Predicting Student Degree Completion using Random Forest

Tatiana A. Cardona, Missouri University of Science and Technology

Tatiana A. Cardona is a Ph.D. candidate in Systems engineering at Missouri University of Science and Technology (MS&T)from where she also received her M.S. in Engineering Management in 2016. Tatiana completed her B.S. in Industrial Engineering at Technological University of Pereira, Colombia in 2009. Her research interests include statistical modeling, Operations research and Data Science. She has served as a head teaching assistant for four semesters in operations management and project management in the MS&T.

Dr. Elizabeth A Cudney, Missouri University of Science and Technology

Dr. Elizabeth Cudney is an Associate Professor in the Engineering Management and Systems Engineering Department at Missouri University of Science and Technology. She received her B.S. in Industrial Engineering from North Carolina State University, Master of Engineering in Mechanical Engineering and MBA from the University of Hartford, and doctorate in Engineering Management from the University of Missouri – Rolla. In 2018, Dr. Cudney received the ASQ Crosby Medal for her book on Design for Six Sigma. Dr. Cudney received the 2018 IISE Fellow Award. She also received the 2017 Yoshio Kondo Academic Research Prize from the International Academy for Quality for sustained performance in exceptional published works. In 2014, Dr. Cudney was elected as an ASEM Fellow. In 2013, Dr. Cudney was elected as an ASQ Fellow. In 2010, Dr. Cudney was inducted into the International Academy for Quality. She received the 2008 ASQ A.V. Feigenbaum Medal and the 2006 SME Outstanding Young Manufacturing Engineering Award. She has published eight books and over 85 journal papers. Dr. Cudney is a certified Lean Six Sigma Master Black Belt. She holds eight ASQ certifications, which include ASQ Certified Quality Engineer, Manager of Quality/Operational Excellence, and Certified Six Sigma Black Belt, amongst others.

Dr. Jennifer Snyder, Valencia College

Jennifer Snyder received a PhD in Engineering Management at Missouri University of Science and Technology. She received her B.S. and M.S. in Chemistry from Missouri State University in Springfield, Missouri. She is a dean of science for Valencia College in Orlando, Florida.

Dr. Roger Wesley Hoerl, Union College

Dr. Roger W. Hoerl is the Brate-Peschel Associate Professor of Statistics at Union College in Schenectady, NY. Previously, he led the Applied Statistics Lab at GE Global Research. While at GE, Dr. Hoerl led a team of statisticians, applied mathematicians, and computational financial analysts who worked on some of GE's most challenging research problems, such as developing personalized medicine protocols, enhancing the reliability of aircraft engines, and management of risk for a half-trillion dollar portfolio. Dr. Hoerl has been named a Fellow of the American Statistical Association and the American Society for Quality, and has been elected to the International Statistical Institute and the International Academy for Quality. He has received the Brumbaugh and Hunter Awards, as well as the Shewhart Medal, from the American Society for Quality, and the Founders Award and Deming Lectureship Award from the American Statistical Association. While at GE Global Research, he received the Coolidge Fellowship, honoring one scientist a year from among the four global GE Research and Development sites for lifetime technical achievement.

Predicting Student Degree Completion using Random Forests

Abstract

Recent reports indicate that 40 percent of freshman at four-year public colleges will not graduate. Further, the average completion rate for two-year community colleges is less than 40 percent. Therefore, increasing student retention rates in higher education is of great importance. Student retention is a measure of students' continued enrollment until graduation. To improve retention rates, colleges and universities require strategies for intentional advising to ensure that students are able to complete their majors in a timely manner. Currently, efforts have been made to adjust admission requirements; however, retention rates are still considered low and these strategies have reduced access to higher education for students from different economic sectors. Thus, institutions have recognized the need to understand the factors that impact retention to better focus their efforts. To this end, this research presents the application of random forests to predict degree completion within three years, which represents 150 percent time to completion, and identify the variables that impact student retention at a large community college in the Midwest. The random forest algorithm consists of bagging (combining) decision trees created randomly from the training sample, thus creating a "forest". The model in this study was developed using data on 282 students with 14 variables. The variables included student details such as age, gender, degree, and college GPA. The model results, which include prediction and variable ranking, offer an important understanding about how to develop a more efficient and responsive system to support students.

Introduction

Research indicates there is a skill gap in our workforce that will only continue to widen without corrective action in higher education. At the same time, reports indicate that 40 percent of college freshman will not graduate. Therefore, increasing student retention rates in higher education is critical. Also, the ability of these institutions to prepare and graduate students with specific skills is an indicator of institutional performance, making it one of the main focus areas for universities and colleges [1]. This is perhaps more important to community colleges as they are a growing entry point for higher education [2].

In terms of retention improvement, efforts have been made to adjust admission requirements; however, the retention rates remain low with a national average of 62% for four-year colleges and 60% for universities [3] and many of these strategies have reduced access from different economic sectors to higher education [4]. Thus, many institutions have recognized the need to understand the factors that contribute to retention to better focus their efforts. While universities and colleges collect considerable student data, their ability to process the available information does not occur at the same pace as the collection [5]. There needs to be a method allowing for data utilization and timely implementation to improve student retention. For instance, the creation of predictive models that allow for the recognition of students at risk for attrition will enable timely interventions. By identifying the factors through a prediction model, universities and college can provide intentional student advising and planning. Further, higher education institutions can develop retention strategies that focus on identified student needs that meet their specific campus needs [6].

According to the literature, machine learning techniques have been applied to predict student

success with high confidence [7]. [8] conducted several studies to compare methodologies such as neural networks (NN), support vector machines (SVM), decision trees (DT), and random forests (RF), among others. The results indicated that these machine learning techniques had better prediction results than other statistical techniques such as logistic regression (LR) and discriminant analysis.

The purpose of this research was to develop a prediction model using the RF technique to predict student success by science, technology, engineering, and mathematics (STEM) students in a Midwest community college. RF was selected for three main reasons:

1. RF has consistently performed at or near the top of machine learning modeling approaches in a wide range of applications, similar to multilayer NN (i.e., deep learning) [9].
2. RF also provides insight into the contributions of specific variables to the accuracy of the final model, something that is lacking with most machine learning approaches.
3. The RF algorithm is very stable computationally, more so than NN or SVM, for example.

The time considered for successful degree completion was 150% of normal time for completion. This time was employed for the study in order to be consistent with the 1990 Student Right-to-Know Act, which requires postsecondary institutions to report the rate of students graduating in 150% of the time the program was designed [10]. As the data was from a community college, student success was measured as student completion within three years. A student pursuing an associate's degree should complete the degree program in two years. Therefore, a student is considered successful if they complete the program in three years or less. The following research question was investigated in this study:

> *Does the RF technique, based in its classification accuracy, provide a good resource for the prediction of student success at the Midwest community college for students in STEM majors? If so, what variables that have a higher impact in the prediction of student success?*

The remainder of this paper is structured as follows. First, a literature review provides background on RF applications for student success prediction. The research methodology is described next. The results of the model are then analyzed and discussed. Finally, the conclusions, research limitations, and future work are presented.

Literature Review

A majority of the literature on the application of machine learning techniques in education focuses on the use of an individual machine learning technique. Ensemble machine learning techniques combine several machine learning techniques and are commonly used to improve prediction models. However, the number of studies in the literature that use ensemble machine learning techniques such as RF, Boosted Trees (BT), and stacking of other techniques is low with only four journal papers published from 2010 to 2017. The results of ensemble machine learning show consistently high overall classification accuracy that ranges between 79.36% and 81.67%. Thus, it is important to develop models that can nurture the body of knowledge on how ensemble machine learning techniques can improve current models.

Research by [8] focused on prediction models for retention prior to sophomore year. The study applied classification methods such as NN, DT specifically the C5 algorithm, SVM, and LR.

The results were compared to the use of different ensembles including RF, BT, and information fusion, which stack different predictors. The dataset for analysis was comprised of 16,066 students enrolled as freshmen during 2004 and 2008. A well-balanced dataset was developed such that the classes to predict dropout were equally represented. When using the ensemble with the well-balanced data set, the accuracy of the predictions improved to approximately 80%, which was higher than using the standing alone techniques of SVM and DT. A sensitivity analysis showed the variables that impact at-risk student prediction for this study were student scholarships, loans, and fall GPA.

A comparison of models was proposed by [11] to predict student retention at St. Cloud State University. Principal component analysis (PCA) was used to select linear combinations of the variables that were not correlated with one another. Then, the original database and database after applying PCA were used to compare performance. The study applied six prediction models: k-nearest neighbor (KNN), DT, RF, LR, NN, and Bayesian Belief Networks (BBN). The results showed that the models using the PCA filtered dataset yielded better results. For example, the RF technique presented improvement in all evaluation factors and, together with LR, had the highest accuracy results of 84.77% and 83.07%, respectively.

[12] considered the importance of predicting students' grades in the courses they will enroll in during the next semester. The methodology employed factorization machines (FM), which is an adaptation of second order polynomial regression, along with other regression techniques such as RF, stochastic gradient Descent regression (SGD), KNN, and personalized multiple linear regression (PMLP). The model was used with information for each student or course. The dataset was collected during five years from George Mason University, with a total of 15 terms including summer terms. The model results indicate that PMLP had the lowest error from the individual techniques; however, RF provided more accurate predictions when the data lacked prior student information (i.e., first semester or cold start students).

Machine learning techniques were employed by [13] to predict at-risk students. The dataset used was obtained from the learning management system (LMS) during the first semester of 2015, which was comprised of records from 202 students. The methodology consisted of using LR, SVM, and RF to predict GPA. Classes for prediction were defined as a 1 if their GPA was greater than the average minus one standard deviation and 0 otherwise, meaning the student was at risk. The models were evaluated on the weekly change of the comparative importance of explanatory variables. Prediction from RF showed more stable behavior in terms of precision and sensitivity. With the weekly analysis, the model was able to identify a ranking of important variables depending on the point in time (i.e., number of weeks after the semester started) that was analyzed.

Research Methodology

The research process was conducted according to the main steps of data mining, which include the collection of the data to the reporting and use of it [14]. Although the data utilized in this study was not specifically collected for the purpose of predicting retention, the data mining steps were applied as represented in Figure 1. The research process is presented in the following segments: 1) data description and preparation, 2) data modeling and application of RF, and 3) model assessment.

Data description and preparation

The data for this research was collected from a community college in the Midwest that offers associate degrees in STEM majors. The dataset provided by the institution was comprised of 904 students pursuing degrees in chemistry, biology, and engineering. The data collected included information on students registered from spring 2013 through fall 2017.

The raw dataset contained a considerable amount of missing and inconsistent data. The reason behind this is that the institution is an open-admission institution; thus, information such as high school GPA and standardized exam scores are not required for admission. Therefore, it was reasonable to remove students that did not report high school GPA and standardized exam scores, as the missing information would highly impact the application of the classifier algorithm for predicting student success. Also, cases with inaccurately reported data (for example, scores out of the normal score range) were not taken in account. Table 1 presents a summary of the descriptive statistics for the numerical variables in the initial dataset.

Table 1. Raw data descriptive statistics for numerical variables

Variable	N	Mean	Median	Min	Max
Age	904	24.85	21	16	65
ACT Comp	428	22.64	22	11	34
ACT English	436	22.01	21	7	35
ACT Math	436	22.835	22	13	35
ACT Reading	435	23.13	22	9	36
High School GPA	605	4.13	3.51	1	91.38
College GPA	814	2.775	2.95	0	4.93

Removing the incomplete records resulted in a final dataset of 282 students, which consisted of 51 completers and 231 non-completers. For this research, completers were defined as the students that completed their associate's degree in three years or less. Conversely, non-completers did not finish their associate's degree within three years. The resulting dataset contained a moderate number of variables (14 variables) for developing the RF model. The input variables are presented in Table 2.

Variables as age, gender, first generation student, plan to work, high school GPA, and ACT scores were self-reported when the student applied for admission. College GPA was the overall GPA of the student as of fall 2017 or their GPA upon graduation if the student had completed their studies. The degree was the student's current degree as of fall 2017 or their awarded degree if the student had graduated.

Initial experiments suggested that it was beneficial to generate a subsample to balance the number of instances of the prediction classes (i.e., completers and non-completers). The initial results provided high overall classification accuracy but low precision (correct predictions out of total predictions of the class). These results are consistent with other studies such as [15]. Their research focused on imbalanced data and identified several reasons why learning algorithms work better with balanced data. For example, for the DT algorithm the findings indicated that

successive partitioning left even fewer examples of the minority class, which reduces the confidence estimates. In addition, the sparseness can blurry characteristics that may result in reducing classification performance.

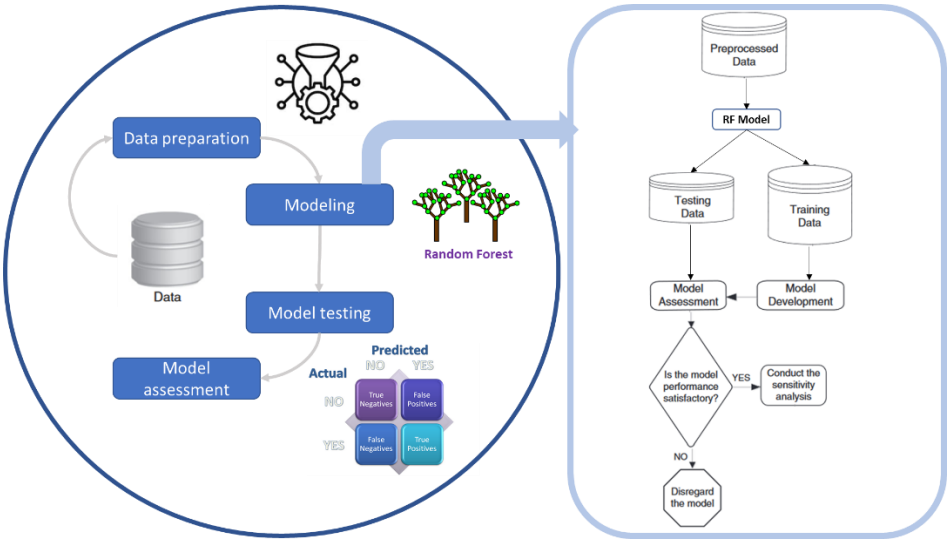

Figure 1. Data analytic methodology

Table 2. Variables used in the study

Variable	**Type**
Complete (Target variable)	Yes/No
Degree	Chemistry, Biology, Engineering
Age	Numerical
Gender	Female/Male
Full Time Student	Yes/No
1st Generation Student	Yes/No
Plan to work	Yes/No
ACT comprehensive	Numerical
ACT English	Numerical
ACT mathematics	Numerical
ACT reading	Numerical
High school GPA	Numerical
College GPA	Numerical

As RF are a collection of DT, they are sensitive to imbalanced data [16]. Therefore, the initial performance results in the experimental phase of this study were attributed to the imbalanced data as there were more non-completers (231) compared to completers (51) as shown in Figure *2*, where 1 indicates completion in three years or less and 0 indicates the student did not complete the program in three years or less. Then, a balanced subsample to continue the modeling process was generated using the stratified sampling function in STATISTICA 12 that allowed a user-defined proportion of the minority class to be over sampled in this specific case. Random under sampling and oversampling techniques to balance datasets has been widely used and have shown to improve classifier accuracy [8, 15, 17].

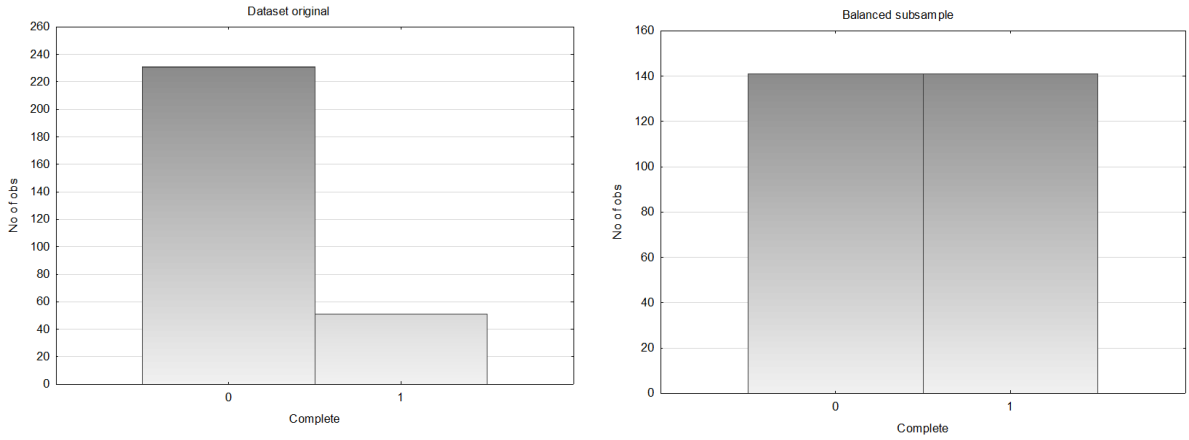

Figure 2. Initial data distribution (left side) vs balanced data distribution (right side)

Data modeling

The RF algorithm is an ensemble of decision trees created randomly from a given dataset. Each tree is created with a different data set chosen randomly (with replacement) from the original data set, a technique known as "bootstrapping." Then, at each branch of each tree, a subset of variables is chosen randomly, and the tree is forced to select from this subset of variables. The intent of this approach is to force the model to consider other variables, besides the most dominant, which might provide greater predictive power with the new data set. The final tree produces a classification response (class prediction) for each observation. This approach is then replicated for numerous trees, producing a "forest." Each tree generates a vote that enables the classification of the input variables into expected classes, completer and non-completer. The forest then classifies by "majority vote." The variables that are important for class prediction are also determined based on measures of internal errors (on the tree nodes), tree strength in the forest (classification accuracy), and correlation between the trees. Thus, a more accurate classification is obtained than if analyzing a standing alone DT [18]. Another advantage of this technique is it is not as prone to overfitting as most machine learning algorithms due to the law of large numbers, which states that performing an experiment a large number of times will provide a stable result long term. In other words, the average of the results will be closer to the expected value as more trials are performed.

The model was implemented using STATISTICA 12. The parameters used in the training were set as shown in Table *3*. Several experiments were run using different combinations of the variable parameters to identify the model with the highest overall classification accuracy. In order to test the model, a subset comprised of 30% of the original dataset was randomly selected and held until the training was concluded.

Model assessment

It is important for the model to be precise at predicting non-completers as the results are intended to improve and develop retention strategies. A retention strategy based on a false negative for completion risk could result in incurred costs for the institution and may not help students. Therefore, the assessment metrics were selected based on the classification accuracy for non-completers precision and recall measures for the testing set and overall classification accuracy for training and testing sets.

Table 3. Modeling parameters

Parameter type	**Parameter**		**Selection**
Fixed	Misclassification cost		Equal
	Prior probabilities		Estimated
	Stopping parameters/each tree	Max *n* of nodes	7
		Max *n* of levels	10
		Min *n* of cases	7
		Min *n* in child node	5
Variable	Number of trees		100, 150, and 250
	Model stopping condition	Percentage decrease in training error (evaluated every 10 cycles)	5%, 1%, and non-stopping condition

The level of importance of the factors that impact the prediction in the model were also identified. Recall that this is a key advantage of RF. STATISTICA calculates the drop in the node impurity and adds the result from every node for each variable. The largest sum represents the most important variable. The ranking score is scaled and presented on a range of 0-100. This measures how often the individual trees split on this variable, and also the additional discriminatory power these splits provided.

Results

Different parameter combinations were tested including the number of trees with a stopping condition of 5% then with a 1% decrease in error. The results are presented in Figure 3 for the scenarios with 100 and stopping condition of 5% decrease in error (stopped at 70 trees) on the left side and 250 trees with non-stopping condition on the right. As shown in Figure 3, the misclassification for the testing data started to be stable (no significant increase or decrease) after approximately 40 trees. This finding was consistent when using a total of 250 trees. Note that Figure 3 shows both classification accuracy with the original "training" data, used to fit or train the model, and also with test data that was held out from fitting the model.

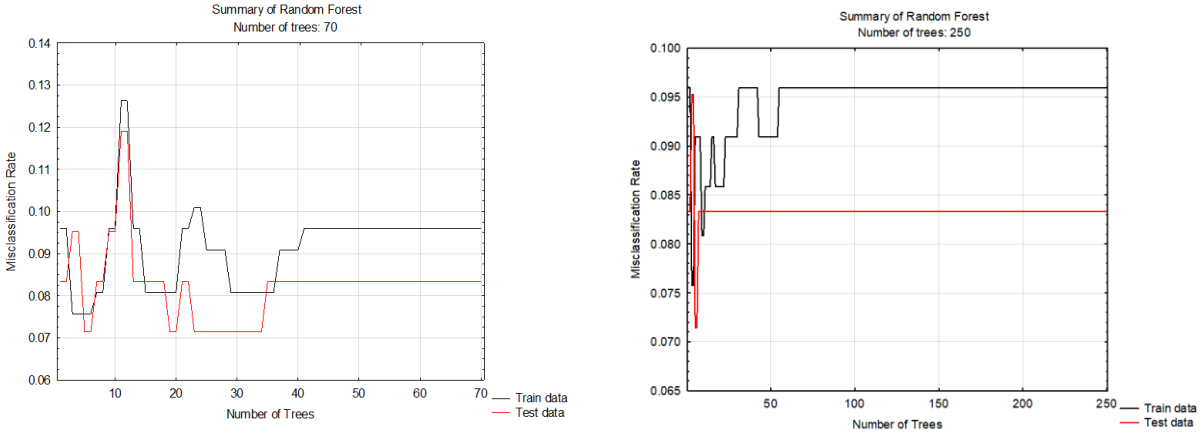

Figure 3. Misclassification rate 70 trees (left), 250 trees (right)

The overall accuracy of the model for the training and test subsets is displayed in Table 4. There is not a significant difference between the overall accuracy performance for the training and testing subset. The results indicate that RF offers a good prediction model for STEM degree completion for the Midwest community college students with a validation performance of approximately 91%. For higher education institutions, this classification accuracy for predicting retention rates supports the development of strategic endeavors to increase student success.

Table 4. Model Accuracy

Subset	Overall accuracy
Train	0.904
Test	0.917

The misclassification ("confusion") matrix is provided in Table 5 and recall and precision measures are presented in Table 6. Both results are indicative of high prediction performance for the classification of non-completers. Specifically, for the test subsample precision (95.2%) and recall (88.9%) shows a risk of misclassification under 11%.

Table 5. Misclassification matrix. Training subsample (left). Testing subsample (right)

Training				
		Predicted		
	Class	0	1	Total
Observed	0	79	17	96
	1	2	100	102
Total		81	117	198

Test				
		Predicted		
	Class	0	1	Total
Observed	0	40	5	45
	1	2	37	39
Total		42	42	84

Table 6. Assessment measures for training and test subsamples

	Training	Test
Recall	0.975	0.889
Precision	0.823	0.952

After evaluating the classification accuracy of the model, it was important to identify the variables that impact the prediction. The information gain (Gini factor for classification models) is used to define the rank of the variables. Each tree is partitioned by choosing the variable that offers a higher information gain [19]. To determine the importance of each variable in the tree, STATISTICA uses the sum of the information gain from the overall nodes to find the variable overall information gain. The rank of the variables is determined by adding the information gain of each variable for all the trees and, scaling it in such way that the highest value will be 100. When the resulting value is less than or equal to zero, the variable does not impact the model and can be removed.

Table 7 and Figure 5 present the rank of importance of the different variables used. The results showed that the most significant variables are age, college GPA, ACT composite, and ACT math. Age is shown as a key variable that can be useful to administrators in predicting completion. Further, of the various academic metrics available, college GPA is the most useful,

at least with this data. Although this information could be helpful in understanding the variable interaction of age with success, as a standalone variable it is not a variable that can govern the student success behavior.

Table 7. Variable importance rank

Variable	**Rank**
Age	100
College GPA	60
ACT Comp	38
ACT Math	33
High School GPA	33
ACT English	32
ACT Reading	27
Degree	17
Part time student	17
Full time student	13
Plans to work	9
Gender	8
First generation	7

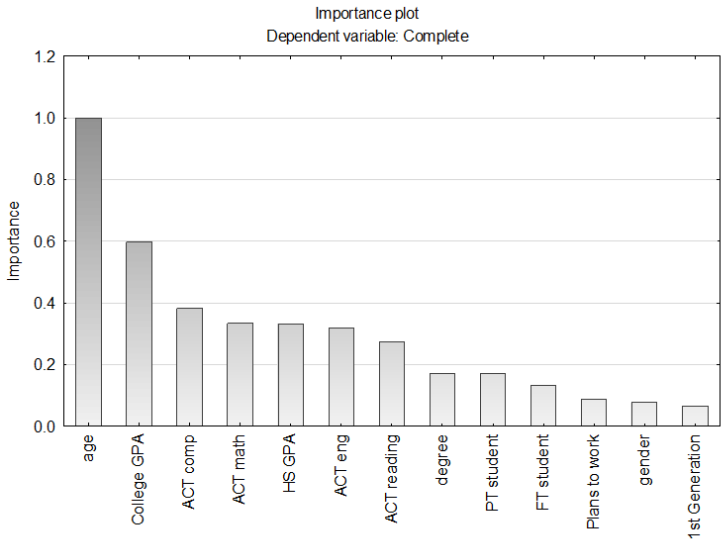

Figure 5. Predictor importance

Conclusions, Limitations, and Future Work

This research presented a complete case of the application of RF for predicting degree completion. The model results showed a good performance with precision rates over 80% and testing overall accuracy also over 80%. Therefore, RF technique provides a good resource for the prediction of student success at the Midwest community college for students in STEM majors. Further, this case study contributes in creating evidence of the application of models specifically to community college data, as most of previous literature of machine learning applications for student success is focused on data collected from universities.

RF can also be used to identify the level of importance of the factors impacting students successfully completing a degree program. Although GPA is a common factor found in prior literature as important for predicting student success, variables such as ACT math and ACT English are not commonly found as variables of high impact in other studies. In addition, age is also a key variable, which was a similar finding to other studies. Further, the findings suggest that the level of importance of those factors depended on the methodology used; however, further investigation should be performed.

Several limitations were present during the development of the described model. In this case, the dataset was not collected specifically for the current research. The number of variables and data points had to be reduced to generate a more adequate sample. This increased the risk of overfitting the model; thus, several combinations of the initial model parameters where tested in order to determine the most adequate combination. Also, it is important to highlight that, while the study achieved a high classification performance, the data is only representative of one community. Therefore, the results are not generalizable. However, the methodology can be used by other higher education institutions to determine the factors of importance.

Further research should be conducted to include other factors such as financial status and other demographic characteristics. This will enable the development of retention strategies and intentional advising that will better address and improve student success. Also, different machine learning techniques should be employed to offer a comparison in performance and a better understanding of the benefits of each approach.

Finally, it would also be interesting to analyze the general behavior of student completion for community colleges by collecting information from different institutions. This may help identify factors that vary by institution which may later become retention issues.

References

[1] A. M. Williford and J. Y. Schaller, "All retention all the time: How institutional research can synthesize information and influence retention practices," In *Proceedings of the 45th Annual Forum of the Association for Institutional Research, May 2005*.

[2] J. Snyder and E. A. Cudney, "Retention models for STEM majors and alignment to community colleges: A review of the literature," in *Journal of STEM Education*, vol. 18(3), pp. 30-39, 2017.

[3] J. Snyder and E. A. Cudney, "A retention model for community college STEM students," in *ASEE Annual Conference & Exposition, Salt Lake City, Utah. June 2018.*

[4] D. Kirp, "The college dropout scandal," *The Chronicle Review*, 2019, https://www.chronicle.com/interactives/20190726-dropout-scandal

[5] L. V. Morris, "Mining data for student success," in *Innovative Higher Education*, vol. 41(3), pp. 183- 185, 2016.

[6] A. Slim, G. L. Heileman, J. Kozlick and C. T. Abdallah, "Predicting student success based on prior performance," in *Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on, pp. 410-415, December 2014.*

[7] T. Cardona, E. A. Cudney and J. Snyder, "*Predicting degree completion through data mining,"* in *ASEE Annual Conference & Exposit*ion, Tampa, FL, June 2019.

[8] D. Delen, “A comparative analysis of machine learning techniques for student retention management,” in *Decision Support Systems*, vol. 49(4), pp. 498-506, 2010.
[9] G. James, D. Witten, T. Hastie and R. Tibshirani, *An Introduction to Statistical Learning*. Springer, NY, 2017.
[10] Undergraduate Retention and Graduation Rates. The condition of education (2018). Retrieved December 16, 2018, from https://nces.ed.gov/programs/coe/indicator_ctr.asp
[11] H. Dissanayake, D. Robinson and O. Al-Azzam, “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 (WorldComp), January 2016.*
[12] M. Sweeney, J. Lester, H. Rangwala and A. Johri, “Next-term student performance prediction: A recommender systems approach,” in *Journal of Educational Data Mining*, vol. 8(1), pp. 22-51, 2016.
[13] N. Kondo, M. Okubo and T. Hatanaka, “Early detection of at-risk students using machine learning based on LMS log data,” in *Advanced Applied Informatics (IIAI-AAI), 2017 6th IIAI International Congress on (pp. 198-201). IEEE. July 2017.*
[14] A. Feelders, H. Daniels and M. Holsheimer, “Methodological and practical aspects of data mining,” in *Information & Management*, vol. 37(5), pp. 271-281, 2000.
[15] H. He and E. A. Garcia, “Learning from imbalanced data,” in *IEEE Transactions on Knowledge and Data Engineering*, vol. 21(9), pp. 1263-1284, 2009.
[16] C. Chen, A. Liaw and L. Breiman, “Using random forest to learn imbalanced data,” University of California, Berkeley, Technical Report 110(1-12), pp. 24, 2004.
[17] K. Millard and M. Richardson, “On the importance of training data sample selection in random forest image classification: A case study in peatland ecosystem mapping,” in *Remote Sensing*, vol. 7(7), pp. 8489-8515, 2015.
[18] L. Breiman, “Random forests,” in *Machine learning*, vol. 45(1), pp. 5-32, 2001.
[19] S. Chakrabarti, E. Cox, E. Frank, R. H. Güting, J. Han, X. Jiang, ... and D. PyleD. (2008). *Data mining: know it all*. Morgan Kaufmann.