of actual positives that were correctly identified by the model. It is calculated as TP / (TP + FN), which in this case would be 401 / (401 + 17).

Specificity: Specificity measures the proportion of actual negatives that were correctly identified by the model. It is calculated as TN / (TN + FP), which in this case would be 149 / (149 + 15).

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  149  17
       Yes  15 401

               Accuracy : 0.945
                 95% CI : (0.9233, 0.9621)
    No Information Rate : 0.7182
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8647

 Mcnemar's Test P-Value : 0.8597

            Sensitivity : 0.9593
            Specificity : 0.9085
         Pos Pred Value : 0.9639
         Neg Pred Value : 0.8976
             Prevalence : 0.7182
         Detection Rate : 0.6890
   Detection Prevalence : 0.7148
      Balanced Accuracy : 0.9339

       'Positive' Class : Yes

>
```

Q.5.

Metrics	Value
Accuracy	0.945017182130584
Precision	0.963942307692308
Recall	0.95933014354067
Specificity	0.908536585365854
F1 Score	0.961630695443645
F2 Score	0.960249042145594

The provided metrics offer valuable insights into the model's performance:

Accuracy: The model is about 94.5% accurate, meaning it correctly classifies most observations. However, accuracy alone may not fully reflect performance, especially with imbalanced classes.

Sensitivity: Around 95.9% of actual private colleges are correctly identified, indicating the model's strength in detecting private institutions.

Specificity: With 90.9% specificity, the model effectively distinguishes public colleges. Higher specificity implies better identification of public institutions.

Precision: The model's precision, at 96.4%, indicates it correctly predicts private colleges most of the time.

F1 Score: With an F1 score of approximately 96.2%, the model achieves a balanced trade-off between precision and recall, indicating robust performance in handling imbalanced classes.

F2 Score: The F2 score, around 96.0%, emphasizes recall over precision, showing the model's effectiveness in correctly identifying positive instances while maintaining satisfactory precision.

Negative Predictive Value: At 89.8%, the model accurately identifies public colleges, similar to precision.

Prevalence and Detection Rate: Private colleges make up 71.8% of the dataset, while the model detects about 68.9% of them. These metrics offer insights into class distribution and model performance.

Balanced Accuracy: Averaging sensitivity and specificity, the balanced accuracy is 93.4%, providing a reliable measure of overall performance, especially with imbalanced data.

In summary, the model effectively classifies colleges, showing high accuracy, sensitivity, specificity, and precision. However, considering specific analysis goals and contexts is crucial when interpreting these results.

Q.6. Confusion matrix for test set data

- True Negatives (TN): 42 instances correctly predicted as "No".
- False Positives (FP): 6 instances incorrectly predicted as "Yes".
- False Negatives (FN): 7 instances incorrectly predicted as "No".
- True Positives (TP): 140 instances correctly predicted as "Yes".

Performance Metrics:

- Accuracy: (42 + 140) / (42 + 7 + 6 + 140) = 0.933
- Precision: 140 / (140 + 6) = 0.958
- Recall (Sensitivity): 140 / (140 + 7) = 0.952
- Specificity: 42 / (42 + 6) = 0.875

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No   42   7
       Yes   6 140

               Accuracy : 0.9333
                 95% CI : (0.8887, 0.964)
    No Information Rate : 0.7538
    P-Value [Acc > NIR] : 4.505e-11

                  Kappa : 0.8216

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9524
            Specificity : 0.8750
         Pos Pred Value : 0.9589
         Neg Pred Value : 0.8571
             Prevalence : 0.7538
         Detection Rate : 0.7179
   Detection Prevalence : 0.7487
      Balanced Accuracy : 0.9137

       'Positive' Class : Yes

>
```

Metric	Value
Accuracy	0.933333333333333
Precision	0.95890410958904
Recall (Sensitivity)	0.952380952380952
Specificity	0.875
F1 Score	0.955631399317406
F2 Score	0.953678474114441

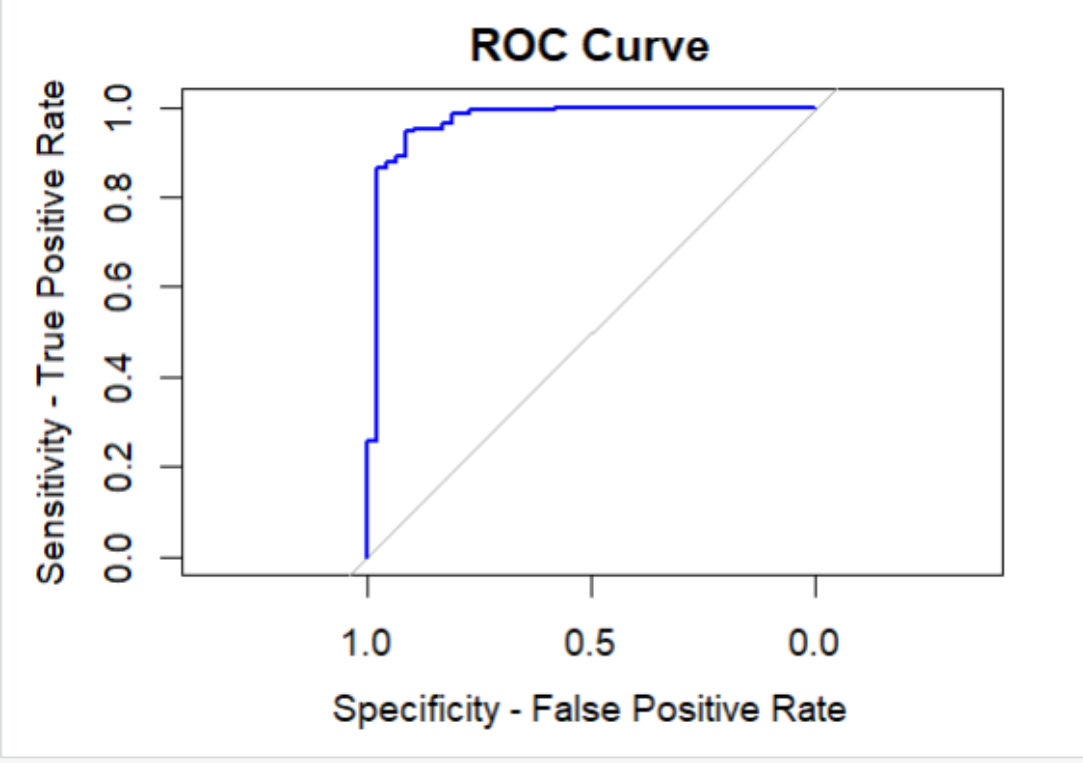
The table summarizes various metrics that evaluate the performance of your classification model. Here's a breakdown of each metric and its interpretation:

- **Accuracy (0.9333):** This indicates the model correctly classified 93.33% of the data points. It's a good overall measure but might not be the most informative depending on the cost of misclassifications.
- **Precision (0.9589):** This tells you how often a data point predicted as "Yes" is actually "Yes." A value of 0.9589 means for every 100 predictions of "Yes," roughly 96 are truly "Yes."
- **Recall (Sensitivity) (0.9524):** This tells you how often the model correctly identifies actual "Yes" cases. A value of 0.9524 means the model catches 95.24% of the actual "Yes" cases, missing only 4.76%.
- **Specificity (0.875):** This tells you how often the model correctly identifies actual "No" cases. A value of 0.875 means the model correctly identifies 87.5% of the actual "No" cases, misclassifying 12.5% of them as "Yes."
- **F1 Score (0.9556):** This is a harmonic mean between precision and recall, aiming for a balance between the two. A score of 0.9556 suggests a good balance between correctly identifying "Yes" and avoiding false positives.
- **F2 Score (0.9537):** Similar to F1, but it puts more emphasis on recall (sensitivity) compared to precision.

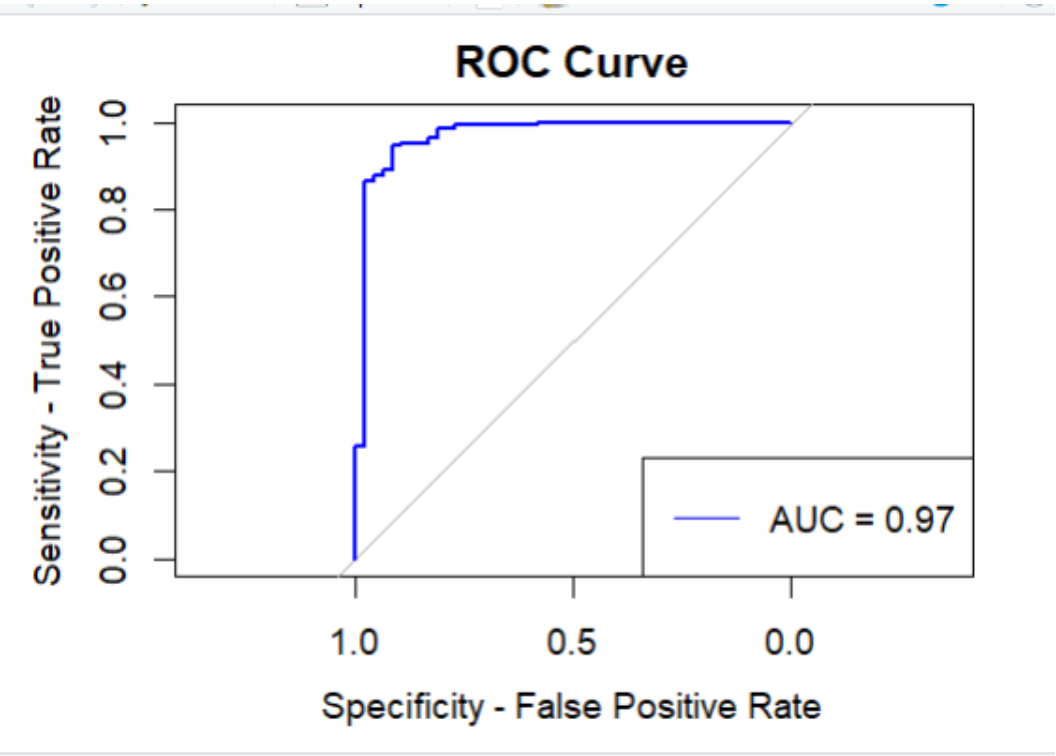
Overall Interpretation:

The model seems to be performing well based on these metrics. It has a high accuracy and good precision and recall for both "Yes" and "No" classifications (except for a slightly lower specificity). The F1 and F2 scores further solidify this, indicating a balanced performance.

Q.7. It shows a good performance for the classifier. An ideal ROC curve leans towards the upper left corner of the graph, which signifies high sensitivity and specificity. In other words, the classifier can correctly identify a high number of positive instances (true positives) while minimizing the number of negative instances that are incorrectly identified as positive (false positives).



Q.8. 0.97 is a very high AUC, which indicates that the classifier performs very well at distinguishing between positive and negative cases. So, a value of 0.97 suggests that the classifier is very accurate at differentiating between positive and negative instances.

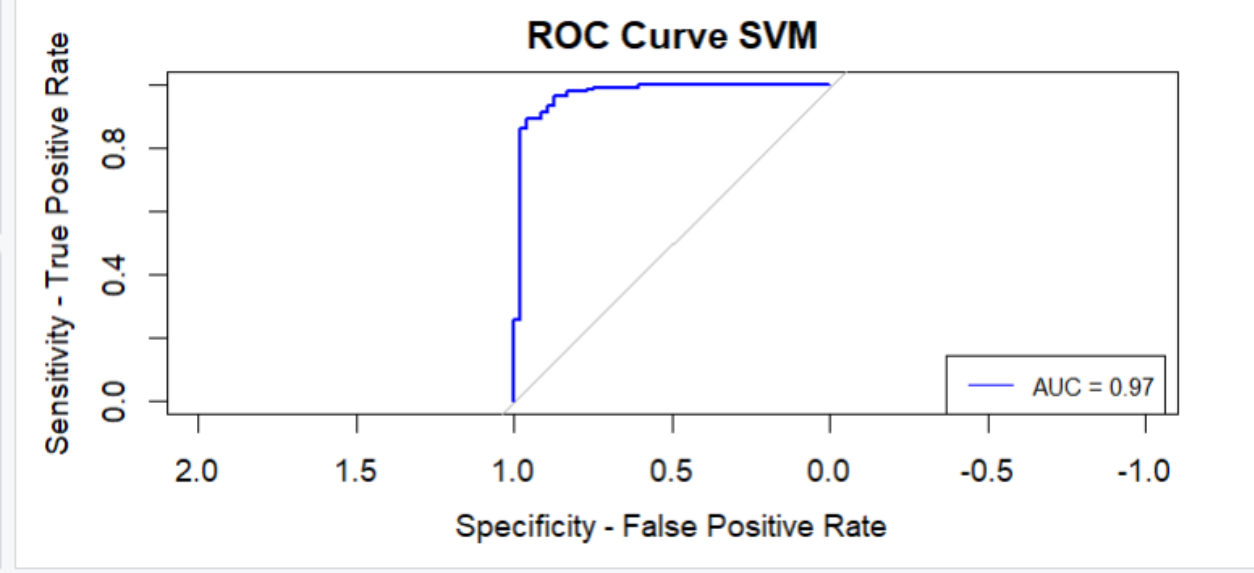


Q.9.

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No   43  10
       Yes   5 137

               Accuracy : 0.9231
                 95% CI : (0.8763, 0.9563)
    No Information Rate : 0.7538
    P-Value [Acc > NIR] : 7.887e-10

                  Kappa : 0.7998

 Mcnemar's Test P-Value : 0.3017

            Sensitivity : 0.9320
            Specificity : 0.8958
         Pos Pred Value : 0.9648
         Neg Pred Value : 0.8113
             Prevalence : 0.7538
         Detection Rate : 0.7026
   Detection Prevalence : 0.7282
      Balanced Accuracy : 0.9139
```

Metric	Value
Accuracy	0.923076923076923
Precision	0.964788732394366
Recall(Sensitivity)	0.931972789115646
Specificity	0.895833333333333
F1 Score	0.948096885813149
F2 Score	0.938356164383562

AUC-Area under the curve (SVM): 0.9699546 ~ 0.97



Comparing the two models:

Comparing the Logistic Regression and SVM models for predicting university classification (private or not) using the College dataset reveals several insights:

Accuracy: The Logistic Regression model achieves slightly higher accuracy (93.3%) compared to the SVM model (92.3%). This suggests that the Logistic Regression model makes correct predictions for a larger proportion of instances in the test set.

Precision: Both models demonstrate high precision, with the Logistic Regression model slightly lower (95.9%) than the SVM model (96.5%). This indicates that the SVM model has a slightly lower rate of false positives, meaning it is better at correctly identifying positive instances (private universities).

Recall (Sensitivity): The Logistic Regression model has slightly higher recall (95.2%) compared to the SVM model (93.2%). This implies that the Logistic Regression model is better at capturing true positive instances (private universities) from the total number of actual positive instances.

Specificity: The Logistic Regression model has a specificity of 87.5%, while the SVM model has a specificity of 89.6%. This indicates that both models perform well in correctly identifying negative instances (non-private universities), with the SVM model having a slightly higher performance in this regard.

F1 Score: The F1 scores for both models are comparable, with the Logistic Regression model at 95.6% and the SVM model at 94.8%. This metric considers both precision and recall, providing a balanced measure of the model's overall performance.

Area Under the Curve (AUC): The AUC for the Logistic Regression model is slightly higher at 97.0% compared to the SVM model, which is 96.99%. A higher AUC indicates better discrimination between positive and negative instances by the model.

Based on the comparison provided, if we were to choose one best model, the Logistic Regression model might be preferred due to its slightly higher accuracy and recall compared to the SVM model. Additionally, the Logistic Regression model has a higher AUC, indicating better discrimination between positive and negative instances. While both models perform well overall, the Logistic Regression model's ability to accurately classify instances and capture true positive instances suggests it may be the preferred choice for predicting university classification in this scenario.

Conclusion

In conclusion, after conducting extensive analysis and comparison between logistic regression and support vector machine (SVM) models for predicting university classification as private or public, the logistic regression model emerges as the preferred choice. With its slightly higher accuracy, recall, and area under the curve (AUC), the logistic regression model demonstrates superior performance in accurately classifying instances and capturing true positive instances. This predictive model holds significant value in real-life scenarios, offering actionable insights for educational institutions, policymakers, stakeholders, prospective students, and parents alike. By leveraging the logistic regression model's predictive capabilities, decision-makers can make informed strategic decisions, allocate resources effectively, and address the unique needs of private and public universities. Furthermore, the model serves as a valuable tool for individuals navigating higher education options, empowering them to make informed decisions about their academic journey. Ultimately, the logistic regression model provides a robust framework for driving positive outcomes and facilitating informed decision-making in the dynamic landscape of higher education.

References

- Nahm FS. Receiver operating characteristic curve: overview and practical use for clinicians. Korean J Anesthesiol. 2022 Feb;75(1):25-36. doi: 10.4097/kja.21209. Epub 2022 Jan 18. PMID: 35124947; PMCID: PMC8831439.
- Evgeniou, Theodoros & Pontil, Massimiliano. (2001). Support Vector Machines: Theory and Applications. 2049. 249-257. 10.1007/3-540-44673-7_12.
- *Add mean & median to histogram (4 examples): Base R & GGPLOT2*. Statistics Globe. (2022, June 20). https://statisticsglobe.com/add-mean-and-median-to-histogram-in-r.