The Use of Semester Course Data for Machine Learning Prediction of College Dropout Rates

Viktor Kiss
Metropolitan State University of Denver

Edgar Maldonado
Metropolitan State University of Denver

Mark Segall
Metropolitan State University of Denver

Predicting those at-risk of dropping out allows schools to assist students before it happens. Machine learning (ML) techniques can predict the likelihood of students completing a course, enrolling in future semesters, or graduating from college. This study compares four ML techniques to predict dropout rates using a student's demographic information and performance in individual courses over all semesters enrolled. Using ten semester models the logistic regression method had the best accuracy of 84.8% versus decision trees (82.2%), neural networks (80.8%), and support vector machines (72.5%). The semester course performance data is a useful input for predicting dropout rates.

Keywords: retention, machine learning, logistic regression, predictive analysis, semester-wide analysis

INTRODUCTION

Retention in higher education is an everlasting battle [NCES, Barbera et al. 2020], just as churn is for companies with subscriptions. Students come from a variety of personal and professional backgrounds. Understanding when a student could be in danger of dropping out is a quintessential problem for colleges. There is not just a monetary incentive, but also a fulfillment of obligation in providing continuing education for those who choose to attend. Identifying students who are in danger of dropping out and taking steps towards their retention could be key in this battle.

In recent years, machine learning methods have become more easily applicable due to wider availability of data and sufficient computing power, which has enabled higher education institutions to develop their own models. Due to the importance of the area, several studies aimed at student retention have been published at dozens of institutions, a lot of them taking advantage of the recent boom in machine learning and relying on its learning methods [Cardona and Cudney, 2020]. What these institutions are predicting with the help of their models can be broadly categorized into three types of studies.

Studies that predict whether a student will:

A.) register for a following semester (i.e. the next semester or the next academic year), B.) graduate from the institutions sometime in the future,

C.) finish a specific class within a semester.

These types of studies have significant overlap, but each operates with different goals and are not interchangeable. The main goal of predicting whether students register for a close-by semester is to make sure that the registration numbers at a given institution do not drop significantly. Predicting graduation focuses more on whether an individual student is on the right path towards getting the degree. The latter also accounts for patterns where a student does not register for every semester, but graduates. Depending on the institution one method could be more desirable than the other. For example, several European universities do not allow for skipping a semester, or only allow it under exceptional circumstances. In these cases, type A might be more desirable. For institutions where students have a lot of flexibility on which classes they take and when, it is common to create an academic plan, where someone does not take any classes for given semesters. In these cases, type B might be more useful.

Type C is mainly for making sure that a student completes a class, which directly contributes to their success at an institution. This type of study works on a short time scale and is used mostly for difficult required courses, where failing could result in students not being able to finish their studies.

Review of Machine Learning Methods Used

The area of Machine Learning methods suited for retention studies are supervised machine learning methods. Supervised machine learning methods are best suited for the kinds of analyses where the algorithm aims to learn a function that maps inputs (in this study: information about a student) to an output (whether a student graduates). Predicting retention is a classification problem (i.e.: It aims to classify students into distinct groups of dropped-out vs. graduated), where there is a mixture of continuous and categorical variables that are used as predictors. The algorithms chosen for this study will handle this type of problem.

According to the literature, out of the handful of algorithms that do meet these specifications, some are more popular than others when it comes to higher education retention studies, namely Neural Networks, Decision Trees, Logistic Regression and Support Vector Machines. This following section gives a brief technical overview of these methods with their pros and cons.

A Neural Network is a current state of the art method used for prediction. A network is made up of a set of decision-making units, neurons, which each contribute individually to a final prediction based on input values. The network can adjust the contribution of each neuron gradually, based on how it performs on data, and can result in highly accurate models after numerous iterations. These networks need numerical data to operate but can handle numerical representations (e.g.: dummy variables) of categorical variables. They output the classification category which has the highest probability but can also output the probability itself which can be useful if one would like to order students by their likelihood of dropping out.

A Decision Tree model is a supervised learning method, where the algorithm creates a system of sequential decisions based on input variables and eventually arrives at a classification designation for each input value. These sequential decisions are organized in a tree-format, hence the name of the method. It can handle both continuous and categorical variables. By default, Decision Trees output classification categories, but no probabilities. Some advanced methods based on Decision Trees can output probabilities as well, such as Random Forests, where several Decision Trees are developed for the same problem, and the probability of dropping out would be the proportion of these trees that predicted a student to drop-out. Another advanced version is the Gradient Boosted Decision Trees method, where a sequence of small, single decision trees is developed, and each tree aims to predict the part of the target variable the previous tree did not account for.

Logistic Regression is a classification method which uses a linear combination of the input variables and turns their results into probabilities with the help of a non-linear function. The resulting outputs are probabilities of belonging to a given class. The method can handle both continuous and categorical variables. The execution of the algorithm is relatively fast even on large datasets, and it is possible to determine relative importance between input variables. In our list of methods, Logistic Regression is the oldest and most basic method, but remains highly effective in several areas.

Support Vector Machines are a relatively new method used primarily for classifications problems in machine learning. The method they employ aims to find the best separating line between different classes based on their characteristics. It can separate very complex patterns of data using non-linear transformations of the input space. Similarly, to Neural Networks, it needs numerical input, but is capable of handling numerical representations of categorical data. This model outputs classification membership, but no probabilities, although the application of certain techniques can generate probabilities.

Review of Evaluation Metrics

Several metrics exist for Classification tasks in Machine Learning. The ones most prominently used in the literature for student retention are briefly discussed in the following section. They are discussed from the point of view of what research questions they answer. Please note that 'positive cases' refers to the class of interest.

- True Positive Rate (TPR, also termed Sensitivity or Recall) answers the question: What proportion of positive cases did the model identify?
- False Positive Rate (FPR) answers the question: For all cases predicted as positive cases, what proportion of them are false positives?
- Precision answers the question of when a model predicts a positive case, relatively how often is it correct?
- Accuracy answers the question: How often are the predictions correct?

The following section will give a brief overview of the published articles with emphasis on the general methodology (i.e.: machine learning algorithm used) and the end goal of the study (retention/Type A or graduation/Type B).

Literature Review

Fall-to-fall retention (part of a type A study) is a very dominant field in retention studies. Delen used a handful of Machine Learning algorithms in two studies (2010 & 2011) to try to predict whether freshman students will register for classes on the fall of their sophomore year. The methods used (Neural Networks, Decision Trees, Support Vector Machines and Logistic Regression) all yielded similar results in the 2010 study (around the accuracy of 86%-87% using their strongest setup), but significantly differed in the 2011 study, where Neural Networks outperformed other methods (81% accuracy compared to 74% - 78% accuracy). In a recent study, Patacsil (2020) used different versions and ensemble methods of decision trees to predict fall-to-fall retention for freshman and ultimately achieved overall accuracy rates in the range of 69% - 70%. Dissanayake et al. (2016) had a similar goal but used a range of fundamentally different algorithms (KNN, Decision Trees, Binomial GLMs and Neural Networks) to obtain results. Their best models performed at an overall accuracy of 83% - 85%. Bogard et al. developed several models (Neural Networks, Decision Trees, Ensemble models, etc.) to predict fall-to-fall retention and subsequently reached 79% overall accuracy.

Oztekin (2016) used several machine learning algorithms to predict degree completion (type B) for freshman students, and subsequently found that in his study Support Vector machines (77% accuracy) outperformed Decision Trees, Neural Networks and Logistic Regression (74%, 72% and 50% accuracy, respectively). Berka and Marek (2021) also had degree completion as a target variable in their recent paper, where they developed several models incorporating a wide range of information. They predicted separately for students finishing and not-finishing their studies, and achieved results in the 75% - 92% range, depending on the Machine learning algorithm used, and the number of semesters taken into consideration. Cardona and Cudney (2020) used Support Vector Machines to predict degree completion for STEM major students in their study and achieved about 78% accuracy rates when their model was used on separate test data. Cardona et al (2019) also used Decision Trees for similar purposes (predicting degree completion), with the goal of not only prediction accuracy (80%), but also to be able to identify which variables contribute most to degree completion.

Several studies focused on type C outcomes, predicting whether students will finish a given class within a semester. Howard et al (2018) developed several models (Random Forests, K Nearest Neighbors, Support Vector Machines…) and ultimately decided to use a type of Decision Tree (Bayesian Additive Regressive Trees) to predict final mark for students based on their in-class performance. They predicted final grades, with mean absolute error to 6.5 percentage points, after 6 weeks of classes. Tsao et al. (2017) used Decision Trees to try to predict whether a student will pass the class and reached accuracy scores in the low 70% depending on the amount of data used for prediction. Kondo et al (2017) developed several models (Logistic Regression, Support Vector Machines and Random Forests) and evaluated their performance on a weekly basis during a semester. Their Random Forest model yielded the most stable performance, being able to detect 40% of at-risk students by the end of the third week. Costa et al (2017) tried several methods (Support Vector Machines, Decision Trees, Neural Networks, and the Naïve Bayes method) and found that a Support Vector Machine was the most successful methods for predicting student success in a class.

Description of Institution

The higher education institution studied in this paper is no exception when it comes to difficulties retaining students. The institution analyzed in this research is a university located in a large urban center in the Western United States. The four-year undergraduate programs at this university provide education to around 20,000 students, and it is considered a modified open enrollment institution. Approximately 38% of students earn their degree at the institution after taking at least one class. Although this calculation includes students who take classes without ultimately seeking a degree, making sure that the institution creates an environment where more students can be helped to earn a degree is essential. The University's student population is not a traditional one: The proportions of older (average age of 25) and first-generation students is over 51%, higher than most other higher education institutions. This results in different patterns of how students take classes, especially regarding course load. A full-course load is not necessarily feasible for a significant proportion of students, due to time and financial constraints.

The study of students' retention in non-traditional institutions may need different approaches that those used for student the problem on traditional universities. The goal of this study is to develop a model that effectively predicts whether a student will not finish their studies, by using a timeless analysis of students and their individual academic load by semester (i.e.: dropping out before graduating).

The following section gives a description of the dataset used to build the model.

Approach Used in This Study

The current study will create a model that tries to predict whether students will graduate (type B study). The primary reason behind it is that semester by semester retention studies already exist at the institution and creating a different type of model ultimately will complement the currently ran models to gain more insight into student retention. Also, due to the non-traditional student body, it is possible that several unique paths to degree completions exist, where several of them include skipping semesters, something that type A studies generally do not consider.

METHODOLOGY

The problem at hand is a binary classification problem, where the target variable has two values: Dropped-out or Graduated. Seven predictor variables were used (please see Table 1. for details). The dataset is divided for the study: 70% training data, 15% validation data and 15% testing data. The models` hyperparameters were adjusted based on their performance on the validation data. The results reported for each model are from a single run on the testing data. Table 1 includes all the variables used in this research.

Our goal is to identify students who are in danger of eventually dropping out and not finishing their program, and to minimize false positive results, which would drive up costs and result in less focus on actual in-danger students. In line with this goal, the results will be evaluated based on two metrics: True Positive Rate and False Positive Rate. Ultimately, an ideal model will have a high TPR and a low FPR.

TABLE 1
VARIABLES USED IN ALL MODELS

Type of Variable	Type of Data	Variable	Description
Predictor	Demographic	Gender	Binary
		Race	Categorical
		Age	Numerical
		Major	Categorical
		Student Type	Categorical
	Performance	Transferred GPA	Numerical
		Grade at Every Course	Categorical
Target	Outcome	Retention	Binary

Dataset

The acquired dataset has 29,282 individual students, enrolled in the College of Business at the institution, with demographic data and their performance metrics at the institution. The date of the first active1 semester ranges from Fall 2007 to Fall 2019. There are 3 different semester types: Fall, Spring and Summer. Although students tend to take less courses in the summer semester, it was included in the dataset with the same considerations as Fall and Spring semesters.

Preprocessing

Since completing a full degree ideally takes anywhere from 2 to 4 years (depending on potential transfer credits, associate degree transfer agreements, and other circumstances), the training of the model only considers students with their first active semester before or on Fall 2015, so everyone in the study has had a fair chance of obtaining their degree. The number of unique courses students took in the dataset is 4,651, most of them were taken only by a negligible fraction of the students. As a result, courses that were taken by less than 100 students in the dataset are dropped, keeping 276 unique courses in the study. The resulting dataset has 21,079 students.

Demographic Description of Data

The following plots describe the basic demographics (age, gender, and race) of our dataset.

FIGURE 1
AGE DISTRIBUTION OF STUDENTS

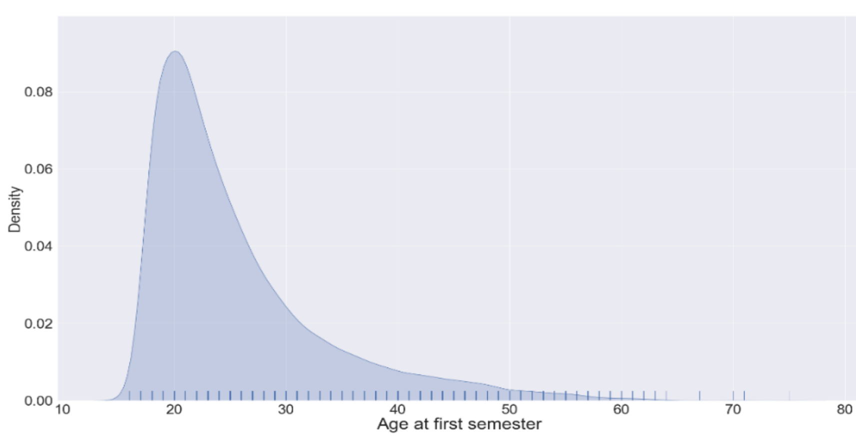
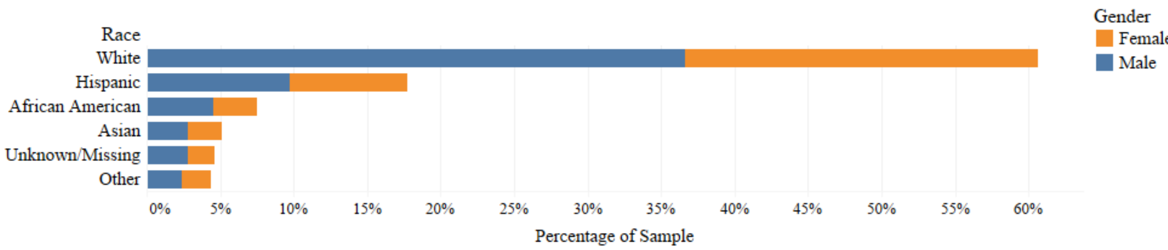


FIGURE 2
GENDER AND RACE DISTRIBUTION OF STUDENTS



The age distribution is skewed to the right (as expected), with the mean and median age being 25.64 and 23 years respectively). Almost 20% of the students started their first semester after their 30th birthday. There are significantly more male students (59% versus 41%). Approximately 30% of the students are of Hispanic, African American, or Asian descent.

Structure of Data Used for Model Building

The study aims to capture as much information as possible from the dataset obtained from the institution. The dataset was transformed in a way that each student has information for every semester about every one of the 276 unique courses, regardless of whether they took it or not. That way the model captures information on not just the performance of students on a given course, but also differentiates between students who took or did not take a course in a given semester. The semesters are considered as consecutive timeless periods. The timing component is not considered, but instead the order of the semester for a given student is important. For example, the first semester of a student could be Fall 2007 but for another student was Fall 2005; our representation considers both as "Semester 1". This representation of the data is an essential part of the study, and it was first introduced as "semester-wise" representation by Amari (2015). That results in the following, very sparse matrix.

TABLE 2
SEMESTER-WISE REPRESENTATION OF STUDENT DATA

	Demographic data					Semester 1				Semester 2					Semester 40				
	Gender	Race	Age	⋮	Major	Course_1	Course_2	⋮	Course_276	Course_1	Course_2	⋮	Course_276	⋮	Course_1	Course_2	⋮	Course_276	Dropped out
Student_1	F	A	25	...	MKT	C	-	...	-	-	B	...	-	...	-	-	...	-	0
Student_2	M	W	28	...	CIS	-	A	...	-	-	-	...	B	...	A	-	...	-	1
Student_3	F	H	19	...	MGT	-	-	...	-	D	-	...	-	...	-	-	...	-	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Student_29281	F	H	34	...	CIS	B	-	-	F	-	-	...	A	...	-	D	...	-	1
Student_29282	M	A	21	...	MKT	-	A	-	-	-	-	...	-	...	-	-	...	-	0

Since evaluation of in-danger students happens on a continual basis, a model is developed for each of the semesters in the dataset, meaning that there is a separate model only containing data from the first semesters of the students, one for up to the second semesters, and so on. This way the model is applicable in practice, where a practitioner can look at active students who completed their first n semesters, and identify which students are in danger of dropping out. The first active semester is a baseline for all students independent of the calendar year they started.

Software Used for Analysis

The analysis was done using Python, a high-level, general purpose programming language. It`s main machine learning library, Scikit-learn, which specializes in running the previously described algorithms, was used extensively for this project.

RESULTS

Benchmark Scores

In order to create a benchmark we can compare our model against the null error rate, which was 42.4%, meaning we would be wrong 42.4% of the times if we predicted the majority class (dropped-out, 57.6%).

Comparison of Candidate Algorithms

As discussed in the introduction section, the most popular algorithms for retention studies are Neural Networks, Decision Trees, Logistic Regression and Support Vector Machines. After running several different algorithms (Logistic Regression, Neural Networks, Decision trees (Random Forests & Gradient Boosted Decision Trees), Support Vector Machines), the study settled on the best performing algorithm: Logistic Regression. Although Gradient Boosted Decision Trees had comparable performance the Logistic Regression was chosen because it had the best overall accuracy (See Table 3) and is easy to interpret, with its predictions being given in percentage form (0 – 1). That enables the decision makers to focus on a given tier of students (e.g.: Students over 70% chance of dropping out) at a time. It is less complicated than some of the other algorithms, with no hyperparameters to set to adjust the model.

TABLE 3
VALIDATION ACCURACY OF MODELS FOR NINE SEMESTERS

Validation Accuracy				
# of semesters	Neural Networks	Decision Trees	Logistic Regression	Support Vector Machines
1	67.5%	70.0%	72.5%	64.4%
2	70.5%	72.8%	75.3%	67.4%
3	70.9%	74.7%	77.3%	71.1%
4	74.9%	77.6%	80.0%	72.7%
5	74.8%	78.1%	80.9%	73.9%
6	76.3%	78.6%	81.8%	73.4%
7	79.9%	80.6%	83.1%	73.7%
8	78.3%	81.9%	84.0%	73.4%
9	80.8%	82.2%	84.8%	72.5%

Logistic Regression Results

The summary of the results of the Logistic Regression Classifier can be seen in Table 4. In order to make sure our models do not suffer from high variance, each of the models were rerun 50 times with different randomly selected rows split between the training and validation data. Variance of models trained on different samples from the same population are an ever-present problem in machine learning. Models with a high variance are less dependable since their predictions highly depend on the idiosyncrasies of the training data. It can be seen that the benchmark score was surpassed by a large margin, implying that the model in-fact adds value to prediction.

TABLE 4
SUMMARY OF LOGISTIC REGRESSION MODEL ACCURACY MEASURES

LOGISTIC REGRESSION						
Stability checked for n = 50			Single test Run			
Number of semesters	Training Accuracy	Standard Deviation	Testing Accuracy	Sensitivity	False Positive Rate	AUC
1	72.3%	0.17%	71.7%	85.2%	46.8%	77.1%
2	75.7%	0.16%	74.5%	86.2%	41.7%	81.2%
3	78.1%	0.16%	76.9%	87.6%	37.9%	82.8%
4	81.2%	0.18%	79.0%	89.6%	35.7%	86.0%
5	82.6%	0.14%	79.4%	89.9%	35.1%	86.6%
6	83.8%	0.13%	80.9%	91.2%	33.4%	87.6%
7	85.3%	0.14%	82.4%	92.8%	31.9%	88.6%
8	86.4%	0.27%	83.7%	93.8%	30.2%	89.1%
9	87.2%	0.14%	84.5%	93.8%	28.5%	89.4%
10	88.1%	0.12%	85.1%	94.7%	28.1%	90.0%
11	88.8%	0.12%	85.6%	94.7%	26.9%	90.3%
12	89.1%	0.12%	85.9%	94.7%	26.2%	90.8%

For every semester, the training accuracy is reported as an average of 50 different runs with different training / validation splits to check stability. The standard deviation of the accuracies is under 0.2 percentage points for each number of semesters, indicating that the results are stable and not prone to variability because of sampling. The Accuracy, Sensitivity, and False Positive rates are reported for a single run on the testing set. The results indicate that the model developed already catches over 85% of students who ultimately won`t graduate after their first semester, and does so at a relatively low cost, at under 50% false positive rate. Although the 50% might seem high compared to some other applications of classification methods, one has to remember that in academic settings, the first semester of a student has less information, since a lot of students are just dipping their toes into higher education. The results of the model gradually get better with time (i.e.: there is data from more semesters), as expected. By the end of the 4th semester, the sensitivity is almost at 90% with the false positives around 35%.

Logistic Regression - ROC AUC

Since the Logistic Regression model predicts probabilities, it is possible to construct a Receiving Operating Characteristics (ROC) curve and calculate the AUC (Area Under the Curve) as a performance metric. ROC-AUC is an evaluation metric that measures the performance of a classification task with various thresholds used as a cut-off for positive and negative classification. At each threshold from 0 to 1, it identifies and plots True Positive Rate and the False positive rate, ultimately ending up in a curve. The better the model the higher it deviates from the baseline prediction which is random guessing. The performance is measured with the area under the curve, the closer to 1, the better the model.

The AUC scores can be seen in the results table. For demonstration purposes please see the ROC curve for the Logistic Regression model for 4 and 10 semesters (See Figures 3 and 4). Decision makers can use each of the 12 ROC curves - for the different models for each of the 12 semesters – to identify which threshold could be the most useful for their purposes.

FIGURE 3
LOGISTIC MODEL: FOUR SEMESTER ROC CURVE

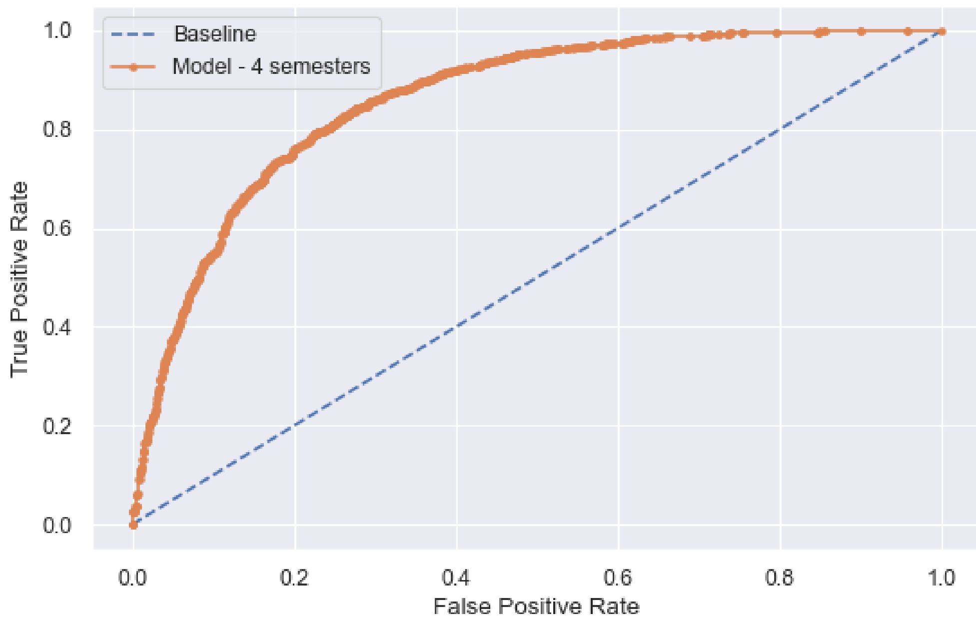
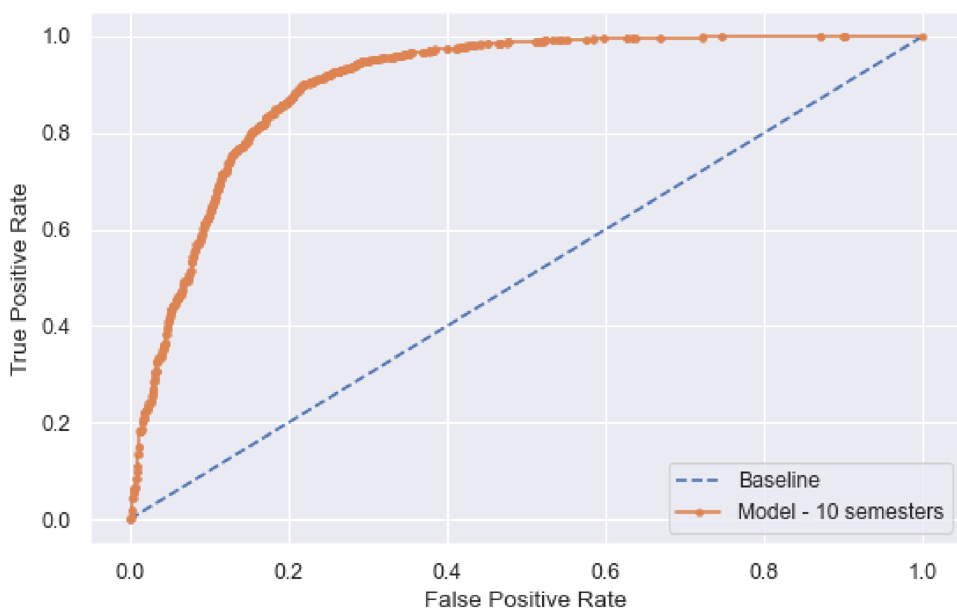


FIGURE 4
LOGISTIC MODEL: TEN SEMESTER ROC CURVE



In the case of logistic regression, a practitioner aiming to identify students who are in danger of dropping-out can preprocess the data in the same way as the study (code will be made available), and run a logistic regression, which will give the probabilities of dropping-out for each student.

CONCLUSION AND FUTURE RESEARCH

This research used a timeless approach to modeling student retention for non-traditional institutions. This method yields comparable results of prior research on traditional institutions. Therefore, a semester-wise model of academic load adds a better understanding to prediction of retention in non-traditional students. The presented model may be used by advisors in non-traditional academic institutions to intervene and provide support to the students.

This study serves as a baseline study for retention at an urban institution serving a large population of non-traditional students. It was specifically developed to follow current trends in retention studies and show that similar results can be obtained compared to peer traditional institutions. The next steps will include the effort to develop a model based on more sophisticated algorithms, such as deep learning, and ultimately surpass the results of this current paper. Because this study looked at students in a business college additional models can be developed for other colleges in the institution. Also, the combination of the currently used year-on-year retention model and our graduation focused model could yield beneficial results for the institution.

REFERENCES

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

Ameri, S. (2015). *Survival analysis approach for early prediction of student dropout*. Wayne State University Theses. 443. Retrieved from https://digitalcommons.wayne.edu/oa_theses/443

Barbera, S.A., Berkshire, S.D., Boronat, C.B., & Kennedy, M.H. (2020). Review of Undergraduate Student Retention and Graduation Since 2010: Patterns, Predictions, and Recommendations for 2020. *Journal of College Student Retention*, *22*(2), 227–250.

Berka, P., & Marek, L. (2021). Bachelor's degree student dropouts: Who tend to stay and who tend to leave? *Studies in Educational Evaluation*, *70*, 100999. https://doi.org/10.1016/j.stueduc.2021.100999

Bogard, M., Helbig, T., Huff, G., & James, C. (2011). *A Comparison of Empirical Models for Predicting Student Retention*. Retrieved from https://www.wku.edu/instres/documents/comparison_of_empirical_models.pdf

Cardona, T., Cudney, E., & Snyder, J. (2019). Predicting Degree Completion through Data Mining. *2019 ASEE Annual Conference & Exposition Proceedings*. https://doi.org/10.18260/1-2--33183

Cardona, T.A. & Cudney, E.A. (2019). Predicting Student Retention Using Support Vector Machines. *Procedia Manufacturing*, *39*, 1827–1833. https://doi.org/10.1016/j.promfg.2020.01.256

Costa, E.B., Fonseca, B., Santana, M.A., de Araújo, F.F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior*, *73*, 247–256. https://doi.org/10.1016/j.chb.2017.01.047

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

Delen, D. (2011–2012). Predicting Student Attrition with Data Mining Methods. *J. College Student Retention*, *13*(1), 17–35. https://doi.org/10.2190/CS.13.1.b

Dissanayake, H., Robinson, D., & Al-Azzam, O. (2016). Predictive modeling for student retention at St. Cloud state university. In *Proceedings of the international conference on data mining The Steering Committee of the World Congress in Computer Science*, *Computer Engineering and Applied Computing* (p.215).

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

Kondo, N., Okubo, M., & Hatanaka, T. (2017). Early Detection of At-Risk Students Using Machine Learning Based on LMS Log Data. *2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*. https://doi.org/10.1109/iiai-aai.2017.51

NCES, National Center for Education Statistics. (2021). *Undergraduate Retention and Graduation Rates*. Retrieved from https://nces.ed.gov/programs/coe/indicator/ctr

Oztekin, A. (2016). A Hybrid Data Analytic Approach to Predict College Graduation Status and Its Determinative Factors. *Industrial Management & Data Systems*, *116*(8), 1678–1699. https://doi.org/10.1108/imds-09-2015-0363

Patacsil, F.F. (2020). Predicting University Students' Academic Success Using Different Tree Classifiers and Ensemble Approaches to Suggest Suitable Program. *International Journal of Scientific & Technology Research*, *9*(2), 6001–6009.

Timaran Pereira, R., & Caicedo Zambrano, J. (2017). Application of Decision Trees for Detection of Student Dropout Profiles. *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. https://doi.org/10.1109/icmla.2017.0-107

Tsao, N.L., Kuo, C.H., Guo, T.L., & Sun, T.J. (2017). Data Consideration for At-Risk Students Early Alert. *2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*. https://doi.org/10.1109/iiai-aai.2017.133

Wild, S., & Schulze Heuling, L. (2020). Student dropout and retention: An event history analysis among students in cooperative higher education. *International Journal of Educational Research*, *104*, 101687. https://doi.org/10.1016/j.ijer.2020.101687

Zeineddine, H., Braendle, U., & Farah, A. (2021). Enhancing prediction of student success: Automated machine learning approach. *Computers & Electrical Engineering*, *89*, 106903. https://doi.org/10.1016/j.compeleceng.2020.106903