5th International Conference on Higher Education Advances (HEAd'19)
Universitat Politècnica de València, València, 2019
DOI: http://dx.doi.org/10.4995/HEAd19.2019.9241

The Factors Affecting University Retention/ Attrition by Big Data Analytics

Richard K. Cho and Dongmin Kim
Faculty of Business, University of New Brunswick Saint John, Canada

Abstract

Using the enrollment data at the Faculty of Business (FOB) in the University of New Brunswick, Saint John (UNBSJ), we perform the big data analytics to examine the cause of attrition: a) the existence of potential risk groups and b) the potential courses, which can be the predictors of student attrition in the first few years in the university. The logistic regression was used to find the potential predictors for students' retention in UNBSJ, and the cluster analysis also suggests the existence of inherently high-risk groups in UNBSJ students. By providing institutional support for the high-risk groups to successfully complete the program, the retention rate could be improved..

Keywords: *University Retention / Attrition; Big Data Analytics; Logistic Regression; Cluster Analysis; Identifying High-Risk Group.*

1. Introduction

Student retention, which is keeping students until graduation, is one of the strategic focuses in the Canadian University. In Canada, the average dropout (attrition) rate after first year in University was 14% and the overall post-secondary dropout rate was about 16 % (Freeman, 2009)[1]. According to Bean's (1980) review of the previous retention rates, one research reported the median of 50% loss of students in 4 years in the U.S. and another research showed 41.5% attrition in 1966, and the similar rates were shown in Canada, England and Australia.

There are various dropout reasons: to transfer to other institutions or programs, the financial reasons, RTW (Request to Withdrawal) due to the low GPA (Grade Point Average), the lack of interest or finding the limitation to continue the desired major. Survey says that the dropout students were struggling with meeting deadlines, academic performance and study behaviour in their first year, and many of them thought of leaving in their first year (Freeman, 2009). The same article also mentioned that the less preparedness in first-year students continues more strongly in the internet-oriented age.

While there are numerous studies that examined the causes of the dropout in post-secondary institutes, their main focuses are on the characteristics of the dropout students and the perceptions of the dropout students about the institutional support such as commitment, quality and the university governance styles (Bean, 1980; Tinto, 2000). Instead of attributing the dropouts to the personal characteristics and/or institutional systematic problems, our research focuses on identifying the key courses, which can serve as predictors to student's retention.

In order to find the predictor courses in the University of New Brunswick, Saint John (UNBSJ), we collect the demographic and enrollment data for 7 years and the graduation list in the Faculty of Business (FOB) from the Registrar's office. Our assumption is that the failure or poor performance in a certain course(s) makes students frustrated and results in their dropping out.

Our research questions are (a) What courses in the first year can be a predictor of students' attrition after the first year? and (b) Are there any groups requiring a special attention in the first year?

After figuring out key courses and special groups in need, Gwinnett Education Division developed the support program and made a big success in their program (LaValle *et al.*,

[1] The post-secondary education includes University, College, Polytechnic, Apprenticeships and Private Vocational Colleges.

2011). We expect the similar effect in UNBSJ and wish to implement the support programs for the vulnerable groups to succeed in key courses and untimately to graduate.

2. Literature Review

Gwinnett County, GA, is one of the successful anecdotes that the analytics using the big data increases the effectiveness of education. After allocating more resources to help the students in need in the focused areas predicted by analytics, the academic performance and the graduation rate were remarkably improved. (LaValle *et al*., 2011)

There have been studies about the university retention for the last half century. The theories behind the university attrition or dropout were well discussed in Bean (1980), Braxton (Editor: 2000) and Tinto (2012). Bean (1980) was cited widely because he performed the empirical research about his conceptual and causal model for student attrition. He found the potential causes for attrition: For female students, three variables were statistically significant in explaining dropout: institutional commitment (-.47), institutional quality (-.11), and routinization (.10). The numbers in the parenthesis are the regression coefficients to the dependent variable "dropout". For male students, four variables were significant: institutional commitment (-.29), routinization (.15), satisfaction (.14), and communication (rules) (-.13). The common causes are the low institutional commitment (or loyalty toward organizational membership) and the high routinization (or repetitive role view about students). But, there are some gender differences about dropouts. From the path analysis, he concluded that institutional quality and opportunity (transfer) were the two most important variables influencing institutional commitment.

Tinto (1975) synthesized the previous research and revised a theoretical model later (1993). He explains the effective retention program as the utmost commitment to all their students and the development of supportive social and educational communities. Especially, he emphasizes the first year as the transition period to college in both social and academic structure. For the smooth transition, the university needs to assist the first year students including monitoring and early warning, and counseling and advising. Tinto (2000, 2012) again wrote the book about refining and rethinking of the college education. He emphasizes the first year experience again and he points to the classroom as the center of student education and life, and therefore the primary target for institutional action.

As a matter of fact, many white papers are available about the admission process (that is, selection of students) and how to improve the enrollment rate from admitted students. For instance, Henschen (2013) and Information Builders (2013) report that Taylor University in Indiana analyzed the 12 years student data and found the strategy to maintain 85 % student retention rate.

3. Methodology

3.1. General procedure

The data for the Faculty of Business in UNBSJ was collected directly from the Registrar's office. The data includes the following information for each semester from 2006 to 2012:

- Personal: Student ID, Gender, Birth Date
- Academic: Degree, Major, Load (Fulltime?), Current Year of program
- Demographic: Citizenship Country and Province, last High School
- Credit-related: CGPA (Cumulative Grade Point Average), Registered Credit Hours, Transferred credits
- Course-related: Course ID, Course level, Credit hours, Final Grade

After several steps of refining the data, we filtered out 7 key courses with at least 90 data. For 483 students remained after removing students who didn't take those 7 courses[2], we applied the retention/attrition result to each student. Using this database, we perform the Correlation Analyiss and the Logistic Analysis for those courses to figure out the potential impacting courses and do the Cluster Analysis to find the potential subgroups in each course.

3.2. (Binary) Logistic Regression

The logistic regression is a tool to analyze the binary dependent variable (p) with respect to the continuous (interval) independent variables. The multiple regression predicts the success probability of p in [0,1] range. To guarantee the predicted value of p in [0,1], we need to modify the regression formula as follows (it is also called a sigmoid curve): $p = e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)}/[1+e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)}]$.

The cut-off value, c, is used to classify the observation into 0 or 1. If the predicted value from an observation is above c, it would be classified to 1. Otherwise, it would be 0. The cut-off value is determined to maximize the overall accuracy of prediction. In our analysis, the retention status is a dependent variable, and course GPA taken during year 1 and 2 and other demographic data are independent variables (or predictors). We will divide valid dataset into the model (70%) and evaluation (30%) to find the predication power, as mostly used in big data analytics.

3.3. Cluster Analysis

This study also uses to figure out the similar subgroups (called a cluster) in the students taking the same course(s). The purpose of this cluster analysis is to find the subgroup in high risk

[2] The course number was modified to shield the confidentiality.

and implement to a support program to help those student groups. The hierarchical cluster also helps to find the important variables to cluster the data into groups.

4. Results

4.1. Correlation Analysis

For all the data including RTW students, we performed three diffrent correlation analyses: (a) Retension or CGPA with key demographic variables, (b) CGPA vs. key courses, and (c) correlations among course GPAs. Those results are summarized in Table 1(a). Note that the CGPA is the major factor to associate with all courses and with most of demographic variables.

In order to figure out the relationship among variables and courses for retained students only, we performed similar analysis after excluding the RTW students, whose results are in Table 1(b). Since RTW is for less than 2.0 CGPA, we eliminate 111 (or 23%) of those students and have a reduced data set (n = 372) at this stage. Although the correlations of Retention to CGPA or NoMajor are reduced from the value at the previous stage, they are still significant at 0.01 level. The variable “Male” and “International” become more important to Retention.

Table 1. Results from Correlation Analysis

(a) Data including RTW students (n = 483)

	CGPA	**NoMajor**	**International**	**FullTime**	**Male**	**Trans Credit**	**StartAge**
Retention	.599**	-.385**	0.032	-0.054	-0.073	.205**	.097*
CGPA		-.308**	-.129**	-0.067	-.221**	.250**	.220**
	BA-131	**BA-170**	**BA-160**	**BA-260**	**BA-270**	**BA-222**	**BA-231**
CGPA	.711**	.607**	.570**	.583**	.562**	.706**	.611**
	BA-131	**BA-170**	**BA-160**	**BA-260**	**BA-270**	**BA-222**	**BA-231**
BA-131	1						
BA-170	.551**	1					
BA-160	.451**	0.103	1				
BA-260	.566**	.395**	0.135	1			
BA-270	.573**	.547**	0.327	.501**	1		
BA-222	.630**	.787**	0.503	.619**	.582**	1	
BA-231	.577**	.421**	.587*	.723**	.424**	.638**	1

(b) Data without RTW students (n = 372)

	CGPA	**NoMajor**	**International**	**FullTime**	**Male**	**Trans Credit**	**StartAge**
Retention	.204**	-.280**	.115*	0.024	.103*	.108*	-0.006
CGPA	1	-.152**	-.147**	-0.026	-.161**	.229**	.209**
	BA-131	**BA-170**	**BA-160**	**BA-260**	**BA-270**	**BA-222**	**BA-231**
CGPA	.686**	.585**	.536**	.613**	.496**	.766**	.657**

(** 0.01, * 0.05 statistical significance)

The CGPA is still very clasely related to the key courses. However, The logistics analysis and cluster analysis are not useful tools in the stage of no RTW students, because the sample is already filtered only to passed students, and thus the retention rate is all high.

4.2. Logistic Analysis to predict retention (with including RTW students)

Because of the dominant effect of CGPA for all models, we have a difficulty in analyzing the effect of each course GPA. As a matter of fact, the CGPA has the highest correlation with retention and it absorbs the effect of each course on retention. Hence, we need to remove the variable "CGPA" to see the effect of each course. The results are shown in Table 2.

Note that the "NoMajor" variable is now important because this variable is another proxy for CGPA, but it is better to keep the "categorical" variable in the model, instead of numerical variable "CGPA". Compared to other course models, the one with "CGPA" shows the better prediction power and the higher coefficient of GPA part.

The coefficients of logistic regression equation are found from $p = \frac{e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)}}{1+e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)}}$.

For example of BA-131, logit(p) $= ln\left(\frac{p}{1-p}\right)$ = - 0.126 + 0.866 * [Course_GPA] + 1.582 * [International] - 2.665 * [NoMajor] + 0.003 * [TransCredit]. The risky students group here is students with low Course_GPA, domestic and no_major students. While the international student with Course_GPA=3.0 having a "Major" have a retention probability (that is, p) of 98.3%, a domestic student with Course_GPA=2.3 and "No Major" have a 31.0%. Although the CGPA is critical part of retention from correlation analysis, a new finding from this analysis is the importance of "Declaring Major" in early stage of students' university life. Hence, we can find the risk group from the analyzed coefficient for each course.

Table 2. Logistic Analysis (with including RTW students)

Model	BA-131	BA-170	BA-160	BA-260	BA-270	BA-222	BA-231
Constant	-0.126	0.187	-2.21	0.129	19.406	-4.978 *	16.79
Course GPA	0.866 **	0.559 **	1.449 **	0.913 **	0.607 *	2.084 **	
FullTime						1.832 *	
International	1.582 **	1.297 **	1.58 **	1.888 **		2.927 **	
Male							
NoMajor	-2.665 **	-2.485 **	-2.503 **	-2.77 *	-20.311	-3.195 *	-20.816
StartAge							0.216 *
TransCredit	0.03	0.03 **				0.042 *	-0.035
n	209	189	171	138	111	101	98
% correct selected	81.4	74.6	75.8	81.4	79	82.7	76.5
% correct unselected	75	72.9	69.8	73.2	76.7	80.8	50

Tabla 3. Cluster Analysis (with including RTW students)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Size	0.288	0.244	0.238	0.23
Description	International students with low CGPA	Domestic students with (very) low CGPA	Domestic students with high CGPA, no-major	students with high CGPA, and major
Have major	0.014	0	0	1
International	1	0	0.009	0.234
CGPA	2.26	1.91	3.02	2.99
StartAge	22.5	21	24.9	22.8
Retention	0.568	0	0.965	0.973

4.3. Cluster Analysis (with including RTW students)

The high-risk group is also identified by the cluster analysis with all the demographic data and CGPA without specific course GPA's. Table 3 shows four different groups, which can be explained in the decision tree as Figure 1.

We also did the cluster analysis for each course, and somewhat different sub-groups were verified, with respect to different thresholds of NoMajor, International, FullTime and Retention. Although the detailed group information for each course was not presented in this paper, we can figure the high-risk groups out from students population, and develop the supportive program to help those groups.

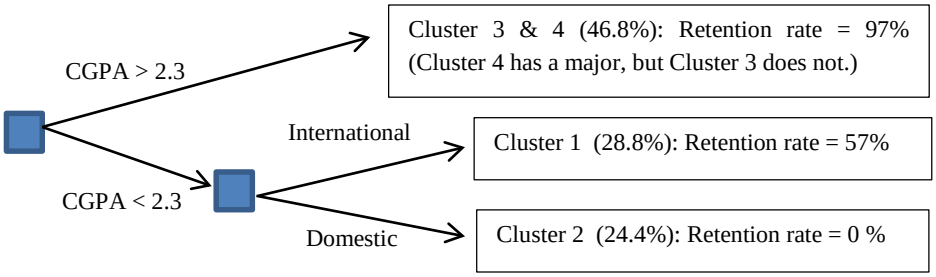

Figure 1. Decision Tree from Cluster Analysis

5. Conclusions

From the logistic regression and the cluster analysis using the FOB enrollment data, we figure out the predictor for retention. We expected to find the courses to predict the student's retention, but the results showed no valid courses to predict. We found one course, BA170, with 10% significance level, but the power to predict is now doubtful due to the high correlation with CGPA. Hence, we perform the logistic regression analysis without CGPA, which shows the importance of declaring his/her major in early stage of university life.

From the cluster analysis, we successfully identify the high-risk groups: the low GPA domestic students directly from high school, and the low GPA international students. By carefully designing the sequence of courses, we may find the right time to help those high-risk students develop their potential abilities to continue their university studies.

It may be worthwhile to say the difference between dropouts of Bean (1980) and UNBSJ case: whether it is voluntary dropouts or RTW (Request to Withdraw) due to the low CGPA: in other words, attitude or study habits. However, by reinforcing the institutional effort, it may be possible to motivate students study hard and to increase their CGPA from their first or second year in university. By focusing on helping the students of CGPA between 1.6 and

2.4 (out of 4.3 scale) and on setting up the retention target for specific CGPA groups, we can improve the retention rate by 6%.

References

Bean, J. P. (1980), Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Research in Higher Education*, 12(2), pp. 155–187.

Braxton, J.M. (Ed.) (2000), *Reworking the student departure puzzle*. Nashville, TN: Vanderbilt University Press. (https://books.google.ca/books?id=WF8itWof7aIC)

Freeman S. (2009), 1 in 6 first-year university students won't make the grade, *Toronto Star*, Sep. 20 2009 (accessed from http://www.thestar.com/news/canada/2009/09/20/1_in_6_firstyear_university_students_wont_make_the_grade.html).

Henschen, D. (2013), Small University Graduates To Advanced Analytics, *Information Builders Magazine*, 22(1), p. 43.

Information Builders (2013), *Taylor University Drives Student Retention With Predictive Analytics*, (accessed from http://www.informationbuilders.com/applications/taylor-university).

LaValle, S., E. Lesser, R. Shockley, M.S. Hopkins and N. Kruschwitz, (2011) "Big Data, Analytics And The Path From Insights To Value", *MIT Sloan Management Review*, Winter 2011, pp. 20-31.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, 45(1), 89–125.

Tinto, V. (1993). *Leaving College: Rethinking the Causes and Cures of Student Attrition*, University of Chicago Press.

Tinto, V. (2000) *Linking Learning and Leaving: Exploring the Role of the College Classroom in Student Departure. Reworking the Student Departure Puzzle* (Ed. by J. Braxton), Vanderbilt University Press, pp. 81-94.

Tinto, V. (2012). *Completing College: Rethinking Institutional Action*, University of Chicago Press. (https://books.google.ca/books?id=zMEy9V4BqDAC)