

Undergraduate Student Retention, Can we catch them before they fall?

Lohit Bhandari, Oklahoma State University

ABSTRACT

Educational institutions are very concerned about the retention of students in their undergraduate programs. Until few years ago, the institutions focused only on storing and maintaining student data without utilizing it for making data informed decisions. With the widespread availability of advanced analytics tools, we can derive useful insights from such datasets. The datasets used in this study include variables relating to demographic, admissions, financial, student employee and academic information. The Institutional Research and Information Management department at the Oklahoma State University has provided the required datasets. The study focuses on the factors that impact retention of first year freshman students from first to second semester. The study takes a particular interest in how financial support influences students' decision to continue their studies in the undergraduate programs. The financial support includes the amount of financial aid provided to the student as loan or a scholarship grant, and the availability of the on-campus employment opportunities to those without the financial aid. The technical approach is to cleanse, transform and consolidate the datasets in Base SAS®. The final output dataset is imported into the SAS® Enterprise Miner. The significance of the above-mentioned financial indicators are assessed using the predictive modeling techniques available in SAS Enterprise Miner. This would help in improving the quality of the factors that affect the retention of students in higher education. Students, faculty, and program administrators all stand to benefit from such insights.

INTRODUCTION

The paper focuses on the first-time freshmen students in the Oklahoma State University. First-time freshmen are the students who have chosen Oklahoma as their first choice for an undergraduate college degree, without attending any other universities prior to this. The other types of students could be those transferred from other universities. The college would like to retain every student, but by focusing on this subset of students, it can minimize the other comparison factors that could play a role in a student's decision to stay in or drop out of school. The study would like to understand what specific factors prevent student from completing their degrees and dropping out or transferring to other universities.

The school has tracked such students for only the Fall-2017 and we have used this list to compare it with the enrollment data in Spring-2018 to see if the student has enrolled again or not. Students with no enrollment in the Spring-2018 are considered as dropouts. The focus is on understanding the significant factors that can help account for such dropouts.

DATA PREPARATION

The Institutional Research at Oklahoma State University maintains records of the students through different applications. Any personally identifiable information was de-sensitized from the provided data sets by masking. Since first-time freshmen indicator were tracked as part of a recent effort, the list of such students has been provided as a separate list manually. Each student can be identified by the masked Student ID value, and will serve as the key field to join the several data sets to obtain the factors used to build the predictive model.

The data sets provided include first-time freshmen list, enrollments data – fall2017 & spring2018, GPA data – fall2017, financial aid – fall2017 & spring2018 and student employee hours – fall2017. is given below.

Figure 1 shows an overview of the datasets used and the master table that is used to join with them.

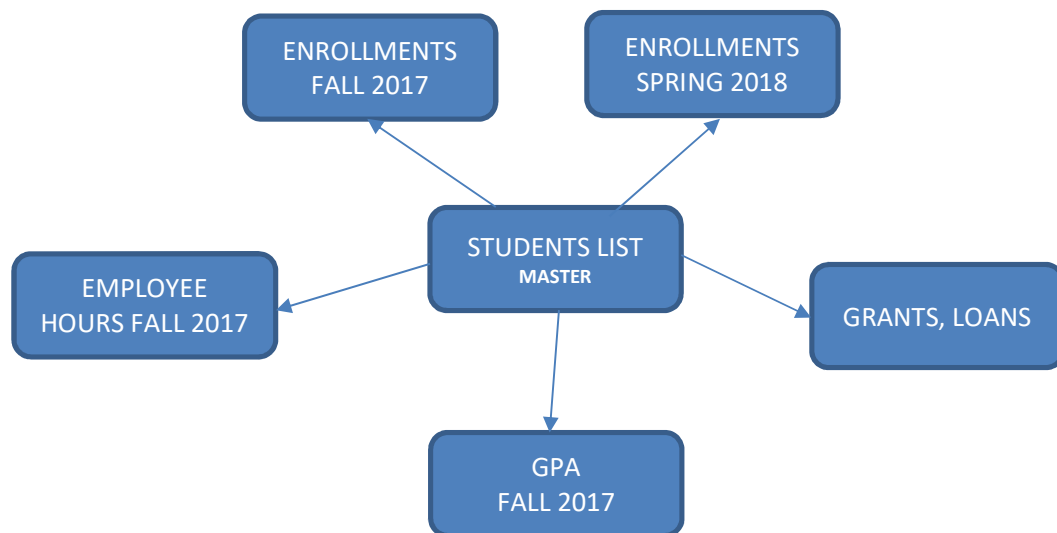


Figure 1. Overview of the Data Sets Used

The tables are joined and the required columns are taken from each of the datasets. An overview of the variable list is provided below.

DATA OVERVIEW

Table 1 shows the list of variables used.

Variable	Description
STUDENT_ID	Unique Identifier for each Student
GENDER	Gender
ACT_ENGLISH	ACT Score in English
ACT_MATH	ACT Score in Math
ACT_READING	ACT Score in Reading
ACT_SCIENCE	ACT Score in Science
ACT_COMP	Composite ACT Score
IRIM_CITIZENSHIP	Country of Citizenship
IRIM_FIRST_GEN	First generation student flag
IRIM_MINORITY	Minority status flag
IRIM_STATE_OF_RESIDENCE	State of Residence in USA
MARITAL_STATUS	Marital Status
CURRENT_AGE	Age of the student
RETAINED	Binary Target Variable. Indicating Student Retention
GPA	Overall GPA for Fall 2017
GPA_CREDITS	Total Credits for Fall 2017
QLTY_POINTS	Quality points for Fall 2017

Variable	Description
GRANT_SCHOLARSHIP.FALL	Amount of Grants/Scholarship in Fall 2017
LOAN.FALL	Amount of Loans in Fall 2017
GRANT_SCHOLARSHIP.SPRING	Amount of Grants/Scholarship in Spring 2018
LOAN.SPRING	Amount of Loans in Fall 2018
TOT_HRS_WORKED	Total Hours worked in the Fall 2017

Table 1. Variables used from the source datasets

MISSING DATA IMPUTATION

Some of the predictive models can only use complete data, hence the observations with missing data is excluded from modelling. To avoid this and make use of all the available observations we need to impute the missing values with appropriate values.

Some of the observations have the values missing for grant, scholarship or loan fields, since some students were not awarded any financial aid. So the missing values for these variables is imputed to zero.

The variables GPA and ACT scores have some missing values. We can impute them with the median values of the respective column. Median values are more reasonable to use compared to mean or any other value. Assigning median values will also not skew the distribution.

The citizenship values for 65 entries are missing in the data set. The value 'XX' is assigned.

FEATURE ENGINEERING

Feature engineering is an important step in the model. It is the process of introducing new columns based on the existing columns with the use of subject knowledge. This can sometimes help to reduce the dimensionality, which can greatly improve the model performance.

In the source files, we have the data pertaining to Grants, Scholarships and Loans. The primary difference between a grant/scholarship and a loan is repayment. A loan requires you to repay the money you borrow, whereas a grant does not. Grant and scholarships are free money disbursed to students by the institution or the government. Instead of introducing these values as is, we would like to look at the change in the loan amount, an increase or same amount of money in both the semesters may be an added incentive for a student to continue his studies. On the other hand, a decrease in the grant/scholarship amount might cause the student to rethink his options to continue his studies.

A column called Change_In_Total_Aid is created. This field captures the cumulative change in grant and loan amount in spring compared to fall. If the sum amount of grant and loan increases for a student this column will be a positive value and if the grant value decreases in spring then it will be a negative value and if the grant amount remains same then it is equal to zero.

Four more flag variables called Has_Spring_Loan, Has_Spring_Grant, Has_Fall_Loan and Has_Fall_Grant are created. A value of one would indicate that the student has a loan or grant provided in the respective semester. This columns' significance in a model will indicate that a student can be retained with a nominal loan or grant for a semester, irrespective of the change being higher than earlier semesters or not.

Next, we have data for the GPA and the total number of credit hours taken by each student. The GPA signifies the capability of the student in all the courses he registered for the fall semester. Moreover, we can also use Quality_Points column. Quality points are the cumulative points used to calculate your GPA or grade point averages. Only classes receiving grades of A, B, C, D, or F are used to calculate the GPA. The quality points will take into account the total number of credits and GPA for those credits.

There are many distinct Age values in the data set. We expect that the normal student age would be 18-19 years in the first year of undergraduate. We can create a flag variable for this purpose, indicating

whether the age is less than or equal to 19 or not. “LE 19” indicates 19 or less and “NE 19” indicates greater than 19 years.

The study would like to focus on the availability of on-campus employment opportunities as a financial aid to students, but with the unavailability of employee hours for majority of the students, it has been removed from the analysis.

MODEL BUILDING IN ENTERPRISE MINER

The final data set created above is imported into SAS Enterprise Miner. There are many predictive models available to model the binary target variable. However, to make the results interpretable, only the decision tree and the logistic regression models were considered for analysis in this study. There are 3835 observations available in the data set. 70% of the data is used for training the model and 30% of the data will be used to validate the model.

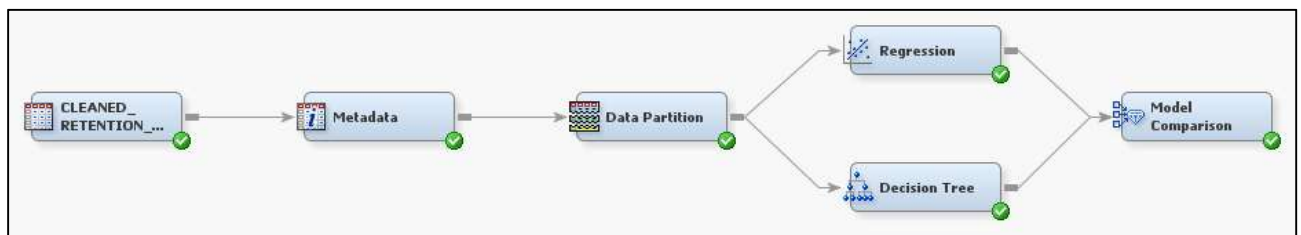


Figure 2. Screenshot of the Enterprise Miner Diagram

Then this partitioned data is given as input to the Regression Node and Decision Tree node. The regression node is modified to use stepwise model selection and the selection criterion is based on the validation misclassification rate. The Decision tree node setting is also updated to select Misclassification as the assessment measure.

In the initial run all the demographic variables, ACT scores and financial aid factors were included. However, the variables Grant_Scholarship_Fall, Grant_Scholarship_Spring, Loan_Fall and Loan_Spring were removed since new columns, Change_In_Total_Aid , was created as mentioned above and captures the information of the source columns.

The output of the Regression node and the Decision Tree node is given to the Model Comparison node. The Decision Tree Node has been selected as the best model based on it's lower misclassification rate as shown below.

Fit Statistics			
Selected Model	Model Node	Model Description	Selection Criterion: Valid: Misclassification Rate
Y	Tree	Decision Tree	0.026933
	Reg	Regression	0.030408

Figure 3. Results of Model Comparison Node

The Table 2 shows the confusion matrix on the validation data set based on the decision tree model.

	Outcome_0	Outcome_1
Target_0	60	29
Target_1	2	1060

Table 2. Decision Tree Confusion Matrix

The misclassification rate is calculated as the sum of false positives (29) and false negatives (2) for the validation data set of 1151 observations. Thus in this case it is $(29+2)/(60+29+2+1060) = 2.7\%$

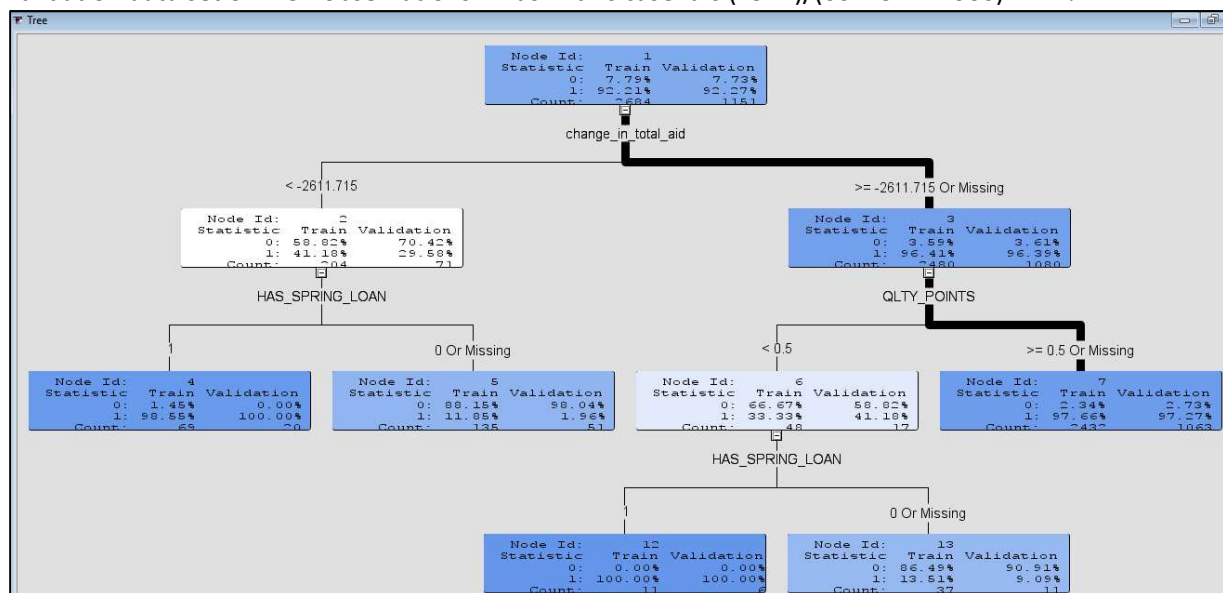


Figure 4. Decision Tree Model Result

The decision tree reveals the importance of LOAN_CHANGE, QLTY_POINTS and ACCT_BALANCE indicator

Variable Importance						
Variable Name	Label ▼	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance	
change_in_total_aid		1	1.0000	1.0000	1.0000	
HAS_SPRING_LOAN		2	0.8409	0.7809	0.9286	
QLTY_POINTS		1	0.5819	0.4228	0.7265	

Figure 4. Variable Importance in the Decision Tree Model

The decision tree clearly indicates that the most important variable is Change_In_Total_Aid. This confirms that the financial aids such as grants, loans and scholarships do effect the retention of the freshman students. The next significant variable is Has_Spring_Loan, which suggests that the availability of loans to fund their studies is very important. It can be a decisive factor for a student to continue studying at the university. The next significant variable is Qlty_Points. We can observe that students who do not take many credits at the university are not invested in the program and are more likely to not continue with school.

CONCLUSION

The paper discusses the process of building a binary classification model for student retention. The results of the model can be improved by collecting more data for the upcoming semesters and comparing the results to find insightful patterns. Predictive models are powerful in helping us better understand patterns from complex data. The decision tree model has proved useful in identifying financial aid as significant factor in retaining students at the university. It has shown that freshmen students can be retained with not just grant and scholarship money but also providing loan opportunities. Future research can look into the magnitude of financial aid that affects retention, that can help in better management of student loan and grant disbursement.

ACKNOWLEDGMENTS

I would like to thank my professors who have guided and motivated me for completing this paper:

Dr. Goutam Chakraborty
Director, Master of Science in Business Analytics
Oklahoma State University

Dr. Miriam McGaugh
Professor, Oklahoma State University

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Lohit Bhanadari
Oklahoma State University
Lohit.bhandari@okstate.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.