The Added Value of Student Affairs in Retention Prediction Models

Shaun Boren, Ed.D.
University of Florida
Director, Student Life Assessment and Research
shaun.boren@ufl.edu | LinkedIn

Qichen Li, M.S.
University of Florida
Assistant Director, Student Life Assessment & Research
LinkedIn

Maddy Trudeau, M.A.
University of Kentucky
Director, Student Success Assessment
LinkedIn

Abstract: For this inquiry we supplemented typical academic variables to evaluate the degree to which several student affairs variables add value in a machine learning model predicting first-year retention. Findings indicated that living on campus, being Greek, as well as engaging in recreational sports all had positive contributions to predicting retention. Higher education leaders should use this study to advocate for and enact the inclusion of student affairs variables into predictive models of student success.

Keywords: machine learning, student success, retention, student affairs

---

During the pandemic, our institution developed an early alert system to identify students at risk for dropping out of the institution in order to offer more targeted and timely support. The system primarily utilized academic data, including mid-semester grades and utilization of the learning management system. Working in a student affairs assessment office, we were interested in whether adding student affairs variables would improve the power of the early alert system. This study was conducted by student affairs assessment professionals, involved permission from data custodians of several student affairs departments, and included collaboration with analysts involved in our university's early alert program. It was pitched to student affairs leadership as a way to build the case for including student affairs data in future models.

Machine Learning

Our institution's early alert system utilized machine learning predictive models. Machine learning is a type of artificial intelligence that uses a computer (machine) to improve predictive capabilities (learning) without human programming at each iteration. Machine learning uses algorithms - such as logistic regression - to analyze data, and the resulting output is the model, used to make predictions. Our study involved a supervised machine learning model, where algorithms are trained on a dataset where the actual outcome is known, so we could then test the predictive capability of the resulting model.

The literature underscores the potential of machine learning for identifying at-risk students through early warning systems, where insights drawn from institutional data enable

institutions to take proactive measures to support vulnerable students (Hoffait & Schyns, 2017; Howard et al., 2018). Machine learning models are emerging as transformative tools in this area due to their adaptability and ability to handle complex datasets. Unlike traditional statistical models, which can be restricted by assumptions or limited tolerance for incomplete data, machine learning models perform effectively even with missing information and have been shown to enhance predictive accuracy for retention outcomes (Alkhasawneh & Hargraves, 2014; Delen, 2011).

However, machine learning model accuracy is highly variable and context-dependent, with few studies directly comparing the effectiveness of different predictive models in this domain. Scholars such as Cardona et al. (2020) recommend combining various models to identify which approaches yield the highest accuracy across institutional contexts. Addressing specific challenges like the *cold start* problem, where limited data is available on new students, has led to advancements in prediction techniques, with strategies designed to extend model accuracy for students who lack historical data (Sweeney et al., 2016).

Retention

The outcome we selected for this study was first-year retention, specifically a student's continuous enrollment from the first fall semester to the following fall. Using retention as an outcome measure offers institutions the advantage of identifying and addressing potential risks early, rather than waiting for graduation data, making it a foundational metric in student success tracking (DesJardines et al., 2003). Many institutions have implemented data-driven approaches to better understand student retention and predict potential dropouts (Cui et al., 2019).

Traditional retention prediction models often include a range of variables spanning academics, demographics, and socioeconomics. Studies consistently find that academic metrics, such as freshman grades, are highly significant in predicting retention, with pre-college variables like high school performance also playing a critical role (Cardona et al., 2020; Delen, 2010; Raju & Schumacker, 2015). In addition to academic performance, demographic and socioeconomic variables—such as age, gender, financial status, and residency—are frequently included in models. However, their impact on retention is inconsistent, with studies reporting mixed findings depending on institutional type or discipline. For example, while demographics like race and gender are relevant predictors in STEM fields, other studies find these variables to have minimal impact in different academic settings (Alkhasawneh & Hargraves, 2014; McAleer & Szakas, 2010).

Financial and socioeconomic variables also hold predictive value, although, as some studies indicate, results are inconsistent due to variances in how family socioeconomic status is measured. For instance, Oztekin (2016) found monetary variables are less relevant predictors of graduation rates. However, Marquez-Vera et al. (2016) identified 'mother's level of education' as a predictor for dropout rates among Mexican high school students, which aligns with findings in the United States showing that family socio-economic status is a strong correlate of academic performance (Sirin, 2005).

Researchers have increasingly called for an expansion in the types of variables included in retention prediction models, arguing that non-academic variables—such as emotional well-being, family background, and cultural variables—play a significant role in students' academic decisions (Delen, 2010; Slim et al., 2014). For example, Delen (2010) emphasizes that student retention is often higher when students perceive their university environment as aligned with their personal values and social interests. Others have since found that social integration and engagement variables, such as living on campus, involvement in campus organizations, or engaging in campus recreation, are influential predictors of student persistence (Graham et al., 2021; Milton et al., 2020; Oztekin, 2016).

Methods

Data Sources and Measurement Plan

This study utilized multiple institutional data sources spanning three academic years (2017–2020) of freshman data to predict retention into their second fall semester. Primary data sources included the university's student information system for academic and demographic data, as well as data from the systems of several student affairs units: on/off-campus status from Housing, facility visits from Recreational Sports, greek affiliation from Greek Life, and program engagement from the Career Center.

The measurement approach was developed by the authors, who were staff members conducting this inquiry in their roles in an assessment office within a student affairs division at a southeast university. Authors have various levels of experience in higher education, from 1 to 20+ years. The authors also vary in their education level, ranging from masters to doctoral degrees. The selection of variables was informed by previous retention studies (Delen, 2010; Oztekin, 2016) and input from colleagues in student affairs at the same institution regarding available data points that could indicate student engagement.

While machine learning models perform well with missing data, they benefit from a more even distribution of data across outcomes (Abd Elrahman & Abraham, 2013). At this institution, first-year retention is around 95%, leaving only around 5% in the non-retention group. To decrease this difference, the data set used in this study focused on the 2,054 students with a cumulative first-year GPA of <=3.0. This cutoff was selected because <=2.0 would have been too small a sample, and <=3.0 is in alignment with the threshold to maintain scholarships such as Florida Bright Futures (Florida Department of Education, 2024). With this group, retention was closer to 75%, as detailed in Table 1.

Table 1. *Number of Students Retained by GPA*

Cumulative GPA	Retained	Not Retained
GPA<=3.0	2,040	504
Any GPA	20,898	929

Table 2 details a list of variables included in the analysis, noting those from student affairs. A Greek Student is a student that was an active member of a university-approved fraternity or sorority in their first year. Live on Campus is a student who lives in university-run Housing. Recreational Sports visits in Spring or Fall is the total number of times a student swipes into a Recreational Sports facility that semester. Career Center visits in Spring or Fall is the total number of times a student uses a menu of Career Center services that semester, including various 1:1 appointment types as well as workshops and larger events such as career fairs.

Table 2. *Variables Included in Analysis*

Description	Data Type
* Career Center visits in Fall	Number
* Career Center visits in Spring	Number
Carried Hours in Fall	Number
Carried Hours in Spring	Number
Classification	Number
College	Category
County Code	Category
Ethnicity	Category
First Generation Flag	Binary nominal
Gender	Binary nominal
* Greek Student	Binary nominal
* Live on Campus	Binary nominal
Major	Category
Term Registered Hours in Fall	Number
Transfer Hours	Number
* Recreational Sports visits in Fall	Number
* Recreational Sports visits in Spring	Number
Residency	Category
Second fall registered (Y/N)	Binary nominal
Term GPA in Fall	Number
Term GPA in Spring	Number

* Variables from Student Affairs

Implementation and Analysis Process

We used the Cross-Industry Standard Process for Data Mining (CRISP-DM), which provides a systematic and structured way of conducting data mining studies, and hence increasing the likelihood of obtaining accurate and reliable results (Delen, 2010). This method contains six steps:

1. Understand the business needs and develop the goal for study
2. Identify, collect, and understand the relevant data for the study
3. Select attributes, clean, and transform the data for modeling
4. Use various modeling techniques to develop models
5. Evaluate and assess if the results of models are valid and meet the goal of study
6. Deploy the model and use the result in a decision-making process

Before applying the machine learning models, several preprocessing steps were necessary to prepare the data. Categorical variables, such as student major, residency status, and Greek life membership, were transformed using one-hot encoding, a technique that converts categorical variables into a binary format suitable for machine learning algorithms (Potdar et al., 2017). For example, the 'Greek Student' variable was converted into two binary columns: 'Greek_Yes' and 'Greek_No', where each column would contain a value of 1 or 0 for each student.

Numerical variables, including GPA and visit counts, were standardized using z-score normalization to ensure all features were on the same scale. This transformation is particularly important for the logistic regression and neural network algorithms, which are sensitive to the scale of input features (Zhang et al., 2019). The standardization process involved subtracting the mean and dividing by the standard deviation for each numerical feature:

$$Z = (X - \mu) / \sigma$$

During implementation, the primary challenge was data imbalance in the original dataset, which led to our focus on students with ≤3.0 GPA and using K-fold cross-validation.

K-fold cross-validation was applied to minimize the bias and estimate model performances. This procedure involves dividing the data into k groups of samples, which are called folds. As k gets larger, the difference in size between the training set and the resampling subsets gets smaller. As this difference decreases, the bias of the technique becomes smaller (Kuhn & Johnson, 2013). In this study, we set k to 10. This meant the dataset was divided into 10 folds, with nine folds used to train the model and one fold used to test the predictive performance of the model.

The technical implementation utilized Python for data preprocessing and model development, employing the Scikit-learn library for machine learning implementations, Pandas for data manipulation, and NumPy for numerical computations. Data visualization was accomplished using Matplotlib and Seaborn libraries.

Classification Models

In this study, we used three supervised machine learning algorithms: Logistic regression, random forest, and artificial neural network. Logistic regression, which can be considered a machine learning as well as general data analysis technique, is powerful at solving classification problems (Ray, 2019). It predicts the chances of a categorical outcome given one or more ranked input variables. At the most basic level, logistic regression has a binary outcome, such as whether a student is retained, but it can also extend to multi-class outcomes (Delen, 2010; Karsmakers et al., 2007).

Random forest combines a set of decision trees, which are branching choices of how to predict outcomes. The model is refined using bootstrapping, which involves picking random samples to test against each other and average across the entire dataset (Gislason et al., 2004). Accuracy is then verified against a final sample from the original dataset that was not part of the bootstrapping.

An artificial neural network mimics the functionality between neurons in the human brain (Zhang et al., 2019). Its basic structure is formed from the input layer, hidden layer, output layer, each of which consists of at least one unit (neuron). Units in the hidden layer receive data from the input unit, adjust the weight through connection and function, and then pass the data to the output layer. (Krose & Smagt, 2011). These hidden layers between input and output use this processing to predict nonlinear relationships (Dreiseitl & Ohno-Mchado, 2003; Lek et al., 1996).

Model Evaluation and Validation

The performance of each model was assessed using metrics organized into what is called a confusion matrix, which is a table containing the counts of predicted and actual values. The rows represent the actual categories, and the columns represent the predicted categories. The subsequent four metrics are defined as follows:

- True positive (TP): number of students correctly predicted as not retained
- False positive (FP): number of records incorrectly predicted as not retained
- False negative (FN): number of records incorrectly predicted as retained
- True negative (TN): number of records correctly predicted as retained

These metrics are in turn combined in the following equations to calculate accuracy, sensitivity, and specificity:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

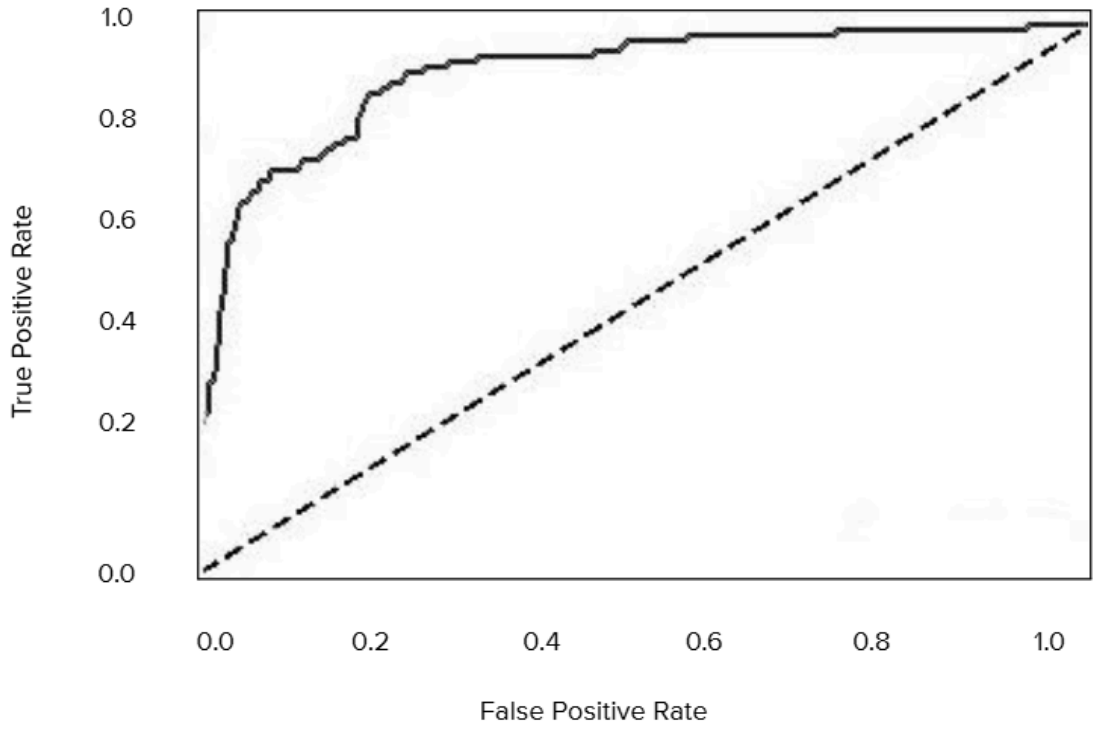
$$\text{Specificity} = \frac{TN}{TN + FP}$$

In this study, sensitivity is the ability of the model to predict which students will drop out, and specificity is the ability of the model to predict which students will come back. Accuracy is the percentage of students correctly classified out of all students in the model.

Sensitivity and specificity are inversely related. As a model is adjusted to increase its ability to predict true positives, this comes with more false positives, so its ability to predict true negatives decreases. This relationship between sensitivity and specificity is often visualized in a graph summarizing all possible adjustments to the model, as shown in Figure 1. This is called the Receiver Operating Characteristics Curve (ROC Curve). The y-axis is the true positive rate, while the x-axis is the false positive rate. A model with no predictive power would yield a 50% split, which is represented by a dotted line across the graph. The solid line contains a point for the performance of each model's variation in threshold between sensitivity and specificity. The higher this line is above the 50% default, the better that model's performance.

Area under the ROC curve, often abbreviated as AUC, is the two-dimensional space under the ROC curve. This serves as an aggregate measure of performance across all thresholds of a model. A perfect model would have an area of 1, meaning that all cases are correctly identified as positives or negatives (Hosmer et al., 2013). Once a model is trained and identified as meeting an acceptable AUC, permutation feature importance is calculated to examine how important each input variable is to the model. Feature importance is calculated by quantifying the decrease in accuracy after randomly shuffling the value of a single feature.

Figure 1. *Example ROC Plot with Area Under the Curve.*



Results

Model Performance

The aim of this study was to supplement academic predictors to evaluate the degree to which several student affairs variables add value to classification models predicting student retention. Across all three models tested, the inclusion of student affairs data demonstrated improved predictive capability (Table 3). The random forest model achieved the highest AUC value (0.910), while the neural network showed the highest accuracy (88%) and specificity (0.904). Both logistic regression and random forest models demonstrated equal sensitivity (0.822). Notably, all three models resulted in AUC values above 0.9, indicating an outstanding level of discrimination between positive and negative instances (Hosmer & Lemeshow, 2000).

Table 3. *Model Results*

Measure	Random Forest	Logistic Regression	Neural Network
Models Including SA Variables			
Accuracy	83%	85%	88%
Sensitivity	0.822	0.822	0.772
Specificity	0.838	0.858	0.904
AUC	0.910	0.900	0.903
Models Not Including SA Variables			
Accuracy	81%	85%	84%
Sensitivity	0.891	0.832	0.822
Specificity	0.784	0.865	0.840
AUC	0.889	0.899	0.897

When comparing classification models with and without student affairs data, the models incorporating student affairs variables consistently showed higher accuracy and AUC scores. The logistic regression and neural network models specifically demonstrated improved specificity scores with the inclusion of student affairs data. While sensitivity scores were generally lower in the student affairs group, the random forest model proved an exception, showing enhanced sensitivity with the inclusion of these variables. These results suggest that student affairs data can contribute meaningfully to the overall performance of retention prediction models.

Feature Importance and Variable Relationships

Analysis of feature importance scores across the three models revealed that while traditional academic metrics maintained high importance (Spring term GPA, Spring Total Carried Hours, Fall term GPA), several student affairs variables emerged as significant predictors. Specifically, Recreational Sports visits in Spring, Live on Campus status, and

Greek Student affiliation appeared consistently among the top 10 features across models as noted in Table 4.

Table 4. *Top 10 Model Features Importance Score by Model*

Model Feature	Importance Score
Logistic Regression	
Spring term GPA	0.032711
Spring Total Carried Hours	0.009823
Fall term GPA	0.008350
Spring Total Earned Hours	0.007073
College	0.004028
* Recreational Sports visits in Spring	0.003733
Residency	0.003536
* Live on Campus	0.003340
Term Registered Hours in Fall	0.003143
First Generation Flag	0.002554
Artificial Neural Network	
Fall term GPA	0.079699
* Greek Student	0.003798
Fall term GPA	0.002554
Spring Total Earned Hours	0.001768
* Live on Campus	0.001637
First Generation Flag	0.001375
Term Registered Hours in Fall	0.001244
Gender	0.001244
Fall Carried Hours	0.001179
Major	0.000917
Random Forest	
Spring term GPA	0.045187
Spring Total Carried Hours	0.011788
* Recreational Sports visits in Spring	0.007859
Fall Carried Hours	0.006483
* Recreational Sports visits in Fall	0.006418
College	0.006221
Fall term GPA	0.005697
Transfer Hours	0.004650
* Greek Student	0.003733
Ethnicity	0.002620

* Variables from Student Affairs.

The addition of student affairs data influenced importance scores differently across models: in the logistic regression model, Live on Campus and Recreational Sports visits in Spring ranked among the top ten important variables; the random forest model ranked Greek Student status as the third most important predictor; and the neural network model identified Recreational Sports visits in Spring and Greek Student status as key student affairs predictors.

To understand the directional relationship between student affairs variables and retention, follow-up correlation analyses were conducted in Python. As noted in Table 5, all student affairs variables showed significant (n = 2,054, p < .0001) positive associations with retention, with Recreational Sports visits showing the strongest correlation (Spring: r = 0.333, Fall: r = 0.304), followed by Living on Campus (r = 0.204) and Greek participation (r = 0.133). These findings have implications for both immediate practice and future inquiry, suggesting the value of incorporating broader student affairs metrics into retention prediction models while highlighting opportunities for improved data collection and integration across student affairs units.

Table 5. *Follow-up Correlations*

Variable	Phi Coefficient	Probability
Recreational Sports visits in Spring	0.3330	<.0001
Recreational Sports visits in Fall	0.3038	<.0001
Live on Campus in Fall	0.2042	<.0001
Greek	0.1330	<.0001

Stakeholder Engagement and Implementation

The collaboration with each contributing department helped advance several initiatives: departments have agreed to expand the development of more proactive assessment plans than they have in the past; division leadership has committed to collecting more unique student participation data; once this data is collected, we have also gained buy-in to integrate disparate data platforms into a unified database; and we have more support to standardized metrics for division-wide assessment and storytelling. This study is also helping us pursue an expanded university early alert system that includes student affairs variables.

Recommendations

Results from this study support the value of several student affairs variables when included in models of predicting student retention. Recreational Sports Visits, Living on Campus, and being Greek had positive contributions to predicting retention. These findings confirm other studies reporting the importance of social interaction variables when analyzing retention (Delen, 2010; Oztekin, 2016; Slim et al., 2014).

Results also answered the call from several studies to merge several machine learning models when developing predictions of student success (Cardona et al., 2020; Sweeney et

al., 2016). Future use of machine learning for student success, in inquiry and application, should benefit from similar merging of multiple models.

There are several limitations to acknowledge in this study. The dataset used is from only a single institution and from a single three-year period that happened to lead into a pandemic. In addition to variables that may differ across institutions and years, there is a high retention rate at this institution restricting the amount of data on non-retained students. The methods of this study should therefore be replicated at other institutions with varied retention rates to assess whether the results are generalizable. If exploring replication at other institutions, consider engaging institutional research, information technology, faculty, and other campus partners with expertise in machine learning models.

This study also only uses a subset of student affairs variables that happened to be available. There are a variety of other student affairs variables worth considering for student success models, such as student activities, student organizations, leadership programs, other health programs, and disability resource center services. Future inquiry should continue and expand the incorporation of such student affairs variables as potential contributors to predicting student success. There is also opportunity to apply theoretical frameworks in exploration of which student affairs variables might impact retention as well as understanding the mechanisms of these impacts.

Another limitation is that we only focused on first-year retention as the output of the predictive model. There are other commonly used measures of student success, each with varied gaps between prediction and outcome. For example, term or even course GPA are finalized all within a single semester. On the other end of the spectrum, the graduation rate takes 4-6 years to be realized. The contributions of student affairs to predicting such disparate outcomes requires additional inquiry.

Conclusion

Student affairs units and leadership can use results from this study to advocate for including their data in university efforts to predict and identify targeted support for student success. Furthermore, since most included student affairs data points had a positive association with retention, results also evidence the contribution of student affairs to not just predictive models, but also to the goal of student success. While historical and demographic variables cannot be changed by the time students attend higher education, the student affairs variables in this study—which change as students pursue higher education—offer an important opportunity to support the success of even more students.

References

Abd Elrahman, S. M., & Abraham, A. (2013). A review of class imbalance problem. *Journal of Network and Innovative Computing, 1*(2013), 332–340. https://cspub-jnic.org/index.php/jnic/article/view/42

Alkhasawneh, R., & Hargraves, R. H. (2014). Developing a hybrid model to predict student first year retention and academic success in STEM disciplines using neural network. *Journal of STEM Education: Innovations and Research*, *15*(3), 35–42.

Astin, A. (1993). *What matters in college: Four critical years revisited.* Jossey-Bass.

Cardona, T., Cudney, E. A., Hoerl, R., & Snyder, J. (2020). Data Mining and Machine Learning Retention Models in Higher Education. *Journal of College Student Retention: Research, Theory and Practice*, *25*(1), 1–25. https://doi.org/10.1177/1521025120964920

Cui, Y., Chen, F., Shiri, A., & Fan, Y. (2019). Predictive analytic models of student success in higher education: A review of methodology. *Information and Learning Sciences*, *120*(3/4), 208–227. https://doi.org/10.1108/ILS-10-2018-0104

Delen, D. (2011). Predicting student attrition with data mining methods. *Journal of College Student Retention: Research, Theory and Practice*, *13*(1), 17–35. https://doi.org/10.2190/CS.13.1.b

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, *49*(4), 498–506. https://doi.org/10.1016/j.dss.2010.06.003

DesJardins, S. L., Kim, D. O., & Rzonca, C. S. (2003). A nested analysis of factors affecting bachelor's degree completion. *Journal of College Student Retention: Research, Theory & Practice, 4*(4), 407–435. https://www.proquest.com/docview/196711670

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics, 35*(5-6), 352–359. https://doi.org/10.1016/S1532-0464(03)00034-0

Florida Department of Education. (2024). *Bright Futures Student Handbook*. https://www.floridastudentfinancialaidsg.org/PDF/BFHandbookChapter3.pdf

Gislason, P. O., Benediktsson, J. A., & Sveinsson, J. R. (2004). Random Forest classification of multisource remote sensing and geographic data. *2004 IEEE International Geoscience and Remote Sensing Symposium, 2,* 1049–1052. https://doi.org/10.1109/IGARSS.2004.1368591

Graham, P. A., Gonyea, R. M., Fosnacht, K., & Fassett, K. T. (2021). The case for campus housing: Results from a national study (A Brief for Students, Parents, and Media). *ACUHO-I*. https://www.acuho-i.org/wp-content/uploads/2024/03/2024_caseforcampushousing_parents.pdf

Hoffait, A. S., & Schyns, M. (2017). Early detection of university students with potential difficulties. *Decision Support Systems, 101*, 1–11. https://doi.org/10.1016/j.dss.2017.05.003

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression.* John Wiley & Sons.

Hosmer Jr., D. W., Lemeshow, S., & Sturdivant, R. X. (2013) *Applied Logistic Regression* (3rd ed.). John Wiley & Sons.

Howard, E., Meehan, M., & Parnell, A. (2018). Contrasting prediction methods for early warning systems at undergraduate level. *The Internet and Higher Education, 37*, 66–75. https://doi.org/10.1016/j.iheduc.2018.02.001

Karsmakers, P., Pelckmans, K., & Suykens, J. A. (2007, August). Multi-class kernel logistic regression: a fixed-size implementation. *International Joint Conference on Neural Networks*, 1756–1761. https://doi.org/10.1109/IJCNN.2007.4371223

Krose, B., & Smagt, P. V. D. (2011). *An introduction to neural networks.* The University of Amsterdam.

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., & Aulagnier, S. (1996). Application of neural networks to modelling nonlinear relationships in ecology. *Ecological modelling, 90*(1), 39–52. https://doi.org/10.1016/0304-3800(95)00142-5

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: A case study with high school students. *Expert Systems*, *33*(1), 107–124. https://doi.org/10.1111/exsy.12135

McAleer, B., & Szakas, J. S. (2010). Myth busting: Using data mining to refute link between transfer students and retention risk. *Information Systems Education Journal*, *8*(19), 3–7. https://isedj.org/8/19/

Milton, P. R., Williamson, L. M., Brubaker, K., & Papania, M. (2020). Recreate and retain: How entrance into a campus recreation facility impacts retention. *Recreational Sports Journal, 44*(2), 89–98. https://doi.org/10.1177/1558866120964818

Oztekin, A. (2016). A hybrid data analytic approach to predict college graduation status and its determinative factors. *Industrial Management and Data Systems*, *116*(8), 1678–1699. https://doi.org/10.1108/IMDS-09-2015-0363

Raju, D., & Schumacker, R. (2015). Exploring student characteristics of retention that lead to graduation in higher education using data mining models. *Journal of College Student Retention: Research, Theory and Practice*, *16*(4), 563–591. https://doi.org/10.2190/CS.16.4.e

Ray, S. (2019, February 14–16). *A quick review of machine learning algorithms {Paper presentation]*. 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon, Faridabad, India.). https://doi.org/10.1109/COMITCon.2019.8862451

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, *75*(3), 417–453. https://doi.org/10.3102/00346543075003417

Slim, A., Heileman, G. L., Kozlick, J., & Abdallah, C. T. (2014, December 9–12). *Predicting student success based on prior performance* [Paper presentation]. 2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), Orlando, FL, United States. https://doi.org/10.1109/CIDM.2014.7008697

Sweeney, M., Rangwala, H., Lester, J., & Johri, A. (2016). Next-term student performance prediction: A recommender systems approach. *Journal of Educational Data Mining, 8*(1), 22–51. https://doi.org/10.5281/zenodo.3554603

Zhang, Q., Yu, H., Barbiero, M., Wang, B., & Gu, M. (2019). Artificial neural networks enabled by nanophotonics. *Light: Science & Applications*, *8*(1), 1–14. https://www.nature.com/articles/s41377-019-0151-0

How to cite this article: Boren, S., Li, Q., & Trudeau, M. (2024). The added value of student affairs in retention prediction models. *Journal of Student Affairs Inquiry, Improvement, and Impact*, *7*(1), 171–183. https://doi.org/10.18060/28144