Exploring Four Different Data Mining Models to Predict Community College First-Year Retention

Camille Gasaway Pace, Ed.D.
The College Board
gasaway.camille@gmail.com
Tel: 470-963-0009

Lantry L. Brockmeier, Ph.D.
Leadership, Technology, and Workforce Development
Valdosta State University
1500 N. Patterson St. Valdosta, GA 31698
llbrockmeier@valdosta.edu
Tel: 229-333-5633

Michael J. Bochenko, Ed.D.
Leadership, Technology, and Workforce Development
Valdosta State University
1500 N. Patterson St. Valdosta, GA 31698
mjbochenko@valdosta.edu
Tel: 229-333-5633

Daesang Kim, Ph.D.
Leadership, Technology, and Workforce Development
Valdosta State University
1500 N. Patterson St. Valdosta, GA 31698
daekim@valdosta.edu
Tel: 229-333-5633

Abstract

This study aimed to create a predictive model for student retention using background, academic, and financial factors to guide other community colleges to use when investigating institutional retention. Four different data mining models (neural networks, random forest trees, support vector machines, and logistic regression) identified significant factors for retention. The number of credit hours was consistently the most crucial variable in retention. In addition, the interactions between the number of credit hours, GPA, and financial aid variables were significant in student retention in their first year. There were no consistent variables among the retention models that can predict students' nonretention in the freshman year. Background predictors (age, gender, race, or ethnicity) were not significant in predicting retained or nonretained students. The comparison of the retention models found that the random forest model had the best performance for accurately classifying the non-retained and retained students overall and the retained students individually.

Keywords: data mining, student retention, community college, classification models

1. Introduction

Community colleges play a vital part in the educational landscape serving the needs of nontraditional students and more than half of the minority students in the country. However, the retention of community college students continues to be a concern as the overall enrollment of students continues to decrease, and half of freshmen students do not return to continue their education (Juszkiewicz, 2020). In addition, community college students may face environmental factors such as employment, family obligations, and financial insecurity affecting their ability to remain enrolled. Specialized retention models can help colleges identify important variables for student retention serving as the basis for

new programs and initiatives. Additionally, these models can provide a framework for institutions to understand the needs of specific populations.

Retention frameworks identified students' social and performance factors, including their collegiate relationships in higher education institutions (Aljohani, 2016; Berger et al., 2012). Bean and Metzner (1985) developed the nontraditional undergraduate student attrition model built onto previous retention models and focused primarily on the nontraditional student population (Aljohani, 2016; Bean & Metzner, 1985; Johnson et al., 2014). Their model theorized that student retention depended on the link between high school and college performance, psychological and environmental outcomes playing a more significant role than academic variables, and background variables influencing student persistence (Aljohani, 2016; Bean & Metzner, 1985). The model was tested and found that nontraditional students drop out for academic reasons unrelated to social interaction (Metzner & Bean, 1987).

Metzner and Bean (1987) theorized that "samples of nontraditional students tend to be heterogeneous and probably differ substantially from university to university so that the combination of several schools might not produce additive effects" (p. 34), indicating the need for individualized retention models. Another type of modeling is sector-based retention models using similar schools within a sector or area to identify relationships not detected in institutions with smaller populations (Herzog, 2006). With roughly 31% of community college students transferring to four-year institutions, specialized sector models could help the community colleges where students begin and the institutions to where they move (Shapiro et al., 2017).

Many of the academic, background, and financial factors identified by Bean and Metzner (1985) will be used to answer the research questions. The student background characteristics gathered at enrollment reflected demographic information such as high school performance and other demographic factors (Johnson et al., 2014). Academic performance is an indicator of future performance in the retention and graduation of community college students at their current and future institutions (Pascarella & Terenzini, 2005). Students' financial attitudes slightly impact retention for students in the nontraditional student attrition model and student retention integrated model (Bean & Metzner, 1985; Cabrera et al., 1993).

Data mining methods in retention models have increased over the last decade, with the introduction of these techniques allowing higher education institutions to expand from models whose performance is affected by skewed data or outliers. Instead, institutions can quickly create models to understand the factors currently affecting students and compare them to previous models. (Cardona, Cudney, Hoerl & Snyder, 2020). One type of data mining model (i.e., classification models) can predict the classes to which individual cases of the dependent variables belong and is ideal for retention models (Attewell & Monaghan, 2015; Bharati & Ramageri, 2010; Breiman, 1999; Breiman et al., 1984; Han et al., 2011). Wolpert's No Free Lunch Theorem indicated that no one classifier could handle all data sets (Wolpert, 1996). Kuhn and Johnson (2013) recommend that researchers start with complex models with the most flexibility and less interpretability to give the most accurate results. Fernández-Delgado et al. (2014) found that researchers often use familiar classification methods that are not the most accurate classifier for the problem. They measured the accuracy rates of 179 different classifiers on 121 data sets to determine classifier behavior. They found that random forests, support vector machine (SVM), neural networks, and boosting ensemble models had the most accurate results among the 121 different data sets (Fernández-Delgado et al., 2014).

Among classification models, random forest trees have some of the highest accuracy rates among different data sets and disciplines (Caruana & Niculescu-Mizil, 2006; Dissanayake, Robinson & Al-Azzam, 2016; Fernández-Delgado et al., 2014; He, Levine, Fan, Beemer & Stronach, 2018). Random forest models have predicted student progress, student performance, completion and graduation rates, and licensing rates, but these are less used than decision trees (Goga et al., 2015; Hardman, Paucar-Caceres & Fielding, 2013; He et al., 2018; Hutt, Gardener, Kamentz, Duckworth & D'Mello, 2018; Langan, Harris, Barrett, Hamshire & Wibberley, 2018). For higher education data, SVMs have been used to predict student retention with mixed results compared to other classifier methods in accuracy (Delen, 2010; Lauría, Baron, Devireddy, Sundararaju & Jayaprakash, 2012; Zhang, Oussena, Clark & Kim, 2010). SVMs use different kernel functions to transform the data based on the specific function, with the optimal kernel determination occurring through trial and error (Attewell & Monaghan, 2015, Fernández-Delgado et al., 2014; James et al., 2013). Neural networks have predicted student course selection, institutional application, retention, and graduation times (Delen, 2010; González & DesJardins, 2002; Herzog, 2006; Kardan, Sadeghi, Ghidary & Sani, 2013; Luan, 2002). Even with higher classification accuracy, neural networks can be challenging to interpret the input and output relationship (Attewell & Monaghan, 2015; González & DesJardins, 2002).

A standard classification method used in higher education is logistic regression which dates back to the 1960s (Cabrera, 1994). Higher education research using logistic regression range in topics from student retention, student graduation, and the interactions between students and faculty (Astin & Oseguera, 2005; Chatterjee, Marachi, Natekar, Rai & Yeung, 2018; Delen, 2010; Herzog, 2006; Lauría et al., 2012; Pyke & Sheridan, 1993). With the logistic regression's similarity to linear regression and a more straightforward interpretation of the results versus other data mining techniques, many educational researchers choose logistic regression as their statistical method (Gunu, Lee, Gyasi & Roe, 2017; Peng, So, Stage & John, 2002).

2. Purpose of the Study

Individual community colleges may create retention models to understand student populations but may miss patterns or significant variables due to sample size. The purpose of this study was to develop a predictive model for student retention using background, academic, and financial factors serving as a guide for other community colleges to use when investigating institutional retention. This study expands the impact of a school-specific model to include seven different community colleges to identify important predictors and relationships among the state college sector schools. The models created in this research represented freshmen students for two years and can provide scalability to individual schools and whole sectors of community colleges. The background, academic, and financial predictors in this study aligned with Bean and Metzner's nontraditional undergraduate student attrition model (1985). Additionally, the study aims to identify which of the four types of models produces the most accurate results for classifying student retention. The model selection of random forests, SVMs, and neural networks were based on the recommendations of Fernández-Delgado, Cernadas, Barro & Amorim (2014).

3. Methodology

The methodology section is divided into three subsections. The first part of this section will focus on the research design. This will be followed by a discussion of the participants. Finally, the last part of the section will discuss the data analysis.

3.1 Research Design

This study using archival data is a nonexperimental, correlational classification research design created to predict student retention who completed three consecutive semesters at seven community colleges in Georgia. The first area of research in this study wanted to identify if there were background factors (age, gender, race or ethnicity, and high school GPA), academic factors (college GPA, percentage of courses taken in an online format, number of remedial classes taken, and the number of credits earned during the first academic year), or financial factors (FAFSA completion, amount of financial aid awarded, and amount of financial assistance paid to the student during the first academic year) that were significant in predicting first-year student retention for community college students.

The second area of research focused on creating five classification models (random forest trees, support vector machine with the radial kernel, support vector machine with the polynomial kernel, neural networks, and logistic regression) to determine significant background, academic, and financial factors of retention. The models were compared using respective evaluation metrics and inferential statistics to see if one model outperformed the other models.

3.2 Participants

The target population for the study is community college students attending public institutions in Georgia, beginning with their freshman year. The research participants are past students who attended their respective colleges from the academic years of Fall 2017 through Fall 2019 without dual enrollment or transfer status. First-time freshmen were identified for two different cohorts with 6,834 (51.44%) students in the Fall 2017 cohort and 6,452 (48.56%) students in the Fall 2018 cohort to create a total of 13,286 students used in the data analysis.

3.3 Data Analysis

Each of the two cohorts was analyzed separately and had similar student demographic characteristics and predictor values. The cohorts were combined into one final dataset to answer the research questions. The dataset was divided with a 70% split of data in the training data set (n = 9,301) and the remaining 30% in the test data set (n = 3,985). Numeric transformations of outlier capping, Yeo-Johnson, normalization, and bagImputation were applied to the training data set before model creation. Additional interactions were created to identify possible relationships between academic and financial predictors. Each model was created using the training data set and produced model-specific evaluation metrics and variable importance. The highest ROC_AUC value from each model identified the optimal settings to create the final models using the test data set. The final models yielded the evaluation metrics and significant predictors.

4. Results

This study focused on two research questions to identify predictors of retention and the overall performance of the models. The findings of both questions can provide a starting point for community colleges wanting to create or modify retention models. The significant predictors were chosen based on the nontraditional student population of community colleges for first-year retention. The models allowed for different methods in determining significant predictors and the overall classification rates of the students. The first section of this section will summarize the results of the background factors (age, gender, race or ethnicity, and high school GPA). The second section will outline the results of the academic factors (college GPA, percentage of courses taken in an online format, number of remedial courses taken, and the number of credits earned during the first academic year). The third part will discuss the findings for the financial factors (FAFSA completion, amount of financial aid awarded, and amount of financial aid paid to the student during the first academic year). The fourth part of the section will focus on the inferential tests for the comparison of the training and test data mining models. The final four sections will discuss the individual data mining models' results; random forest, support vector machines, neural networks, and logistic regression.

4.1 Background Factors

The average age of the students in the study was approximately 18.75 years, with the median age being 18. There was an extensive range for ages 15 to 70, with the data being positively skewed. The community college population is diverse, with different types of people enrolling at different stages in their lives, explaining the considerable variation in this variable. Several categorizations of age were tried. None of these improved the normality of the data, including dividing the data set into traditional and nontraditional students. Age was not significant to retention or nonretention in the findings. The SVM training models did identify age as a predictor of nonretention, but it was only identified in the testing phase for SVM with the polynomial kernel with a value close to 0.

Female students accounted for 60% of the students in the dataset, but there was no significant finding that gender influenced retention in any of the models. Among the five models, race or ethnicity was not consistent in the retention or nonretention of community college students. The random forest model found that being a Black or African American student was significant to retention, whereas the SVM with polynomial kernel found being a Black or African American student was important to nonretention. The logistic regression model identified the variable for Hispanic or Latino students as critical to retention, but the variable importance plot did not identify any background variables, including race or ethnicity, as significant.

Both the SVM models and the logistic regression model found no significance in the high school GPA for retention or nonretention. The neural network model showed high school GAP significance for nonretention. The random forest model identified high school GPA as critical to retention after the number of credit hours and the interaction between credit hours and GPA. While the findings were not consistent throughout the models, high school GPAs should still be considered in future retention models.

4.2 Academic Factors

The results showed that GPA had significance on retention in all the models in the first three semesters for significant academic predictors. The interaction of GPA with other variables had higher importance in retention than just GPA alone. In three models, random forest, SVM with the radial kernel, and neural network, the interaction between credit hours and GPA was significant to retention ranking in the first or second position. The SVM with the polynomial kernel, neural network, and the logistic regression model indicates that the interaction between GPA and the amount of financial aid awarded was significant to retention in modeling. The interaction between GPA and the percentage of financial assistance used was substantial in the SVM with polynomial kernel and logistic regression models. The SVM with polynomial kernel model and logistic regression also identified a significant interaction between GPA and the number of remedial courses.

The percentage of courses taken in an online format had little impact on retention in this study. Online courses may not be the best indicator of retention since the median percentage of online courses taken was 0 courses. The average percentage for online courses taken for the population was 12.42%, equivalent to one 3-hour course for the entire three semesters, indicating students are not taking many online courses. While online courses may be necessary for some students, they had minimal significance in the models overall.

Remedial courses represent roughly 10% of all courses earned at community colleges, but students may not get college credit for these courses (Scott-Clayton & Rodriguez, 2015). The number of remedial courses does not appear to be a good indicator of retention since the median number was 0 and the mean number was 0.83.

The variable by itself (with no interaction) was not significant in any of the models. The interaction of the number of remedial courses with other variables was identified in four models necessary for retention. Three of the models found the interaction between GPA and the number of remedial courses critical to the retention of students. The random forest model identified the interaction between credit hours and the number of remedial classes as significant, while three models found the interaction between credit hours, GPA, and the number of remedial courses as critical to retention. The interaction between remedial classes and other academic variables (GPA and the number of credit hours) is significant to retention.

The seven community colleges in the study have been included in the Momentum Year approach, where students are encouraged to take at least 15 credit hours each semester and graduate in four years (What is a Momentum Year, 2019). Every model except for the neural network had the number of credit hours as the top predictor of student retention and supported the theory that the number of credit hours in the first year of attendance is significant. The interaction between credit hours and GPA was significant to retention, ranking in the first or second position for three models. The interaction between the number of credit hours and the different variables for financial aid was significant to retention in every model except for the random forest model.

4.3 Financial Factors

Financial factors can significantly impact students since they may rely on financial aid to persist in college (Hurford et al., 2017). If students' financial assistance and additional resources are unable to pay for their education costs, students may have to work longer hours to pay for their education or drop out (Bound et al., 2010; Scott-Clayton, 2012; Johnson & Rochkind, 2009). For students to qualify for financial aid, students must complete the FASFA application. The FAFSA completion rate for the students in this study was around 92% indicating most students fill out the form. The only model to have FASFA completion as a significant variable was the logistic model and the overall impact on nonretention was very weak. With most students completing the FASFA, the variable may have a minimal effect on the overall model.

With 72% of students receiving financial aid during the 2015-2016 academic year, financial assistance is vital to their retention (NCES, 2018a). Three different financial variables were used to measure the impact of financial aid on student retention; the amount of financial assistance awarded, the amount of financial aid paid, and the amount of financial aid used during the first academic year. The different variables for financial aid were not significant to retention or nonretention in any of the models. However, the financial variables had a more significant impact on retention when applied as part of the interactions with the number of credit hours or GPA. All the models, except for random forest, had these interactions ranked as significant to the retention of students.

4.4 Inferential Tests for Model Comparison

Five models (random forest, support vector machine with the polynomial kernel, support vector machine with the radial kernel, neural network, and logistic regression) were evaluated using ROC curves, confusion matrices, and evaluation metrics (accuracy, ROC_AUC, specificity, sensitivity, and F1-value) for the training and test data sets. A Mann-Whitney U test was used to determine if there was any difference in the models from the training and test data sets to reassure the validity of the results more than visual comparison. For the accuracy metric, there was no difference among the models between the training and test data set, $U(N\_training = .743, N\_test = .750,) = 16.00$, $p = .531$. For the F1-values metric, there was no difference among the models between the training and test data set, $U(N\_training = .701, N\_test = .708,) = 14.00$, $p = .835$, with all the models underfitting the training data. There was no difference among the models between the training and test data set, $U(N\_training = .802, N\_test = .797,) = 10.00$, $p = .676$ for the ROC_AUC scores. For the sensitivity metric, there was no difference among the models between the training and test data set, $U(N\_training = .638, N\_test = .647,) = 15.00$, $p = .676$. There was no difference among the models between the training and test data set for the specificity metric, $U(N\_training = .837, N\_test = .849,) = 16.00$, $p = .531$.

4.5 Random Forest

The random forest model was created using the randomForest engine and determined the optimal values for the final random forest model as a mtry value of 10,1781 trees, and a min_n value of 36 using the highest ROC_AUC value in the training of the model. This final model had an accuracy rate of .766, ROC_AUC value of .821, specificity value of .818, sensitivity value of .707, and F1-value of .741 (Table 1). The combined ROC curves for the test data showed the random forest model having a higher curve than the other models. The Friedman and Wilcoxon signed-rank tests confirmed that the random forest model produced the highest ROC_AUC value, sensitivity value, accuracy value, and F1 value compared to the other models.

Table 1: Train and Test Data Set Evaluation Metrics for Classification Models

Classification Model	Accuracy	F1-Values	ROC_AUC	Sensitivity	Specificity
Training Data Set					
Random Forest	.761	.734	.823	.697	.818
SVM Polynomial	.741	.684	.796	.591	.876
SVM Radial	.734	.675	.795	.585	.868
Neural Networks	.750	.717	.807	.670	.822
Logistic Regression	.743	.701	.802	.638	.837
Test Data Set					
Random Forest	.766	.741	.821	.707	.818
SVM Polynomial	.750	.695	.797	.602	.882
SVM Radial	.738	.681	.797	.592	.868
Neural Networks	.755	.715	.791	.650	.849
Logistic Regression	.748	.708	.802	.647	.839

*Not*e. ROC_AUC is ROC Area Under the Curve

4.6 Support Vector Machines

The SVM model with the polynomial kernel was created using the kernlab engine with the polynomial kernel fitting the support vector classifier in a higher-dimensional space, allowing for more flexibility in the decision boundary (James et al., 2013). The tuning parameter for this SVM model was a cost value of 1.66 and was found by using the highest ROC_AUC value (Attewell & Monaghan, 2015). This final model had an accuracy rate of .750, a ROC_AUC value of .797, a specificity value of .882, a sensitivity value of .602, and an F1 value of .695 (Table 1). The SVM with polynomial kernel had the highest value for the specificity among the models and was supported by the Friedman and Wilcoxon signed-rank tests.

The SVM model with the radial kernel was created using the kernlab engine with the tuning parameters of cost and sigma (Attewell & Monaghan, 2015). The final SVM model with a radial kernel had a cost of 0.024 and a sigma of 0.030 from the highest ROC_AUC value in the training phase. Since both tuning values are low, the model may have under-fitted the data due to its larger, inflexible margins. This final model had an accuracy rate of .738, a ROC_AUC value of .797, a specificity value of .868, a sensitivity value of .592, and an F1 value of .681(Table 1). While the SVM with radial kernel had the most extensive range for accuracy, sensitivity, and specificity, it did not have any significant evaluation metrics in the inferential tests.

4.7 Neural Network

The neural network models were created using the Keras engine with three different parameters: the hidden unit of 4, the penalty was 0.540, and the epochs were 811, which were identified using the highest ROC_AUC value (Attewell & Monaghan, 2015). This final model had an accuracy rate of .755, a ROC_AUC value of .791, a specificity value of .849, a sensitivity value of .650, and an F1 value of .715 (Table 1). The neural network evaluation metrics constantly ranked in third or fourth place in the models.

4.8 Logistic Regression

The logistic regression model was created using the glm engine and did not require tuning throughout the modeling process. The Hosmer and Lemeshow goodness of fit test ($\chi^2(8) = 30.48$, $p < 0.001$) indicated that the logistic regression model does not fit the data well. Nagelkerke's pseudo R squared value of 0.375 and McFadden's pseudo R squared value of 0.239 describe the variation (37.5% and 23.9%) of the academic, background, and financial factors that contribute to the retention status of the students in this model. This final model had an accuracy rate of .748, a ROC_AUC value of .802, a specificity value of .839, a sensitivity value of .647, and an F1 value of .708 (Table 1).

5. Discussion

The study focused on identifying significant factors of first-year student retention in community colleges and if one model could outperform the other models based on evaluation metrics. Academic and financial factors play an essential role in retaining community college students during their first year.

Bean and Metzner (1985) theorized that community college students enroll in their colleges for academic reasons, and these academic interactions serve as their primary integration method. During the first year, the number of credit hours was the most significant variable in any of the models in predicting first-year retention. The workload associated with more credit hours may influence the mindset of students to devote more time and resources to complete the work. Conversely, students who take few credit hours may have external factors limiting the number of classes they can take and the time they can commit to completing coursework.

Three models (random forest, SVM with radial kernel, and neural network) identified the importance of the interaction between the number of credit hours and the first three semesters' GPA. Students who pass their courses have higher GPAs than students who fail courses and accumulate more credit hours. Interactions between credit hours, GPA, and the percentage of online courses or the number of remedial classes were identified as slightly significant to retention. Four models (SVM with the polynomial kernel, SVM with the radial kernel, neural network, and logistic regression) identified the interactions between GPA and financial variables (amount of financial aid awarded, financial aid amount paid, and percentage of financial assistance used) as significant to retention. The logistic model found that GPA and the number of remedial courses were essential to retention. All the academic and financial variables play a role in retaining first-year community college students. The interactions between credit hours and other academic and financial variables were also important to retention signifying that academic and financial factors can impact students' number of credit hours.

There were no consistent variables among the five models predicting first-year students' nonretention. Two models, random forest and logistic regression did not identify any significant variables for nonretention. The SVM with polynomial kernel model identified the following predictors as significant to nonretention: the interaction between the number of credit hours and remedial courses, being a Black or African American student, the interaction between the number of credit hours, percentage of online courses, and remedial courses, the amount of financial aid awarded, and the percentage of financial assistance used. The percentage of online courses and the interaction between GPA and the number of remedial classes were significant to nonretention in the SVM with the radial kernel. The neural network model indicated that high school GPA was significant to nonretention. The absence of significant factors for nonretention could suggest that important variables were not included in these models, such as environmental factors like employment and family obligations.

The different classification models were compared to see if any models would have a higher classifier performance based on the different evaluation metrics and inferential tests. The random forest model performed better than the other models in accuracy, F1-values, ROC_AUC, and sensitivity. Visual inspection of the grouped ROC curves showed that the random forest could be the optimal model for first-year retention as the highest accurate classifier. The SVM with polynomial kernel had the highest value for the specificity of all the models identifying the students who were correctly classified as nonretained.

The research by Fernández-Delgado et al. (2014) aligned with the five classifier models' results for retention, with the random forest model having the highest accuracy value and being significantly higher than the other models. Additionally, Fernández-Delgado et al. (2014) ranked the logistic regression model lower than the SVM and neural network models in overall accuracy, yet this was different from the findings in this study. While the neural network model had a higher accuracy metric for predicting students' retention than both the logistic regression and SVM models, the logistic regression model had a higher accuracy metric than both SVM models. Therefore, the random forest model is the ideal model for predicting community college first-year retention.

6. Conclusion

With community colleges providing educational access to roughly half of all undergraduate students in the United States of America, retention models need to be created to serve their populations better (Horn, Nevill & Griffith, 2006; Mullin, 2012; NCES, 2013; NCES 2018b). Among the seven community colleges, the number of credit hours was consistently the most critical variable in retention. The interactions between the number of credit hours, GPA, and financial aid variables were significant in student retention in their first year. Additionally, the interaction between GPA, financial aid variables, and the number of remedial hours was crucial for first-year retention. Combining these variables shows that academic and financial variables are interconnected and may need multiple messages to reach students. Specific marketing campaigns about the benefits of reaching credit hour milestones, GPA requirements, and FASFA deadlines for financial aid may help impact these variables. Schools with similar populations could work together to share ideas and resources to help gain a broader reach.

No consistent variables among the retention models predicted students' nonretention in the first year of their college career. Many background predictors (age, gender, race, or ethnicity) were not significant in predicting retained or nonretained students. Even though these variables were not significant in these sector-based models, individual community colleges may want to include these variables to understand students' backgrounds better. While FAFSA's completion had no impact on the model, students must complete it for financial aid and should be included in any retention message to understand its importance.

The comparison of the retention models found that the random forest model had the best overall performance for accurately classifying the nonretained and retained students together and the retained students individually. The SVM with a polynomial kernel had the highest value for specificity, which identifies the nonretained students. Logistic regression, a commonly used model for retention analysis, did not perform as well as the other models because of the skewed data and correlated variables. The support vector machine with the radial kernel and neural network evaluation metrics never performed better than the random forest and SVM with the polynomial kernel. While the random forest model may not be commonly used in retention models, it may be an ideal model for identifying the overall retention of community college students since it can handle different data (binary, categorical, ordinal) and not be affected by outliers. The study shows that using more than one model allows for validating the variable's importance and potential patterns. Free statistical software makes model creation affordable with no-cost training and information to learn software such as R or Python. With higher education funding tied to student retention numbers, specialized retention models can help institutions and systems identify and intervene with students at risk of not returning and help stabilize their funding.

7. References

Aljohani, O. (2016). A comprehensive review of the major studies and theoretical models of student retention in higher education. *Higher Education Studies, 6*(2), 1-18.

Astin, A. W., & Oseguera, L. (2005). Pre-college and institutional influences on degree attainment. *College student retention: Formula for student success*, (pp. 245-276). Plymouth: Rowman & Littlefield Publishers.

Attewell, P., & Monaghan, D., (2015). Data mining for the social sciences: An introduction. Oakland: University of California Press.

Bean, J. P., & Metzner, B. S. (1985). A conceptual model of nontraditional undergraduate student attrition. Review of Educational Research, 55(4), 485-540.

Berger, J., Ramirez, G. B., & Lyon, S. (2012). Past to present: A historical look at retention. *College student retention: Formula for student success*, (pp. 7-34). Plymouth: Rowman & Littlefield Publishers.

Bharati, M., & Ramageri, M. (2010). Data mining techniques and applications. *Indian Journal of Computer Science and Engineering, 1*(4), 301-305.

Breiman L. (1999). Random Forests—random features. Statistics Department, University of California, Berkeley, Technical Report 567, September 1999.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees.* Boca Raton: Chapman & Hall/CRC.

Cabrera, A. F. (1994). Logistic regression analysis in higher education: An applied perspective. *Higher education: Handbook of theory and research*, 10, 225-256.

Cabrera, A. F., Nora, A., & Castaneda, M. B. (1993). College persistence: Structural equations modeling test of an integrated model of student retention. *The Journal of Higher Education, 64*(2), 123-139.

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 & Practice*, *13*(1), 17–35

Caruana, R. & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd International Conference on Machine Learning*, 161-168. Retrieved from http://lacam.di.uniba.it:8000/ people/courses /IA/IA0809/sistemi/caruana .icml06.pdf

Chatterjee, A., Marachi, C., Natekar, S., Rai, C., & Yeung, F. (2018). Using logistic regression model to identify student characteristics to tailor graduation initiatives. *College Student Journal, 52*(3), 352-360.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems, 49*(4), 498-506.

Dissanayake, H., Robinson, D., & Al-Azzam, O. (2016). Predictive modeling for student retention at St. Cloud state university. *Proceedings of the International Conference on Data Mining*, 215-221. Retrieved from https://search.proquest.com /docview/ 1806428521?pq-origsite=gscholar

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems?. *The Journal of Machine Learning Research, 15*(1), 3133-3181.

Goga, M., Kuyoro, S., & Goga, N. (2015). A recommender for improving the student academic performance. *Procedia-Social and Behavioral Sciences, 180*, 1481-1488.

González, J. M. B., & DesJardins, S. L. (2002). Artificial neural networks: A new approach to predicting application behavior. *Research in Higher Education, 43*(2), 235-258.

Gunu, E. A., Lee, C., Gyasi, W. K., & Roe, R. M. (2017). Modern predictive models for modeling the college graduation rates. Proceedings of the 15th International Conference on Software Engineering Research, Management and Applications, 39-45. Retrieved from https://ieeexplore.ieee.org/abstract/document/7965705

Han, J., Pei, J., & Kamber, M. (2011). Data mining: Concepts and techniques. Waltham: Elsevier.

Hardman, J., Paucar-Caceres, A., and Fielding, A. (2013). Predicting students' progression in higher education by using the random forest algorithm. *Systems Research and Behavioral Science, 30*(2), 194-203.

He, L., Levine, R. A., Fan, J., Beemer, J., & Stronach, J. (2018). Random forest as a predictive analytics alternative to regression in institutional research. *Practical Assessment, Research & Evaluation, 23*(1), 1-16.

Herzog, S. (2006). Estimating student retention and degree- completion time: Decision trees and neural networks vis- à- vis regression. *New Directions for Institutional Research, 131*, 17-33.

Herzog, S. (2018). Financial aid and college persistence: Do student loans help or hurt?. *Research in Higher Education, 59*(3), 273-301.

Horn, L., Nevill, S., & Griffith, J. (2006). Profile of undergraduates in US postsecondary education institutions, 2003-04: With a special analysis of community college students. *Statistical Analysis Report*. NCES 2006-184. National Center for Education Statistics.

Hurford, D. P., Ivy, W. A., Winters, B., & Eckstein, H. (2017). Examination of the variables that predict freshman retention. *The Midwest Quarterly, 3*, 302.

Hutt, S., Gardener, M., Kamentz, D., Duckworth, A. L., & D'Mello, S. K. (2018). Prospectively predicting 4-year college graduation from student applications. *Proceedings of the 8th International Conference on Learning Analytics and Knowledge,* 280-289. Retrieved from https://dl.acm.org/doi/10.1145 /3170358.3170395

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

Johnson, D. R., Wasserman, T. H., Yildirim, N., & Yonai, B. A. (2014). Examining the effects of stress and campus climate on the persistence of students of color and white students: An application of Bean and Eaton's psychological model of retention. *Research in Higher Education, 55*(1), 75-100.

Juszkiewicz, J. (2020). Trends in community college enrollment and completion data, 2020. *American Association of Community Colleges.*

Kardan, A. A., Sadeghi, H., Ghidary, S. S., & Sani, M. R. F. (2013). Prediction of student course selection in online higher education institutes using neural network. *Computers & Education, 65*, 1-11.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling.* New York: Springer.

Langan, A. M., Harris, W. E., Barrett, N., Hamshire, C., & Wibberley, C. (2018). Benchmarking factor selection and sensitivity: a case study with nursing courses. *Studies in Higher Education, 43*(9), 1586-1596.

Lauría, E. J., Baron, J. D., Devireddy, M., Sundararaju, V., & Jayaprakash, S. M. (2012). Mining academic data to improve college student retention: An open source perspective. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* pp. 139-142. Retrieved from http://cs.colby.edu/courses /S16/cs251-labs/final-lauria-studentRetention-LAK2012.pdf

Luan, J. (2002, June). *Data Mining and Knowledge Management in Higher Education Applications.* Paper presented at the Annual Forum for the Association for Institutional Research, Toronto, Ontario, Canada. Retrieved from http://eric.ed. gov/ERICWebPortal/detail?accno=ED474143

Metzner, B. S., & Bean, J. P. (1987). The estimation of a conceptual model of nontraditional undergraduate student attrition. *Research in Higher Education, 27*(1), 15-38.

Mullin, C. M. (2012). Why Access Matters: The Community College Student Body. AACC Policy Brief 2012-01PBL. *American Association of Community Colleges (NJ1)*

National Center for Educational Statistics [NCES] (2013). 2011–12 National Postsecondary Student Aid Study Retrieved from https://nces.ed.gov/pubs2013/ 2013165.pdf

National Center for Educational Statistics [NCES] (2018b). Enrollment and Employees in Postsecondary Institutions Fall 2016; and Financial Statistics and Academic Libraries, Fiscal Year 2016. Retrieved from https://nces.ed.gov/ pubs2018/2018002.pdf

Pascarella, E. T., & Terenzini, P. T. (2005). *How College Affects Students: A Third Decade of Research. Volume 2*. Indianapolis: Jossey-Bass.

Peng, C. Y. J., So, T. S. H., Stage, F. K., & John, E. P. S. (2002). The use and interpretation of logistic regression in higher education journals: 1988–1999. Research in Higher Education, 43(3), 259-293.

Pyke, S. W., & Sheridan, P. M. (1993). Logistic regression analysis of graduate student retention. *Canadian Journal of Higher Education, 23*(2), 44-64.

Scott-Clayton, J., & Rodriguez, O. (2015). Development, discouragement, or diversion? New evidence on the effects of college remediation policy. *Education Finance and Policy, 10*(1), 4-45.

Shapiro, D., Dundar, A., Huie, F., Wakhungu, P. K., Yuan, X., Nathan, A., & Hwang, Y. (2017). Tracking transfer: Measures of effectiveness in helping community college students to complete bachelor's degrees. *Signature Report,* (13).

What is a Momentum Year? (2019). Retrieved from https://completega.org/what-momentum-year.

Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. Neural computation, 8(7), 1341-1390.

Zhang, Y., Oussena, S., Clark, T., & Kim, H. (2010). Use Data Mining to Improve Student Retention in Higher Education-A Case Study. *ICEIS, 1*, 190-197.