Business Analytics

Student Dropout Analysis Based on Previously Acquired Educational Achievements: An Application to the case of Riga Technical University

Submitted to: Professor Ilze Birzniece

Submitted by: Vânia Silva
Moeid Ahmed
Rafael Fernandes

Contents

PART I – INTRODUCTION

1.1 Problem Statement

Choosing an educational establishment, deciding to attend university, majoring in which academic discipline, leaving one educational establishment for another and early dropouts are few instances of the issues which are being faced by the educational researchers and policy makers while addressing the issues and problems of higher education institutions and universities.

The capability of an institution to deal with problems like that of student retention and enrollment certainly lays in the fact that how well a researcher can answer some of the questions addressed in the beginning.

Educational outcome is a combination of many factors pertaining to that of student characteristics, their mental caliber and of the educational institution itself. Secondly, most of the educational outcomes are of dichotomous nature and no interval or ratio scales are available to quantify them. For example, either a student will attend the university or not, obtains the degree or not, completes his major or not etc.

Although many techniques, tools and frameworks are now available which can help us to identify which factors are more relevant to a particular educational behavior under study, but the challenge of quantifying the effect that these factors may have has been constrained by the nature of the dependent variable under the study. Although many techniques are now available but only few of them can be applied successfully to cater the needs of dichotomous nature of variables such as that of degree completion and enrollment. Few such methods which can be employed in such a state are: structural modeling for binary dependent variables, log linear analysis, probit and logistic regression functions.

1.2 Objective and purpose of study

A study by Lassibille & Gomez (2007), confirmed that enrollment age, previous education attained before entering in university, scores and marks of pre university examinations and whether a student is living in the same town as that of his or her unversity, all have a significant impact on the drouputs.

In apropos with the discussion above, this study examines the situation at Riga Techincal University, Latvia, centered around the scenario of students' previously gained mathematical knowledge because of the vast application of mathematics in technical studies and its effect on the educational quality and student dropouts as part of the reforms introduced in educational sector which makes it obligatory for all the student at secondary level to take centralized exam (CE) in math which also serves as one of university entrance exams. A normal distribution curve of school leaving CE is being shown in Figure 1.

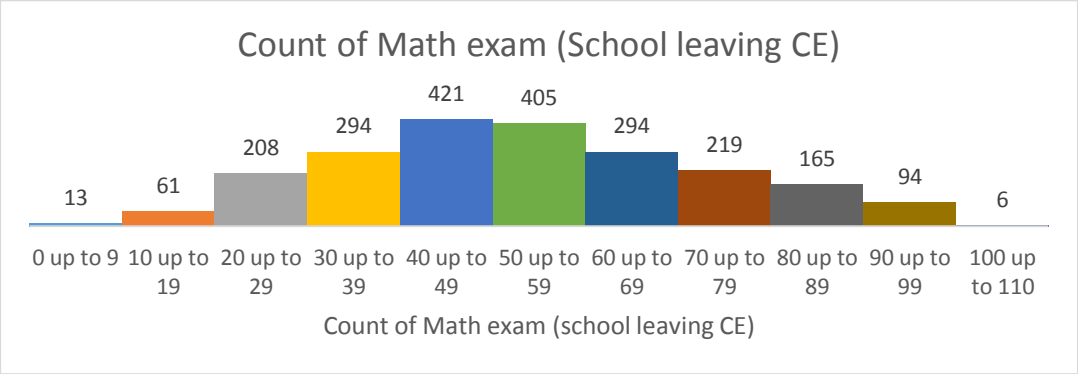

Figure 1. normal distribution curve of school leaving CE

Much of the previous research have showed that investing in the quality of education reaps more benefits than in the quantity alone[1]. Historically many countries have been observed to spend more towards educational quantity than the quality. Generally they are observed to increase their enrollment ratios while giving less importance to improving indicators which contributes towards educational quality such as that of student teacher ratio which ultimately can lead to a better understanding of the concepts and of subject matter at hand by a student[2].

[1] For example, Behrman & N.Birdsall, (1893) ; Card & Krueger, (1992). Quality of schooling in these studies have been quantified by student-teacher ratio, average period of schooling of teachers, amount the teachers are being paid and the length of the term.

[2] Few of them increased their enrollment rates over the period of 1980-95. They also depicted an increase in their allocation of expenditures related to education in GDP and an increase in their pupil teacher ratio. Though Iran showed a decrease in their educational expense in GDP, its enrollment and student-teacher ratio showed an increasing trend.

PART II – PROPOSED FRAMEWORK FOR METRICS AND KPI, DATA EXPLORATORY ANALYSIS AND MODEL BUILDING PREFENCES

2.1 Framework for Metrics and Formulating KPI

2.1.1 Economic, Functional and Political theories on access to Education

Theories regarding increasing access to basic education across different countries and nations from a wider-perspective – sociological, economic, political and religious – have been developed overtime and can provide us with valuable information regarding global expansion of education. First, according to the economic perspective the increase in demand from the labor force to gain new skills for building up of national economies has led to the phenomenon of mass schooling; Clark,(1961); Harbison & Myers,(1964). Second, functional theory suggested that problems and issues of social stratification can be solved with the help of increasing educational quality by providing the elites with a competition; Bowles & Gintis, (1976); Bourdieu & Passeron,(1977); Carnoy,(1982). Third, some political theories debated how class interests have been transformed by political structures into reforms and policies which are more prone towards the expansion of education; Rubinson,(1987). Though these theories were useful but they have some limitations. These basic education theories downplayed the role played by economic conditions because of the fact that mass education was found to be poorly related to industrialization; Meyer et al., (1992).

2.1.2 Human Capital Formation and Signaling Theorists' Perspectives on Education

In recent decades, economic studies of Bils & Klenow (2000), Patrinos (2000) and Friedman (2005) reported a positive correlation among the economic growth and education. Private and government spending on education have been given great importance. It was looked down by them as long term investment in human capital. Human capital theorists are also of the view that increased investment in schooling ultimately leads to benefits such as that of increased productivity of labor force, technological innovation, reduction in crime and a better community participation in a democratic society[3]. The externalities as a result of individual education have also been termed as a rationale for government spending in education. Apart from the human capital theory, signaling theory also discuss investment in education but it is of the view that an investment in education is a signal to the individual's employer about his or her future productivity by looking at the educational levels that he has managed to attain; Connelly et al., (2011)[4].

2.1.3 Student Retention and Improved Mathematics' Grades as a Proxy for Key Performance Indicators

In light of the above discussion, we can safely come up with higher student retention and improved mathematics' result as key performance indicators associated with efforts pertaining to increase in the quality of education in the case of Riga Technical University.

[3] Hall (2006).

[4] Wolpin (1977) in his study discussed that signaling is not a successful phenomenon in case of self-employed individuals. Based on his discussion, Arai (1989) made use of the data from Japan on income gained through self-employment and enrollment rates to test this hypothesis. We did not incorporate these variables in our research because of the lack of data available.

2.2 Data Needs and Preferences

2.2.1 Data Analysis, Pre-processing, Cleaning, and Transformation

This study is based on Riga Technical University's entrance data for the year 2013. The data contained records of 2180 undergraduate students who have started studies at RTU in Fall 2013.

Columns were numbered in the excel sheet 1 through 8 respectively. These columns were "N", "Study programme code", "Math exam (school leaving CE)", "Math (school leaving mark)", "University ranking", "Status", "13/14 - Fall (First; Last; N of attempts)", and "13/14 - Spring (Last; First)".

For the purpose of data analysis, total number of students was obtained by counting the values using the "count" command. To further split the each category according to the student status, following excel commands were used in column 6.

=COUNTIF(F2:F2181,"Academcal_break")

=COUNTIF(F2:F2181,"Dropped")

=COUNTIF(F2:F2181,"Full-time")

=COUNTIF(F2:F2181,"Part-time")

Column no. 7 was the trickiest column to deal with as it had 3 values in each row. To separate these values in 3 different columns, the "text to columns" function was used under Data tab, using ";" as a de-limiter. The columns which were formed were called 7a, 7b, and 7c which corresponded to "1st attempt", "last attempt" and "number of attempts" respectively. To calculate the number of students who passed the first attempt, an additional column was created with the formula =IF(OR(B2>=4, B2="passed"),"1","0"). The values of '1' then counted using the formula =COUNTIF(E2:E2181, 1).

This formula however, also counted the students who did not attend the course. To overcome this problem, we counted the students who did not attend the course in a separate column and subtracted it from the original number.

The number of students who passed after 1st attempt was calculated using the "CountIf" formula on the column 7b. The total number of students who failed the course was calculated by subtracting the number of students who passed the 1st attempt, no. of students who did not attend the first attempt, and the number of students who passed after 1st attempt from the total number of students. The results obtained were as follows.

Passed in 1st attempt	1184
1st attempt not attended	393
Failed in all attempts	369
Passed after 1st attempt	234
Total no. of students	2180

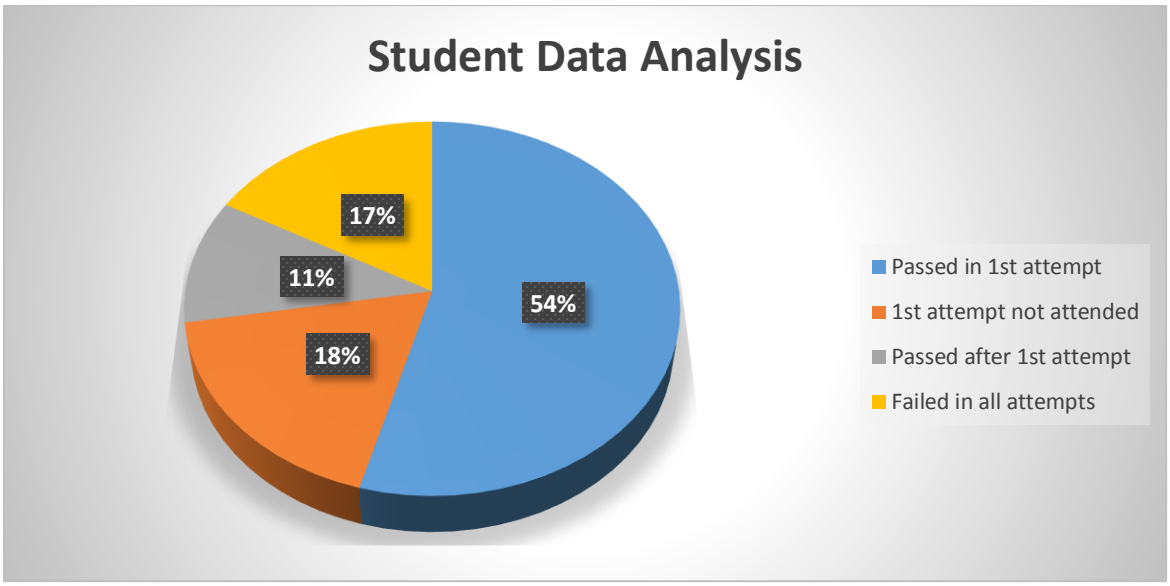

Figure 2. Student Data Analysis

Since this number of students not attempting and failing the tests is a rather large number, therefore it needs to be investigated that how come such a situation can be tackled with.

Furthermore, a correlation analysis showed an extremely weak correlation between the scores of CE exam of mathematics, math (school leaving mark) and whether a student passed during the first attempt in mathematics at RTU or not. This situation makes us ponder over the fact why such a situation is being prevalent, rather the correlation analysis should be showing a high correlation between the previous math scores and whether a student is passing his math exam in the university or not. This ultimately can also cause expulsion of the student from the university due to poor grades.

The results of correlation analysis is summarized in the Table 1.

	mathexamsc~e	*mathschool~k*	*passedorno~t*
mathexamsc~e	1.0000		
mathschool~k	0.7294	1.0000	
passedorno~t	0.2005	0.1870	1.0000

2.2.2 Model Building Preferences

In light of the discussion above, we decided to include student status, CE exam score and math school leaving mark in our analysis to see the effect of the previous math scores on the student's decision to drop out from the university or not and how the previous scores can contribute more towards increasing the quality of education i.e increasing the enrollment rates.

2.3 Specification of Modelling Choices

Gupta, et al.,(2002) proposed an educational production function in the form:

$Y_t = f(X_{1_t}, X_{2_t} \ldots)$

Where Y_t is an indicator measuring the educational attainment of a country and t is the current time period. Consequently the model specified for the current study to see the effect of previous math scores on educational quality or outcome is as follows:

$$Educational\ Outcome_t = f(X_{1_t}, X_{2_t})$$

This production function can be written in the equation form as:

$Status_{Fulltime/Dropped_t} = \alpha_0 + \beta_1 mathexamschoolleaving(CE)_t + \beta_2 mathschoolleavingmark_t$

2.3.1 Estimation Technique

The situation regarding the status of the student is as follows:

Academic Leave	35
Dropped	699
Full Time	1441
Part Time	5
	2180

In order to simplify our estimation, we decided to make our dependent variable as a binary variable having values 0 and 1. Students who are on academic leave and those who have dropped out from the university have been assigned a value 0, whereas full time and part time students have been assigned a value 1.

When a dependent variable is dichotomous, the ordinary least squares regression (OLS) method can no longer produce the best linear unbiased estimator (BLUE); that is, OLS is biased and inefficient.

We cannot use regression models (linear probability model) because of the fact that in the linear probability model, $F(x'\beta) = x'\beta$ (basically a linear function)

and we will have a probability: $P = pr(y = 1|x) = x'\beta$

Now the problem here is that the predicted probability will not be between 0 and 1 because there is no restriction on $x'\beta$ and we can end up having a predicted probability which can be greater than 1 or less than 0, which does not make any sense. Thus we can safely establish this that why regression models cannot be used in case of binary outcome.

One such method which allows us to overcome this problem is the logistic regression. Extending the logic of the simple logistic regression to multiple predictors (X1 = $mathexamschoolleaving(CE)$ and X2 = $mathschoolleavingmark$), one can construct a complex logistic regression for Y (status) as follows:

$$logit(y) = \ln\left(\frac{\pi}{1-\pi}\right) = \alpha_0 + \beta_1 mathexamschoolleaving(CE)_t + \beta_2 mathschoolleavingmark_t$$

Taking the antilog on both sides, one derives an equation to predict the probability of the occurrence of the outcome of interest as follows:

$$\pi = Probability\,(y = outcome\ of\ interest \mid X_1 = x_1\,, X_2 = x_2)$$

$$= \frac{e^{\alpha_0 + \beta_1 mathexamschoolleaving(CE)_t + \beta_2 mathschoolleavingmark_t}}{1 + e^{\alpha_0 + \beta_1 mathexamschoolleaving(CE)_t + \beta_2 mathschoolleavingmark_t}}$$

Where ($X_1 = x_1\,, X_2 = x_2$) are specific values for predictors and π is the probability of the event, α is the Y intercept, βs are regression coefficients, and Xs are a set of predictors.

The advantage of the logit model is the fact that the predicted probability will be between 0 and 1.

PART III - MODEL ESTIMATES, ANALYSIS OF FINDINGS AND CONCLUSION

3.1 Testing for the Significance

```
Logistic regression                                     Number of obs   =       2180
                                                        LR chi2(2)      =     216.22
                                                        Prob > chi2     =     0.0000
Log likelihood = -1287.8808                             Pseudo R2       =     0.0774

------------------------------------------------------------------------------
              Status | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------------------+--------------------------------------------------------------
mathschoolleavingmark |   1.361505   .0503737     8.34   0.000     1.266269    1.463903
     univerityranking |   1.007605   .0017979     4.25   0.000     1.004087    1.011135
                _cons |   .1148731   .0241331   -10.30   0.000     .0761015    .1733976
------------------------------------------------------------------------------
```

This model fits the data very well (p<.0000) and all independent variables are statistically significant at the .01 level.

3.2 Calculating the Coefficients (Log of odds)

```
Iteration 0:   log likelihood = -1395.9931
Iteration 1:   log likelihood = -1296.8345
Iteration 2:   log likelihood = -1295.3707
Iteration 3:   log likelihood = -1295.3706

Logistic regression                                     Number of obs   =       2180
                                                        LR chi2(2)      =     201.25
                                                        Prob > chi2     =     0.0000
Log likelihood = -1295.3706                             Pseudo R2       =     0.0721

------------------------------------------------------------------------------
                Status |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------------+------------------------------------------------------------
mathexamschoolleavingce |   .0103255   .0034911     2.96   0.003     .0034831     .017168
  mathschoolleavingmark |   .3144925   .0422111     7.45   0.000     .2317603    .3972246
                 _cons |   -1.89068   .1989827    -9.50   0.000    -2.280679   -1.500681
------------------------------------------------------------------------------
```

coefficient of variable is the corresponding logarithmic transformed odds ratio calculated in the first step.

For example, coefficient of mathexamschoolleavingce is 0.0103255=log(1.361505) or 1.361505=exp(0.0103255)

The interpretation of the coefficients is that an increase in X increases/decreases the likelihood that y=1 (makes that outcome more/less likely). In other words, an increase in X makes the outcome of 1 (student being a full time student) more or less likely.

3.3 Likelihood Ratio Test

```
( 1)  [Status]mathexamschoolleavingce = 0

           chi2(  1) =    8.75
         Prob > chi2 =    0.0031

( 1)  [Status]mathschoolleavingmark = 0

           chi2(  1) =   55.51
         Prob > chi2 =    0.0000
```

A large chi-squared rejects the null hypothesis that the parameter of mathexamschoolleavingce is zero. mathexamschoolleavingce has a significant positive impact on status of the student.

Similarly, for mathschoolleavingmark large chi-squared rejects the null hypothesis that the parameter of mathschoolleavingmark is zero. mathschoolleavingmark has a significant positive impact on status of the student.

3.4 Marginal Changes

When estimating logit models, it is common to report the marginal effects after calculating the coefficients. The marginal effects will reflect the change in the probability of y=1 given a 1 unit change in an independent variable x. Since the marginal effects depend on X, so we need to estimate the marginal effects at a specific value of X. For our analysis we are calculating the marginal effects at the mean value of our independent variables.

```
Mariginal changes

Marginal effects after logit
      y  = Pr(Status) (predict)
         =  .67739337
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
mathex~e |   .0022565      .00076    2.96   0.003   .000764  .003749   53.2307
mathsc~k |   .0687265      .00916    7.50   0.000   .050766  .086687   6.62294
------------------------------------------------------------------------------
```

The predicted probability of a student being a full time is .677 at the average age of score of 53 in mathexamschoolleavingce and an average score of 6.6 in mathschoolleavingmark.

Marginal effects and discrete changes are listed under dy/dx.

For a unit increase in mathexamschoolleavingce, the predicted probability of student being a full time student will increase by 0.2 percent, holding other independent variables constant at the reference points.

3.5 Conclusion

Significant efforts should be invested by RTU in order to improve its student retention rate. Clear and concise performance oriented targets should be set at the university level as well as on students' group level and the need to achieve improvements in mathematics score should be highlighted in the university's strategic plan.

Furthermore, what a university can do is to conduct a preliminary mathematic exam for everyone without any penalty associated with it and then offer a refresher or remedial math course for the students failing in that exam. This test and these classes should be offered in the beginning of the semester. Such a practice is also being followed at the Kühne Logistics University in Hamburg, Germany.

Also, students should be made aware and encouraged to use massive open online courses available on the internet which are of advanced level. These courses can also be beneficial in preparing students for what lies ahead as well as to raise their mental caliber to the university level mathematics' problems.

References.

Arai, K. (1989). A cross-sectional analysis of the determinants of enrollment in higher education in Japan. *Hitotsubashi Journal of Economics, 30*(2), 101-120.

Behrman, J., & N.Birdsall. (1893). The Quality of Schooling: Quantity Alone is Misleading. *American Economic Review, 66*(2), 928-946.

Bils, M., & Klenow, P. (2000). Does schooling cause growth? *The American Economic Review, 90*(5), 1160-1183.

Bourdieu, P., & Passeron, J. (1977). *Reproduction in Education, Society and Culture.* Beverly Hill,CA: Sage Publications.

Bowles, S., & Gintis, H. (1976). *Schooling in Capitalist America.* New York: Basic Books.

Card, D., & Krueger, A. B. (1992). Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States. *Journal of Political Economy, 100*(1), 1-39.

Carnoy, M. (1982). Education for alternative development. *Comparative Education Review, 26*, 160-177.

Clark, B. (1961). *Educating the Expert Society.* San Francisco: Chandler Publishing.

Connelly, B., Certo, S., Ireland, R., & Reutzel, C. (2011). Signaling theory: a review and assessment. *Journal of Management, 37*(1), 39-67.

Friedman, B. (2005). *The Moral Consequence of Economic Growth.* New York : Vintage.

Gupta, S., Verhoeven, M., & Tiongsan, E. (2002). The Effectiveness of Government Spending on Education and health Care in Developing and Transition Economics. *European Journal of Political Economy, 18*(4), 717-737

Hall, J. (2006). Postivie externalities and government involvement in education. *Journal of Private Enterprise, 21*(2), 165-175.

Lassibille, G., & Gomez, L. N. (2007). Why do higher education students drop out? Evidence from Spain. *Education Economics, 16*(1), 89-105.

Meyer, J., Ramirez, F., & Soysal, Y. (1992). World expansion of mass education, 1870-1980. *Sociology of Education, 65*(2), 128-149.

Patrinos, H. (2000). Market forces in education. *European Journal of Education, 35*(1), 61-80.

Rubinson, R. (1987). Class formation, politics and institutions: schooling in the United States. *American Journal of Sociology, 92*, 519-548.

Wolpin, K. (1977). Education and screening. *American Economic Review, 67*, 949-958.