A Data Mining Approach for Preventing Undergraduate Students Retention

Hadautho Roberto Barros da Silva
Center for Informatics
UFPE – Federal University of Pernambuco
Recife, Brazil
hrbs@cin.ufpe.br

Paulo Jorge Leitão Adeodato
Center for Informatics
UFPE – Federal University of Pernambuco
Recife, Brazil
pjla@cin.ufpe.br

Abstract-The Brazilian government has invested funds via the REUNI program to increase the amount of places in public universities. It has also made another important attempt to increase the number of places by approving regulations to legalize the release of the places occupied by retained students for new students. According to these regulations' retention criteria, there were around 50% of places retained by these students. This Article presents a data mining approach for assessing the risk of undergraduate student retention at the end of the second semester of course aiming at supporting the counseling of students to prevent their retention. The approach has been developed for the Federal University of Pernambuco, focusing on 6 of its major courses which involved data transformation from a database of over 400,000 subjects records, the application of logistic regression for risk assessment, and induction of rules for explaining that risk at the counseling process. The risk estimation solution reached Max_KS=0.51 and AUC_ROC=0.84 in performance. Three scenarios of 30%, 50% and 70% of efficacy in the counseling process show that the procedure should be implemented and is economically viable with decision threshold and return of investment varying according to that efficacy.

Keywords-student retention; student evasion; data mining; logistic regression; a priori

Introduction

According to the 2008's higher education census, carried out by the Brazilian national institute of studies and research, INEP, the completion rate (ratio of graduating students at the proper course duration and the number of students who entered the course that time before graduation) is 67% [1]. It means that one out of each three undergraduate students either does not end their course or takes longer than expected. This inefficiency of the educational system is the result of two phenomena according to education experts [2, 3]: student evasion and student retention. Retention is characterized as a prolongation of the student's stay in the institution beyond the course expected duration while the evasion is defined as the student's dropout. At the Federal University of Pernambuco, these phenomena were higher than the country's average, with 58.4% for retention and 7% for evasion with a completion rate of 61.42% [4].

According to Silva [2], both retention and evasion are complex problems due to different causes, and must be treated through a permanent approach for a complete solution. Campello [3] considers that, since these problems have causes not completely understood, they are difficult to solve and produce very damaging effects in society, such as waste of the education capacity; less productivity and efficiency of the industries; loss of national competitiveness; lack of specialized manpower and others.

For the Federal University of Pernambuco and others Brazilian public universities, that are completely free to the students, the retention and evasion, also known as attrition, are very prejudicial because a retained student costs more than a normal student to the university and an evaded student wastes all the investment made by the government during its academic life.

Since the last decade, these phenomena have been monitored and studied and some actions have already been taken to either solve or reduce them. Normally, these actions are from two types: specific and timely actions, and governmental programs, like the REUNI — the federal universities restructuration and expansion program.

Typically, the specific actions of each institution are focused on the reality of each institution and usually include changes to the curriculum, adaptation of teaching methodologies and evaluation procedures, as specified by UFPE [4].

The governmental programs, different from the specific actions, consist in financial support for improvement in staff quality upgrade at the universities, providing grants for research on these phenomena and infrastructure for the institutions in exchange of targets being reached at key performance indicators.

Only in 2008, the REUNI program forwarded the amount of R$ 491,882,340.00 (around US$ 280 million) for the 53 federal universities that have joined the program that year. The UFPE joined the program on 2008.

One of the two overall goals of the REUNI program states that the completion rate should be 90%. To achieve this goal, UFPE [4] has set a target for 2012, to reduce the evasion rates of 7% in 2007 to 2% and retention of 58.4% in 2007 to 20%, resulting in an increase in the completion rate from 61.92% in 2007 to 90%. Thus, the definition of methodologies and techniques to identify and prevent the retention / dropout of students can significantly impact the efficiency of our universities with the same number of vacancies they have today.

For this work, there has been chosen a single moment at the end of the students' second semester (equivalent of the end of the first year in the United States but in Brazil each semester is independent) of their undergraduate course, for assessing the risk of retention/evasion and deciding whether to provide them counseling or not. This initial scope has been chosen on the fact that there was enough behavioral data on the students' academic performance and there was still time to help them recover in case of detected deficiencies.

Once the risky students are detected, many actions could be taken, such as psychological and educational support, registering orientation, shift change *etc*. to reduce their risk of retention / dropout.

This research followed the CRISP-DM methodology and this paper is organized in four more sections. Section 2 describes the original data and the transformations performed on them. Section 3 describes the data mining process and techniques used in this work. Section 4 discusses the results obtained and the simulation scenarios realized and section 5 summarizes the contributions and presents the conclusions and future research to be conducted.

Data

Original Data

The original data was a collection of the students results on their courses subjects extracted from the University's database and it covers the period from 1998 to 2008. This study refers to six courses selected to represent all knowledge areas of the University. These courses are law, economic sciences, civil engineering, pedagogy, medicine and languages. The database has 415.327 records from 11036 different undergraduate UFPE students. The data structure and this distribution by course are presented below.

TABLE I. ORIGINAL FIELDS

Field	**Description**
ID	Student's ID
Course	Student´s course
Entrance	Year and semester of student course beginning. The students can start their courses in both semesters
Subject Name	The name of the subject
Subject Code	The code of the subject
Subject Shift	The shift of the subject in relation to the recommended semester
Semester	The semester when the student attended the subject
Status	The student's result in the subject (Approved, Failed, Absentee, *etc*.)
Grade	The grade obtained by the student

TABLE II. STUDENT DISTRIBUTION BY COURSE

Course	Records	
	Number	*%*
Economic Sciences	755	13.0
Law	1350	23.3
Civil Engineering	749	12.9
Languages	871	15,0
Medicine	751	13.0
Pedagogy	1317	22.7

The granularity of this database is course/semester/subject/student and, for the purpose of this work, the records had to be transformed to represent the students' behavioral data at each semester. This transformation would provide data both for input the first two semesters to the system and for labeling the target variable according to the University's retention / evasion criteria along the students' academic life. Furthermore, some variables such as "most important subject of the semester" had to be created to embed high level concepts common to all courses to allow the development of a single model thus benefiting from a larger single data set that would be divided by six courses from different knowledge areas otherwise.

Data Pre-Processing

Before the generation of new variables, all outliers have been removed and all missing data were treated in the database. These issues were all related to the status and grade variables.

Other important work was eliminating records from students who had not started and finished all academic life inside the analyzed period (1998 to 2008). Students who had dropped out the course either in first or second semester were eliminated, because they had not reached the decision moment defined as the scope of this work. After this pre-processing the dataset had 5793 records at the student granularity (different student IDs).

Data Transformation

To change the actual data granularity to the desired Course/Student 21 new variables were created plus 3 original ones (ID, course, shift). These new variables were calculated from the original ones and attempted to capture higher level concepts common to all courses:

- The first and second semesters grade average and the relation between these.
- The attendance and grade failure rate from the number of subjects registered in both semesters and the absolute numbers of failures.
- The failure rate variation, indicating if the number of failures stayed the same, increased or decreased from the first semester to the second.
- The final exam approval rate in both semesters.
- The second semester locking indicator.
- The number of registered subjects in the second semester (In the UFPE, the first semester registration is obligatory).
- The subject cancellation rate.
- The coefficient of variation of the grades in the first and second semesters.
- A flag indicating if the student failed in the most difficult subjects in the first and in the second semesters.

To calculate this last item, an analysis was made in each subject, verifying the subject with the highest failure rate from each semester in each course. Despite not being an appropriate measure of the subject relevance for each course, this indicator is effective in capturing an important aspect in students' academic life: the dismay before tough obstacles.

Since the intelligent model will learn a binary decision (retention or not), the target variable was defined based on the following UFPE's criteria, valid back in 2009 [5], when the data for this research was extracted:

- Exceeding the deadline established for finishing the course.
- Having 3 failures, by attendance or grade, on

either the same or equivalent subject.

- Failing, by attendance or grade, in all subjects in a single semester.
- Getting an overall semester performance average less than 30% for any two semesters.

After applying these rules, the retained students' percentage in the dataset was 47%, or 2723 individuals, a little bit lower than the overall University rate. All courses have different retention levels. Table III shows the retention by course below.

TABLE III. RETENTION BY COURSE

Course	Retention (%)
Economic Sciences	71.9
Law	19.0
Civil Engineering	67.7
Languages	66.9
Medicine	0.9
Pedagogy	62.7

After the above variable transformations, all numeric variables were normalized between 0 and 1 and, for the two categorical variables (course and shift), dummy variables that indicate all available courses and shifts were created.

For training the model, the data was partitioned into 2 datasets: the training dataset and the test dataset. For a best and realistic division of the data, the division was used an "Out-Of-Sample" division. In this kind of division, the data are ordered chronologically and the training set is generated from the first records and the test set is from last ones, trying to mimic a realistic scenario where the model is developed to be applied in a future time. The proportion established was about the 2/3 of the records for the training set (period from the year 1998.1 to 2004.1) and the test set with the last 1/3 of the records (from 2004.2 to 2008.2).

Process and Techniques

CRISP-DM

The solution has been developed according to the **CR**oss-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining (CRISP-DM) [6] which is a hierarchical process that consists in a set of tasks. The process conceived with four levels of abstraction aims at standardizing data mining project development across different industry sectors. The CRISP-DM initiative involved some of the biggest companies that execute data mining projects like DaimlerChrysler, IBM SPSS and NCR gathered within a consortium.

Logistic Regression

The modeling technique chosen was logistic regression for several interesting features it possesses. It has been successfully applied to binary classification problems, particularly to credit risk assessment [7], it does not require a validation set for over-fitting prevention and it presents explicitly the knowledge extracted from data in terms of coefficient statistically validated [11].

The logistic regression technique is well-suited to study the behavior of a binary dependent variable based on a set of p independent variables x_p (explanatory features).

The logistic regression model can be expressed by the logit function

$$\log\left\{\frac{\pi(x)}{1-\pi(x)}\right\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p, \qquad (1)$$

where $\pi(x)$ is defined as

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}, \qquad (2)$$

and $x_1, \ldots, x_p$ are the explanatory variables.

The logistic regression was executed with the "Forward Stepwise" method. This method selects the variable entry on the model based on the score statistic and tests the variable removal based on the probability of a likelihood-ratio statistic based on the maximum partial likelihood estimates [8]. After the training process, were selected 13 explanatory variables plus the constant factor to generate the risk estimator. The final variable list is presented below along with their coefficient (β) and the significance value (p). This method starts with one variable and after

TABLE IV. LOGISTIC REGRESSION FINAL VARIABLES

Variable	β	**Sign. (p)**
CourseLaw	-2.53	0.000
CoursePedagogy	-0.97	0.000
CourseMedicine	-5.83	0.000
ShiftMorning	-0.30	0.000
ShilftFull	-1.13	0.000
MainDisc2SemFailure	0.43	0.004
Locking2sem	-3.26	0.000
CancellingRate1Sem	0.83	0.000
AttendFailureRate1Sem	3.42	0.000
AttendFailureRate2Sem	-2.16	0.000
GradefailureRate1Sem	1.93	0.000
FinalExamApprovalRate1Sem	0.51	0.000
CoefVariationGrade2Sem	-5.65	0.000
Constant	4.56	0.000

Rules Induction

Rule induction is one of the most important techniques of machine learning. The rules' importance comes from their easy understanding. Rule induction is one of the fundamental tools of the data mining process [9] and could be used either in the beginning of the process to discover new data relationships or at the end to explain some results.

As the purpose of the classification rules in this work is to explain the risk individually assessed by logistic regression for each student, the best rules are those whose confidence most divert from the sample average. This helps identify the profiles of high performing and poor performing students which encompass each student for supporting their risk recommendation.

The algorithm filters only the classification rules induced which produce a statistically significant difference ($\alpha=0.05$) from the original sample average, assuming a normal approximation of the binomial distribution for the relative frequency (proportion) of the target class in the rule (cube).

The quality of the induced rules is measured by the ratio between the confidence (estimated by the relative frequency of the target class) of the rule and the confidence of the whole sample; a metric called *lift*. The two tables below present some rules that have statistically significant *lift* characterizing the poor performing and the high performing student profiles in relation to the retention problem. These rules are a helpful tool in the counseling process for risky students when they are registering for the third semester onwards.

TABLE V. BEST INDUCED RULES BY LIFT FOR RETENTION/EVASION (POOR PERFORMING STUDENT PROFILE) WITH ONE AND TWO VARIABLES

Rule	**Support**	**Confidence**	**Lift**
Failure in the most important subject of the second semester *and* Second semester cancellation rate between 25% and 50%	0.55%	100.0%	2.12
Second semester grade average < 3	9.48%	99.45%	2.11
Course=Economic Sciences *and* Number of registered subjects in the second semester < 3	0.64%	97.3%	2.06
Course=Economic Sciences	13.03%	71.9%	1.52

TABLE VI. WORST INDUCED RULES BY LIFT FOR RETENTION/EVASION (HIGH PERFORMING STUDENT PROFILE) WITH ONE AND TWO VARIABLES

Rule	**Support**	**Confidence**	**Lift**
Couse=Medicine *and* No failure in the most important subject of the first semester	9.17%	0.00%	0.00
Couse=Medicine *and* Second semester grade average > 8.5	4.59%	0.00%	0.00
Course=Medicine	12.96%	0.93%	0.01
Course=Law	23.30%	19.04%	0.40

Results, Interpretation and Simulation Scenarios

For Decision Support Systems, risk estimators that produce continuous output for binary decision making are very appropriate once that the decision results from the setting of the threshold finely adjusted to the application Key Performance Indicators (KPIs).

The student retention risk estimator was evaluated by both the Kolmogorov-Smirnov curve (KS2) [10] and the area under the ROC curve (AUC_ROC) [11, 12] as technical metrics. The choice of the best operation point has been made based on the trade-off between cost and benefit of the decision outcomes considering three scenarios of counseling efficacy.

These metrics are widely accepted for performance assessment of binary classification based on continuous output and represent similar forms of performance evaluation with slightly different points of view.

The KS statistical method is a traditional non parametric tool used for measuring the adherence of a cumulative distribution function (CDF) to the cumulative representation of the actual data [10]. In binary decision systems, this metric is applied for assessing the lack of adherence between the data sets from the 2 classes, having the score as independent variable. The Kolmogorov-Smirnov Curves are the difference between the CDFs of the data sets of the two classes and the higher the curve, the better the system. The point of maximum value is particularly important for the performance evaluation. The area under the curve metrics (AUC_KS2) is very relevant but can only be used when the horizontal axis is the proportion of the population sorted by the score. The larger the AUC_KS, the better the system is for class separability throughout the whole score range. The ideal decision system would have the AUC_KS2 equal to ½.

The Receiver Operating Characteristic Curve (ROC Curve) [11] is a widely used tool whose plot represents the compromise between the true positive and the false positive example classifications based on a continuous output along all its possible decision threshold values (the score). The closer the ROC curve is to the upper left corner (optimum point), the better the decision system is. In this context, the minimum distance of the curve to this point is an important metric and assessing the performance throughout the whole X-axis range consists of calculating the area under the ROC curve (AUC) [12]. The bigger the area, the closer the system is to the optimum decision. If the ROC curve of a classifier appears above that of another classifier along the entire domain of variation, the former system is better than the latter. The ideal decision system would have the AUC_ROC equal to one (1).

The maximum KS of this model was 0.51 and its curve is depicted in the figure 1 along with the Cumulative Distribution Functions of the two classes (High Performing and Poor Performing students).

The area under the ROC curve reached a value of 0.84 (the maximum value is 1) and is presented in the figure 2.

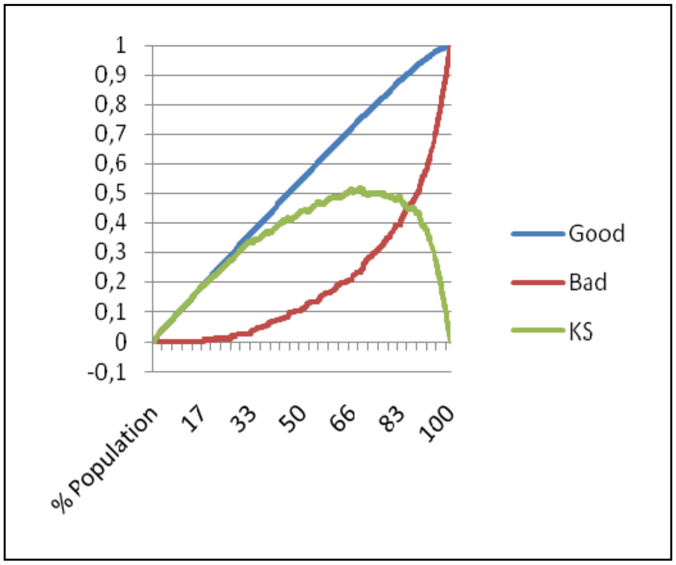
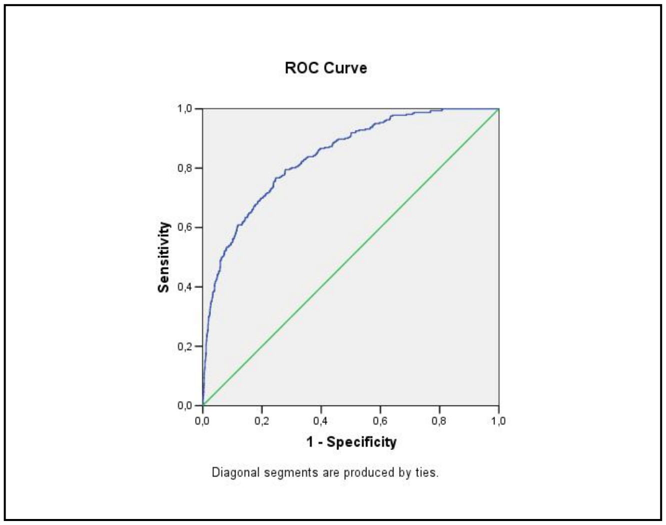


Figure 1. KS Chart



Diagonal segments are produced by ties.

Figure 2. ROC Chart

After calculating these two metrics, three recovering scenarios were estimated to produce a viability analysis of this approach. In 2010, the students' mean individual cost was R$ 14,216.73 (around U$ 8,123.00) to UFPE for their whole course [13]. The six courses focused on by this work have a mean duration of 8.5 semesters resulting in a cost per semester of around R$ 1,672.55 (U$ 955.00). Assuming that lecturers with the counseling attribution were professors that initially earn R$ 4,651.59 (U$ 2,658.00) [14] per month and that individual counseling lasts thirty minutes per student, eight hours of counseling that gives an amount of about 320 students in a month.

Calculating that all 2723 students marked as retained wastes R$ 38,712,155.79 (U$ 22,121,231.00) from UFPE and were necessary 9 lecturers to support them, that costs R$ 41,860.71 (U$ 23,920.00) per month and R$ 544,189.23 (U$ 310,965.00) per year. Assuming that the lecturers will follow the all possible retained students for all their University life, the counseling program will cost R$ 4,625,608.45 (U$ 2,643,204.00) to the university.

For those recovering scenarios, the two first semesters could not be considered as cost to UFPE because the estimator runs after these semesters, so these values could not be considered in the scenarios. To generate the scenarios were chosen 3 different identification and recovery factors. To be very representative, the factors are 30%, 50% and 70%. All simulation scenarios are presented in the table below.

TABLE VII. SIMULATION SCENARIOS

Efficiency	Economy	Lecturers Cost	Balance
30%	R$ 8,881,034.54 (U$ 5,074,876.00)	R$ 4,625,608.45 (U$ 2,643,204.00)	R$ 4,255,426.10 (U$ 2,431,672.00)
50%	R$ 14,801,724.25 (U$ 8,458,128.00)	R$ 4,625,608.45 (U$ 2,643,608.00)	R$10,176,115.80 (U$ 5,814923.00)
70%	R$ 20,722,413.94 (U$ 11,841,379.00)	R$ 4,625,608.45 (U$ 2,643,608.00)	R$16,096,805.49 (U$ 9,198,174.00)

As showed in the balance column, these three scenarios have a positive balance, indicating that suggested approach could be financially viable to UFPE with a recover projection from 47.9% to 77.6% depending of the scenario. Identification and recovery factors below or equal to 15% are not financially viable for the institution.

Conclusions

This paper has presented a data mining solution for the undergraduate students' retention seen as a binary decision problem. The system estimates the students risk of retention with an MLP neural network and a set of classification rules induced via the *a priori* algorithm explain the risk.

The quality of the MLP estimation has been assessed with the AUC_ROC and Max_KS metrics while the induced rules quality has been measured by the *lift* they produce compared to the population average confidence.

The decision support system operates at the end of the students' second school semester and is based on the students' behavior along the course. The purpose of the system is to select the students who would need counseling for the third semester registration and three scenarios based on different levels of effectiveness in the counseling process have shown that only a threshold adjustment is needed for controlling the optimal operating point.

Future research involves the inclusion of the students' socio-economic-cultural data and the lecturers' behavioral data as inputs, the creation of general concepts across all courses as input variables for expanding the solution to the courses with few students and a post-processing stage for analyzing the conditions where the system's performance is unsatisfactory.

References

[1] REUNI, "Reestruturação e Expansão das Universidades Federais – Diretivas Gerais". 2007 (in Portuguese).

[2] Silva, G. E. G., "A Evasão e a retenção". 2010 (in Portuguese).

[3] Campello, A. V. C; Lins, L. N. "Metodologia de análise e tratamento da evasão e retenção em cursos de graduação de instituições federais de ensino superior". 2008 (in Portuguese).

[4] UFPE, "Programa de reestruturação e expansão das universidades federais – Projeto REUNI/UFPE". 2007 (in Portuguese).

[5] UFPE-Federal University of Pernambuco, "Resolução 09/2009/CCEPE", 2009. Retrieved on January 2012, from http://www.ufpe.br/dqf/images/documentos/jubilamento.pdf (in Portuguese).

[6] Chapman, P; Clinton, J;Kerber, R; Khabaza, T; Reinartz, T; Shearer C; Wirth, R, "CRISP-DM 1.0 – Step-by-step data mining guide", 2000.

[7] West, D.: Neural network credit scoring models. Computers and Operations Research, 27, pp. 1131—1152, 2000.Hilbe, J. M.: Logistic Regression Models. Chapman & Hall/CRC Press , 2009.

[8] IBM , SPSS Regression 17.0 Manual2010.

[9] Grzymala-Buse, J. W, "Data Mining and Knowledge Discovery Handbok ", Chapter 13, Springer, 2010.

[10] W. J. Conover, "*Practical Nonparametric Statistics*", Third edition, John Wiley & Sons, NY, USA, 1999.

[11] F. Provost, T. Fawcett, "Robust Classification for Imprecise Environments." *Machine Learning Journal*, vol.42, n.3. pp. 203-231, March 2001.

[12] T. Fawcett. "ROC Graphs: Notes and Practical Considerations for Researchers", 2004. Retrieved on March 2008, from http://www.binf.gmu.edu/mmasso/ROC101.pdf.

[13] Proplan - UFPE: Indicadores TCU 2010. UFPE, 2010 (in Portuguese).

[14] PROACAD-UFPE, "Concursos Públicos para Docentes do Magistério Superior", UFPE, 2011. Retrieved on January 2012, from http://www.ufpe.br/proacad/images/Editais_concursos/edital109/edital_109.pdf (in Portuguese).