Machine Learning in Institutional Research:
Random Forest Model of Undergraduate Transfer Risk

Eric Braun and Dr. Chris Ferland, *GCSU Office of Institutional Research and Effectiveness*

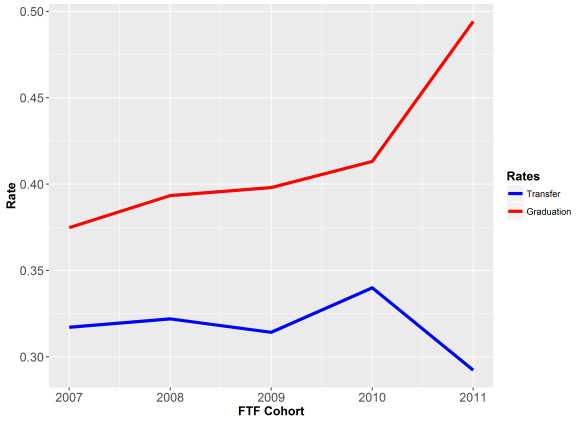
Machine learning comprises a set of cutting edge computational tools for data driven decision making. While machine learning has been widely adopted in industry, it has yet to become a staple of institutional research. The Office of Institutional Research and Effectiveness has developed a machine learning model that predicts student transfer and graduation in order to demonstrate that machine learning can provide insight into an issue of primary importance to higher education administrators. Our model predicts the likelihood a student will transfer or graduate in their next term of enrollment with 88% and 62% accuracy respectively; the predictive strength of the model, especially in regards to transfer risk, yields a powerful tool for both assessing and developing interventions to improve student success. Recommendations for initial implementation of the model in decision making include uses in Academic Advising, Academic Department Chairing and Enrollment Management

I. Background

Machine learning methods have become a tool of choice for leveraging data to assist in decision making. Machine learning has found widespread usage in industry. Examples of machine learning's diverse applications include predicting consumer churn, returning internet search results and identifying fraudulent financial transactions. In order to demonstrate how machine learning can be applied in the higher education setting, the GCSU Office of Institutional Research and Effectiveness has developed a random forest model, a robust machine learning method, that predicts student transfer and graduation risk. The ability to predict retention and graduation as well as assess associated factors allows for more informed development and assessment of retention and graduation interventions.

The current trend in GCSU's four year graduation and transfer rates is positive. The four year graduation rate increased from 39% to 49% from the 2009 to 2011 first time full time freshmen cohorts; the four year transfer rate for those same cohorts decreased from 32% to 29% over the same period. While these trends are positive, the likely result of a number of concurrent interventions such as the broadening of summer course offerings and expansion of the Supplemental Instruction program, there remains significant room for improvement.

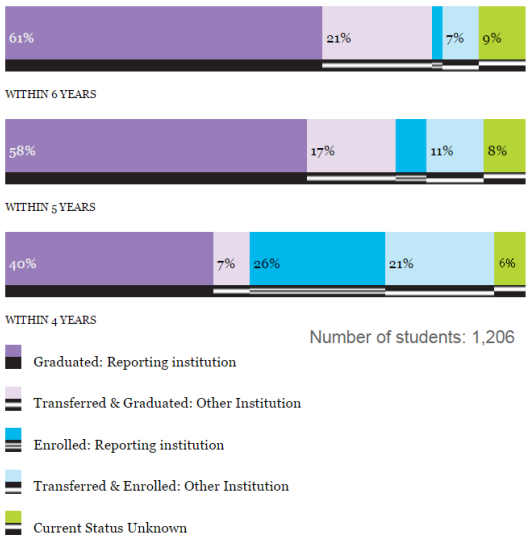
4 Year Graduation and Transfer Rate
2007 - 2011 Trend



If we examine the six year outcomes for the 2009 cohort, we see that over 61% of students graduated from GCSU with another 2% still enrolled. A full 28% either graduated or was enrolled at another institution, while only 9% of the 2009 cohort had stopped out of higher education altogether. Proportionately, retaining more transfer students would have a more significant impact on the graduation rate than retain-

ing more stopouts. Transfers both represent a much greater proportion of the cohort and 75% of transfers subsequently went on to graduate from their transfer institution.

FTF 2009 Cohort Retention and Graduation



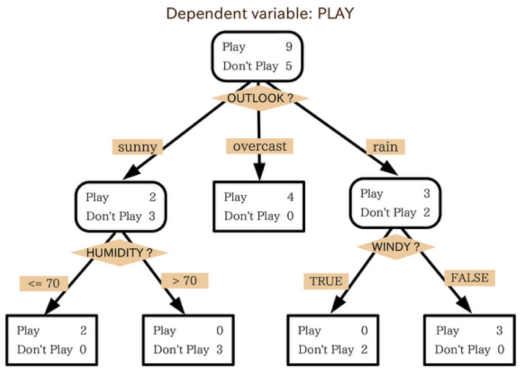
National Student Clearinghouse, 2015

It is difficult to discern what factors motivate the transfer and graduation trends. Many factors changed over the academic tenure of the 2009 to 2011 first time full time freshmen cohorts, including but not limited to new academic support programs, changing student characteristics, and different faculty. In order to better isolate the relationship between the many possible factors at play and predict the future behavior of students, a "random forest" machine learning model was developed and applied to available institutional data on students, faculty and college programs.

II. Methods

The data used for the random forest machine learning model consisted of 8,691 first-time full-time undergraduates, comprising 97% of the full 2007 to 2014 first-time full-time undergraduate cohorts. The data included 5,400 females and 3,291 males, and 7,540 Caucasians, 436 Latinos, 345 African Americans and 370 of other ethnicities.[1] The institutional portion of the data included demographic, academic performance, course, faculty, and financial characteristics. Data from several college programs were able to be included, including the Career Center and the GIVE Center. US Census data on student's home census tracts were also included.

Example: A Decision Tree



CitizenNet, 2012

The particular machine learning model used, random forests, have proven to be amongst the most robust methods available. Machine learning methods, in general, use observed data to 'train' a model to predict a future outcome, though they differ greatly in the approach used to train and predict. The intuition behind the random forest method begins with a basic decision tree. For example, imagine one was at the park and had to determine whether the weather will remain fair enough to play a game of football. One could go through a set of variables, such as cloud cover, precipitation humidity and wind, to make the decision as seen in the adjacent figure. It is difficult, however, to determine a priori which variables should be included and at what point in the tree. Random forests address this issue by repeatedly selecting a random subset of variables from all the variables available, selecting a random subset of observed data from which to train, and constructing a tree based on a chosen splitting rule. When predicting a new outcome, each constructed tree gets a vote, with the majority vote yielding the ultimate prediction.

In the case of predicting student transfer and graduation risk, we chose to use a random forest with a competing risks type splitting rule [2]. Transfer and graduation are event history outcomes, and are thus suited to the event hazard formulation of the splitting rule.

III. Results

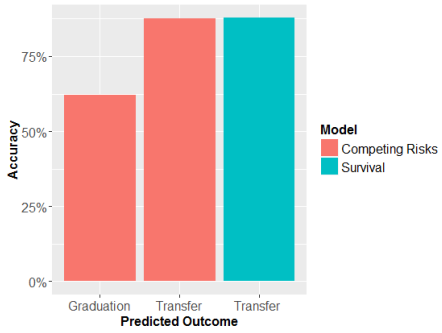
The model was found to predict transfer and graduation with 88% and 62% accuracy respectively on

[1] See GCSU's 2015 Factbook and OIRE's dashboards for greater cohort data detail.

[2] R package 'rfsrc' implementation of competing risks random forests

average. These accuracies were validated both using a random subset data to train and the excluded data to test, as well as predicting the 2014-2015 academic year outcomes with the rest of the data being used to train the model. For methodological comparison, the same data were used to train a survival random forest targeted at predicting transfer outcomes. The considerably higher error rate in predicting graduation is due to two factors. First, in order to predict whether a given student graduates correctly, the model must also predict whether that student transferred since transferring precludes graduation. Second, transferring and graduating students are similar in their characteristics in the model data. It is hoped with additional data and model development the graduation prediction accuracy will be substantially improved.

Model Accuracies



Top 10 Predictors of Transfer Risk

Factor	Relative Importanc
HOPE Scholarship	1.00
Trimester	0.92
Summer Terms Attended	0.86
Matriculation Year	0.71
Loans	0.58
Culm. Credit Hours Earned*	0.21
Full Time Faculty Taught Courses* %	0.15
Course Registration Timeliness*	0.09
Ave. Term Units Withdrawn*	0.07
Ave. Term Units Taken* %	0.07
*: Lagged Variable	

Top 10 Predictors of Graduation Risk

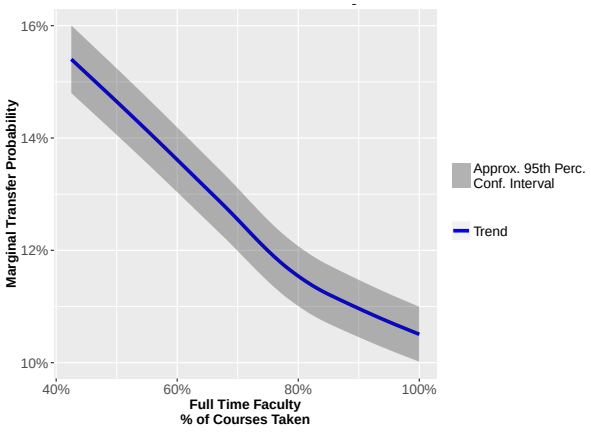
Factor	Relative Importance
Trimester	1.00
Merit Scholarship	0.35
Matriculation Year	0.33
Summer Terms Attended	0.29
Ave. Units Withdrawn*	0.23
Culm. GPA*	0.23
Career Center Event Attendance*	0.21
Course Registration Timeliness	0.16
Ave. Term Hours Attempted*	0.15
Ave. Difficulty of Courses Taken*	0.08
*: Lagged Variable	

The random forest model was used to produce three sets of results that exhibit the variety of information that can be gleaned from the methodology: variable importance, marginal variable impact, and cohort identification. The model could also notably be used to predict outcomes for individual students. A full list of the model variables organized by outcome and variable importance used in the model can be found in the appendix.

Variable importance measures how predictive a given variable is on the outcome. It should be noted that variable importance does not have the same interpretation as a regression coefficient; variable importances are not the individual marginal effects of a linear combination of variables. Rather, variable importance is a measure of the influence of a variable on the random forest. In order to facilitate an intuitive interpretation, variable importances here are calculated relative to each other. Accordingly, the most important variable has a value of 1 with the others listed in decreasing relative proportion.

The two included variable relative importance tables can be used as starting points for developing interventions. The random forest model does not provide an explanation for how these variables tell the story of transfer and graduation risk, only that these variables have been found to be important. Further investigation in the form of field research would be required to develop a theoretical explanation.

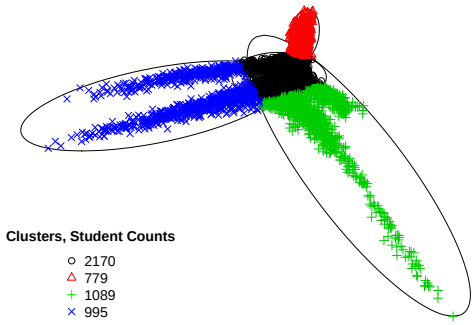
First Year Transfer Risk
% Courses Taught By Full Time Faculty



The random forest model can be used to estimate

the marginal effect of the variables. We have found, for example, that freshmen who have taken 100% of their courses with full time faculty have approximately 1/3 lower percent chance of transferring in the following year than freshmen who have taken only 40% of their courses with full time faculty. The rate of decrease in transfer probability is relatively constant as the percent of courses taught by full time faculty increases, suggesting that there is value in increasing the proportion of courses taught by full time faculty as close to 100% as possible.

FTF Student Clusters
Graduates and Transfers, Fall 2008 - Fall 2015



Another application of the random forest model is cohort identification. The cluster plot visualizes four distinct clusters [3] of students based on all the factors included in the random forest models. Student profiles could be developed based on the characteristics of each student cluster to guide the creation of interventions specific to the unique needs of student these four student sub-populations.

IV. Recommendations

The breadth and accuracy of the possible insights from the random forest model sketch the potential machine learning methods have to assist in furthering university priorities. The implementation of these models, however, requires top down support in order to motivate machine learning assisted decision making and bottom up belief in the change being worthwhile. Wider institutional support for the use of machine learning methods in decision making would thus be facilitated if there is a tangible example of successful application. Given OIRE has already developed a random forest model for transfer and graduation behavior, OIRE could collaborate with appropriate entities such as the Center for Student Success and Enrollment Management to align the model analyses to the needs of current decision making or to assist in the implementation of a current intervention. To this effect, an outreach effort was conducted, resulting in interviews with Academic Advising, an academic department chair and Enrollment Management. The following recommendations for a first application of the model emerged from these conversations.

IV.I. Academic Advising: Student Risk Flags

Advisors have a limited amount of resources to devote to each of their advisees. In addition, advisors have many possible interventions they might to suggest to a student depending on their situation, such as the Learning Center for those struggling academically to the Career Center for those without clear direction in their studies. The model could help assist advisors by providing two different indicators for each of their advisees, one for graduation risk and one for transfer risk. The indicators could be as simple as a green/yellow/red - high/medium/low risk flags or as complex as a specific percentage likelihood and a list of top risk factors for each student. The advisor could then use these indicators, along with all other available information and their own expertise, to determine which students are most likely to benefit from an intervention to improve their chances of graduating and/or not transferring. The model is not meant here to be the final word but, rather, a data-driven perspective that could help highlight in need students.

IV.II. Academic Department Chairs: Declared Majors Intelligence

Department chairs, as advocates for their departments, need to be aware of the performance of their department's majors to make informed decisions about the deployment of academic resources. The student level predictions of graduation and transfer risk from the models could be aggregated up to the department level so department chairs could get an assessment of the likely future success of their department majors as well as a list of the top risk indicators for the full population of a department's majors. This information would then give depart-

[3] The clusters were created by applying pam clustering (a robust variation of k-means clustering) to the proximity matrix of the random forest.

ment chairs a data driven view into the challenges facing their department majors.

IV.III. Enrollment Management: Application Yield

The model could be re-purposed for other uses. With some modification to the baseline data and methodology, the model could be transitioned into a tool that can assess the risk of a application to Georgia College yielding. Given that Georgia College's prestige is contingent on a high percentage yield of admitted students, greater accuracy in this area would be a boon to the entire institution. An indicator for each application could be offered in the same fashion as the indicators for the students: something as simple green/yellow/red - high/medium/low risk flags or as complex as a specific percentage likelihood and a list of the top risk factors. Enrollment Management could then use the indicator as an aid in making decisions on specific applications and prognosticating the likely overall yield and class size.

IV.IV. Appendix: Full Model Relative Variable Importance

Relative Variable Importance, Transfer Risk

Variable	Relative Importance
Loan	1.00
Trimester	0.64
Merit Scholarship	0.63
Summer Terms Attended	0.53
Matriculation Year	0.34
Culm. Credit Hours Earned*	0.11
Full Time Faculty Taught Courses* %	0.10
Ave. Term Hours Attempted*	0.06
Course Registration Timeliness	0.05
Minority Faculty Taught Courses %	0.05
Ave. Units Withdrawn*	0.04
Career Center Event Attendance*	0.04
Ave. Difficulty of Course Taken*	0.03
Ave. GIVE Center Hours*	0.03
Needs Based Scholarship	0.03
Major	0.03
Major Change Count*	0.02
College	0.02
Career Center Appointments*	0.02
Culm. GPA*	0.02
High School GPA	0.01
Female Faculty Taught Courses	0.01
Gender	0.01
AP Credits	0.01
Median Home Price, Home Census Tract	0.01
Aggregate Student Income	0.00
Secondary Edu. Attainment, Home Census Tract	0.00
SAT Score	0.00
App. Submitted Pre-UGA Deadline	0.00
Historical High School GPA	0.00
Ethnicity	0.00
Undecided Major	0.00
Population Density, Home Census Tract	0.00
Parental Edu. Attainment	0.00
*: Lagged Variable	

Relative Variable Importance, Graduation Risk

Variable	Relative Importance
Trimester	1.00
Merit Scholarship	0.35
Matriculation Year	0.33
Summer Terms Attended	0.29
Ave. Units Withdrawn*	0.23
Culm. GPA*	0.23
Career Center Event Attendance*	0.21
Course Registration Timeliness	0.16
Ave. Term Hours Attempted*	0.15
Ave. Difficulty of Courses Taken*	0.08
High School GPA	0.08
Gender	0.07
Full Time Faculty Taught Courses*	0.07
Culm. Credit Hours Earned*	0.05
College	0.05
Major	0.05
Ave. Give Center Hours	0.05
AP Credits	0.04
Loan	0.03
SAT Score	0.02
Female Faculty Taught Courses	0.02
Career Center Appointments*	0.02
App. Submitted Pre-UGA Deadline	0.02
Female Faculty Taught Courses	0.02
Needs Based Scholarship	0.02
Ethnicity	0.01
Historical High Sschool GPA	0.01
Seconday Edu. Attainment, Home Census Tract	0.00
Median Home Price, Home Census Tract	0.00
Undecided Major	0.00
Parental Edu. Attainment	0.00
Major Change Count*	0.00
Aggregate Student Household Income	0.00
Population Density, Home Census Tract	0.00
*: Lagged Variable	