Syed Faizan

College Type Prediction

Main Tool: R Programming Language

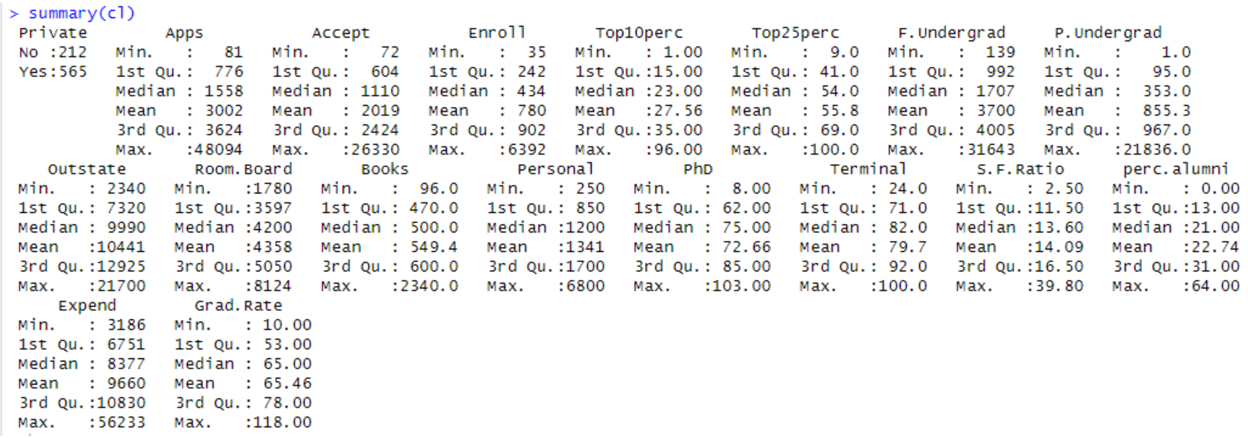
Technique: Logistic Regression

Syed Faizan

Introduction

This report details the analysis, utilizing the glm() function in R, to construct a Logistic Regression model, making use of the College dataset from the ISLR library to determine if a university was designated as private or public (James, Witten, Hastie, & Tibshirani, 2013). This endeavor acted as a hands-on demonstration of logistic regression, an essential statistical technique for binary categorization. The task helped me enhance my abilities in employing R for detailed data analysis, model creation, and data interpretation. Through this project, I acquired practical experience in dataset preparation, conducting exploratory data evaluations, and applying sophisticated modeling strategies to heighten predictive precision. This project not only improved my technical skills but also equipped me to address both strategic and operational queries using generalized linear models, effectively connecting theoretical knowledge with practical application scenarios. I have presented the analysis in the form of answers to the project questions in order to maintain continuity and order.

I imported the data set ‘college’ from the ISLR library into R Studio as ‘cl’ for convenience.

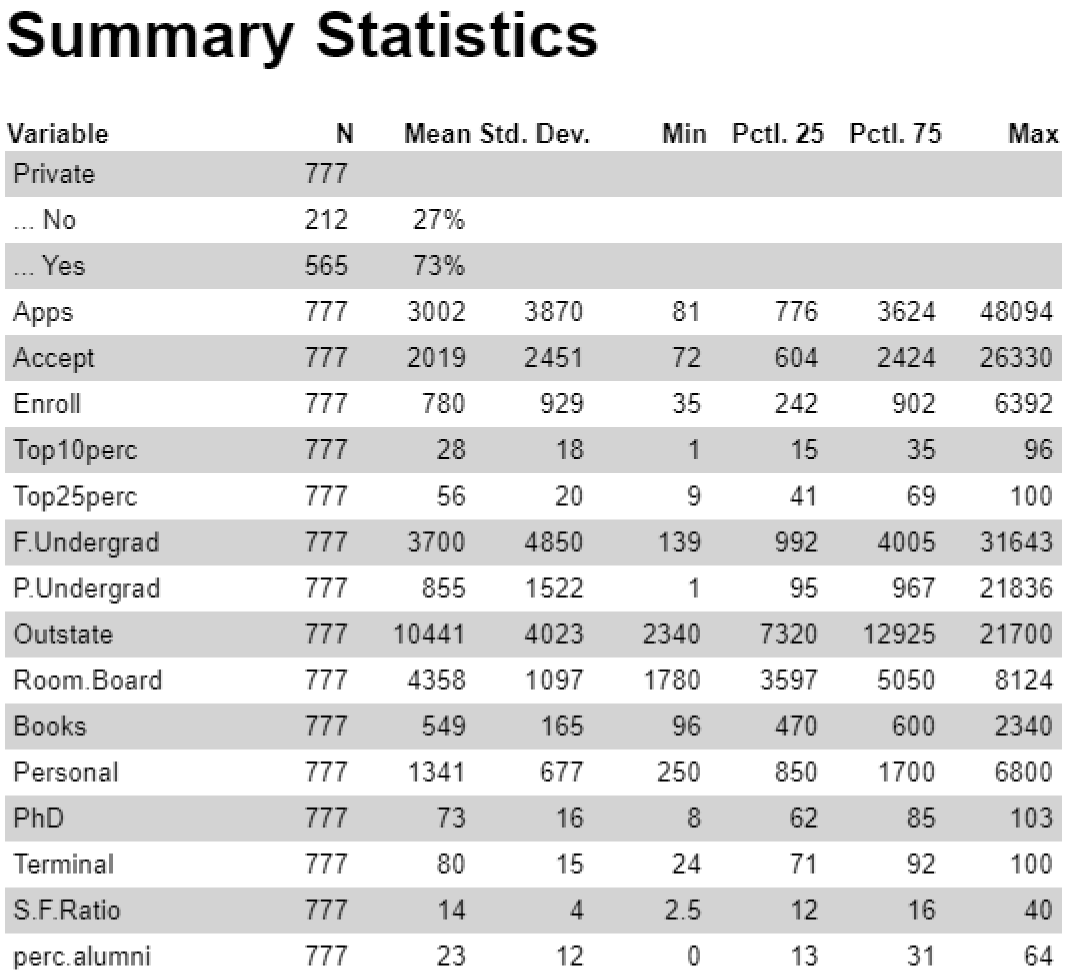


```
> summary(cl)
 Private        Apps           Accept          Enroll        Top10perc       Top25perc      F.Undergrad     P.Undergrad     
 No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00   Min.   :  9.0   Min.   :  139   Min.   :    1.0  
 Yes:565   1st Qu.:  776   1st Qu.: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

Figure 1 Summary of the College dataset in R

The dataset was imported as a data frame. The dataset is part of the famous ISLR package that goes with the classic textbook ‘Introduction to Statistical Learning (ISL) with applications in R’ from Stanford professors Trevor Hastie and Tibshirani (James, Witten, Hastie, & Tibshirani, 2013) who were also pioneers in the development of the S programming language which was a precursor to R.

The dataset contains data pertaining to US Colleges from the 1995 issue of US News and World Report. There are no missing values in the data set.



Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Private	777						
... No	212	27%					
... Yes	565	73%					
Apps	777	3002	3870	81	776	3624	48094
Accept	777	2019	2451	72	604	2424	26330
Enroll	777	780	929	35	242	902	6392
Top10perc	777	28	18	1	15	35	96
Top25perc	777	56	20	9	41	69	100
F.Undergrad	777	3700	4850	139	992	4005	31643
P.Undergrad	777	855	1522	1	95	967	21836
Outstate	777	10441	4023	2340	7320	12925	21700
Room.Board	777	4358	1097	1780	3597	5050	8124
Books	777	549	165	96	470	600	2340
Personal	777	1341	677	250	850	1700	6800
PhD	777	73	16	8	62	85	103
Terminal	777	80	15	24	71	92	100
S.F.Ratio	777	14	4	2.5	12	16	40
perc.alumni	777	23	12	0	13	31	64

Figure 2 Descriptive Statistics for the dataset

The College dataset encompasses 777 observations across 18 variables, providing a comprehensive overview of different aspects of higher education institutions. A significant dichotomy within this dataset is the 'Private' variable, which categorizes universities into private (73%) and public (27%), indicative of a higher prevalence of private institutions within this collection. All variables apart from this one are continuous numeric. The row names contain the names of the institutions. The number of applications received ('Apps') exhibits considerable variability, with a mean of 3002 and a broad range from 81 to 48094, suggesting diversity in the popularity or selectivity of these institutions. (Upon further investigation I found that this institution, which received 48094 applications was an outlier.)

Acceptance numbers ('Accept') follow suit with a mean of 2019, ranging markedly from 72 to 26330, while the enrollment figures ('Enroll') average to 780, spanning from 35 to 6392. This variance may reflect differing institutional capacities and enrollment strategies. Academic stature, as captured by the percentage of new students from the top 10% ('Top10perc') and top 25% ('Top25perc') of high school classes, presents averages of 28% and 56%, respectively, maxing out at 100%, which underscores the competitive edge of certain universities in attracting high-achieving students.

The full-time ('F.Undergrad') and part-time ('P.Undergrad') undergraduate populations mean at USD $3700 and $ 855 USD, with wide ranges suggesting diverse institutional sizes. Fiscal aspects, including out-of-state tuition ('Outstate') and room and board costs ('Room.Board'), present averages of $10441 USD and $4358 USD, respectively, reflecting the financial commitments expected from students. Meanwhile, expenditures per student ('Expend') average $9660 USD, with a maximum value of $56233 USD, possibly indicating a discrepancy in the financial resources and investments made towards educational quality across institutions.

Faculty credentials, quantified by the percentage holding Ph.D.'s ('PhD') and terminal degrees ('Terminal'), show averages of 72.66% and 79.7%, revealing a substantial proportion of highly qualified staff. The student-to-faculty ratio ('S.F.Ratio') averages at 14, while alumni donation rates ('perc.alumni') at 23%, both metrics offering insight into the community and support aspects of these institutions. Finally, the graduation rate ('Grad.Rate') has an average of 65.46%, underscoring the varying success rates of students completing their studies.

Univariate Analysis

It is important to bear in mind the objective of the data analysis while undertaking an EDA. Keeping in mind the anticipated logistic regression I sought to carry out a univariate analysis in pseudo bivariate form, looking at visualizations that would help me gauge how the continuous variables stack against the ‘Private’ column that will serve as our response variable in our predictive logistic regression modelling.

I created a function to automate boxplots for the continuous variables with respect to the private and public colleges.

The first box plot, namely the applications received by college type required special treatment owing to an outlier of 48094 applications that made visualization difficult, so I ventured to remove its presence through zooming in on the plot produced in R Studio.

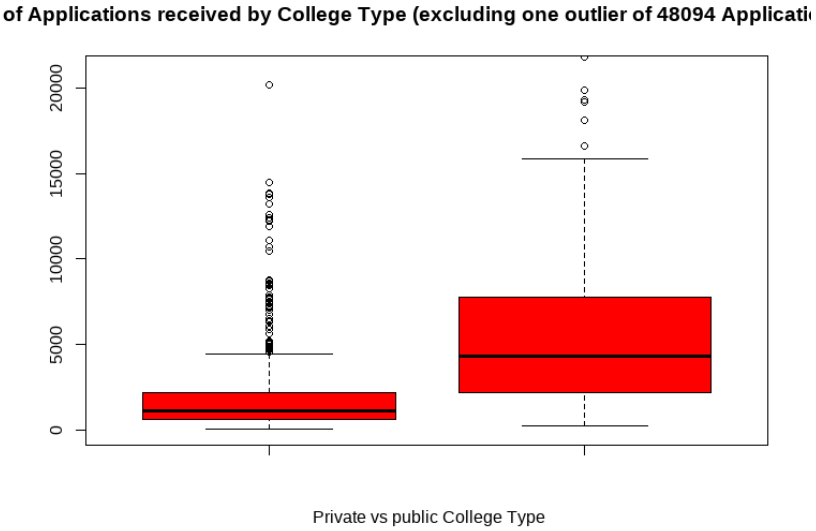


Figure 3 Box plot of applications received by college type presented the challege of a huge outlier

The above boxplot illustrates the distribution of applications received by the two distinct categories of higher education institutions represented in this dataset: private and public colleges, with an identified outlier of 48,094 applications being excluded for enhanced clarity. The median value, highlighted by the horizontal line within each box, is notably higher for public colleges, suggesting that, on average, they receive a greater number of applications compared to private colleges. This disparity could reflect the larger size and potentially lower tuition rates of public institutions, making them accessible to a wider applicant pool.

In both categories, the presence of outliers, as depicted by individual points beyond the whiskers, indicates a substantial variance in application numbers among institutions, with some colleges receiving substantially more applications than their counterparts.

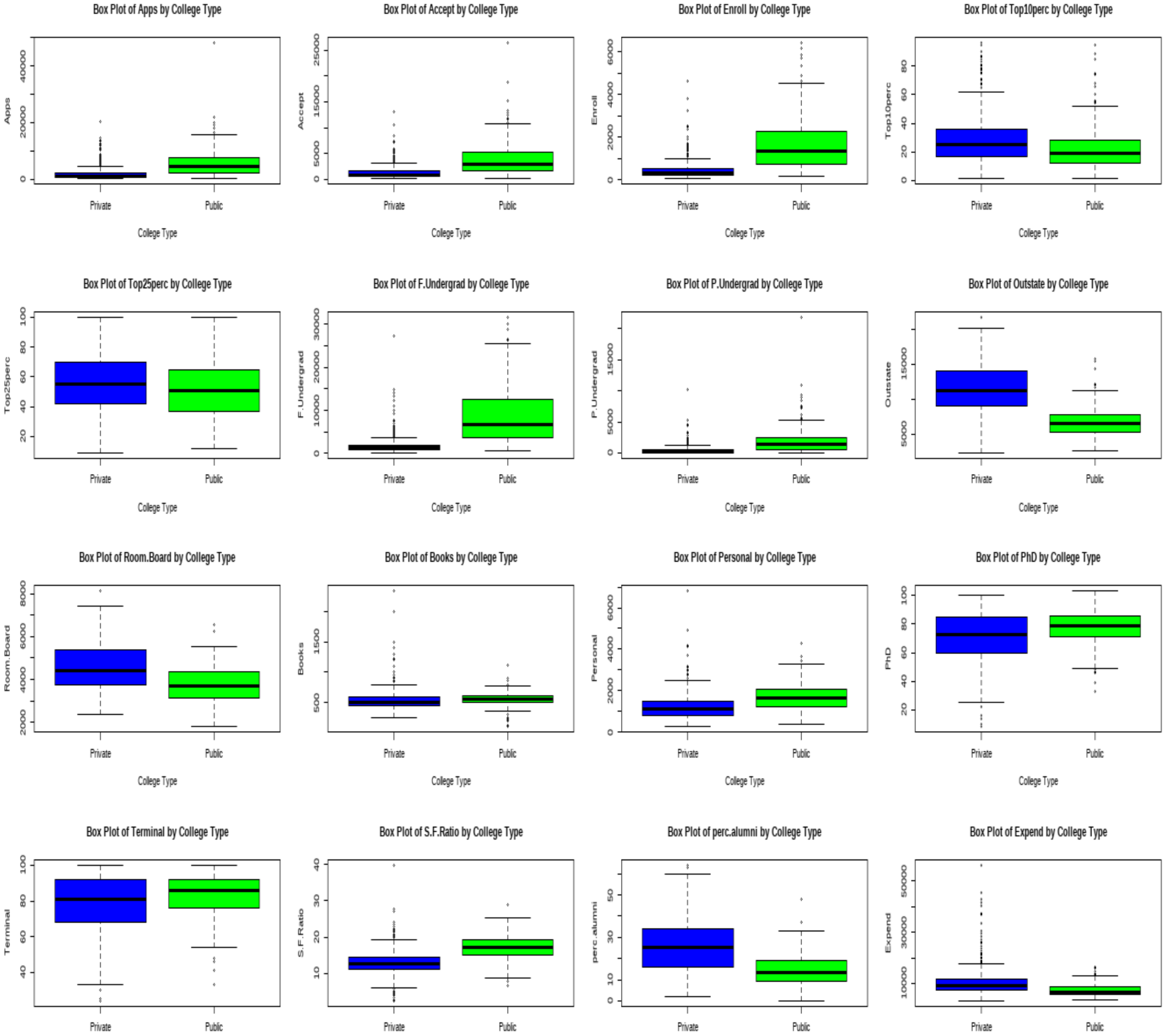


Figure 4 Boxplots of the numerical explanatory variables by the response variable 'college type: Private or Public'.

The above shown collection of box plots provides a visual comparison across various numerical explanatory variables, stratified by the binary response variable delineating 'Private' and 'Public' college types. Each plot represents a distinct variable from the dataset, facilitating an examination of the central tendency, dispersion, and skewness of the distributions, as well as the presence of outliers within each subgroup.

Applications ('Apps') and acceptances ('Accept') demonstrate a higher median and greater variability in public colleges, which could be attributed to their larger size and broader appeal. Enrollment figures ('Enroll') and full-time undergraduate numbers ('F.Undergrad') follow a similar pattern, hinting at the expansive capacity of public institutions. Conversely, private colleges exhibit a higher proportion of students from the top 10% ('Top10perc') and top 25% ('Top25perc') of their high school classes, suggesting a potentially more selective or prestigious cohort.

Part-time undergraduate numbers ('P.Undergrad') and out-of-state tuition ('Outstate') present a pronounced disparity, with public institutions again showcasing a wider range and higher medians, reflecting both their diversity and the broader spectrum of tuition fees. The costs associated with room and board ('Room.Board'), books ('Books'), and personal expenses ('Personal') are similarly distributed, with a noticeable spread among public colleges, indicative of the range of living costs and student lifestyles.

Faculty qualifications, measured by the percentage with Ph.D.'s ('PhD') and terminal degrees ('Terminal'), show a slightly higher median for private colleges, pointing to a concentration of highly qualified faculty within these institutions. The student-to-faculty ratio ('S.F.Ratio') is notably more favorable in private colleges, potentially translating into more personalized attention and smaller class sizes. Alumni donation percentages ('perc.alumni') are significantly higher in private colleges, suggesting a stronger alumni engagement or satisfaction. Lastly, instructional expenditures per student ('Expend') display a pronounced median and range for private colleges, which may correlate with higher investments in educational quality and student services.

Together, these box plots underscore the fundamental differences in the demographic, financial, and academic profiles between private and public colleges that can be leveraged to our advantage as we seek to build a binary logistic prediction model that predicts whether an institution is public or private using these numerical variables. After a preliminary visual examination of these box plots I have found Applications ('Apps'), full-time undergraduate numbers ('F.Undergrad'), out-of-state tuition ('Outstate') and Alumni donation percentages ('perc.alumni') to be potential candidates to serve as explanatory variables in a predictive modelling context.

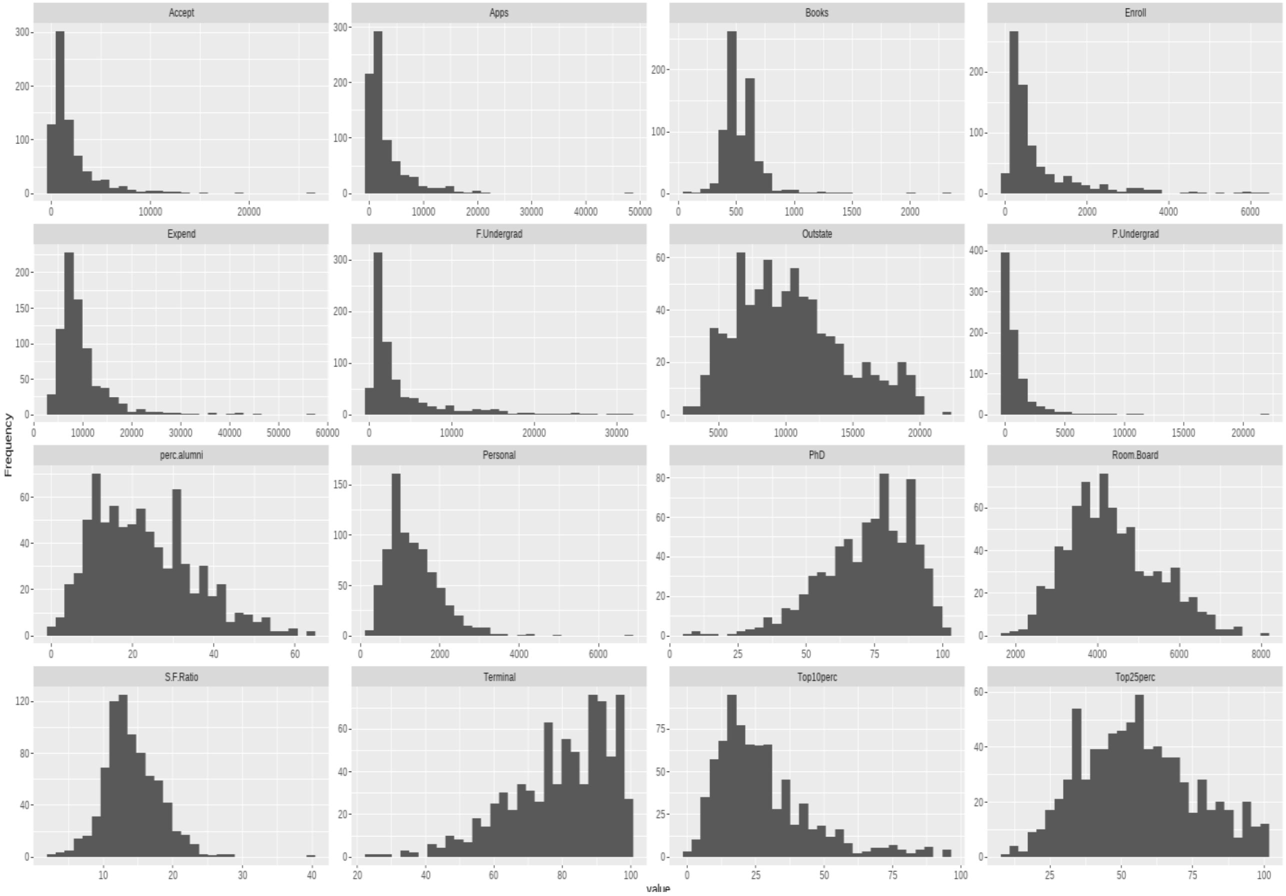


Figure 5 Histograms of the numerical variables

In the above histograms the Applications ('Apps') and acceptances ('Accept') reveal heavily right-skewed distributions, indicating that a majority of institutions receive a relatively low number of applications and acceptances compared to a few colleges with exceptionally high numbers. This skewness may reflect a concentration of applicants among a select group of institutions, possibly due to varying degrees of prestige or program offerings.

Full-time undergraduate enrollment ('F.Undergrad') and part-time undergraduate enrollment ('P.Undergrad') show similar right-skewed patterns, suggestive of the presence of predominantly smaller institutions with a limited number of larger universities that significantly expand the upper range of these variables.

The out-of-state tuition ('Outstate') and room and board costs ('Room.Board') exhibit a somewhat less skewed distribution, yet still lean rightward, underscoring the economic diversity across colleges, with a subset imposing substantially higher costs.

Cost-related variables such as book expenses ('Books'), personal spending estimates ('Personal'), and instructional expenditures per student ('Expend') display right-skewness as expected, indicative of wide-ranging financial demands on students and investment disparities in educational facilities and resources among the surveyed institutions.

The histogram of the percentage of alumni who donate ('perc.alumni') suggests that while a considerable proportion of institutions enjoy modest alumni donation rates, there exists a tail of colleges with exceptionally high rates, potentially reflecting varying degrees of alumni engagement or satisfaction.

Academic quality indicators like the percentage of faculty with Ph.D.'s ('PhD'), the percentage with terminal degrees ('Terminal'), and the percentage of new students from the top 10% ('Top10perc') and top 25% ('Top25perc') of their high school classes present multimodal distributions. This implies the existence of distinct clusters within the institutions, possibly aligning with institutional type or mission.

Finally, the student-to-faculty ratio ('S.F.Ratio') assumes a distribution with a clear central tendency and less pronounced tails, pointing towards a more uniform distribution across colleges in terms of class size and faculty engagement.

It must be borne in mind that unlike in linear regression, normality of distribution is ***not*** an assumption of a generalized regression model using a 'logit' link function.

Bivariate Analysis

Bivariate analysis of the variables was carried out with a view to examining the relationships between the columns identified as having potential explanatory variables after the univariate analysis. Thus after a preliminary viewing of the overall scatterplots using the pairs() function, I focused on Applications ('Apps') , full-time undergraduate numbers ('F.Undergrad'), out-of-state tuition ('Outstate') and Alumni donation percentages ('perc.alumni') along with one additional scatterplot between expenditure related variables.

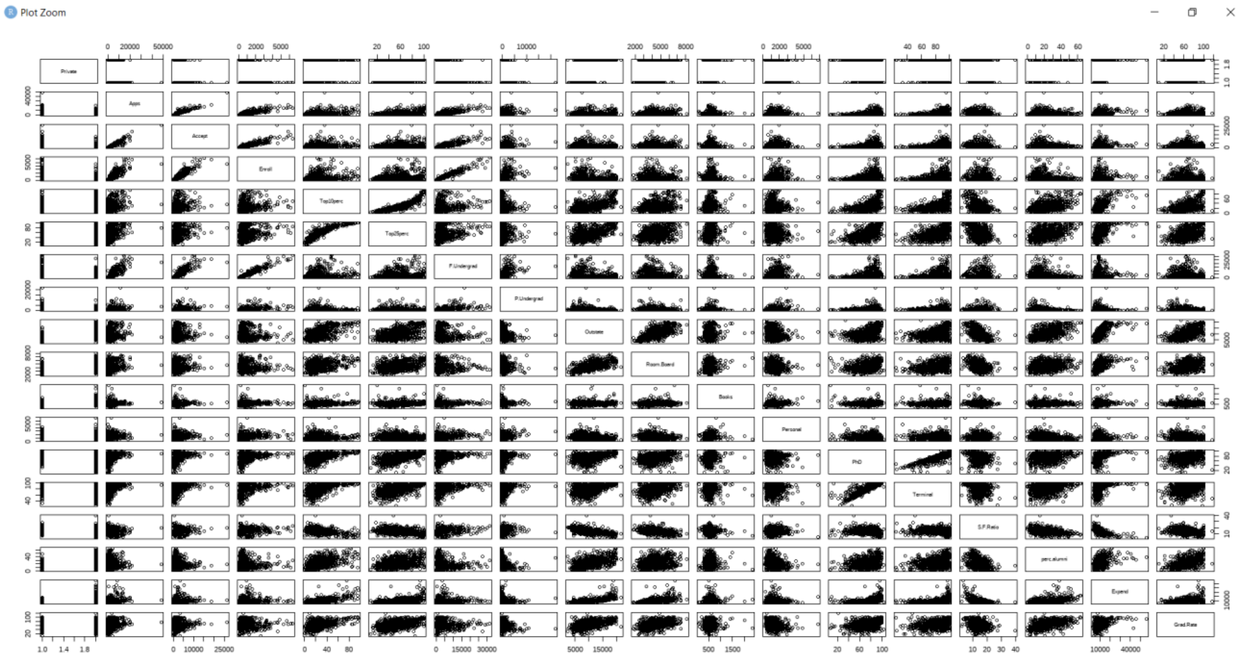


Figure 6 Pairs() function used to obtain an overview of scatterplots between the continuous variables.

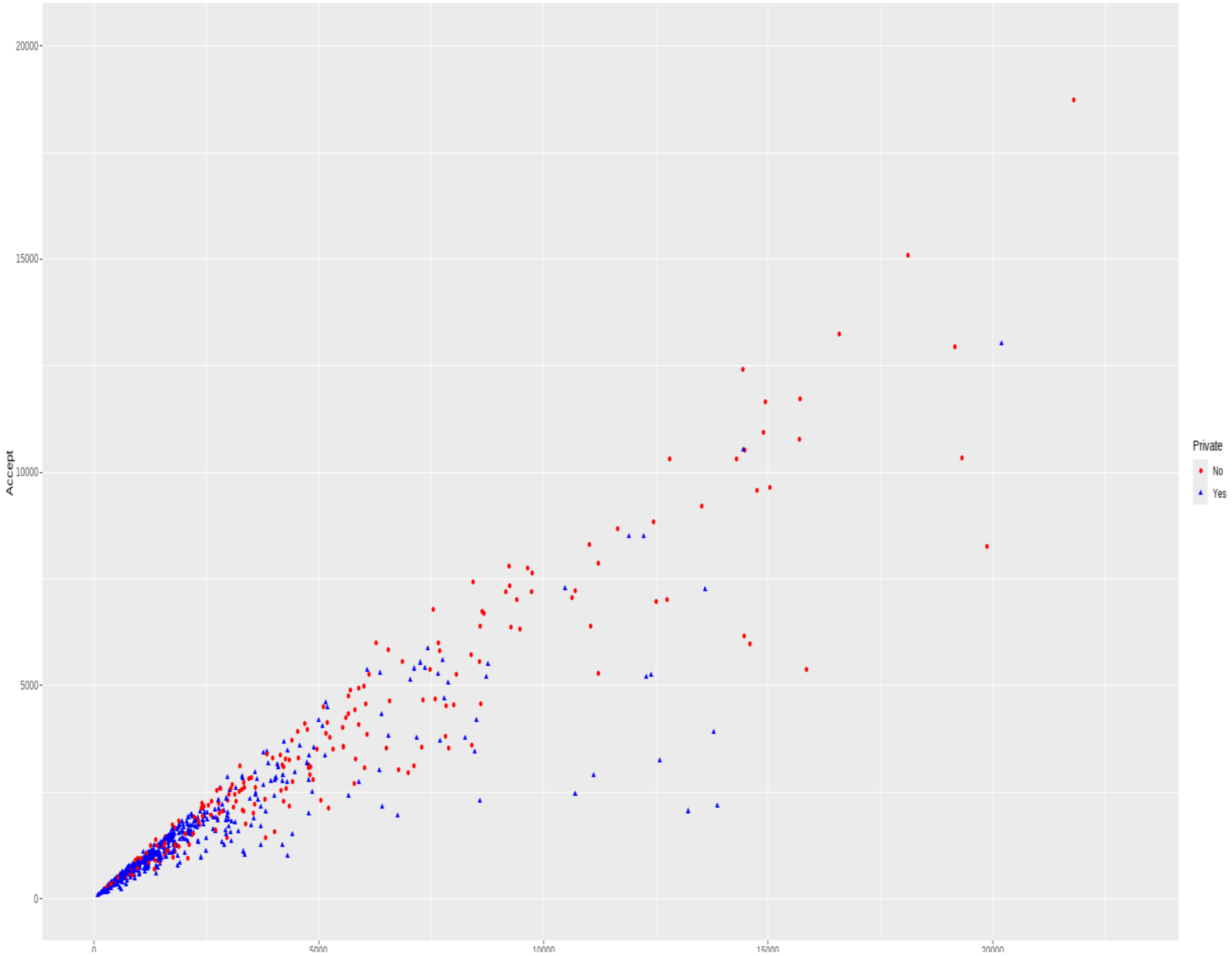


Figure 7 Scatterplot of Applications accepted on y-axis and apppications submitted on the x- axis stratified by color and shape for college type.

The scatterplot visualizes the relationship between the number of applications submitted (x-axis) and the number of applications accepted (y-axis), stratified by college type, with private colleges denoted by blue triangles and public colleges by red circles. The dispersion of points suggests a positive correlation between applications submitted and accepted, which is expected as colleges with larger applicant pools tend to admit more students. The linear relationship counsels against the inclusion of ***both*** these variables in a binary logistic model as multicollinearity and confounding might turn out to be an issue.

A closer inspection reveals that public colleges (red circles) are generally associated with higher values on both axes, indicating that one of these variables must be looked upon favorably while including predictors in our logistic model. Notably, the highest values for both applications submitted and accepted appear to belong predominantly to public colleges, possibly indicating a broader appeal or larger capacity.

The distribution of private colleges shows a tighter cluster, especially towards the lower end of the scale, hinting at a more selective or smaller-scale application process. Also noticeable are prominent outliers on both axes.

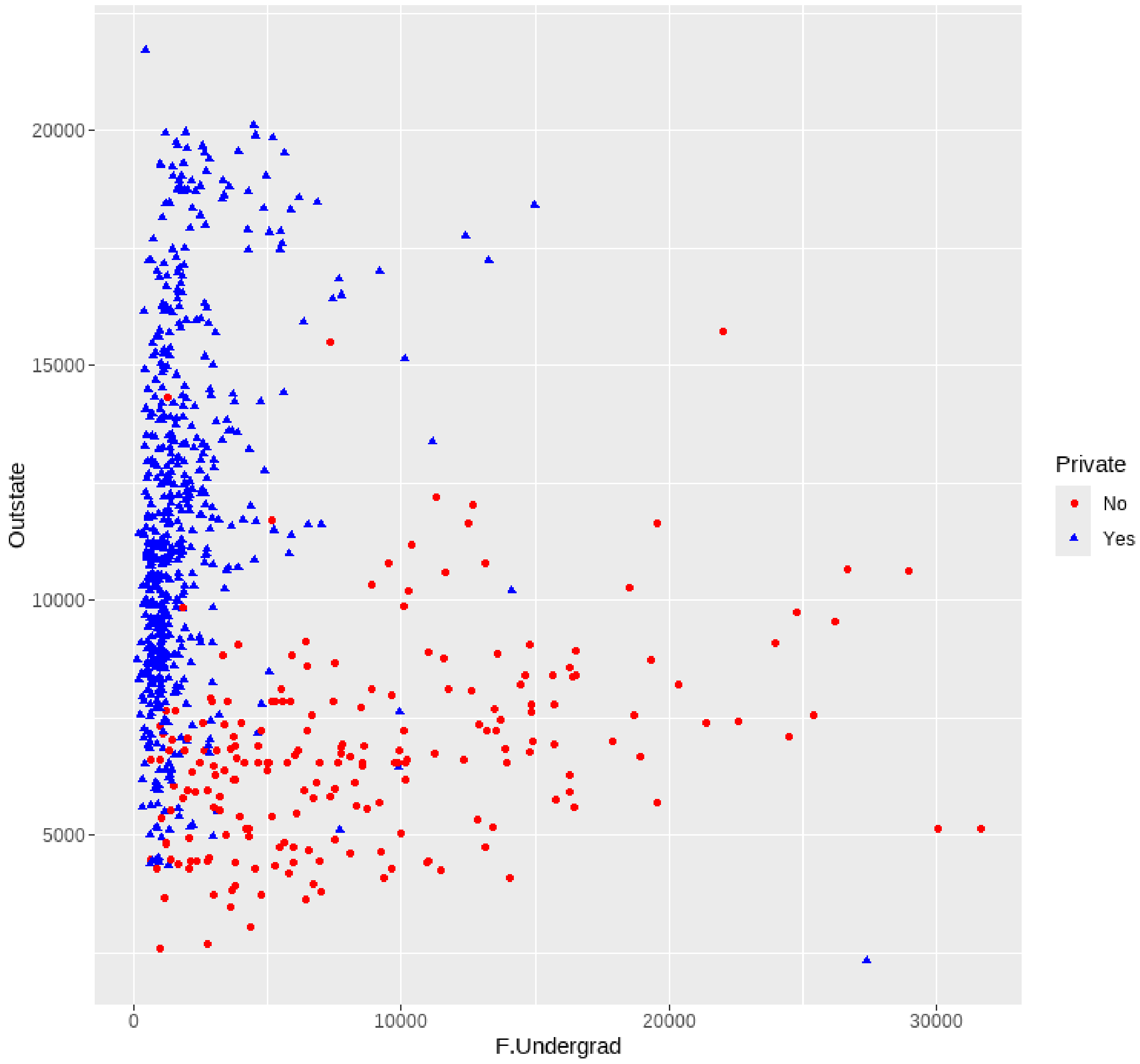


Figure 8 Scatterplot of Out-of-state tuition on y-axis and Full-time undergraduates on the x- axis stratified by color and shape for college type.

The above scatterplot elucidates the distribution of out-of-state tuition against the number of full-time undergraduates, stratified by college type. For a binary logistic model predicting college type, these variables could be potential predictors. The separation in tuition fees suggests that out-of-state tuition might be a strong predictor, with private colleges clustering at the higher end of tuition fees and public colleges at the lower end.

However, the number of full-time undergraduates does not display a clear distinction between private and public colleges, indicating that it might be a weaker predictor of college type. The overlap in the number of full-time undergraduates between private and public colleges may reduce the predictive power of this variable when used alone.

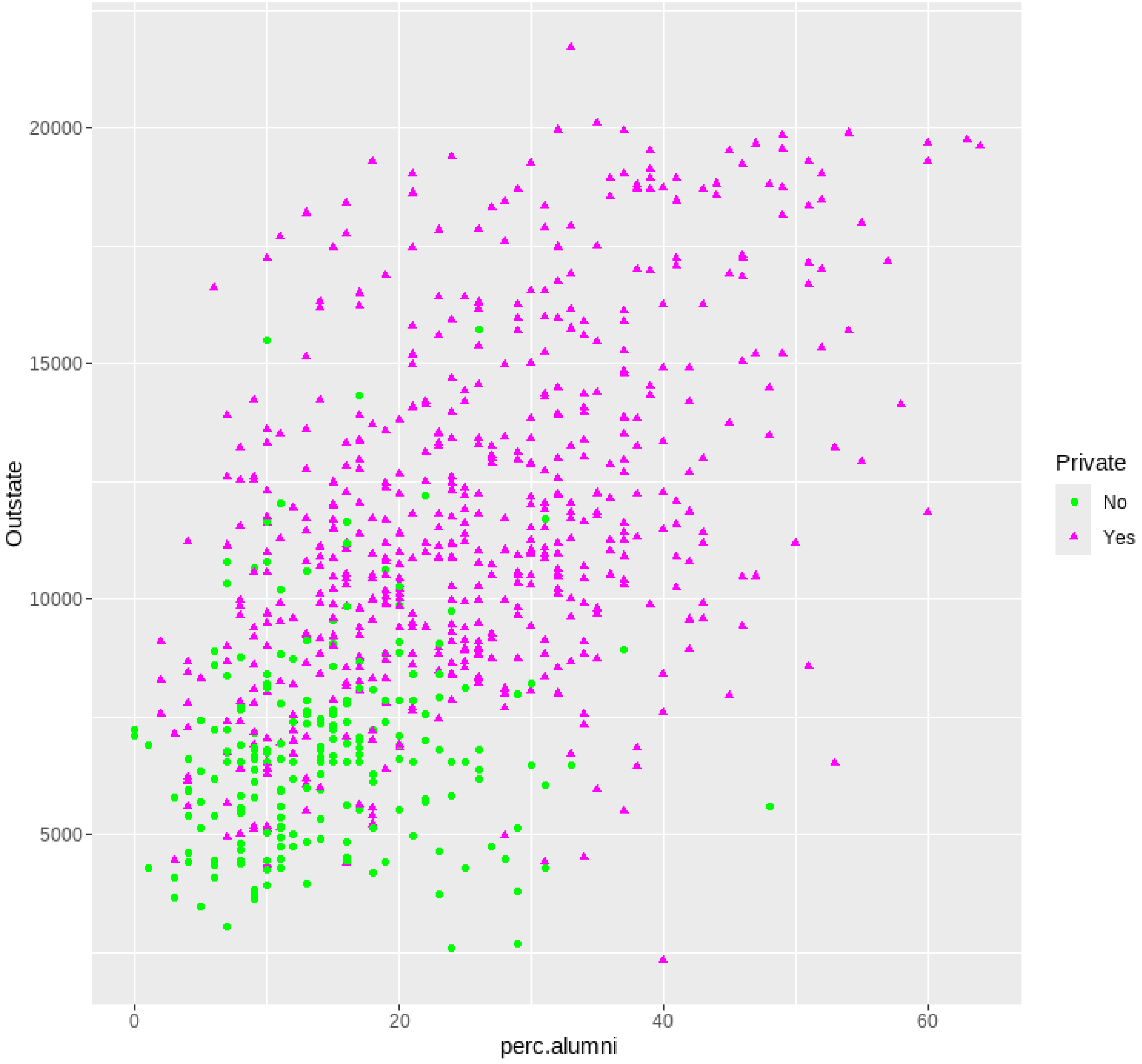


Figure 9 Scatterplot of Out-of-state tuition on y-axis and Percentage of alumni donating on the x- axis stratified by color and shape for college type.

The above displayed scatterplot offers a bivariate analysis of out-of-state tuition fees against the percentage of alumni who donate, color and shape-coded to represent private (purple triangles) and public (green circles) institutions. An initial observation reveals no clear, discernible pattern suggesting a straightforward linear relationship between the percentage of alumni donors and tuition fees across college types.

For our purposes of constructing a binary logistic regression aiming to predict college type, the percentage of alumni donors shows potential as an explanatory variable given the noticeable clustering of private colleges at higher percentages of alumni donations. This suggests that private institutions may have more successful alumni engagement or benefits that encourage giving. Out-of-state tuition also appears to separate college types to some extent, with public colleges displaying a larger spread in tuition fees.

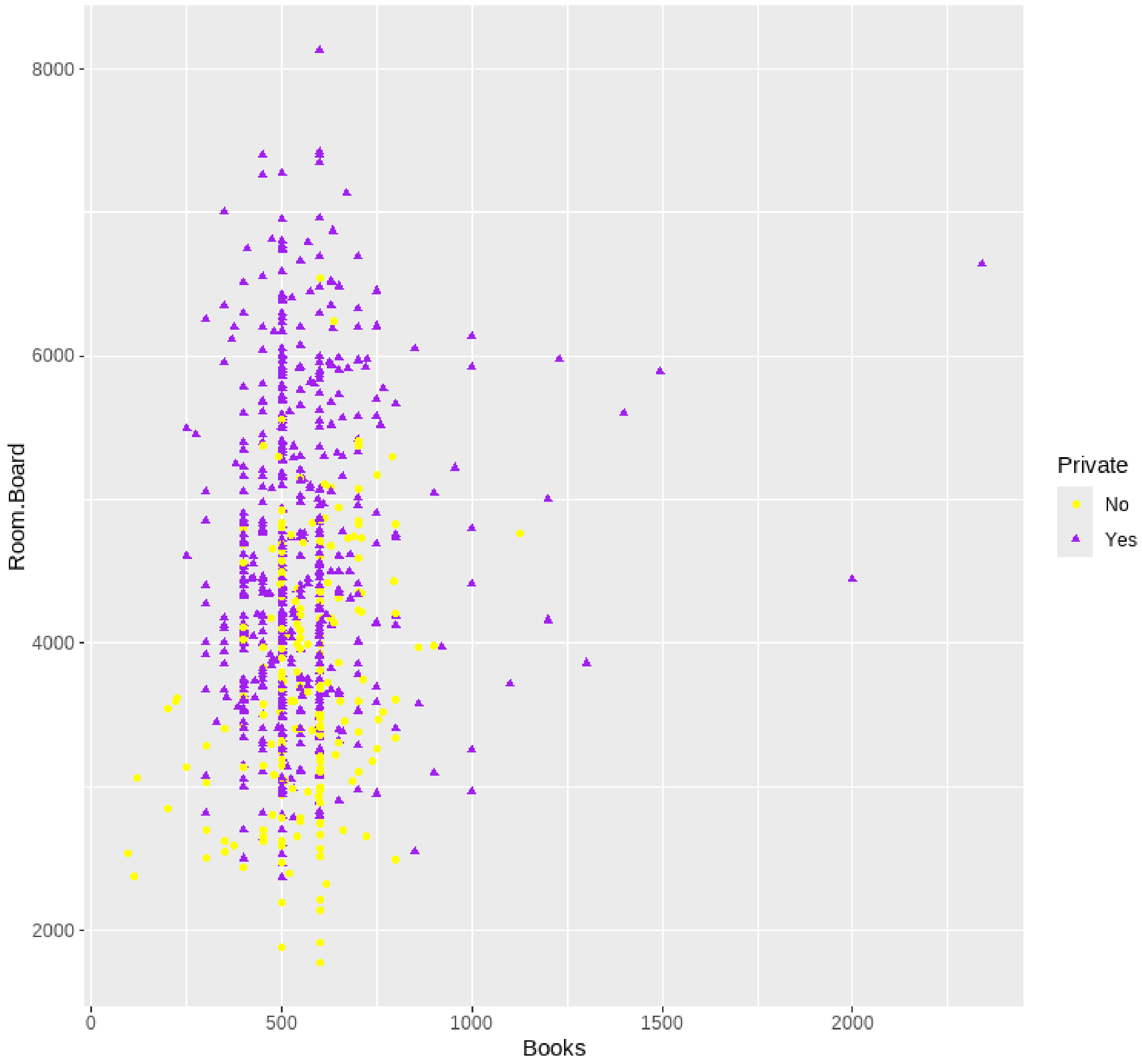


Figure 10 Scatterplot of cost of Room and Board on y-axis and the expenditure on books on the x- axis stratified by color and shape for college type

The above scatterplot maps the cost of room and board against book expenditures for colleges, categorized by type—private institutions (purple triangles) and public institutions (yellow circles). The cost of room and board appears to be higher for private colleges, while book expenditures do not show a clear distinction between college types, suggesting that room and board costs could be a more discriminating factor for predicting college type in a binary logistic model.

The distribution of data points reveals a denser concentration at lower book expenses and moderate room and board costs, with private colleges more frequently occupying the higher end of room and board costs. However, there is considerable overlap in the central mass of data points for both variables across college types, which could limit the predictive utility of these variables when used independently in a logistic model.

Feature Engineering

The feature engineering involved the creation of two novel columns aimed at facilitating subsequent analyses. The first addition entailed generating a unique identifier for each institution, derived from the dataset's row names, thereby enabling a clear demarcation of each college within the dataset.

The second modification addressed the binary nature of the institution type. The variable 'Private', which categorizes colleges into 'Private' and 'Public', was transformed into a numeric binary format in a new column called ‘college binary’ where 'Public' colleges were assigned a value of 0, and 'Private' colleges were assigned a value of 1. This numeric encoding simplifies the use of this categorical variable in correlation analysis aimed at finding which predictors are correlated most with college type, so as to include them in our predictive binary logistic modelling.

Subsequent to these enhancements, the 'Private' variable was converted into a factor, a data structure suitable for representing categorical data in R. Additionally, the category 'Public' was established as the reference level within this factor. **The reference level in a factor is critical as it serves as the baseline against which other categories are compared in logistic regression analyses.**

Choosing the prospective predictors in the logistic regression model:Correlation Analysis

There are several means by which we may choose which predictors to include in our logistic regression model. This selection is determined primarily by the object of the intended logistic regression model.

Logistic models serve primarily two purposes:

1. Predictive modelling.
2. Effect size modelling.

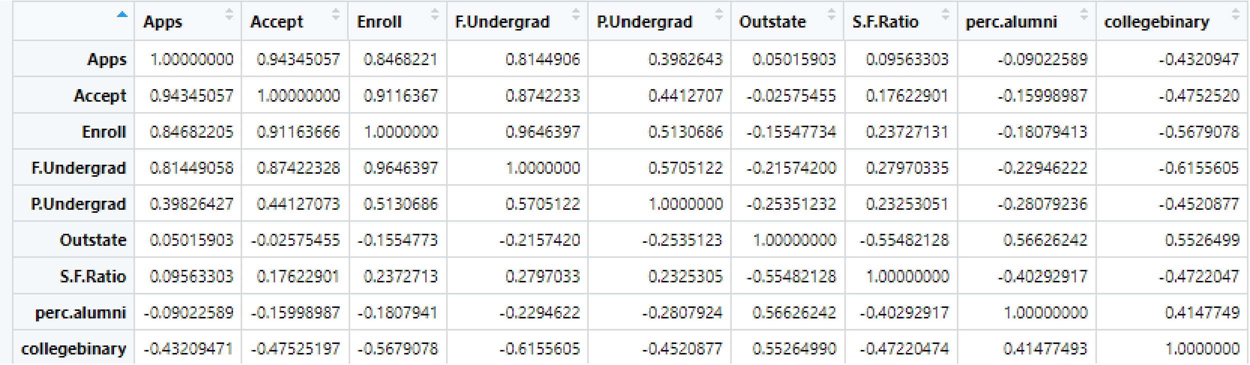
Since our project specifies a predictive model we choose those variables that may serve best as predictors based on the following analyses:

1. Conceptual understanding: Certain predictors might be logically and conceptually expected to play a major role in prediction of college type even without and prior to statistical modelling i.e. a priori based on a conceptual understanding of the domain. For example, a higher ‘out-of-state’ tuition can be expected to, conceptually, be linked with private institutions based on a general comprehension of the education sector in the U.S.A. This may be considered ‘face validation’.
2. Bivariate Analyses: Hypothesis testing and correlation analysis are important to choose proper predictors. Since we have already carried out a bivariate analysis we proceed to correlation.

Biserial Correlation Coefficient analysis:

I initially included all the continuous variables in the correlation analysis. However, in order to render the correlation plot less busy and more pleasing to the eye I included only those variables that had an absolute value of the Pearson's correlation coefficient higher than 0.4 .

Biserial Correlation analysis refers to a specific type of correlation between continuous numerical predictors and a binary coded categorical variable. In our case we sought to understand if college type (encoded to 0 = 'Public' and 1 = 'Private') can be found to correlate with continuous variable predictors.



	Apps	Accept	Enroll	F.Undergrad	P.Undergrad	Outstate	S.F.Ratio	perc.alumni	collegebinary
Apps	1.00000000	0.94345057	0.8468221	0.8144906	0.3982643	0.05015903	0.09563303	-0.09022589	-0.4320947
Accept	0.94345057	1.00000000	0.9116367	0.8742233	0.4412707	-0.02575455	0.17622901	-0.15998987	-0.4752520
Enroll	0.84682205	0.91163666	1.0000000	0.9646397	0.5130686	-0.15547734	0.23727131	-0.18079413	-0.5679078
F.Undergrad	0.81449058	0.87422328	0.9646397	1.0000000	0.5705122	-0.21574200	0.27970335	-0.22946222	-0.6155605
P.Undergrad	0.39826427	0.44127073	0.5130686	0.5705122	1.0000000	-0.25351232	0.23253051	-0.28079236	-0.4520877
Outstate	0.05015903	-0.02575455	-0.1554773	-0.2157420	-0.2535123	1.00000000	-0.55482128	0.56626242	0.5526499
S.F.Ratio	0.09563303	0.17622901	0.2372713	0.2797033	0.2325305	-0.55482128	1.00000000	-0.40292917	-0.4722047
perc.alumni	-0.09022589	-0.15998987	-0.1807941	-0.2294622	-0.2807924	0.56626242	-0.40292917	1.00000000	0.4147749
collegebinary	-0.43209471	-0.47525197	-0.5679078	-0.6155605	-0.4520877	0.55264990	-0.47220474	0.41477493	1.0000000

Figure 11 Correlation matrix refined to include only those potential predictors that have an absolute value of r more than 0.4

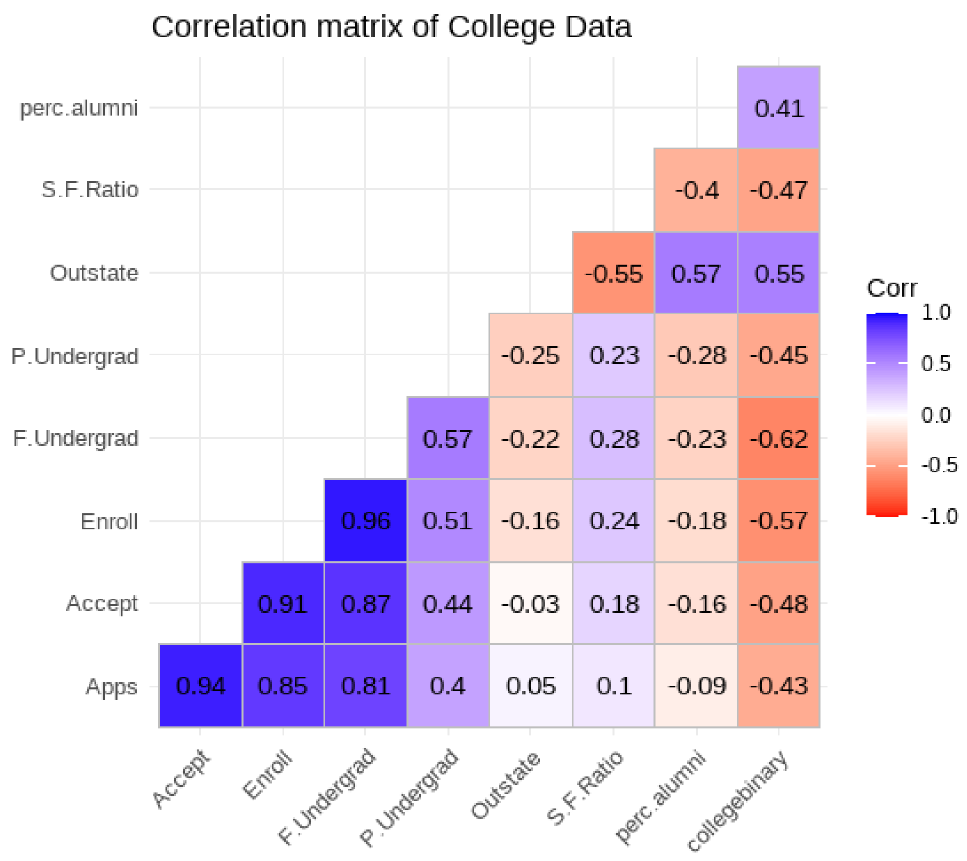


Figure 12 Plot showing the correlation matrix between the binary encoded college type response variable and potential predictors

The above correlation matrix reveals that the variable 'Outstate' shows a relatively strong positive correlation with the college binary variable (r ≈ 0.55). This implies that as the 'Outstate' tuition increases, the likelihood of the college being private rather than public also increases. This may be reflective of the general trend that private colleges often charge higher out-of-state tuition fees compared to public colleges.

Conversely, the variable 'perc.alumni' displays a negative correlation with the college binary variable (r ≈ -0.45). This suggests that colleges with a higher percentage of alumni donors are more likely to be public institutions. This could be interpreted as an indicator of the alumni's connection to the institution, potentially influenced by the scale of the institution and the nature of its funding.

The 'F.Undergrad' and 'P.Undergrad' variables, representing the number of full-time and part-time undergraduates, respectively, show negligible correlation with the college binary variable (r ≈ -0.22 and r ≈ -0.45). This indicates that the size of the undergraduate population, whether full-time or part-time, is not a strong determinant of whether a college is public or private.

It is important to recognize that **correlation does not imply causation**, and these relationships merely indicate the degree to which these variables linearly co-vary with the college type designation. The presence of correlation does allow us to posit hypotheses about the nature of these institutions, but further statistical analysis through logistic regression would be necessary to draw more robust inferences about the factors that influence college type.

The strength of the correlations presented, particularly for 'Outstate' and 'perc.alumni', indicate potential avenues for further investigation and could be considered for inclusion in predictive models for college type classification based on continuous variable predictors.

Splitting the data into a train and test set

Train (70%)

Private- 396 records

Public – 149 records

Test (30%)

Private- 169 records

Public- 63 records

As described above dataset was partitioned into training and test subsets. This adheres to standard practice in machine learning to evaluate the generalizability of predictive models. The training subset, comprising 70% of the data, is used to fit or train the model, allowing the algorithm to identify patterns and relationships within the data. The test subset, constituting the remaining 30%, is employed to assess the model's performance on unseen data, thus providing an estimate of its predictive accuracy in real-world scenarios.

In the training subset, there are a total of 545 records, of which 396 are labeled as 'Private' and 149 as 'Public'. This distribution indicates a substantial imbalance, with approximately 72.7% of the records in the training set belonging to the 'Private' category and 27.3% to the 'Public' category. Such an imbalance may necessitate special consideration during model training to ensure that the algorithm does not become biased towards the majority class. A **stratified sampling** approach could have been considered but was omitted as it was beyond the purview of the tasks assigned in this Module 3 project.

The test subset contains 232 records, with 169 labeled as 'Private' and 63 as 'Public'. The proportion of 'Private' to 'Public' records in the test set is roughly consistent with that of the training set, with 'Private' constituting about 72.8% and 'Public' about 27.2% of the data. This consistency in distribution between training and test sets is crucial for maintaining the validity of the model evaluation process.

I used the glm() function from the ‘stats’ package in base R to fit a logistic regression model using the predictors derived from the bivariate and correlation analysis and after conceptual face validation. I attach the summaries of the three models I refined prior to arrive at the final model and cursorily describe them. Since these models were ‘nested models’ I used Likelihood Ratio Tests (LRT) and Akaike Information Criterion (AIC) to evaluate whether I might be more parsimonious with respect to the number of predictors. After three steps of refinement I arrived at the final model that I decided to use on the test data set. The below schema describes this feature selction.(Faizan, 2024)

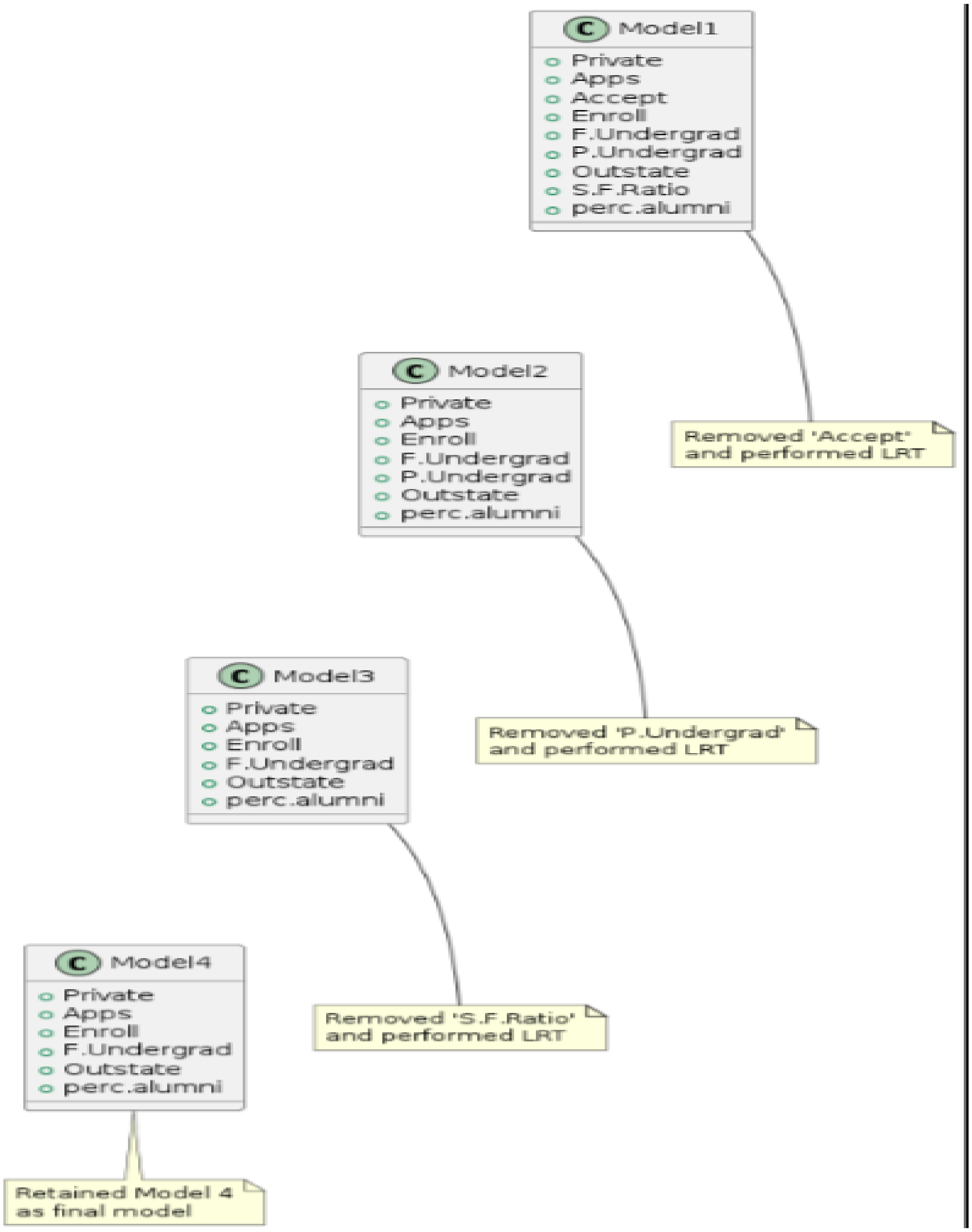
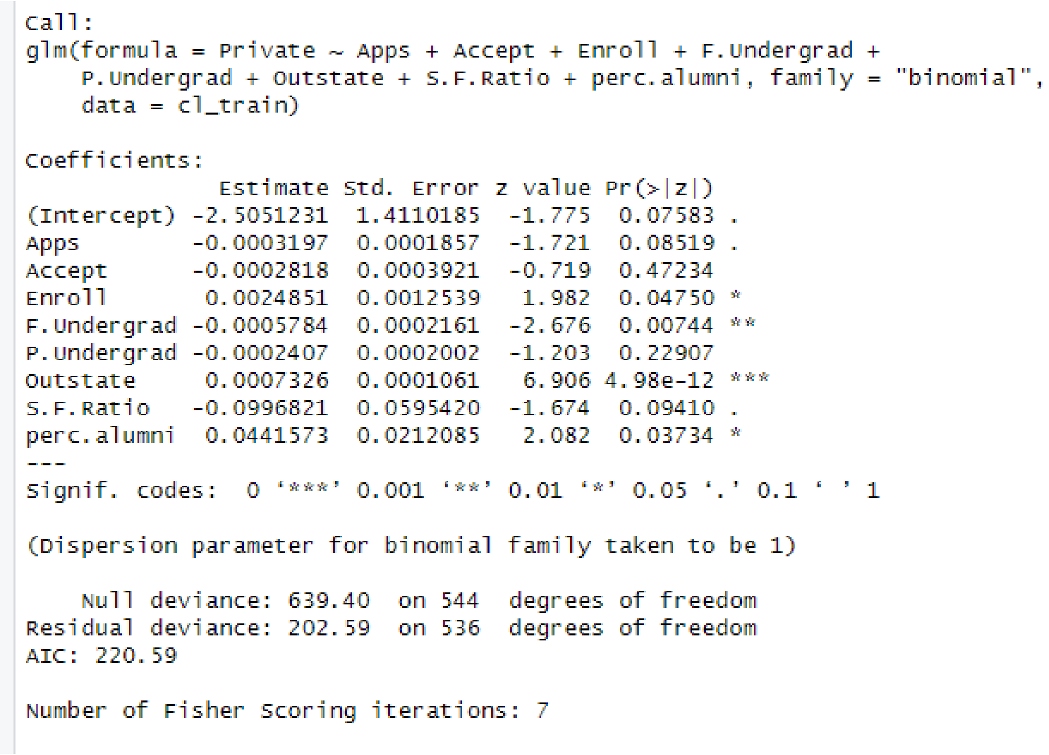


Figure 13 Feature selection to arrive at the final model

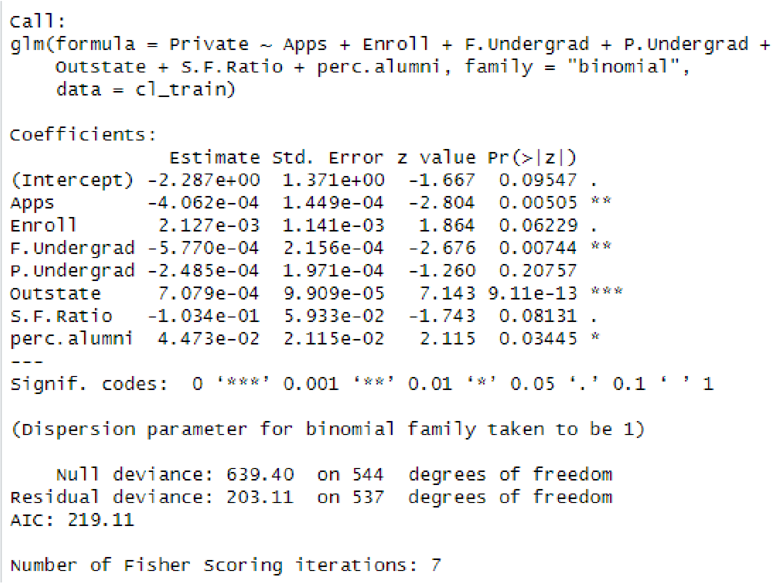
Feature selection

Please note that I only describe the final model in detail by interpreting the coefficients and the log-odds and odds-ratio and give the summary of the intermediate models (Model 1 , Model 2 , Model 3 before the final Model 4), which were only steps to arrive at the final model by way of refinement, briefly in order to avoid verbosity.



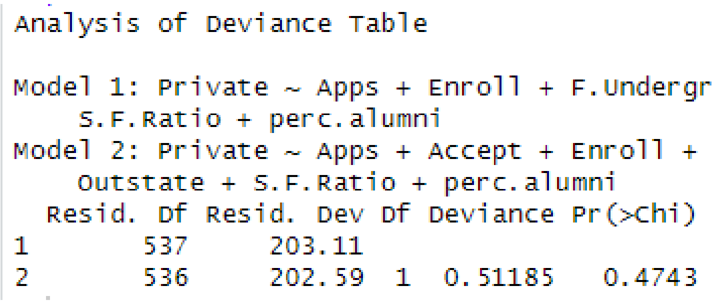
```
Call:
glm(formula = Private ~ Apps + Accept + Enroll + F.Undergrad + 
    P.Undergrad + Outstate + S.F.Ratio + perc.alumni, family = "binomial", 
    data = cl_train)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -2.5051231  1.4110185  -1.775  0.07583 .  
Apps        -0.0003197  0.0001857  -1.721  0.08519 .  
Accept      -0.0002818  0.0003921  -0.719  0.47234    
Enroll       0.0024851  0.0012539   1.982  0.04750 *  
F.Undergrad -0.0005784  0.0002161  -2.676  0.00744 ** 
P.Undergrad -0.0002407  0.0002002  -1.203  0.22907    
Outstate     0.0007326  0.0001061   6.906 4.98e-12 ***
S.F.Ratio   -0.0996821  0.0595420  -1.674  0.09410 .  
perc.alumni  0.0441573  0.0212085   2.082  0.03734 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 639.40  on 544  degrees of freedom
Residual deviance: 202.59  on 536  degrees of freedom
AIC: 220.59

Number of Fisher Scoring iterations: 7
```

Figure 14 Model 1 with 8 predictors. Further refinement was deemed necessary.



```
Call:
glm(formula = Private ~ Apps + Enroll + F.Undergrad + P.Undergrad + 
    Outstate + S.F.Ratio + perc.alumni, family = "binomial", 
    data = cl_train)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -2.287e+00  1.371e+00  -1.667  0.09547 .  
Apps        -4.062e-04  1.449e-04  -2.804  0.00505 ** 
Enroll       2.127e-03  1.141e-03   1.864  0.06229 .  
F.Undergrad -5.770e-04  2.156e-04  -2.676  0.00744 ** 
P.Undergrad -2.485e-04  1.971e-04  -1.260  0.20757    
Outstate     7.079e-04  9.909e-05   7.143 9.11e-13 ***
S.F.Ratio   -1.034e-01  5.933e-02  -1.743  0.08131 .  
perc.alumni  4.473e-02  2.115e-02   2.115  0.03445 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 639.40  on 544  degrees of freedom
Residual deviance: 203.11  on 537  degrees of freedom
AIC: 219.11

Number of Fisher Scoring iterations: 7
```

Figure 15 Model 1 and Model 2. In model 2 I removed 'Applications accepted' as a predictor.



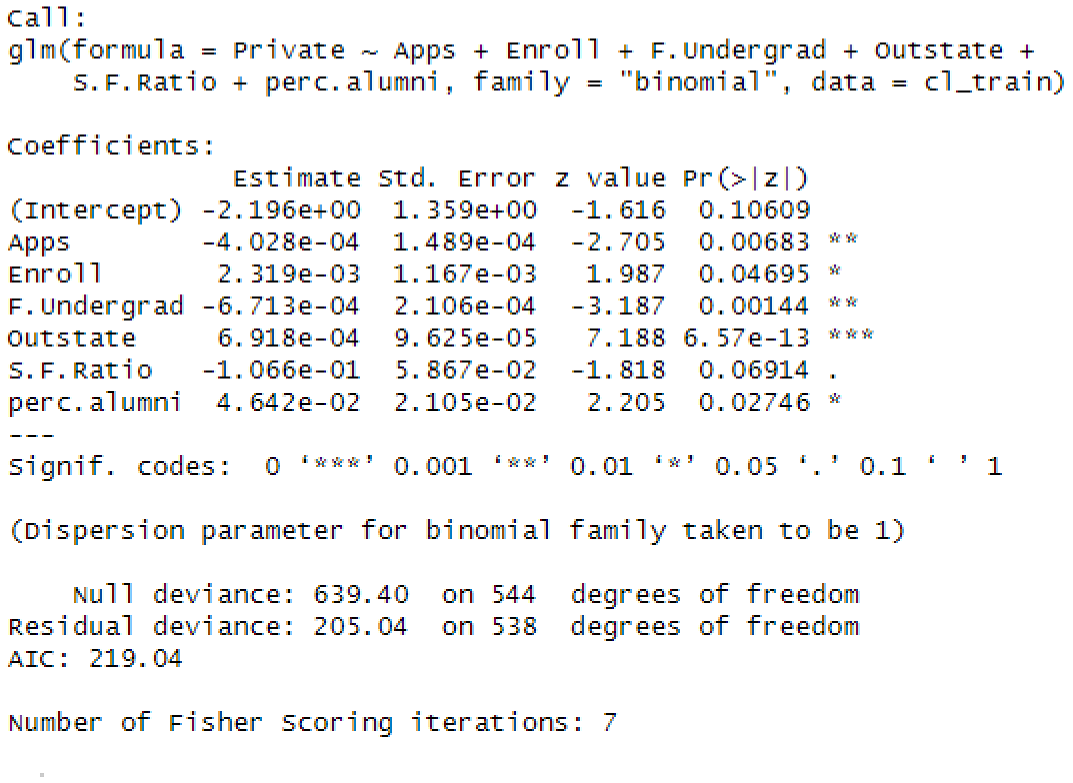
```
Analysis of Deviance Table

Model 1: Private ~ Apps + Enroll + F.Undergrad + P.Undergrad + Outstate +
    S.F.Ratio + perc.alumni
Model 2: Private ~ Apps + Accept + Enroll + F.Undergrad + P.Undergrad +
    Outstate + S.F.Ratio + perc.alumni
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       537     203.11
2       536     202.59  1  0.51185   0.4743
```

Figure 16 Likelihood Ratio Test between model 2 and 1.

The following sequence of feature selection using significance levels and the Likelihood Ratio Test (LRT) was followed:

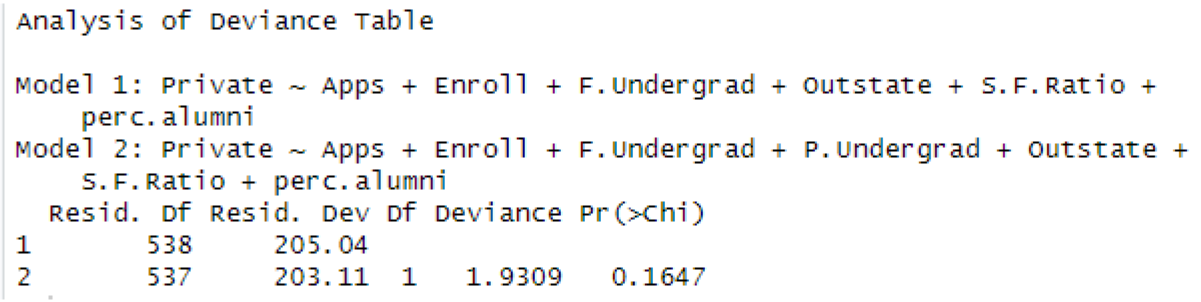
From Model 1 to Model 2:

- The variable 'Accept' was removed from Model 2 as it was deemed not statistically significant enough to be retained in the model because its p value in the model summary came out to be 0.47 at an alpha = 0.05 for significance.
- The LRT comparing Model 1 to Model 2 resulted in a chi-square value of 0.51185 with a p-value of 0.4743. This high p-value suggests that the removal of the 'Accept' variable does not significantly reduce the explanatory power of the model, justifying its exclusion.



```
Call:
glm(formula = Private ~ Apps + Enroll + F.Undergrad + Outstate +
    S.F.Ratio + perc.alumni, family = "binomial", data = cl_train)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.196e+00  1.359e+00  -1.616  0.10609
Apps        -4.028e-04  1.489e-04  -2.705  0.00683 **
Enroll       2.319e-03  1.167e-03   1.987  0.04695 *
F.Undergrad -6.713e-04  2.106e-04  -3.187  0.00144 **
Outstate     6.918e-04  9.625e-05   7.188 6.57e-13 ***
S.F.Ratio   -1.066e-01  5.867e-02  -1.818  0.06914 .
perc.alumni  4.642e-02  2.105e-02   2.205  0.02746 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 639.40  on 544  degrees of freedom
Residual deviance: 205.04  on 538  degrees of freedom
AIC: 219.04

Number of Fisher Scoring iterations: 7
```

Figure 17. Model 3 after removing the part-time undergraduates from the model 2.

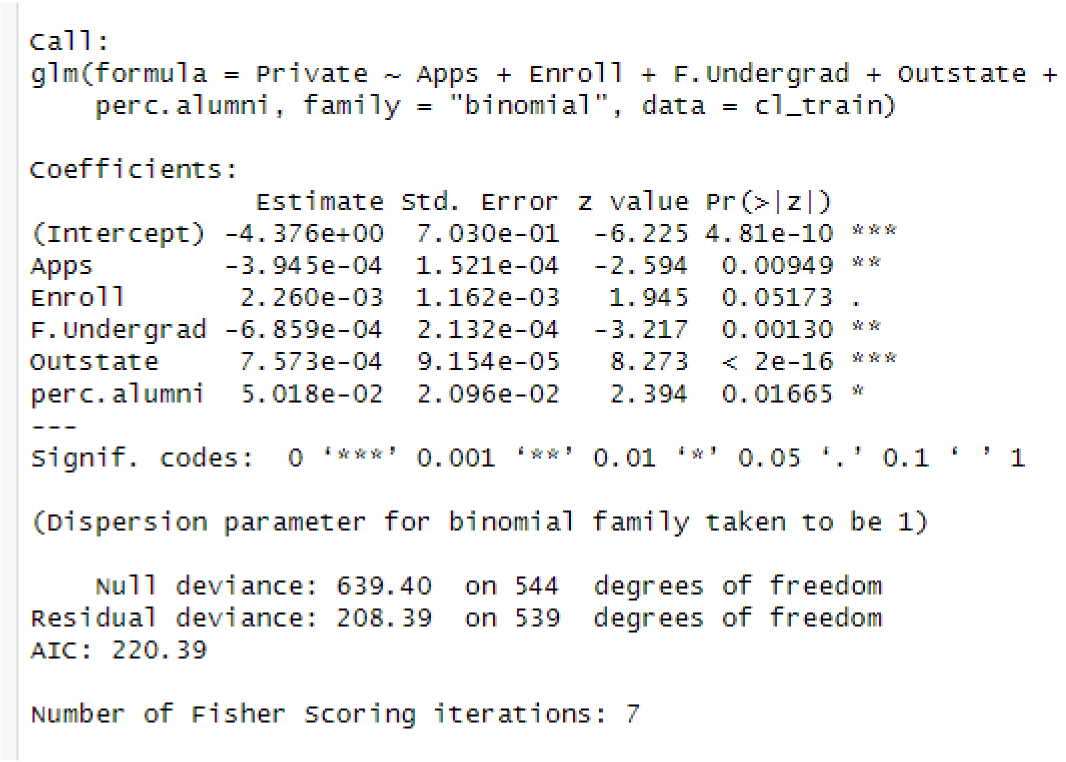


```
Analysis of Deviance Table

Model 1: Private ~ Apps + Enroll + F.Undergrad + Outstate + S.F.Ratio + 
    perc.alumni
Model 2: Private ~ Apps + Enroll + F.Undergrad + P.Undergrad + Outstate + 
    S.F.Ratio + perc.alumni
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       538     205.04                     
2       537     203.11  1   1.9309   0.1647
```

Figure 18 Likelihood ratio test indicating no significant change.

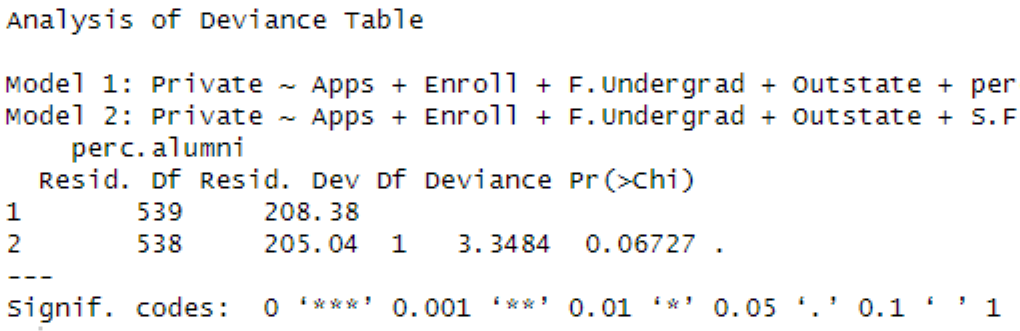
From Model 2 to Model 3:

- The variable 'P.Undergrad' was removed in Model 3 as it had a p value of 0.207 in model 2 ot indicating sufficient significance. Also note that the co-efficients in the model were not significantly altered after this removal except a decrease in the co-efficient of full-time undergraduates, suggesting some confounding was taking place.
- The LRT comparing Model 2 to Model 3 yielded a chi-square value of 1.9309 with a p-value of 0.1647. Again, the non-significant p-value supports the decision to remove 'P.Undergrad' from the model.



```
Call:
glm(formula = Private ~ Apps + Enroll + F.Undergrad + Outstate + 
    perc.alumni, family = "binomial", data = cl_train)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -4.376e+00  7.030e-01  -6.225 4.81e-10 ***
Apps        -3.945e-04  1.521e-04  -2.594  0.00949 ** 
Enroll       2.260e-03  1.162e-03   1.945  0.05173 .  
F.Undergrad -6.859e-04  2.132e-04  -3.217  0.00130 ** 
Outstate     7.573e-04  9.154e-05   8.273  < 2e-16 ***
perc.alumni  5.018e-02  2.096e-02   2.394  0.01665 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 639.40  on 544  degrees of freedom
Residual deviance: 208.39  on 539  degrees of freedom
AIC: 220.39

Number of Fisher Scoring iterations: 7
```

Figure 19 Final model after removing the Student Faculty Ratio.



```
Analysis of Deviance Table

Model 1: Private ~ Apps + Enroll + F.Undergrad + Outstate + perc.alumni
Model 2: Private ~ Apps + Enroll + F.Undergrad + Outstate + S.F.Ratio + 
    perc.alumni
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
1       539     208.38                       
2       538     205.04  1   3.3484  0.06727 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

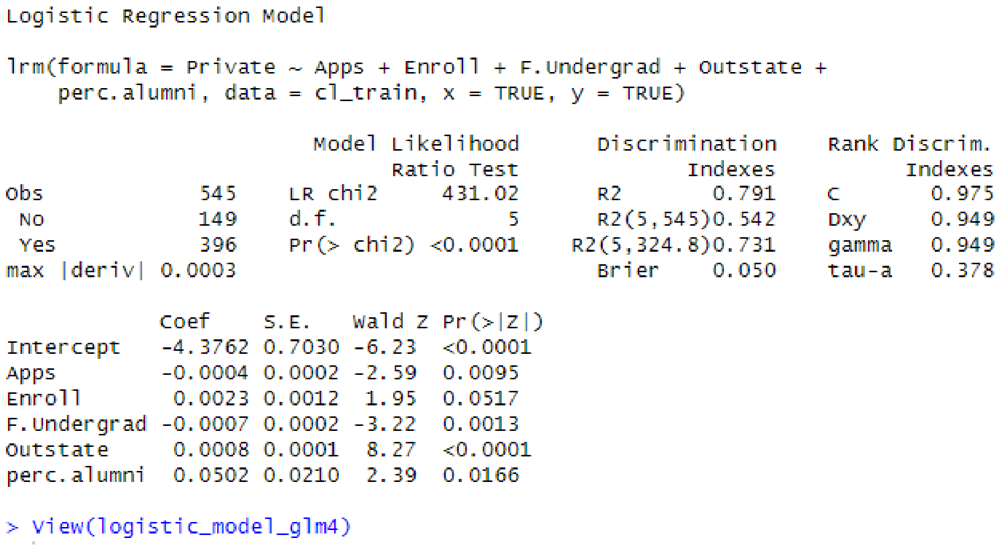
From Model 3 to Model 4, the final model:

- The 'S.F.Ratio' variable was removed to create Model 4, the final model, as it had a p value of 0.06 which , while hovering just above the alpha value for significance was still deemed removable.
- The LRT for this transition showed a chi-square value of 3.3484 with a p-value of 0.06727. While this p-value is lower than the previous tests, indicating a more substantial change in the model fit, it still does not meet the typical alpha level of 0.05 for statistical significance. Consequently, the removal of 'S.F.Ratio' can be considered as not significantly worsening the model's fit.

The feature selection process described above illustrates a stepwise regression approach where variables are removed one at a time based on their statistical significance and the overall impact on the model. By considering both the p-values of individual coefficients and the chi-square values from LRTs, through the process I sought to retain only those predictors that provide meaningful contributions to the model.

Detailed description and interpretation of the final model as fitted on the training set

In addition to the ‘stats’ package function glm() I also used the ‘rms’ package (Harrell Jr, 2024) to get a more detailed summary.



```
Logistic Regression Model

lrm(formula = Private ~ Apps + Enroll + F.Undergrad + Outstate + 
    perc.alumni, data = cl_train, x = TRUE, y = TRUE)

                      Model Likelihood      Discrimination    Rank Discrim.    
                            Ratio Test             Indexes          Indexes    
Obs            545    LR chi2     431.02      R2       0.791    C       0.975    
 No            149    d.f.             5    R2(5,545)0.542    Dxy     0.949    
 Yes           396    Pr(> chi2) <0.0001    R2(5,324.8)0.731    gamma   0.949    
max |deriv| 0.0003                            Brier    0.050    tau-a   0.378    

            Coef    S.E.   Wald Z Pr(>|Z|)
Intercept   -4.3762 0.7030 -6.23  <0.0001 
Apps        -0.0004 0.0002 -2.59  0.0095  
Enroll       0.0023 0.0012  1.95  0.0517  
F.Undergrad -0.0007 0.0002 -3.22  0.0013  
Outstate     0.0008 0.0001  8.27  <0.0001 
perc.alumni  0.0502 0.0210  2.39  0.0166  

> View(logistic_model_glm4)
```

Figure 20 Final model summary using the 'rms' package.

We have constructed a predictive logistic regression model to associate continuous variables with the binary ‘college type’ outcome, where the reference category is 'Public College' and the contrast is 'Private College'. The model's coefficients represent the log odds of a college being 'Private' versus 'Public', controlling for other variables in the model.

1. **Coefficient Interpretation:**
   - **Applications (Apps):** The coefficient for applications is -0.0004, indicating that for each unit increase in the number of applications, the log odds of a college being 'Private' as compared to ‘Public’ decrease by the amount of 0.0004, holding other variables constant.
   - **Enrollment (Enroll):** Enrollment has a coefficient of 0.0023, which suggests that higher enrollment numbers are associated with an increased log odds of the college being 'Private' as opposed to ‘Public’. Each enrollment increases log odds of college being private by an amount of 0.0023.
   - **Full-time Undergraduates (F.Undergrad):** The coefficient of -0.0007 for full-time undergraduates implies that for each full-time undergraduate , the log odds of a college being 'Private' decrease by that amount.
   - **Out-of-state Tuition (Outstate):** The positive coefficient of 0.0008 for out-of-state tuition indicates that colleges with higher out-of-state tuition costs are more likely to be 'Private'. Each US dollar increase in Out-of-State Tuition increases log-odds of ‘Private’ college as opposed to ‘Public’ college by 0.0008.
   - **Percentage of Alumni Who Donate (perc.alumni):** With a coefficient of 0.0502, a higher percentage of alumni donors is significantly associated with an increased log odds of a college being 'Private'.
2. **Intercept (y-intercept):**
   - The intercept of -4.3762 can be interpreted as the log odds of a college being 'Private' when all the predictors are held at zero. In the context of this model, this value represents the log odds of the baseline category ('Public') when all other variables are absent, which is typically not a realistic scenario in the natural setting of the data.
3. **Equation of the Model:**
   - The logistic regression model can be formulated as:

$$\log(P(\text{'Private'})/1-P(\text{'Private'}))=-4.3762-0.0004\times\text{Apps}+0.0023\times\text{Enroll}-0.0007\times\text{F.Undergrad}+0.0008\times\text{Outstate}+0.0502\times\text{perc.alumni}$$

4. **Odds Ratios:**
   - By exponentiating the coefficients, we derive the odds ratios (OR), which express the multiplicative change in odds for a one-unit increase in the predictor:
     - Apps: OR = exp(-0.0004) ≈ 0.9996
     - Enroll: OR = exp(0.0023) ≈ 1.0023

- F.Undergrad: OR = exp(-0.0007) ≈ 0.9993
- Outstate: OR = exp(0.0008) ≈ 1.0008
- perc.alumni: OR = exp(0.0502) ≈ 1.0515

5. **Measures of the efficacy of the model.**
Unlike in Linear Regression while Logistic Regression does not have a measure like R-Squared to reflect the efficacy of the model, several R-like measures have been proposed to shed light on the efficacy of the predictive modelling. Some of these are summarised in the output and mentioned below.

1. **Model Likelihood Ratio Test**:
   - The Likelihood Ratio (LR) chi-squared test statistic is 431.02 with 5 degrees of freedom (d.f.), which is highly significant (Pr(>chi2) < 0.0001). This test evaluates whether the null hypothesis that all coefficients of the model (except the intercept) are equal to zero can be rejected. The p-value indicates that the model with predictors fits the data significantly better than the model with only the intercept.
2. **Discrimination Indexes**:
   - The **R2** indicates the Nagelkerke's R-squared, which is a pseudo-R-squared measure for logistic regression. It has a value of 0.791 for the 'No' category and 0.731 for the 'Yes' category. These values suggest a strong relationship between the predictors and the response variable since they can be interpreted as the proportion of variance explained by the model.
   - The **C** statistic or concordance index (also known as the Area Under the Curve or AUC for ROC curve) is 0.975, which is close to 1, indicating excellent discriminatory ability of the model. This means the model is very good at distinguishing between the two response categories.
   - The **Dxy rank** discriminant index is 0.949, further indicating the model's strong discriminative power.
   - The **gamma statistic** is also 0.949, consistent with the other discrimination indices, showing a strong predictive association.
   - The **tau-a index** is 0.378, indicating a moderate to strong relationship between predicted probabilities and observed responses.
3. **Brier Score**: The Brier score is 0.050, which is a measure of the accuracy of probabilistic predictions. A Brier score ranges from 0 for a perfect model to 0.25 for a no-skill model. In this context, a score of 0.050 indicates a high level of accuracy.

Since I have already described the model above, I proceed to discuss the Confusion Matrix.

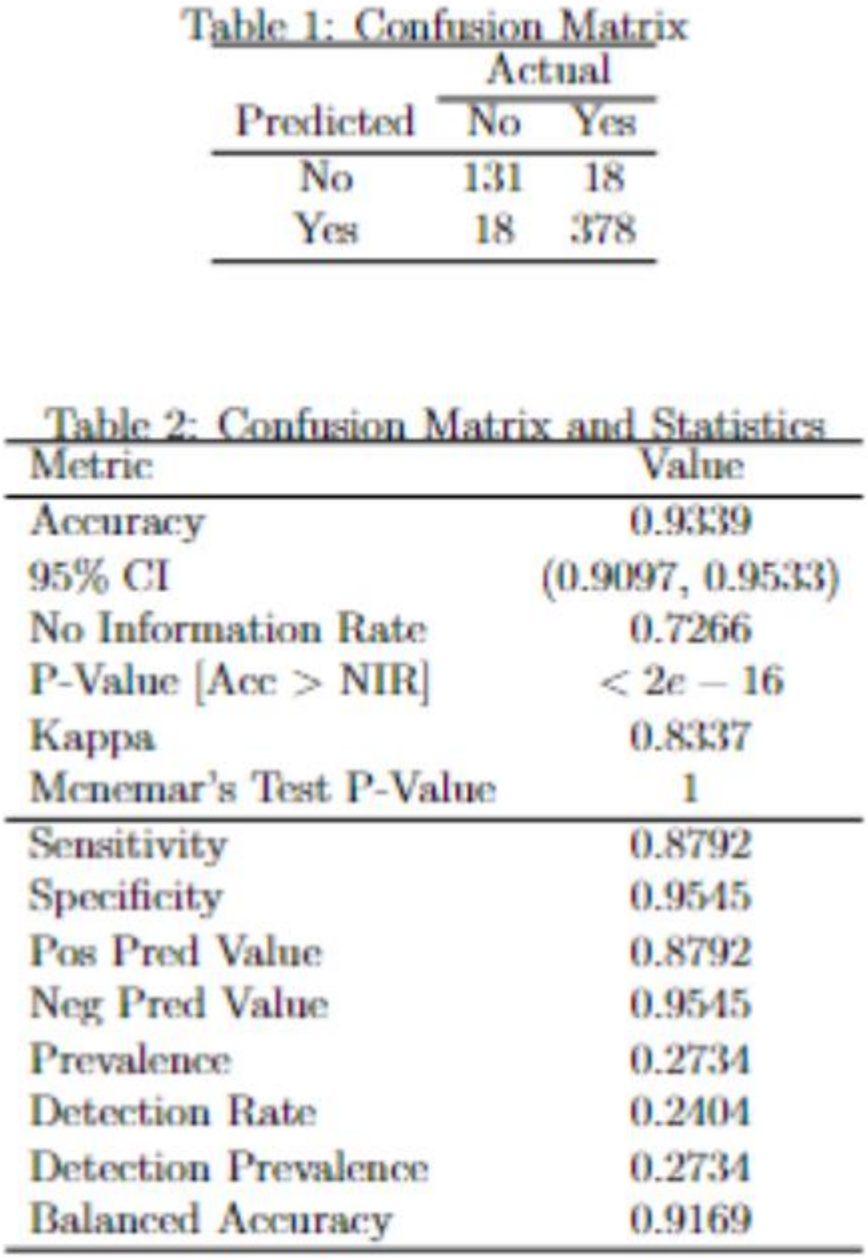


Table 1: Confusion Matrix

Predicted	Actual No	Actual Yes
No	131	18
Yes	18	378

Table 2: Confusion Matrix and Statistics

Metric	Value
Accuracy	0.9339
95% CI	(0.9097, 0.9533)
No Information Rate	0.7266
P-Value [Acc > NIR]	$< 2e - 16$
Kappa	0.8337
Mcnemar's Test P-Value	1
Sensitivity	0.8792
Specificity	0.9545
Pos Pred Value	0.8792
Neg Pred Value	0.9545
Prevalence	0.2734
Detection Rate	0.2404
Detection Prevalence	0.2734
Balanced Accuracy	0.9169

Figure 21 Confusion Matrix for the Model fit on the training data set.

Confusion Matrix

- The matrix shows four entries:
  - True Negatives (TN): The model correctly predicted 131 instances of the actual 'Private College' class.
  - False Negatives (FN): The model incorrectly predicted 18 instances of the actual 'Public College' class as 'Private College'.

- False Positives (FP): The model incorrectly predicted 18 instances of the actual 'Private College' class as 'Public College'.
- True Positives (TP): The model correctly predicted 378 instances of the actual 'Public College' class.

Table 2: Confusion Matrix and Statistics

- **Accuracy**: The accuracy of the model is 93.93%, indicating a high overall rate of correctly predicted instances out of all predictions made. The 95% Confidence Interval (CI) for accuracy ranges from 90.97% to 95.33%.
- **No Information Rate (NIR)**: This rate is the accuracy that could be achieved by always predicting the most frequent class. Here, it is 72.66% for 'Public College', which is significantly lower than the model's accuracy.
- **P-Value (Acc > NIR)**: The p-value suggests that the model's accuracy is significantly better than the no information rate, indicating meaningful predictive ability.
- **Kappa**: The Kappa statistic of 0.837 indicates almost perfect agreement between the actual and predicted classifications, corrected for chance agreement.
- **McNemar's Test P-Value**: The value of 1 suggests no significant difference between the numbers of false positives and false negatives.
- **Sensitivity (Recall)**: The sensitivity of the model is 87.92%, indicating it correctly identifies 87.92% of the actual 'Private College' instances.
- **Specificity**: The model's specificity is 95.15%, showing it correctly identifies 95.15% of the actual 'Public College' instances.
- **Positive Predictive Value (Precision)**: The positive predictive value is 87.92%, which means that when the model predicts 'Private College', it is correct about 87.92% of the time.
- **Negative Predictive Value**: The negative predictive value is 87.95%, indicating that when the model predicts 'Public College', it is correct about 87.95% of the time.
- **Balanced Accuracy**: The balanced accuracy is 91.69%, indicating that the model performs well across both the 'Public College' and 'Private College' classifications.

The determination of which type of misclassification is more damaging depends on the specific implications of each error within the study's aims and the potential impact of these errors on subsequent decisions or policies. The question of ‘why’ we are carrying out this study to differentiate and classify colleges into public and private will dictate whether false positives are more damaging or false negatives.

False Positives (Predicting 'Private College' when it's actually 'Public College'):

- **Resource Allocation**: If the model is used for allocating funds that are meant for public institutions, falsely identifying a public college as private could result in a misallocation of government resources.
- **Policy Implementation**: Policies designed specifically for public colleges, such as regulations or subsidies, could incorrectly be applied to private institutions.
- **Reporting and Accountability**: Public colleges might have different reporting requirements or standards of accountability. A false positive could lead to an incorrect assessment of compliance.

False Negatives (Predicting 'Public College' when it's actually 'Private College'):

- **Funding Opportunities**: Private colleges could miss out on funding opportunities, grants, or programs aimed at private institutions if they are incorrectly labeled as public.
- **Regulatory Oversight**: Private colleges might be subject to less stringent regulations in certain areas. Incorrectly classifying them could subject them to inappropriate regulatory scrutiny.
- **Market Positioning**: Misclassification could affect the market positioning of private colleges, as public perception and reputation could be influenced by their classification.

In scenarios where public colleges are under governmental scrutiny for quality control or resource allocation, false positives could be more damaging as they could lead to private institutions unfairly benefitting from public resources. Conversely, if private colleges rely on their classification to attract a certain student demographic or qualify for private funding, false negatives could be more harmful.

The study must evaluate the consequences of each type of error primarily in relation to its objectives.

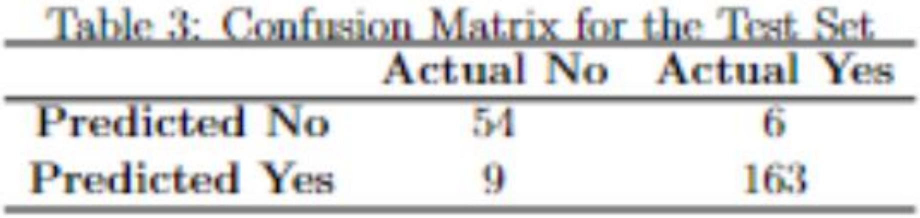


Table 3: Confusion Matrix for the Test Set

	Actual No	Actual Yes
Predicted No	54	6
Predicted Yes	9	163

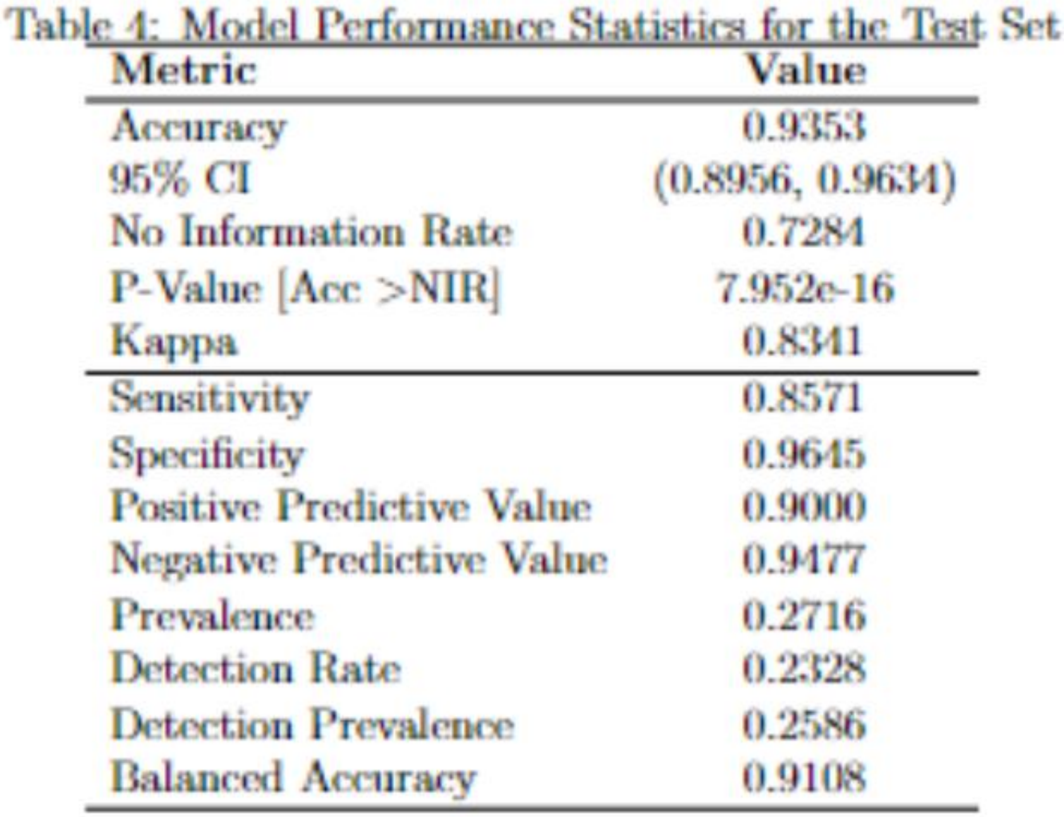


Table 4: Model Performance Statistics for the Test Set

Metric	Value
Accuracy	0.9353
95% CI	(0.8956, 0.9634)
No Information Rate	0.7284
P-Value [Acc >NIR]	7.952e-16
Kappa	0.8341
Sensitivity	0.8571
Specificity	0.9645
Positive Predictive Value	0.9000
Negative Predictive Value	0.9477
Prevalence	0.2716
Detection Rate	0.2328
Detection Prevalence	0.2586
Balanced Accuracy	0.9108

Figure 22 Confusion Matrix for the Test set.

Confusion Matrix for the Test Set

- **Predicted vs. Actual**: The confusion matrix outlines the number of correct and incorrect predictions made by the model when classifying colleges as either 'Public' or 'Private'.
- **True Negatives (TN)**: The model correctly identified 51 instances where colleges were 'Private College'.
- **False Negatives (FN)**: The model incorrectly predicted 6 instances of 'Public College' as 'Private College'.
- **False Positives (FP)**: The model incorrectly predicted 9 instances of 'Private College' as 'Public College'.

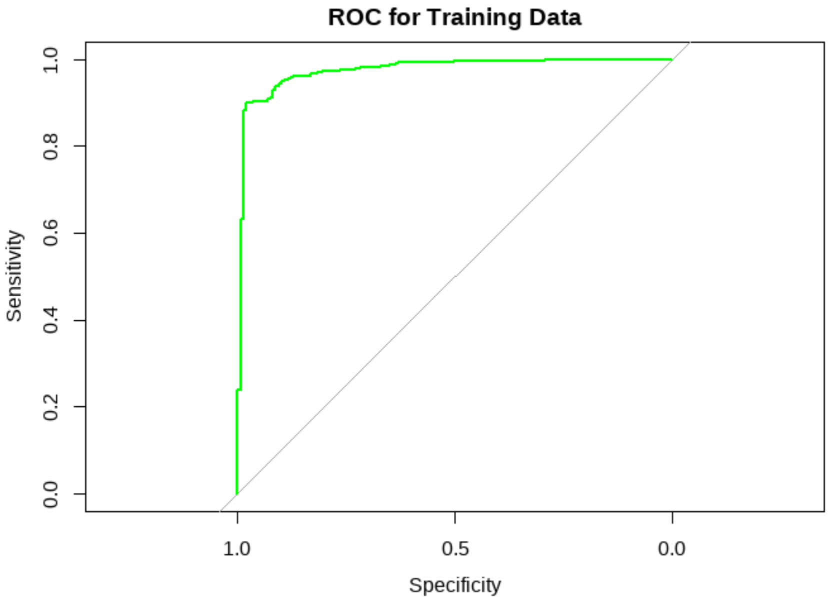
- **True Positives (TP)**: The model correctly identified 163 instances as 'Public College'.

Table 4: Model Performance Statistics for the Test Set

- **Accuracy**: The accuracy of the model is 93.53%, which suggests a high level of overall predictive success. The provided 95% confidence interval (95% CI) from 89.56% to 96.31% indicates the range within which the true accuracy of the model is expected to lie with 95% confidence.
- **No Information Rate (NIR)**: The NIR is 72.81%, representing the baseline accuracy that would be achieved by always predicting the most frequent class. The model's accuracy substantially exceeds this rate.
- **P-Value (Acc > NIR)**: The p-value is extremely small (7.952e-16), indicating that the model's accuracy is significantly better than the no information rate, and the result is unlikely to be due to random chance.
- **Kappa**: The Kappa statistic of 0.831 shows substantial agreement, suggesting that the model's predictions are not only better than chance but also demonstrate a high degree of reliability.
- **Sensitivity**: Also known as recall, the sensitivity of 85.71% indicates that the model is able to correctly identify 85.71% of the 'Public College' instances.
- **Specificity**: The specificity of 96.15% indicates that the model correctly identifies 96.15% of the 'Public College' instances.
- **Positive Predictive Value (PPV)**: Also known as precision, the PPV of 90.00% indicates that when the model predicts a college as 'Public College', it is correct 90% of the time.
- **Negative Predictive Value (NPV)**: The NPV of 91.77% suggests that when the model predicts a college as 'Public College', it is correct 91.77% of the time.
- **Balanced Accuracy**: The balanced accuracy of 91.08% indicates that the model has a high predictive performance that is balanced between sensitivity and specificity for both 'Private College' and 'Public College' predictions.

The ROC curve.

The below ROC curves are graphical plots that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. In an ideal Receiver Operating Characteristic curve the area under the curve approaches 1.



ROC for Training Data:

- The green curve represents the model performance on the training data.
- Similar to the testing data ROC curve, the training data curve shows a steep ascent towards the top-left, indicating high sensitivity, and then follows the border of the ROC space, denoting high specificity across varying thresholds.
- The curve for the training data sits even closer to the top-left corner of the plot compared to the testing data curve, suggesting an even higher sensitivity and specificity, which is typical given that models tend to perform better on the data they were trained on.

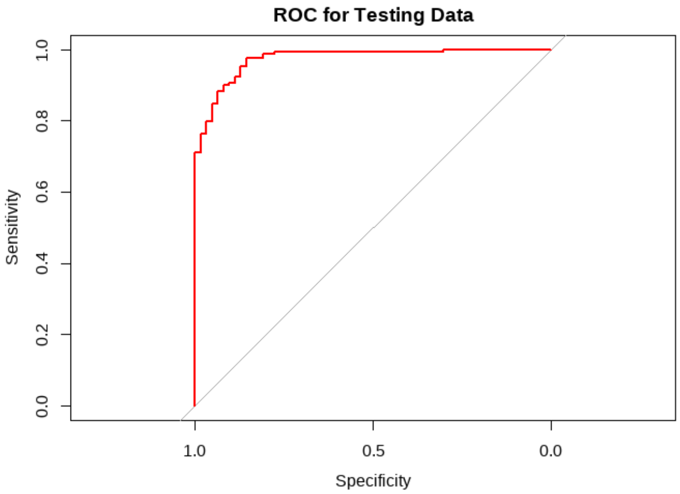
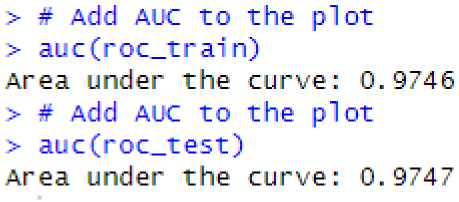


Figure 23 ROC for both Training and Testing Data

ROC for Testing Data:

- The red curve represents the relationship between sensitivity (True Positive Rate) and 1-specificity (False Positive Rate) for the testing data set.
- The curve rapidly rises towards the top-left corner, which indicates that the model has a high sensitivity, meaning it correctly identifies a high proportion of actual 'Private College' instances as such.
- The curve then follows closely along the left border and the top border of the ROC space, which indicates that the model maintains high sensitivity and high specificity across different thresholds; in other words, it correctly classifies both 'Private' and 'Public' colleges with high accuracy.

Calculate and interpret the AUC



```
> # Add AUC to the plot
> auc(roc_train)
Area under the curve: 0.9746
> # Add AUC to the plot
> auc(roc_test)
Area under the curve: 0.9747
```

Figure 24 The AUC as depicted in the console of the R Studio

AUC for Training Data:

- The AUC value is 0.9746, which indicates that the model has excellent predictive ability on the training dataset. An AUC value close to 1 implies that the model has a high probability of distinguishing between 'Private College' and 'Public College' correctly. In this case, the model's predictions on the training data are very accurate, and there is a 97.46%

chance that the model will be able to distinguish a randomly chosen 'Private College' from a 'Public College'.

AUC for Testing Data:

- The AUC value is 0.9747, virtually identical to the training data's AUC. This extraordinary consistency suggests that the model generalizes extremely well from the training data to unseen testing data. It further implies a 97.47% chance that the model will correctly distinguish between 'Private College' and 'Public College' in the testing dataset.

Interpretation with AUC and ROC:

- The similarity in AUC values between training and testing suggests that the model is robust, with no overfitting. Overfitting is often indicated by a high AUC for training data with a significantly lower AUC for testing data.
- The high AUC values complement the visual analysis of the ROC curves. For both training and testing datasets, the ROC curves rapidly rise towards the top-left corner and hug the left and top borders of the ROC space, indicating both high sensitivity and specificity.
- Such high AUC values in both training and testing data sets are indicative of an excellent predictive model. The model demonstrates high reliability in classifying colleges into 'Private College' and 'Public College', which is critical for decision-making processes relying on such predictions.

Conlusion/Interpretations

- In this Project I developed a logistic regression model using the College dataset from the ISLR package in R, enhanced by the 'rms' package.
- Model predicts the likelihood of a college being 'Private' vs. 'Public' with key predictors:
    - Applications: More applications slightly decrease the likelihood of being private.
    - Enrollment: Higher enrollment increases odds of being private.
    - Full-time Undergraduates: Larger numbers reduce the odds of being private.
    - Out-of-state Tuition: Higher tuition suggests a private institution.
    - Alumni Donation Percentage: Greater donation rates increase the odds of being private.
- Efficacy Indicators:
    - Likelihood Ratio Test significant, high pseudo-R2 values, AUC near 1 (0.9746 for training, 0.9747 for testing), low Brier score (0.050), indicating excellent model accuracy and predictive ability.
- Confusion Matrix (Training Data): High accuracy (93.93%), balanced false positives and negatives; impact varies by context (resource allocation, policy, funding opportunities).

- ROC Analysis: Sharp ascent and high AUC values demonstrate model's strong sensitivity and specificity, confirming its effectiveness in classifying colleges.
- Conclusion: Model shows high precision in predicting college type, valuable for educational policy and decision-making.

In summary, the logistic regression model evaluated here, as indicated by the AUC values and the ROC curves, is an excellent classifier for the given task using the College dataset from the ISLR package in R.

References

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) *An Introduction to Statistical Learning with applications in R*, https://www.statlearning.com, Springer-Verlag, New York

Harrell Jr, F. E. (2024). rms: Regression Modeling Strategies [R package version 6.8-0]. Retrieved from https://CRAN.R-project.org/package=rms