Supporting Minority Student Success by using Machine Learning to Identify At-Risk Students

J.D Jayaraman
New Jersey City University and Teachers College, Columbia University
Jersey City, New Jersey
+1 (201) 200 2297
jjayaraman@njcu.edu

Sue Gerber
New Jersey City University
Jersey City, New Jersey
+1 (201) 200 3042
sgerber@njcu.edu

Julian Garcia
New Jersey City University
Jersey City, New Jersey
+1 (201) 200 3463
jgarcia8@njcu.edu

ABSTRACT
Student retention is a major challenge at American universities with the average six year graduation rate hovering around 59%. Among minority students the graduation rate drops to 46% for Blacks and 55% for Hispanics. Low graduation rates not only impact the financial well-being of individuals but the economy as a whole. Thus, improving student retention, in particular, minority student retention, is of paramount importance at institutions of higher education. This paper describes a machine learning approach to predicting minority native and transfer student dropout using a dataset from a four year Hispanic serving institution in the north eastern region of the United States with a large percentage of minority students. The results of the study show that standard machine learning models can predict minority transfer student dropout with a high degree of accuracy of 97% and minority native student dropout with an accuracy of 81%. The features that were most important in predicting minority transfer student dropout were SAT scores, and college cumulative GPA, while high school GPA and college cumulative GPA were the top predictors for minority native student dropout. This study demonstrates that educational institutions can use cost effective off-the-shelf standard machine learning models to achieve a high degree of accuracy in predicting minority student dropout. The high prediction accuracy achieved helps in reliably identifying at-risk minority students and providing them with necessary interventions to support their academic success.

Keywords
Dropout prediction, Minority student retention, Machine learning models, At-risk students

1. INTRODUCTION
Student retention is a major challenge at American universities with the average 6 year graduation rate hovering around 59% [14]. Graduation rates vary with institutional selectivity [18]; the situation being particularly grave at institutions with open admission policies where the 6 year average graduation rate is at a meager 32% [14]. There is substantial variation in the graduation rates by race and ethnicity. African American students had the lowest six year graduation rate at 46% while Hispanic students' graduation rate is at 55% [14]. Transfer student graduation rates are also low at 42% [14]. Thus, improving student retention, particularly for minorities, is of paramount importance at institutions of higher education.

A critical factor in increasing student retention is the ability to accurately identify at-risk students, so that relevant interventions can be provided. But, there is a paucity of literature focused on predicting minority student dropout and even less literature that considers native and transfer students separately, taking into account their individual characteristics. Minority students' graduation rates are among the lowest and they have different characteristics, needs and face different challenges in college. Hence analyzing and modeling their dropout rate separately is warranted. It would also be informative to model the transfer and native students separately as these students have different characteristics and different graduation rates. Such an analysis could be used to identify, target and customize interventions differently to minority transfer and native students and hence lend better support for their success in college.

This paper describes a machine learning approach to predicting minority college student dropout, treating native and transfer students separately. The objective of this paper is not to build an esoteric novel machine learning model hitherto not seen in the literature, instead, we aim to show the effectiveness of standard off-the-shelf machine learning models in predicting minority student dropout. To the best of the authors' knowledge this study is one of the first to employ machine learning techniques to build separate models to predict minority transfer student and native student dropout taking into account the unique characteristics of each. Thus, this study contributes to the literature by demonstrating that standard machine learning models can achieve a high degree of accuracy in predicting minority student dropout by taking into account the different characteristics of native and transfer students. This high prediction accuracy can then be used to reliably identify at-risk students and provide them with the necessary intervention to support their success. This study also contributes by giving readers an idea of which machine learning models work well in this domain, what data preprocessing is needed and what features have good predictive power. Further, this study contributes to student retention practice as it demonstrates that easy to implement and cost effective off-the-shelf machine learning models can achieve a high degree of accuracy in identifying minority students at risk of dropping out.

2. LITERATURE REVIEW
Research on student attrition has traditionally been based on surveying student cohorts and following them to assess drop out. These surveys contributed to the building of theoretical models of student retention, the most famous of them being the Tinto model

J.D Jayaraman, Sue Gerber and Julian Garcia "Supporting Minority Student Success by using Machine Learning to Identify At-Risk Students" In: *The 12th International Conference on Educational Data Mining*, Michel Desmarais, Collin F. Lynch, Agathe Merceron, & Roger Nkambou (eds.) 2019, pp. 584 - 587

[17]. These survey based research have been criticized for being too specific to an institution and hence not generalizable [6]. An alternative to survey based research is to use the data that most higher education institutions routinely collect about their students. This type of research based on institutional databases has been shown to be comparable to survey based research [4].

There are numerous factors that affect student retention. Tinto [17] highlights academic difficulty, adjustment problems, lack of clear academic goals, lack of commitment, inability to integrate with the college community, uncertainty, incongruence and isolation as factors involved in student dropout. Tinto's theory of student integration posits that past and current academic success are crucial factors in determining student attrition and many studies have found high school GPA and SAT scores to have a strong effect on student retention [18]. Declaration of major and number of credit hours taken during the first semester have been used as proxies for institutional and goal commitment.

Transferring from one educational institution to another also has an impact on retention rates. Factors affecting transfer student academic success and persistence include issues regarding institution support [9, 23], financial factors [7], student goals [11] and familial support [1, 10].

When it comes to minority students Moffat [13] found that SAT scores were not a strong predictor of student success for Black students. A key challenge to success minority students face in predominantly white institutions is a sense of alienation due to underrepresentation [15]. Hoffman [10] found that student satisfaction and success can be strongly linked to cocurricular involvement for minority students.

Research on using machine learning techniques to predict student attrition is still in its infancy. Delen [8] and Thammasiri [16] used several machine learning methods such as support vector machines and neural networks to model freshmen student attrition and found that support vector machines performed best, reaching a prediction accuracy close to 80%. Lauria, Baron, Devireddy, Sundararaju, and Jayaprakash [12] used demographic and course related data to show that support vector machines performed better than decision trees at predicting at-risk students.

This study uses many of the factors impacting student retention identified in the literature to build machine learning models to predict student attrition. Factors unique to transfer students, as identified in the literature, are used to build specific models to predict transfer student dropouts.

3. CONCEPTUAL FRAMEWORK

The study is broadly based on the models of student retention developed by several researchers, one of the earliest and popular being that of Tinto [41]. Tinto's model suggested that student success is determined by the degree of academic and social integration. Other popular models of student retention include Bean's student attrition model [2, 3] which takes the employee turnover approach suggesting that students dropout for similar reasons as employees leave an organization and the Cabrera, Nora & Castenada [6] model which integrates the Tinto and Bean models. Based on these models several studies have identified factors that impact student dropout. High school GPA, SAT scores, number of credit hours taken during the first semester, declaration of major, aid based on academic achievement and student loans are among the factors that are predictive of student dropout.

Our study attempted to collect data on various factors based on the theoretical models and the predictive factors that have been identified based on them and use them as features in our machine learning models.

4. METHODOLOGY

4.1 Data

This study used five years (2011 – 2015) of minority student data from a regional Hispanic serving four year college in the north eastern region of the United States. The university is an urban university catering to a largely minority population (80% minorities). The university accepts a large number of transfer students from the local community colleges and other institutions. The overall four year graduation rate at the university was around 55%, with the graduation rate among transfer students at 61% and native students at 45%. Table 1 presents some descriptive statistics of the dataset.

Table 1: Descriptive Statistics

Native Students		**Transfer Students**	
Number of students	3897	Number of students	4703
Female	58%	Female	65%
Male	42%	Male	35%
Hispanic	55%	Hispanic	48%
African American	30%	African American	34%
Asian	11%	Asian	15%
Other race	4%	Other race	2%
Mean age	19	Mean age	28
Mean GPA	2.50	Mean GPA	2.99
Mean SAT Math	450	Mean SAT Math	413
Mean SAT English	430	Mean SAT English	390

We define a student to have dropped out if he/she does not enroll in the year following the last semester of enrollment. Based on this definition we constructed a binary indicator variable to indicate whether a student has dropped out or not. Both the transfer student and native student dataset had about 35% dropouts.

4.2 Features

Table 2 shows the features that were used in the machine learning models.

Table 2: Features

Features used for both native and transfer students
Age
Gender
Race
High School GPA
Gateway Math Status (gateway math course is required or not)
Gateway English Status (gateway English course is required or not)
Math placement (Has student completed the developmental math requirement)
English placement (Has student completed the developmental English requirement)
Cumulative GPA
Trend in the GPA over the semesters enrolled (Increasing, Decreasing, Stable)
SAT Math
SAT English
Difference between credits taken and credits earned
Community involvement (Student belongs to clubs and other student organizations or not)

Has student declared major (yes or no)
Additional features used only for transfer students
Difference between number of college credits applied for transfer and accepted for transfer.
Marital status
Highest degree earned prior to transfer

The features we used broadly fell into the following categories: student achievement, performance and progress, community engagement and demographics. We engineered several of the features from the raw data. From the raw data on GPA we extracted a feature indicating the trend in the GPA. From the credits taken and credits passed data by semester, we computed the difference of total credits taken and passed. The raw data also had information on clubs and other community activities that the student participated in. From this we created a binary variable indicating whether the student participated in community activities or not. For the transfer students we used additional features more relevant to them. Since the mean age of the transfer students (28 years) was much larger we used marital status and highest degree earned as features in the transfer student models.

4.3 Analysis

Student retention data sets are typically imbalanced as the number of dropouts is usually much less than the number that don't. Our dataset was imbalanced with around 35% of dropouts. If the data is imbalanced the standard classifiers have a bias towards the larger majority class. One approach to correcting this imbalance is to preprocess the data in order to balance it out and then build the model. This approach uses various techniques to either oversample the minority class or undersample the majority class or a combination of both. Synthetic Minority Oversampling Technique (SMOTE) is a popular and robust technique that uses a combination of oversampling the minority class and undersampling the majority class which results in better classifier performance [6]. We tried various techniques to correct the imbalance and found the SMOTE technique to yield the best results. Hence our study used SMOTE to correct the imbalance.

We used the features described above and imbalance corrected data to build the following machine learning models: Naïve Bayes, Logistic Regression, Decision Tree, Support Vector Machine, Random Forest, Extreme Gradient Boosting (XGBoost), and a Voting Ensemble model. We chose these models as they are popular classification models some of which have been used in the extant literature. The ensemble model XGBoost has been shown to have high performance in numerous datasets in different domains. The Voting Ensemble is a stacked ensemble model consisting of all of the other models mentioned above. All of the above models were built using 75% of the data for training and 25% for testing and using a 10 fold cross validation to avoid overfitting.

5. RESULTS

The prediction results of the models we trained above are shown in Table 3. Accuracy is a good and intuitive performance measure for balanced datasets. Since we have corrected the imbalance in our dataset prior to building our models, we have reported Accuracy as our measure of performance for the models. The AUC measure and the F1 score were very close to the accuracy measure, so, we have not reported them here, but the information is available on request.

Table 3: Dropout Prediction Results

Model	Accuracy	
	Native Student	**Transfer Student**
Naïve Bayes	71.3%	40.2%
Support Vector Machine	79.9%	81.2%
Logistic Regression	78.8%	81.4%
Random Forest	79.0%	97.0%
Extreme Gradient Boosting	81.3%	97.1%
Voting Ensemble	80.9%	97.4%

Table 4 shows the features ranked by importance. We have only reported the features that had a score of greater than 1% as the rest have negligible predictive power.

Table 4: Features Ordered by Importance

Native Students		Transfer Students	
Feature	**Score**	**Feature**	**Score**
Cumulative GPA	0.135	SAT English	0.261
High School GPA	0.103	SAT Math	0.191
SAT English	0.100	Cumulative GPA	0.149
SAT Math	0.097	Age	0.093
Difference credits taken and earned	0.095	Difference credits taken and earned	0.063
Age	0.074	Difference credits accepted and applied for transfer	0.047
Gateway English Status	0.051	GPA trend	0.043
English placement	0.046	Community involvement	0.031
Community involvement	0.046	Declared major	0.025
GPA trend	0.046	Developmental Math	0.021
Developmental Math	0.031	GPA trend	0.021
Gender	0.021	Gender	0.013
Race	0.013	Highest degree	0.010

6. DISCUSSION

This study demonstrates how standard machine learning models can be used to predict, with a high degree of accuracy, students at-risk of dropping out. To the best of our knowledge this is one of the first studies to focus on predicting minority student dropout and in particular minority transfer and native student dropout using standard machine learning techniques. The high prediction accuracies achieved in this study demonstrates the effectiveness of standard machine learning models for predicting undergraduate minority student dropout. Reliable identification of at-risk students can help in providing timely interventions to support the student's success and thus increase student retention. Any educational institution can adopt the approach demonstrated in this study with relative ease. A discussion of the various classifiers and the feature importance follows.

The Naïve Bayes classifier, as expected, performed the worst on both the native student and transfer student dataset. In the native student dataset the Extreme Gradient Boosting (XGBoost) model performed the best reaching an 81% accuracy. In the case of transfer students again the XGBoost, Random Forest and the Stacked Voting Ensemble all reached a very high 97% accuracy rate. This very high accuracy rate was indeed surprising as we did not expect to be able to achieve such high performance. All of these models performed much worse without imbalance correction, thus, pointing to the importance of correcting for imbalance in this domain. The results also show that different models perform better between native and transfer students with different features being

important, stressing the importance of modeling these student bodies separately.

For native students their college GPA and high school GPA seemed to be the strongest predictors of dropout. This is consistent with prior literature and the conceptual framework we based the study on. The other factors that had reasonably good predictive power were SAT scores and the difference between credits taken and earned. For transfer students the SAT scores were the strongest predictors with college GPA being the next strongest. Surprisingly High school GPA did not seem to have any impact on transfer student dropout prediction, unlike in the case of native students. This is inconsistent with the literature that has found high school grades to be good predictors of college success [5,15]. One potential explanation could be that the transfer students in our sample are older and thus far removed from high school and hence the high school GPA does not have much predictive power.

Another surprising result was that race did not have much predictive power for both the native students and transfer students. Given the racial gap in graduation rates, even among Black, Hispanic and Asian students, we expected to find some relationship between race and dropout, but found none. Similarly gender did not have much predictive power either. Community involvement was a fairly strong predictor of dropout for both the native and transfer students. This is consistent with the theoretical models of student retention used in our conceptual framework that highlight student engagement as a key factor in college success.

There are several limitations to our current study. The performance results that we have obtained are specific to our sample and caution should be exercised in generalizing them. However, our results do give readers and researchers an idea of the possible accuracy that can be achieved by using standard machine learning models to predict minority student dropout and what features are important. Also, our results can be very informative for institutions with a similar minority student profile as ours. Another limitation is that we have not considered any financial aid data, and other factors not related to academic achievement such as personality traits etc. These are all avenues for further research that we are currently pursuing.

7. CONCLUSION

The results of our study demonstrate that standard machine learning techniques can be effective in predicting minority student dropout with a high degree of accuracy. We also show that different models perform better between transfer students and native students thus reinforcing the importance of modeling these two student bodies separately.

The practical implication of this study is that it demonstrates that educational institutions can use their existing databases that contain routine student data and standard off-the-shelf machine learning models to accurately predict at-risk minority students. This is a cost effective way for institutions to identify at-risk students so that they can devote their resources to offering interventions aimed at retaining them.

8. REFERENCES

[1] Anglin, L. W., Davis, J. W., & Mooradian, P. W. (1995). Do transfer students graduate? A comparative study of transfer students and native university students. Community College Journal of Research and Practice, 19(4), 321-330.

[2] Bean, J. P. (1980). Dropouts and turnover: The synthesis and test of a causal model of student attrition. Research in higher education, 12(2), 155-187.

[3] Bean, J. P. (1982). Student attrition, intentions, and confidence: Interaction effects in a path model. Research in Higher Education, 17(4), 291-320.

[4] Caison, A. L. (2007). Analysis of institutionally specific retention research: A comparison between survey and institutional database methods. Research in Higher Education, 48(4), 435-451.

[5] Camara, W. J., & Echternacht, G. (2000). The SAT I and high school grades: Utility in predicting success in college. The College Board Research Notes, RN-10, 1–12.

[6] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.

[7] Davies, T. G., & Casey, K. (1999). Transfer student experiences: Comparing their academic and social lives at the community college and university. College Student Journal, 33, 60-71.

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

[9] Dougherty, K. (1994). The contradictory college: The conflicting origins, impacts and futures of the community college. Albany, NY: State University of New York.

[10] Hoffman, J. L. (2002). The impact of student cocurricular involvement on student success: Racial and religious differences. Journal of College Student Development, 43 (5), 712–739.

[11] Kinnick, M. K., & Kempner, K. (1988). Beyond "front door" access: attaining the bachelor's degree. Research in Higher Education, 29(4), 299-318.

[12] Lauría, E. J., Baron, J. D., Devireddy, M., Sundararaju, V., & Jayaprakash, S. M. (2012, April). Mining academic data to improve college student retention: An open source perspective. In Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (pp. 139-142). ACM.

[13] Moffat, G. K. (1993, February). The validity of the SAT as a predictor of grade point average for nontraditional college students. Paper presented at the annual meeting of the Eastern Educational Research Association, Clearwater Beach, FL. (ERIC Document Reproduction Service No. ED 356 252)

[14] Shapiro, D., Dundar, A., Huie, F., Wakhungu, P., Yuan, X., Nathan, A & Hwang, Y., A. (2017, April). Completing College: A National View of Student Attainment Rates by Race and Ethnicity – Fall 2010 Cohort (Signature Report No. 12b). Herndon, VA: National Student Clearinghouse Research Center.

[15] Schwitzer, A. M., Griffin, O. T., Ancis, J. R., & Thomas, C. R. (1999). Social adjustment experiences of African American college students. Journal of Counseling and Development, 77, 189–197.

[16] Thammasiri, D., Delen, D., Meesad, P., & Kasap, N. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. Expert Systems with Applications, 41(2), 321-330.

[17] Tinto, V. (1993). Building community. *Liberal Education*, *79*(4), 16-21.

[18] Wetzel, J. N., O'Toole, D., & Peterson, S. (1999). Factors affecting student retention probabilities: A case study. Journal of Economics and Finance, 23(1), 45-55.