Predicting Fall-to-Spring Community College Retention with Classification Modeling

Matthew Fikes

DSC550: Data Mining
Spring 2021

INTRODUCTION

For ten years and counting, the State University of New York system has been facing declining enrollment (Giacomelli, 2020). This has been forcing schools to work harder to not only increase enrollment numbers but to improve retention. Approximately between 20-30% of our students simply never return after their first Fall semester, losses felt both immediately and in subsequent semesters of lost aid. Figure 1 shows the values for students lost and retained for the semester represented in the data. This constant attrition necessitates a model for predicting those students likely to drop out. By identifying at-risk students, they may be given additional help or intervention measures devised at a later point.

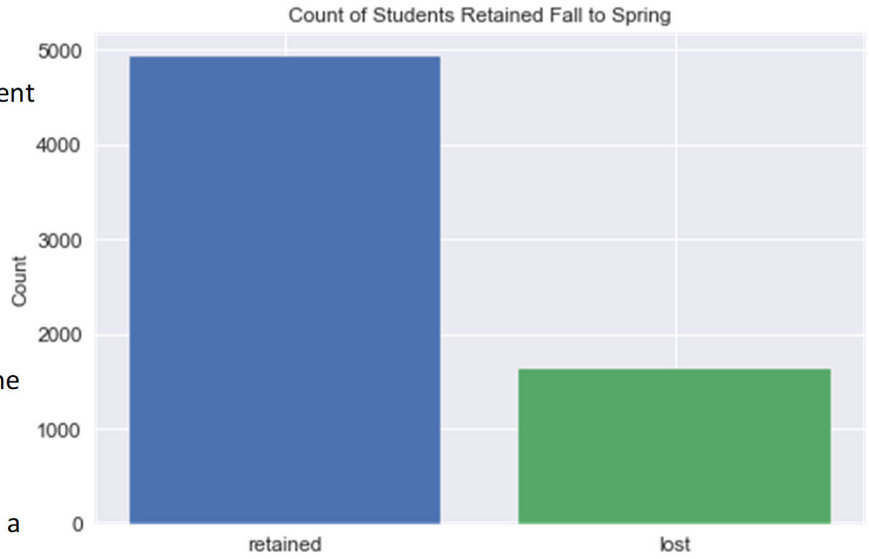

Figure 1

FEATURES

The features in this case study come from three sources:

1. Free Application for Federal Student Aid (FAFSA) responses
2. High School Transcripts
3. Student Data System

These records are all protected by The Family Educational Rights and Privacy Act (FERPA) and as such any personally identifiable information has been altered with the Python Faker package to obfuscate any connection to real individuals. Not every field was easily altered, so some data points have been left out for security issues. I am also leaving out definitions for some categorical variables for privacy reasons.

During the study, the features were split into two sets – one with only the early-obtained features and the second with all the selected features. These features are outlined in more detail in Tables 1-2. The aim of this was to address attrition prior to the end of term. When analyzing those students in real-world conditions those values will not be known until after the student is already gone.

Early-Obtainable Features	
Feature	*Description*
ADM_STATUS	Admission Status
DAD_ED	Highest level of education - Father
DISADVANTAGED	If a student is financially and/or educationally disadvantaged
EDUCATION_GOAL	Education goal stated by student
ETHNICITY	Ethnicity code
HAS_KIDS	Indicator of student having children
HDEG	Highest degree awarded
HEH	Higher Education History
HS_GRADE	High school GPA
HS_STATUS	High School Status
MOM_ED	Highest level of education - Mother
PELL	Pell Grant eligibility
TAP	TAP Financial Award eligibility
ZIP	Student Permanent Home Zipcode

Table 1

Late-Obtainable Features	
Feature	*Description*
A_CRED	Credits accumulated in the current semester.
FINAL	Overall semester grade
GPA	Overall GPA
HOUSING	Dorm Student indicator
REPEAD_IND	Indicator for courses repeated
SS_REMEDIAL	Remedial Credits
SS_REMEDIAL_CE	Remedial Credits – Continuing Education
SUB_CAMPUS	Primary Campus student is registered with
T_CRED	Total Credits

Table 2

These features were selected from a larger set by utilizing exploratory data analysis, including various plots generated with Python packages **pandas-profiling** and **autoviz**. Figure 2 illustrates a correlation matrix for the continuous variables with the target of single-semester retention generated from the autoviz package.

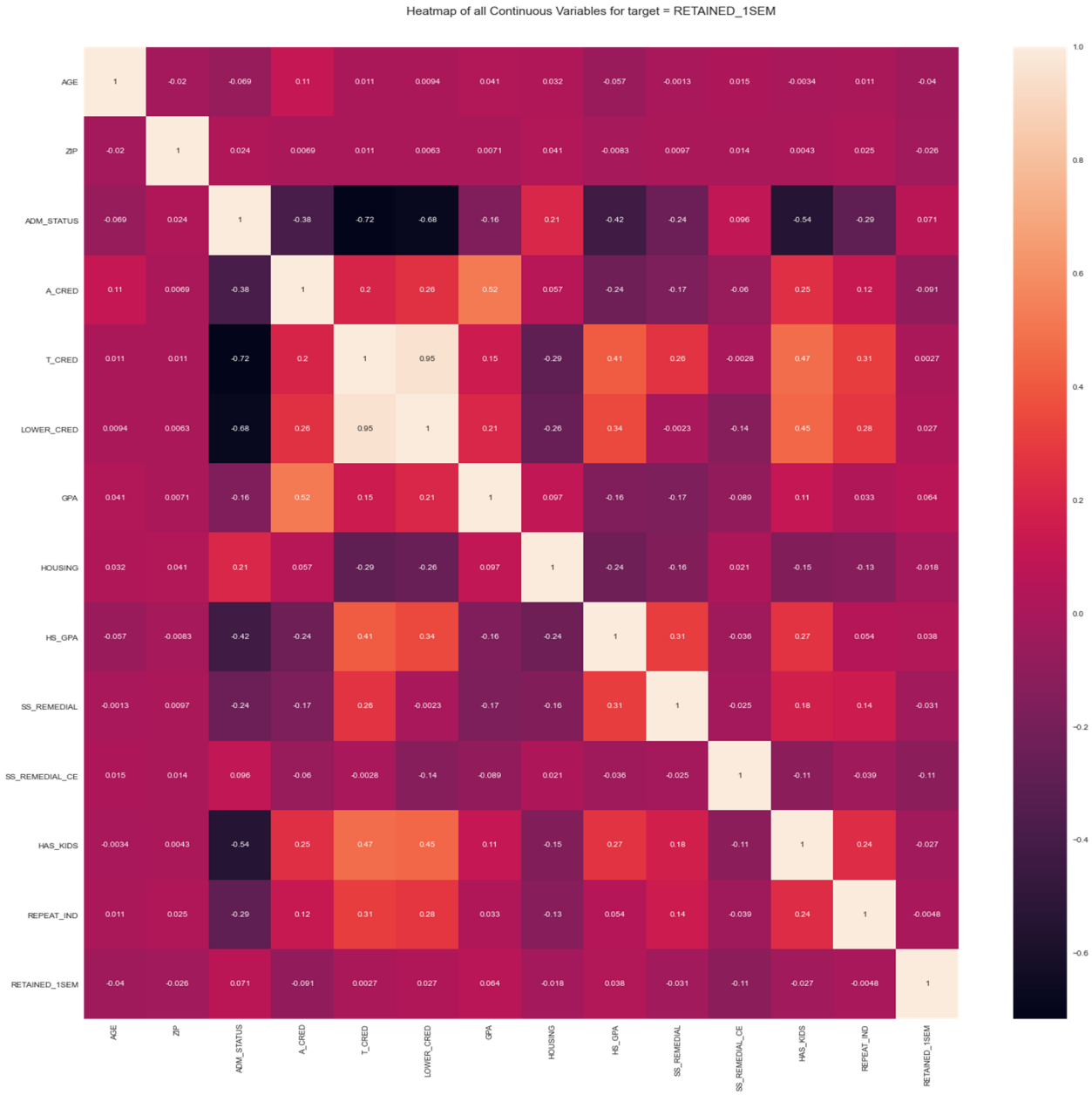

Figure 2

In addition to the correlation matrix, histograms of other categorical variables indicated certain values were associated with non-returns. The inversions shown in these charts and even the distance between the peaks may be meaningful clues to a classification model. Figures 3-6 are good examples of this phenomenon. For reference, all future charts categorize retained as 1 and lost as 0.

Figure 3

GPA – Retention more common with GPA >=3 or value zero, which indicates the value is missing. Students with no GPA are most likely taking non-credit courses through the department of Corporate and Community Education. A low GPA between 0 and approximately 2.8 dropping out is more frequent.

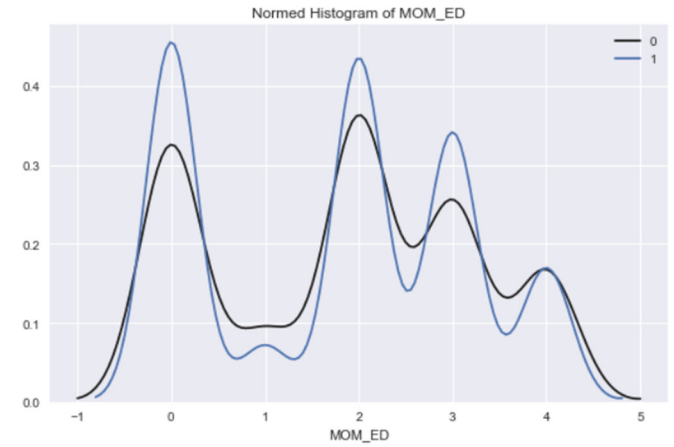

Figure 4

Mother's highest level of education, value of 1 indicates high school degree. Value 4 is Unknown, and has an even chance between retained and lost.

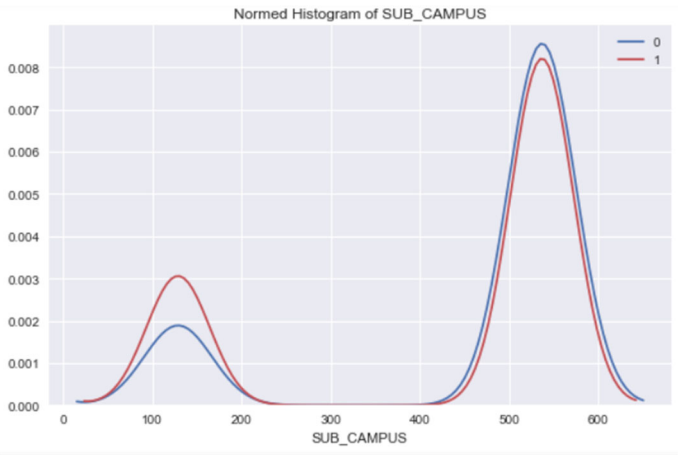

Figure 5

Sub-campus indicates which branch student is enrolled with. Many programs are only at one sub-campus, indicating that some programs are more likely to retain their students.

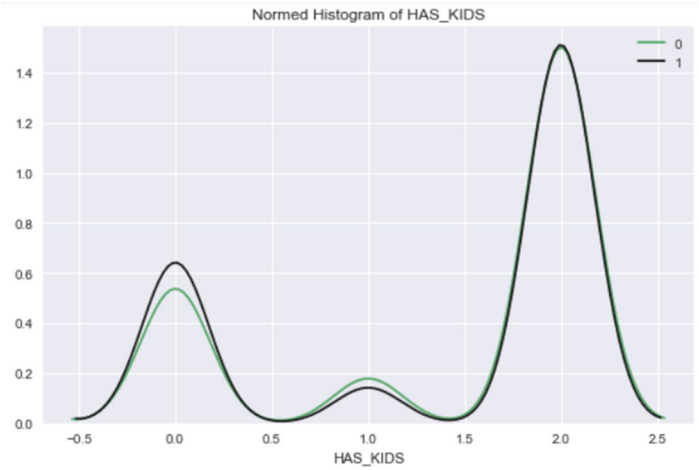

Figure 6

The HAS_KIDS value of 1 means they have one or more children, and a slightly lower rate of retention. Value 2 is Unknown and balanced between retained and lost.

With the features selected, they still required cleaning. For instance, high school GPA was stored on both the 100-point scales and 4.00 scales. For this analysis the values were all converted to the 4.00 scale. Many features were coded as integers and had to be converted to categories for analysis.

Null values were found in the following features: HDEG, HS_STATUS, HAS_KIDS, DISADVANTAGED, FINAL, MOM_ED, DAD_ED, HS_GRADE. The classification models do not accept null values as part of the input and need to be handled. Instead of dropping rows with null values and reducing an already small dataset even further I chose to impute the values using Multivariate Imputation by Chained Equations (MICE). This method will treat each missing value as the target in a linear regression. The method can create multiple imputations to account for statistical uncertainty but at this point I am using the results of a single imputation to determine the value to replace nulls with.

Categorical variables were all one-hot encoded for processing with the classification models and the data was split into training and test sets using StratifiedKFold to ensure the splits were representative of the data as there was imbalance to the classes.

MODEL SELECTION AND ANALYSIS

I chose binary classification as the method for identifying at-risk students. Predicting probability would also work for this task, but it is advantageous to have clear classifications at this point. For binary classification I selected several model types – Logistic Classification, Multilayer Perceptron, and Random Forest Classification. Tables 3-5 list the metrics for performance of each of the models when using both the early and full datasets. The models were all tuned with hyperparameter testing by using GridSearchCV to test parameter sets and choose the best=performing result.

Logistic Classification Model				
Early Dataset			Full Dataset	
accuracy	0.722222222		accuracy	0.398021309
precision	0.682042514		precision	0.623419446
recall	0.722222222		recall	0.398021309
f1	0.693464356		f1	0.411493011
f0.5	0.683858190		f0.5	0.502459310

Table 3

Multilayer Perceptron Model				
Early Dataset			Full Dataset	
accuracy	0.290715373		accuracy	0.750380518
precision	0.692163405		precision	0.563070921
recall	0.290715373		recall	0.750380518
f1	0.195004299		f1	0.857391304
f0.5	0.267165603		f0.5	0.592658755

Table 4

Random Forest Classification Model				
Early Dataset			Full Dataset	
accuracy	0.7747336377		accuracy	0.805936073
precision	0.7986161539		precision	0.802075248
recall	0.7747336377		recall	0.805936073
f1	0.7019374373		f1	0.772577866
f0.5	0.6990665759		f0.5	0.777077272

Table 5

Of particular interest in assessing the performance of model is the F0.5-measure. This allows a balance of precision and recall compared to f1. Setting the fbeta to 0.5 puts more weight on precision and less on recall. This results in a harsher penalty for a false positive. This is of value because a false negative only means we managed to keep an extra student, where a false positive is a shortfall in the budget.

The results of the model testing show Random Forest Classification to be the best predictor of student retention based on the data given to the model. The confusion matrices for the early set (Figure 7) and the full dataset (Figure 8) indicate the best outcomes from the model tuning. Both models have the most difficulty classifying

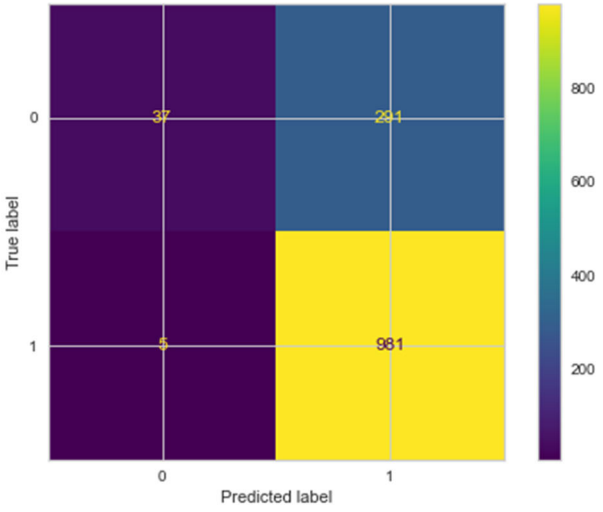

Figure 7

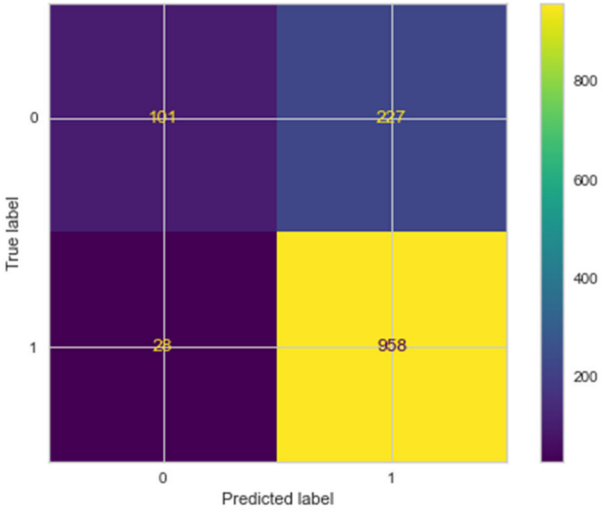

Figure 8

The classifiers also identified the importance of each feature to the models, shown in Figures 9 and 10 with plots from the Seaborn Python library.

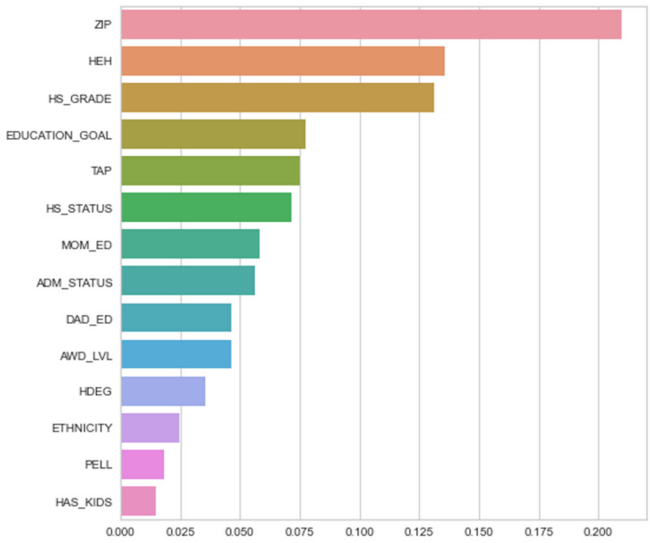

Figure 9: Early Dataset Feature Importance

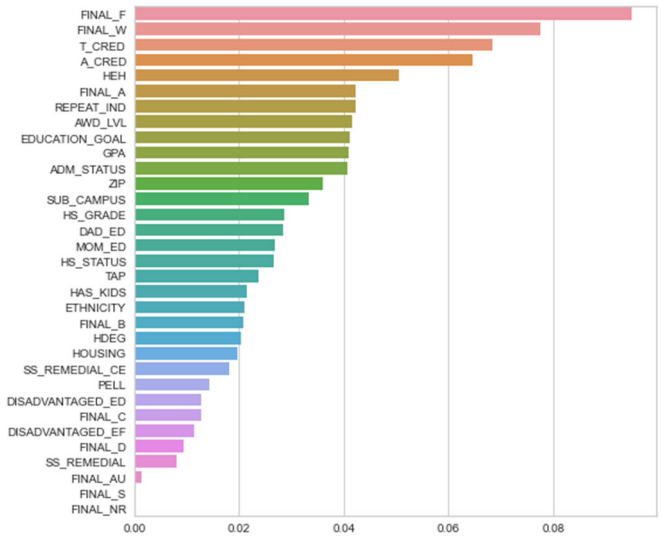

Figure 10: Full Dataset Feature Importance

SUMMARY

The model is fairly accurate, although false positives are more common than false negatives. An accuracy rate of ~80% might be sufficient for identifying students that need additional assistance. The features identified make sense in both sets. HEH and EDUCATION_GOAL relate to how long a student has been in school and how long they plan to stay. High school grade based on GPA and mother and father's highest level of education are also features with some value in the early data. Several of these variables indicate academic readiness to be an important component in predicting retention.

The important features in the early set are joined with information about grades and credits in the full set. Both high and low grades have an impact on retention, and the amount of credits obtained may indicate the anticipated graduation terms of students as they reach their credit requirements.

I recommend checking the model against real-world data early in the semester to identify possible at-risk students so advisors can keep an eye on their individual needs. This early intervention may lead to decreased attrition and outcomes can later be tested against baseline predictions for accuracy.

This model is a good place to start, but it can only classify students. At this stage there is no indication as to what intervention methods may change outcomes. I suggest the data is explored further, as there are hundreds of features in the FAFSA set alone that may help with measuring socioeconomic impacts on students that may be otherwise hard to capture. There may also be value in looking at the sequence that courses are taken and even which instructors were assigned to the course. Looking at student types separately might also improve predictions. First-time students are unlikely to have their total credit count affect their return for a subsequent semester, while a student with nearly enough to graduate is unlikely to come back for a term beyond that necessary to finish. Clustering methods may be able to find commonalities in students mis-classified as being retained that can in turn fine-tune the model and identify points where students need more help.

Understanding the causes of attrition and finding ways to ensure student success will not only support our mission but improve financial outcomes in a time of depressed enrollment. Identifying students most at need of assistance may help with handling the resources available in a budget that necessitates a scarcity mindset.

CHALLENGES AND LIMITATIONS

The data only comes from a single Fall semester and would not capture students who are taking classes only offered in Spring semesters. Additional data may improve the accuracy of the model but studying the data while maintaining FERPA compliance required obfuscation steps that added to collection and transformation time. There was also a certain amount of imprecision required when describing some of the variables and their values for the purposes of protecting student privacy. Future research performed internally would face less restrictions and allow more flexible feature selection and evaluation.

REFERENCES:

Giacomelli, E. (2020, October 11). *SUNY-wide enrollment down for 10th year, but figures 'far better' than expected amid pandemic*. NNY360. https://www.nny360.com/communitynews/education/suny-wide-enrollment-down-for-10th-year-but-figures-far-better-than-expected-amid-pandemic/article_a26f2cb6-5aad-5328-a769-f67d30704a0c.html.