Predicting Texas Public Universities Retention Rate With Multi Variable Linear Regression and Neural Networks in R

James Pavlicek
Department of Information Systems & Analytics and Finance, McCoy College of Business, Texas State University

Problem Statement

Research Questions:
What factors contribute most to student retention rates at Public Universities in Texas, and how can interventions be tailored to improve retention?

Significance and Relevance:
This research is important because it helps us figure out what helps students stay and succeed in Texas public universities. Increased retention will help both the universities and the students. Universities will generate more income and can use those profits to further develop the school and their programs. Additionally, Students will be more likely to graduate with a high retention rate. With more income and more satisfied students the University will grow as a whole.

Data Sources

Data Source

Listed in References/Data Cite

Data Content

This full data set included all data imaginable for public universities. For 1 School over 1 year there is over 1,000 parameters or variables to look at. With such a large dataset I had to narrow down to factors that I believed may contribute to retention significantly either positively or negatively. With a larger scale project, I would consider pulling all the data to account for unseen factors.

Key Variables

Institution_Name
Year

Independent Variables

- Student_to_faculty_ratio
- Full_Time_Staff_per_Student
- Average_salary_of_full_time_professors
- Percent_of_undergraduate_enrollment_Age_18_to_24
- Percent_of_full_time_first_time_undergraduates_awarded_any_financial_aid
- Total_price_for_in_state_students_living_on_campus
- Percent_of_undergraduate_students_enrolled_exclusively_in_distance_education_courses
- Percent_admitted
- Published_in_state_tuition_and_fees
- Books_and_supplies
- Undergraduate_application_fee
- SAT_Math_50th_percentile_score
- SAT_Reading_and_Writing_50th_percentile_score

Dependent: Full_time_retention_rate

Methods

Multi Variable Linear Regression

In my study, I utilized linear regression analysis to explore the relationship between the variables and university retention rates, dividing the dataset into a 70% training set and a 30% testing set for model development and validation. This methodology provided a systematic approach to identify and quantify the impact of key variables on retention rates, setting the stage for in-depth analysis and prediction.

Artificial Neural Networks

Additionally, I applied neural networks to predict university retention rates, using a processed and normalized dataset divided into training (70%) and testing (30%) sets. This approach enabled me to explore complex patterns and relationships between factors affecting retention rates. By training and evaluating the neural network model, I gained insights into its predictive power and the impact of specific variables on student retention.

Results

Multi Variable Linear Regression

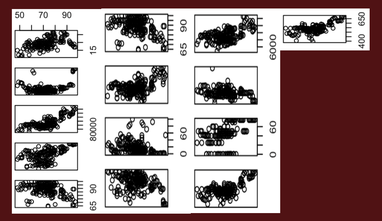
I found that the average salary of full-time professors was the most significant factor affecting full-time retention rates, highlighting a strong correlation between faculty compensation and student retention. Additionally, higher SAT Reading and Writing, as well as Math 50th percentile scores, were significantly associated with improved retention rates, suggesting the importance of academic preparedness. Furthermore, the percentage of undergraduates aged 18 to 24 also proved to be a significant factor, indicating that institutions with a larger proportion of traditional-aged students tend to achieve higher retention rates.

Artificial Neural Networks

In my project, the neural network model exhibited strong predictive performance with a Mean Absolute Error (MAE) of 0.3526 and an R-squared value of 0.7872, effectively capturing the variance in full-time retention rates. This accuracy, highlighted by the low Mean Squared Error (MSE) of 0.2095 and Root Mean Squared Error (RMSE) of 0.4578, underscores the model's robustness in utilizing selected factors to predict outcomes.

Findings

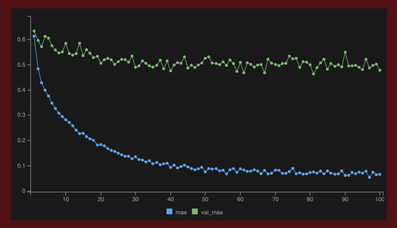
The findings from my research reveal critical insights into improving student success. Through linear regression analysis, I identified the average salary of full-time professors, SAT Reading and Writing, Math 50th percentile scores, and the percentage of undergraduates aged 18 to 24 as significant predictors of full-time retention rates, underscoring the importance of faculty compensation and academic preparedness. The application of an artificial neural network further validated these findings, demonstrating strong predictive performance with an R-squared value of 0.7872, thus highlighting the model's capability to accurately predict retention rates based on selected variables. This comprehensive analysis suggests that strategic interventions focusing on enhancing faculty salaries, student academic readiness, and catering to the needs of traditional-aged undergraduates could significantly bolster retention at Texas public universities.



Full_time_retention_rate on the X Axis

All Independent Variables (Data Sources pane) on the Y Axis

(Multi Variable Linear Regression)



Progress of the training model

(Artificial Neural Network)

Correlation Between Full_time_retention_rate (Dependent) and All Independent Variables.

(Multi Variable Linear Regression)

	Full_time_retention_rate
Student_to_faculty_ratio	0.4749923
Full_Time_Staff_per_Student	0.2780218
Average_salary_of_full_time_professors	0.7628461
Percent_of_undergraduate_enrollment_Age_18_to_24	0.5285648
Percent_of_full_time_first_time_undergraduates_awarded_any_financial_aid	-0.4285287
Total_price_for_in_state_students_living_on_campus	0.4921869
Percent_of_undergraduate_students_enrolled_exclusively_in_distance_education_courses	-0.2307981
Percent_admitted	-0.3437749
Published_in_state_tuition_and_fees	0.5241315
Books_and_supplies	-0.2783890
Undergraduate_application_fee	0.4502599
SAT_Math_50th_percentile_score	0.6709899
SAT_Reading_and_Writing_50th_percentile_score	0.6744942
Full_time_retention_rate	1.0000000

Predicted Values vs Actual Values

(Artificial Neural Network)

```
      [,1]
[1,] 68.03050
[2,] 69.64173
[3,] 76.31854
[4,] 67.75601
[5,] 70.64388
[6,] 70.90386
> head(actual_values)
[1] 73 70 80 73 69 76
```

```
Residuals:
      Min       1Q   Median       3Q      Max
-17.4985  -3.3490   0.2206   4.1845  13.2742

Coefficients:
                                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                                     4.373e+01  2.088e+01   2.095   0.0383 *
Student_to_faculty_ratio                                       -5.891e-02  3.043e-01  -0.194   0.8468
Full_Time_Staff_per_Student                                    -2.181e+01  2.302e+01  -0.947   0.3454
Average_salary_of_full_time_professors                          2.324e-04  5.333e-05   4.358 2.83e-05 ***
Percent_of_undergraduate_enrollment_Age_18_to_24                1.421e-01  6.651e-02   2.137   0.0347 *
Percent_of_full_time_first_time_undergraduates_awarded_any_financial_aid -9.005e-02  9.170e-02  -0.982   0.3281
Total_price_for_in_state_students_living_on_campus             -1.178e-04  4.792e-04  -0.246   0.8063
Percent_of_undergraduate_students_enrolled_exclusively_in_distance_education_courses -3.483e-02  2.994e-02  -1.163   0.2470
Percent_admitted                                               -9.246e-03  4.302e-02  -0.215   0.8302
Published_in_state_tuition_and_fees                            -5.551e-05  7.557e-04  -0.073   0.9416
Books_and_supplies                                             -4.763e-03  2.486e-03  -1.916   0.0578
Undergraduate_application_fee                                   1.386e-02  3.126e-02   0.443   0.6584
SAT_Math_50th_percentile_score                                 -4.091e-02  4.500e-02  -0.909   0.3651
SAT_Reading_and_Writing_50th_percentile_score                   6.624e-02  4.837e-02   1.369   0.1735
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.255 on 117 degrees of freedom
Multiple R-squared:  0.6504,    Adjusted R-squared:  0.6115
F-statistic: 16.74 on 13 and 117 DF,  p-value: < 2.2e-16
```

Linear Regression Output

```
> print(paste("Neural Network MAE:", nn_mae))
[1] "Neural Network MAE: 0.352609217166901"
> print(paste("Neural Network MSE:", nn_mse))
[1] "Neural Network MSE: 0.209542825818062"
> print(paste("Neural Network RMSE:", nn_rmse))
[1] "Neural Network RMSE: 0.457758479788263"
> print(paste("Neural Network R-squared:", r_squared))
[1] "Neural Network R-squared: 0.787230397150182"
```

Testing Model's Statistical Outputs (Artificial Neural Network)

Implications

This study's results underscore the importance of faculty compensation, academic preparedness, and demographic considerations in student retention at Texas public universities, providing actionable insights for targeted interventions. Enhancing professor salaries, bolstering student readiness, and focusing on the needs of traditional-aged students can lead to significant improvements in retention rates. These findings offer a direct pathway to addressing the research questions posed, with significant implications for policy and practice aimed at boosting student success and institutional development.

References/Data Cite

International Postsecondary Education Data System
Overall Website: https://nces.ed.gov/ipeds/
Data Source: https://nces.ed.gov/ipeds/datacenter/InstitutionByName.aspx?goToReportId=5&sid=6d91ee2e-4f45-4b44-bca7-7fb33c561d87&rtid=5

TEXAS STATE UNIVERSITY

MEMBER THE TEXAS STATE UNIVERSITY SYSTEM