**Machine Learning for Higher Education:**

**Analyzing and Forecasting Student Enrollment Statistics**

Thomas Simmons  
 Department of Computer Science and Information Technology   
Austin Peay State University  
 Clarksville, TN 37044, USA  
 tsimmons24@my.apsu.edu, Thomas.Simmons.cs@gmail.com

ABSTRACT

In recent years, higher education institutions have faced mounting challenges in maintaining and forecasting student enrollment and degree completion rates. Traditional statistical methods often fall short in capturing the complex, nonlinear relationships between institutional metrics and student outcomes. This paper explores the application of machine learning, specifically multiple linear regression and ridge regression, to analyze and forecast enrollment and degrees awarded across higher education institutions in Tennessee, with a focused case study on Austin Peay State University. Leveraging institutional data such as student credit hours, faculty workload, and retention rates, this research demonstrates the utility of data-driven modeling in identifying key predictors and improving forecasting accuracy. Models were developed and evaluated using techniques such as standardization, cross-validation, and ridge regression. Findings suggest that enrollment and faculty-related metrics may be strong indicators of degrees awarded. This work contributes to the growing field of educational data mining and highlights the potential of machine learning to support data-informed decision-making in academic administration.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author.

CCS CONCEPTS

• Probability and Statistics • Machine Learning   • Modeling and Simulation • Education

KEYWORDS

Educational Data Mining, Machine Learning Applications, Student Enrollment Forecasting, Ridge Regression, Linear Regression Models, Institutional Research, Higher Education Analytics, Predictive Modeling, Cross-Validation

ACM Reference format:

Thomas Simmons. 2025. Machine Learning for Higher Education Analyzing and Forecasting Student Enrollment Statistics.

1 Introduction

Higher education institutions rely heavily on enrollment and degrees awarded to maintain and increase funding. Public universities receive funding through federal government grants and locally levied taxes, and through tuition collected from enrolled students. There are many socioeconomic, demographic, and situational factors that fall outside a higher education institution’s control and many of these have already been studied, however more research can be done to study how the institution operates internally. This paper seeks to identify key performance indicators such as student to faculty ratios, credit hours to faculty ratios, degrees awarded, and other student statistics.

Accurate predictions for enrollment and the awarding of degrees are of high interest to higher education management as they grapple to maintain high enrollment, especially in a post-COVID environment. The primary challenge in college enrollment and degree analysis is producing accurate predictions while considering multiple factors. Traditional analysis methods use historical trends and statistical modeling, therefore failing to capture complex, nonlinear patterns between enrollment or degrees awarded and higher education parameters. Without accurate predictions, institutions may face challenges such as underutilized resources, budgetary inefficiencies, and ineffective strategies.

Advancements in Machine Learning and Data Mining in the last decade, and particularly in the last few years, provide strong analysis tools to detect these complex, nonlinear patterns that may have been missed before using traditional techniques. A newer data-driven approach enables higher education institutions to anticipate enrollment fluctuations, implement targeted interventions, and enhance decision-making processes.

2  Literature Review

There are not many papers currently that provide analysis on higher education institutional policy and decision-making using machine learning. Current papers on the subject that do use machine learning focus heavily on external factors outside a university’s control such as a student’s socioeconomic status, race, and previous test scores. [1-3] This project seeks to contribute to the extensive Educational Data Mining community by filling gaps in where machine learning has not been utilized.

As stated above, at the time of writing there are not many academic papers and journals that showcase the use of machine learning principles in the scope of internal enrollment statistics. However, it is important to note the critical findings of papers that focus on many important factors outside the scope of internal enrollment and thus, this paper and research. Such factors include race, gender, and socioeconomic status.

3  Statistics Overview

This research paper covers several different kinds of measurements collected by higher education institutions. These measurements range from student enrollment to full time faculty and more. Before jumping into the results of this research it is important to understand what each statistic represents.

3.1 Student Statistics

*3.1.1 Enrollment.* In higher education, enrollment refers to the total number of students attending that institution during a specific year or term. In the case of this research, enrollment is kept to yearly estimates instead of by semester. This was done because not all universities reported enrollment by semester and only reported by year instead. Universities reporting by year were reporting what their enrollment was in the fall semester.

*3.1.2 Student Credit Hours (SCH).* When looking at the total courseload a student is taking, it is common to say how many credit hours they are taking, instead of how many classes they have. Courses can range in difficulty and how much work they require. Certain courses could be one credit hour, meaning it is a relatively small workload, while some may be four or more credit hours, meaning it is a pretty heavy workload. At most Tennessee institutions, as it is with Austin Peay State University, it is common for an average class to be three credit hours. Students generally take around twelve to fifteen credit hours in a given semester. Universities report the total number of student credit hours across all students for a given semester. Universities reporting by year were reporting what their total student credit hours were in the fall semester. A similar statistic to SCH would be Full Time Equivalent Students, which divides total SCH to equate to a number of students.

*3.1.3 Full Time Equivalent (FTE) Students.* Many students in college only take one or two courses, which doesn’t classify them as a full-time student. This can lead to misleading interpretations on how much teaching an institution is doing. By totaling by full time equivalents, a better understanding of student throughput can be gained. For every twelve student credit hours, there is one FTE student.

*3.1.4 Full Time Equivalent (FTE) Faculty.* Not all faculty at an organization are full-time, and this is the case at higher education institutions as well. Faculty are split into three categories to calculate the total FTE: full-time faculty, half-time faculty, and one-third faculty. All full-time faculty are counted as one FTE, half-time faculty are counted as half of an FTE, and one-third faculty are counted as one-third of an FTE. Doing this allows a fair comparison of departments regardless of how many full-time or part-time employees they have.

*3.1.5 Degrees Awarded.* Universities award academic credentials known as degrees upon completion of all the requirements of that degree. How many degrees a higher education institution awards shows how many students are completing their degree path and graduating from that institution. The total is a good indicator of the institution’s throughput and is considered a key indicator of success.

*3.1.6 Student Credit Hours (SCH) / Full Time Equivalent (FTE) Faculty.* SCH / FTE faculty has become an important statistic regarding faculty workload for higher education institutions. The ratio represents the amount of credit hours each faculty member is assigned on average. A high SCH / FTE value means that professors are handling a large number of student credits, while a low value points towards the opposite. A high or low value is not necessarily an indication of inefficiency or failure, although it could point in certain directions. A high value could suggest a professor teaching too many difficult classes or could depict high teaching efficiency.

*3.1.7 Enrollment / Full Time Equivalent (FTE) Faculty.* Enrollment / FTE faculty is very similar to SCH / FTE faculty, with the main difference being the whole students instead of student credits. The SCH to FTE faculty ratio can explain workloads and teaching harder classes, while the enrollment to FTE faculty ratio can account for class sizes a professor is in charge of. Although they are similar statistics, they can each provide unique insights on faculty workloads.

*3.1.8 First-Time Full-Time Freshmen (FTFTF).* This is the count and share of first-time, full-time freshmen from the Retention Rate Cohort who remained enrolled in public postsecondary education for a second fall term. Cohorts of FTFTF are derived from fall end-of-term enrollment data and include first-time freshmen who began their enrollment in the previous summer.

4  System Design

This project uses Google Colab as the Integrated Development Environment (IDE). Google Colab is a free online python IDE that allows code to be run in the web browser. It’s especially useful for data science, machine learning, and artificial intelligence projects because it comes with many popular libraries already installed. It also gives you access to powerful hardware like Graphics Processing Units and Tensor Processing Units, which can help run programs faster.

As mentioned above, Google Colab’s programming language is python so that powerful machine learning and statistical libraries such as Scikit-learn, Keras, TensorFlow, NumPy, and Pandas could be leveraged. The main library used in this project is Scikit-learn. Some other libraries included matplotlib and seaborn which were used for creating visual representations of the results.

Scikit-learn, or as it is popularly known skearn, is an open-source library that is used for machine learning tasks in python. This library comes with many prebuilt functions such as those for machine learning supervised and unsupervised learning, hyperparameter tuning, normalization, and standardization. Having access to these functions ready for use made this research possible, as creating these from scratch would have taken far too long.

Data collected was cleaned and formatted properly in excel. Some exploratory data analysis was done as well to make initial inferences. This was done by using simple tools in excel such as filters.

To replicate the models presented in this paper. Download all the files found on the GitHub link found in the resources. Then upload the .ipynb python files to a Google Drive and open them with Google Colab, which you may need to add as an extension. Once the python file you want to run is open in Colab, you can add the required dataset to the Files folder in Colab, and then click “Runtime” 🡪 “Run all” to run the code.

5  Procedures of Data Collection & Analysis

The scope of data collected was initially going to be all universities within the state of Tennessee. However, due to the lack of a central public data repository of important statistics like SCH or FTE Faculty for Tennessee universities, especially private, this proved to be quite difficult. Some universities had their own public facing websites where important data such as student enrollment and other statistics, while other universities did not. Another major roadblock to data collection proved to be the lack of standardization to how this data was stored. A lot of the data was stored across years and was not combined. This meant that much time was spent on preprocessing and cleaning the collected data into a consolidated dataset to be read by the linear regression models.

5.1  Datasets

*5.1.1 APSU\_Final\_Dataset.* The first, and largest by total records, dataset used in this project is the “APSU\_Final\_Dataset.csv” which was created and cleaned by the author. This dataset encompasses student enrollment, degrees awarded, and other important statistics from Austin Peay State University (APSU) for each academic department in the institution. This dataset consists of nine columns and 168 unique rows. The columns from left to right are named Index, Department, Year, SCH/FTE, Enrollment, SCH, FTE Faculty, Degrees Awarded, and Enrollment /FTE Faculty. This data was collected from 2018 to 2023, which represents 5 years of student enrollment.

The data initially was not in an easily readable format for machine learning, and extensive cleansing and preparation had to be done. Degrees had to be totaled for each department, as each department has multiple kinds of degrees. Other statistics could not be drilled down by kind of degree, so this was the reasoning for totaling them by department instead. Another issue at first was that the data was split by year for each statistic into separate tables. The data had to be collected into one table where each statistic were columns alongside a new “year” column that contained the year the data point was collected. By doing so, the data was transformed into a long format, where each row is uniquely identifiable by a composite key of the year the data was collected and the department it was collected from. A record that contained the totals was removed from the dataset as this would have thrown off the model’s predictions.

*5.1.1 TN\_Enrollment\_Awards\_Dataset.* The second, and largest by number of students enrolled and columns, dataset is the “TN\_Enrollment\_Awards\_Dataset.csv” which was gathered from spreadsheets provided by the Tennessee Higher Education Commission (THEC). The commission was created to achieve coordination between higher education institutions in Tennessee and provides some data metrics for public universities in the state. Some of these were gathered, cleaned, and combined by the author into the aforementioned dataset. The dataset includes all Tennessee Board of Regents institutions, locally governed institutions, and most[[1]](#footnote-2) University of Tennessee institutions, totaling twenty-two institutions. There are thirteen columns named, left to right, Year, Institution, Total Undergraduate Students, Total Graduate Students Enrollment, Total Undergraduate FTE, Total Graduate FTE, Total FTE Students, First-Time Full-Time Freshmen (FTFTF), Retention Rate, Total Undergraduate Awards, Total Graduate Awards, and Degrees Awarded. Similar to the previous dataset, each row is uniquely identifiable by a composite key of the year the data was collected and the institution it was collected from.

*5*.2  Linear Regression

Linear regression is a type of machine learning algorithm, more specifically a supervised algorithm, that learns from observed data and maps the data into optimal linear functions that can be used for prediction on new data. The model estimates the relationship between a dependent scalar response variable and one or multiple explanatory regressor variables. In this research, multiple regressor variables were used, therefore classifying this as multiple linear regression. Linear regression is a heavily researched topic that has been proven to be effective in computing easily interpretable results. [4] The formulation can be explained as follows

where *y* is the dependent variable, β0, β1, β2, ⋯, βp are regression coefficients, and x1, x2, ⋯, xn are independent variables in the model. The relationship is modeled by a random variable or error variable ε, which is an unobserved random variable that adds noise to the linear relationship between the regressors β and the dependent variable *y*. [5] It is common for the formulation to be expressed in matrix notation as

Using Linear regression requires the target and regressor variables to be beholden to certain assumptions for the model’s results to be considered valid and reliable. Probably the most important assumption is linearity, which assumes that the relationship between the target and regressors is linear. If this assumption is not met, the model is intrinsically mis specified, and predictions can be wrongly interpreted. The errors that result from predictions are also important in verifying whether or not the model should be trusted in its predictive capabilities. Assumptions on the homoscedasticity and independence between the errors help strengthen confidence in predictions and can be verified by looking at plots of the errors. Homoscedastic errors are defined by the fact that their variance is not dependent on the values of the predictor variables. Independent errors are not correlated with each other and are important to have if the data is clustered. The last big assumption to be discussed is multicollinearity, which refers to the correlation between predictor variables, should not be perfect or even close to perfect. multicollinearity can lead to misleading predictor weightings and inflated standard errors.

A breach of the linear regression assumptions does not immediately destroy the chances of a model being useful in making predictions. In fact, there are many different methods of loosening the assumptions. This research ended up leveraging several of these techniques when the possibility of breaking assumptions arose.

*5.2.1 Ridge Regression.* Early on in the research, it became likely that the no multicollinearity assumption would be broken. This is due to the predictor variables being collected from similar sources. To maintain the assumption, ridge regression was leveraged to shrink the coefficients and reduce any multicollinearity present between the predictor variables. Ridge regression works by introducing a penalty hyperparameter to prevent the overfitting problem caused by correlated features. This type of regression was chosen in lieu of other regression methods to ensure none of the features were eliminated, or reduced to zero, which can happen in other methods like lasso regression. Ridge regression can be formulated as [6]

where lambda is the ridge regression hyperparameter and *y* is the regressor. Instead of guessing and experimenting with different values, this research finetuned the hyperparameter using cross-validation.

*5.2.2 Cross-validation*. Cross-validation is a technique that can be leveraged to prevent overfitting and improve predictions by tuning the hyperparameter of a model. In the case of this research, it was used to tune the alpha hyperparameter of the ridge regression model. Cross-validation can be formulated in the case of ridge regression as

where *a* and are parameter values and MSE is short for Mean Square Error.

*5.2.3 StandardScaler*. One common challenge with machine learning models is that the data values can vary wildly from feature to feature. To combat this problem, as is used in this research, Scikit-learn’s preprocessing tool StandardScaler standardizes the dataset. Standardization is a process of transforming a dataset so that the mean is 0 and the standard deviation is 1. Doing this improves the overall performance of the linear regression model. Standardization done by StandardScaler can be formulated as

*5*.3  Metrics

*5.3.2 R-Squared.* The coefficient of determination, or r-squared is defined as the proportion of variation in the target variable that is predictable from the predictor variables. The formulation for R-squared is [7]

where SS stands for sum of squares, and the result will be between 0 and 1, with 1 representing perfect predictions. A 1 may indicate issues with the model though.

*5.3.2 Mean Squared Error (MSE)*. The Mean Square Error is the computed measure of the average square of errors. In linear regression, MSE can work as a cost function. MSE is useful in determining the quality of a model’s predictive capabilities. MSE is formulated as [8]

6 Results

Results are split into two headings, Austin Peay State University and Public Tennessee Institutions for each dataset and complementary python file. AustinPeayLinearRegression.ipynb was used in calculating predictions for only Austin Peay State University. TennesseeSchoolsLinearRegression.ipynb was used in calculating predictions for Tennessee higher education institutions such as Middle Tennessee State University, University of Tennessee, Knoxville, and Pellissippi State Community College.

6.1 Austin Peay State University

*6.1.1 Predicting APSU Degrees Awarded.* The first target variable selected to train a linear regression model on was Degrees Awarded and was predicted using the independent variables SCH/FTE, Enrollment, Enrollment/FTE Faculty, and Year. Before training, the predictor variables were scaled using StandardScaler. The data was split into 75% training data and 25% test data. The linear model produced an R-squared value of 0.8067, which means 80.67% of the variance in degrees awarded can be explained by the linear regression model. The MSE produced was 615, which is very high and indicated model inaccuracy. The highest predictor coefficient was enrollment with 48.11 and the lowest slope was FTE Faculty with -8.18. This was quite obviously not realistic, as every student enrolled wouldn’t cause degrees awarded to increase by almost 50. By checking the Variance Inflation Factors (VIF) of the predictors, some came out to be very high, which meant they may have collinearity.

To combat the high VIF values between features, Ridge Regression was used to reduce overfitting and handle multicollinearity. A random test hyperparameter of 1 was chosen. The ridge model produced an R-squared value of 0.6964, a decrease, and an MSE of 618, a slight increase from the previous model. The coefficients did not change much with Enrollment and FTE Faculty keeping their highest and lowest ranks respectively. By looking at some plots of the model, issues of the model could be more easily visualized.

A graph with a red line and blue dots

AI-generated content may be incorrect.

Figure 1: APSU Ridge Regression Actual Vs. Predicted Degrees Awarded with reference line.

Figure 1 shows that the model predicts fairly well when dealing with seemingly smaller departments with a smaller number of degrees being awarded to students. However, as actual degrees increase, the model struggles to keep up with making correct predictions. The residuals plot gives further insight into the problem.

A graph with blue dots

AI-generated content may be incorrect.

Figure 2: APSU Degrees Awarded Residual plot.

The residual plot shows a sort of funnel shape, with residuals getting worse as predicted degrees increase. This may indicate slight Heteroscedasticity, variance of the errors. Cross Validation was used to tune the ridge hyperparameter, but this did little to fix the MSE or improve the R-squared but was worth it to reduce multicollinearity. Ultimately, the model interpreted the following coefficients for predicting degrees awarded in figure 3. Coefficients in blue indicate positive correlations while the one in red indicates a negative correlation.

A graph with blue and red squares

AI-generated content may be incorrect.

Figure 3: APSU Degrees Awarded Feature Coefficients.

Enrollment seems to be a pretty strong indicator of degrees awarded, which makes sense as the more students in a department means the department is larger. Larger departments usually gain more funding and have more resources than other departments. It was surprising to see the FTE Faculty have a negative correlation with Degrees Awarded, but this may be due to inefficiencies in faculty workloads.

*6.1.2 Predicting APSU Enrollment.* The second target variable selected to predict was enrollment using the independent variables SCH/FTE, Degrees Awarded, SCH, FTE Faculty, and Year. The R-Squared value produced was 0.8044 meaning that 80.44% of the variance in enrollment can be explained by the linear regression model. The MSE produced was 8201, which is fairly high and indicates some model inaccuracy. VIF scores were calculated and indicated collinearity, so Cross-Validation Ridge regression was used again. The R-squared dropped slightly to 0.7781 and the MSE rose to 8337. As can be seen in figure 4, a linear model appears to be slightly better at predicting enrollment than degrees awarded. Although the residuals plot in Figure 5 shows a stronger funnel shape.

A graph with blue dots and a red line

AI-generated content may be incorrect.

Figure 4: APSU Cross Validation Ridge Regression Actual vs predicted enrollment with reference line.

A graph with blue dots and a red line

AI-generated content may be incorrect.

Figure 5: APSU Degrees Awarded Residual plot.

The feature coefficient results were not too surprising as Degrees Awarded takes the top spot. This makes sense as enrollment was previously a good predictor in turn. One difference though is that SCH/FTE Faculty proved to be a stronger predictor for enrollment than degrees awarded. This may indicate that more faculty increase enrollment but do not necessarily increase degrees awarded. The full graph can be found in Figure 6:

A graph with red and blue squares

AI-generated content may be incorrect.

Figure 6: APSU Enrollment Feature Coefficients.

6.2 Public State Institutions

*6.2.1 Predicting State Degrees Awarded*. The first target variable for the Tennessee public institutions linear regression model to be trained on was also Degrees awarded like the APSU dataset. This time there were far more independent variables including Term, Total Undergraduate Students, Total Graduate Students, Total Undergraduate FTE, Total Graduate FTE, FTFTF, and Retention Rate. Before training, the predictor variables were scaled using StandardScaler. The data was split into 75% training data and 25% test data. The linear model produced an R-squared value of 0.9739, which means 97.39% of the variance in degrees awarded can be explained by the linear regression model. The MSE produced was 110911, which was expected with the number of students in this dataset. High VIF scores were computed, which was also expected as many of the features were similar in nature. Cross-Validation Ridge Regression was used to negate multicollinearity. The Cross-Validation hyper tuning chart can be seen in Figure 7.

A graph with a line

AI-generated content may be incorrect.

Figure 7: Cross Validation Error Vs. Alpha.

The tuned hyperparameter was used in the ridge regression and the results were promising. The computed R-Squared value only decreased to 0.9537 and MSE actually decreased to 108341. The plots were even more promising and appealing to the eyes. Figure 8 showcases a strong linear relationship between the predicted and actual awarded degrees.

A graph with a red line and blue dots

AI-generated content may be incorrect.

Figure 8: Tennessee Universities Cross Validation Ridge Regression Actual Vs. Predicted Degrees Awarded with reference line.

The residuals plot showed a fairly random spread of points, with only a few outliers to really cause any concern. The plot can be found in Figure 9.

A graph with blue dots and a red line

AI-generated content may be incorrect.

Figure 9: Tennessee Universities Degrees Awarded Residual plot.

Looking at the Coefficients bar chart found in figure 10, we can see that the number of undergraduates, particularly FTE undergraduates, provides a strong indication of how many degrees will be awarded. This is understandable as the more students in a university, the more degrees will be awarded. Interestingly, the total number of grad students had a positive effect on degrees awarded, but FTE grad students had a negative effect. This may be due to the fact that many grad students are taking fewer classes as they may be in the workforce or have other obligations. Another surprising result was that retention metrics had a negative correlation with degrees awarded. This one is harder to explain and requires further analysis to fully understand. A best guess is that programs that weed out students early, can actually lead to more graduates in the future.

A graph with red and blue squares

AI-generated content may be incorrect.

Figure 10: Tennessee Universities Degrees Awarded Feature Coefficients

*6.2.2 Predicting State Enrollment*. Following the trend, the second target variable that a linear regression model was trained for was enrollment. Less features were used for this model due to many of the features being fairly derivative of enrollment, which may have caused some of the issues with the Austin Peay models. The independent variables were Term, Total Awards, FTFTF, and Retention Rate. Before training, the predictor variables were scaled using StandardScaler. The data was split into 75% training data and 25% test data. The linear model produced an R-squared value of 0.9730, which means 97.3% of the variance in enrollment can be explained by the linear regression model. The MSE produced was 2185678, which was much larger than the MSE from predicting degrees awarded. VIF scores were not computed, but to keep the theme, Cross-Validation Ridge Regression was still done. The R-Squared decreased slightly to 0.9483 and unfortunately the MSE stayed about the same. Looking at the enrollment prediction plot in figure 11, it is very similar to the prediction plot for degrees awarded. There is a weird cluster noticeable in the midway point that was not present in the degrees awarded plot, but overall appears to be a bit more linear, especially at the beginning of the plot.

A graph with a red line and blue dots

AI-generated content may be incorrect.

Figure 11: Tennessee Universities Cross Validation Ridge Regression Actual Vs. Predicted Enrollment with reference line.

The residuals plot for this model did bring some cause for concern, as the scatter was not entirely random. The plot shown in Figure 12 may indicate that the model cannot predict well when actual enrollment is higher than usual. This is probably due to the variance in enrollment numbers across Tennessee universities.

A graph with blue dots

AI-generated content may be incorrect.

Figure 12: Tennessee Universities Enrollment Residual plot.

Looking at the feature coefficients, Total degrees awarded has the biggest positive correlation with increasing enrollment. This was predictable, sense total undergraduates had the biggest weight for predicting degrees awarded. FTFTF had a fairly positive weight as well, suggesting that an increase in new freshman can explain rises in enrollment. It was surprising to see the retention rate was almost 0, but slightly negative correlation. One would think retention rates being high would maintain and support increased enrollment. One likely scenario is that universities rely heavily on incoming freshmen to outweigh the loss of current students. The coefficients bar chart for enrollment can be found in Figure 13.

A graph with blue and red squares

AI-generated content may be incorrect.

Figure 13: Tennessee Universities Enrollment Feature Coefficients

7 Conclusion, Limitations, and Future Work

Overall, the models performed adequately with the amount of data they had. For Austin Peay State University, it is possible that maintaining an increase in student enrollment will continue to increase the number of degrees awarded. SCH/FTE Faculty may be an important factor in determining enrollment, but more research needs to be done to verify this.

The models trained on public Tennessee universities provide evidence that universities may be focusing to hard on bringing in new freshman, when it may be more beneficial to work towards awarding more degrees instead. More research needs to be done on how graduate and undergraduate student totals affect degrees enrollment differently.

This research was heavily limited by the amount of data that was able to be collected. The models could prove more reliable if previous years’ data was made available. The number of independent variables also seemed to have a negative impact on the models, highlighting the need for more kinds of data to be collected, or for more and better features to be engineered. It is also possible to test other machine learning models and methods for cleaning the data and fixing assumptions.

As more enrollment data becomes available year by year. The models created from this research can be bolstered and improved so that better predictions are made.

ACKNOWLEDGMENTS

The author thanks Dr. Jiang Li from Austin Peay State University for his help and instruction during the course of this research. He was pivotal in the success of this research. Dr. Li provided valuable direction in his role as Graduate Coordinator for Data Science.

The author thanks Dr. Leong Lee for his leadership and guidance as Austin Peay State University Department Chair for the Department of Computer Science and Information Technology. Dr. Lee also provided the raw enrollment data from Ausitn Peay State University.

The author thanks their parents, David and Teresa Simmons, for their continued support through the author’s academic career.

REFERENCES

[1] Mohammed Nasiru Yakubu and A. Mohammed Abubakar. 2021. Applying machine learning approach to predict students’ performance in higher educational institutions. Kybernetes 51, 2 (2021), 916–934. DOI: https://doi.org/10.1108/K-12-2020-0865.

[2] Ghazi Al-Naymat and Mohammed Azmi Al-Betar. 2024. University student enrollment prediction: A machine learning framework. In Proceedings of the Third International Conference on Innovations in Computing Research (ICR’24), Lecture Notes in Networks and Systems, Vol. 1058. Springer, Cham, 51–62. DOI: https://doi.org/10.1007/978-3-031-65522-7\_5.

[3] John A. Centra. 1978. Financial aid and student enrollment. The Journal of Higher Education 49, 5 (1978), 548–574. DOI: https://doi.org/10.1080/00221546.1978.11780411.

[4] David A. Freedman. 2009. Statistical Models: Theory and Practice. Cambridge University Press, 26.

[5] Xin Yan. 2009. Linear Regression Analysis: Theory and Computing. World Scientific, 1–2. ISBN 9789812834119. Archived from the original on October 4, 2024.

[6] Peter Kennedy. 2003. A Guide to Econometrics (5th ed.). The MIT Press, Cambridge, 205–206. ISBN 0-262-61183-X.

[7] Jay L. Devore. 2011. *Probability and Statistics for Engineering and the Sciences* (8th ed.). Cengage Learning, Boston, MA, 508–510. ISBN 978-0-538-73352-6.

[8] Peter J. Bickel and Kjell A. Doksum. 2015. Mathematical Statistics: Basic Ideas and Selected Topics, Vol. I (2nd ed.). Chapman and Hall/CRC, 20.

1. The University of Tennessee Southern was not included as their enrollment data was only included in THEC’s report for 2023. [↑](#footnote-ref-2)