Predicting Evasion Candidates in Higher Education Institutions

Conference Paper · September 2011

DOI: 10.1007/978-3-642-24443-8_16 · Source: DBLP

5 authors, including:

Hércules Antonio do Prado
Brazilian Agricultural Research Corporation…
59 PUBLICATIONS **54** CITATIONS
SEE PROFILE

Edilson Ferneda
Universidade Católica de Brasília
34 PUBLICATIONS **59** CITATIONS
SEE PROFILE

Paulo Roberto Cobbe
Universidade Católica de Brasília
4 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Predicting evasion candidates in higher education institutions

Remis Balaniuk[1], Hercules Antonio do Prado[1,2], Renato da Veiga Guadagnin[1], Edilson Ferneda[1], Paulo Roberto Cobbe[1,3]

[1] Graduate Program on Knowledge and IT Management, Catholic University of Brasilia, SGAN 916 Avenida W5, 70790-160, Brasília, DF, Brazil

[2]Embrapa - Management and Strategy Secretariat, Parque Estação Biológica - PqEB s/n°, 70770-90, Brasília, DF, Brazil

[3]Information Technology Department, UniCEUB College, SEPN 707/907 Campus do UniCEUB - Bloco 1, 70790-075, Brasília, DF, Brazil

{remis,hercules}@ucb.br, {renatov,eferneda}@pos.ucb.br, paulocobbe@yahoo.com

Abstract: Since the nineties, Data Mining (DM) has shown to be a privileged partner in business by providing the organizations a rich set of tools to extract novel and useful knowledge from databases. In this paper, a DM application in the highly competitive market of educational services is presented. A model was built by combining a set of classifiers into a committee machine to predict the likelihood that a student who completed his/her second term will remain in the institution until graduation. The model was applied to undergraduate student records in a higher education institution in Brasília, the capital of Brazil, and has shown to be predictive for evasion in a high accuracy. The unbiased selection of students with elevated evasion risk affords the institution the opportunity to devise mitigation strategies and preempt a decision by the student to evade.

Keywords: Knowledge Discovery in Databases, Data Mining, Committee Machines, Higher Education Institutions, Student Retention.

1 Introduction

Information systems permeate every facet of modern business, and produce massive amounts of data that are stored in increasingly larger databases. According to some estimations, the volume of data in these databases double in size every 20 months [1]. This growing mass of data holds many layers of information, including patterns or relations that are difficult to be identified through simple analysis.

Techniques like data mining make it possible for organizations to wade through masses of data and identify these relationships, producing meaning, and ultimately value, from previously unintelligible data.

These techniques have been successfully used by many modern businesses in competitive markets, such as credit card companies, investment firms, and retailers, to produce advantage.

Modern economy strongly depends on skilled labor force. Seeking to take advantage from this reality, a growing number of colleges and universities are being created. In Brazil, in particular, there has been tremendous growth in the number of higher education institutions in the last decade, resulting in steadily increasing competitive pressures within the education services market. Studies, in fact, indicate that this trend should continue in the foreseeable future.

To thrive in this market it is essential that higher education service providers seek to gain competitive advantage. In this context, data mining has a role to play, by giving institutions the ability to predict possible future student intent, allowing these institutions the opportunity to devise appropriate mitigation strategies.

Gaioso [9] shows that there are several reasons for students to abandon their undergraduate studies, among them, financial strain, lack of vocational orientation, deficiencies in basic education and schedule conflicts. These reasons, for the most part, remain undetected by HEIs until the moment a student initiates a transfer request, leave of absence request, or drops out.

Institutions that are able to identify students with high evasion risk, and manage to successfully overcome student grievances early on, may establish an environment of cooperation between the school and the student and foster those factors that promote student loyalty, and by so doing, may gain significant advantage over competing HEIs. In addition, by fulfilling its business objectives the education service provider also fulfills its social charter, increasing its graduation rates and contributing, ultimately, to the progress of society as a whole.

This article presents the results of a data mining experiment that demonstrates a method for early detection of students with elevated evasion risk. It uses data mining tools to analyze historic student records and predict whether students who completed their second term of undergraduate studies will go on to graduate, or will abandon his or hers studies prematurely.

In the following sections, this paper provides a brief background of the Brazilian higher education services market, describes the methodology used for the selection of attributes, data extraction, preparation and processing and presents the experiment results.

2 The Brazilian higher education market

After years of limited growth due to sparse public investments and restrictive regulations, that limited the interest of private education service providers, changes in the so called Law of Bases and Directives, signed in 1996, sparked renewed interest in this sector.

After these changes, the number of Higher Education Institutions (HEI) grew from 900 in 1997 to 2,252 in 2008, a two and a half fold increase. During this period, the number of undergraduate level students enrolled in HEIs grew by a similar ratio, from 1.9 million in 1997 to just over 5 million in 2008. Of these 5 million students,

approximately three quarters were enrolled in private higher education institutions [2],[3].

In spite of the growth in the undergraduate student population, Brazil still lags behind other Latin American countries in college enrollment rates. Data from UNESCO [4] shows that, in 2007, 30% of Brazilian college-age students were enrolled in an HEI, compared to 52% and 68% of the college-age population in Chile and Argentina, respectively. Also, the sustained growth of emerging economies such as Brazil, with GDP growth for 2010 estimated at 7.5% [5], point to an increasing demand for a skilled workforce. Consequently, an education market that is already significant, worth roughly US$ 8.2 billion in 2005 [6] using current exchange rates [7], has significant growth potential, suggesting increasing competition among higher education institutions.

Competition for students is already fierce, as evidenced by open vacancy ratios of around 50% [3] in Brazilian higher education institutions. In order to remain viable, HEIs are forced to engage in large and very expensive recruiting campaigns that even include mass media advertisements. But these efforts have no effect to retain students, who abandon their studies before graduation at an estimated rate of 44.7% [3]. The Brazilian National Institute for Educational Research Anísio Teixeira (INEP) does not make clear what percentage of these students transfer to other HEIs or drop out altogether. It is well known, however, that college transfers are common practice among students of private colleges and universities.

With excess capacity and high rates of evasion, it becomes critical for HEIs to maximize retention of enrolled students. Bergamo et al. [8] indicate that student retention is a key issue for the survival of private colleges and universities, and student loyalty is directly linked to factors such as trust, emotional commitment, satisfaction, expectations management and school reputation. HEIs that foster student loyalty gain competitive advantage through a solid revenue stream, increased reputation and reduced marketing and recruiting budgets.

3 Methodology

3.1 Boosting prediction accuracy

Predictions produced by machine learning algorithms vary in performance due to many factors, such as the nature of the data being analyzed, the size of the dataset, the number of patterns contained in the dataset, etc. Some algorithms may misclassify some items that other algorithms may classify correctly. No algorithm is perfect for every situation, each having strengths and weaknesses [10].

Committee machines is a method of combining the prediction results from various learning algorithms, leveraging the strength of each algorithm in order to achieve a combined result superior to any other that could be reached by using a single learning algorithm alone [11].

3.2 The experiment

The main goal of the experiment was to identify students with high evasion risk in order to provide college Deans and Bursar officials with the means to locate and interact with these students, and develop individual mitigation strategies that could preempt a decision to drop out.

This prediction experiment using machine learning algorithms was conducted with data gathered from student records of 11,495 undergraduate students of a top-tier private higher education institution in Brasília, the capital of Brazil.

Data was collected based on factors identified by Gaioso [9] as being important reasons behind student evasion. A number of student attributes were studied, extracted and transformed, resulting in a database that combined the socio-economic and academic information for each student. The attributes examined were the following: *(i)* age group, *(ii)* gender, *(iii)* neighborhood of residence, *(iv)* work status, *(v)* type of high school attended, *(vi)* family income, *(vii)* overall grade point average (GPA) for all classes attended, *(viii)*GPA in second semester classes, *(ix)* overall class attendance average, *(x)* attendance average in second semester classes, and *(xi)* number of failed classes in the two initial semesters.

CRISP-DM [12] was the method adopted as a guide to drive the data mining application. To conduct the evasion prediction experiment, the WEKA (Waikato Environment for Knowledge and Analysis) workbench software for machine learning [13] was selected. This software is open source, and contains a collection of machine learning algorithms suitable for data mining projects

After constructing prediction models using regression, decision tree and neural network algorithms, the results were combined in a committee machine, which produced a report identifying each student and the probability him or her would graduate normally or evade.

3.3 Data preparation

This phase strongly relies on the expertise from the HEI managers. To help diagnose economic strain, family income and neighborhood of residence were selected as indicators of economic health. Anecdotal evidence shows that students, occasionally, misrepresent their family income in fear of having his or her enrollment rejected by the institution, so the neighborhood of residence is a particularly relevant indicator of family income. In general, family income drops the farther from downtown in Brasília a person resides, with residents of the outer suburbs having lower incomes, and residents of the South and North Lake neighborhoods having higher incomes on average. Neighborhoods were divided in six large regions which were North/South Lake, Brasília-Proper (Plano Piloto), Near-Suburbs, Far-Suburbs, Outer Suburbs and Other States.

To identify possible scheduling conflicts, the age group, gender, and work status of the student were selected. It is expected that younger students, or those who do not work, are less likely to experience sustained scheduling conflicts. On the other hand, for female students, pregnancy and child rearing could lead to scheduling conflicts, and, consequently, evasion.

The type of high school attended (public, private or military academy) was selected as an indicator that may point to basic education deficiencies. The Brazilian public school system is known for its structural problems and overall poor performance when compared with private or military schools.

Grade and attendance records, along with number of failed classes, were selected as measures of current academic success. Since the HEI studied grades students using a subjective grading scheme, the GPA had to be calculated using a grade conversion convention, which was adapted from that used by North American HEIs. The grading scheme used by the Institution was converted to numerical values. Passing grades, SS, MS and MM were converted to 4.0, 3.0 and 2.0 respectively. MI and SR, failing grades, were converted to 1.0 and 0, respectively. GPA calculation does not take in consideration canceled or withdrawn classes. The weighted mathematical average for the converted grades was calculated using the number of credits for the class, and transformed in a grade scale of A, B, C and D, with A equivalent to a GPA greater or equal to 3.4, B representing a GPA between 2.7 and 3.4, C, between 2 and 2.7, and D, less than 2 points. The GPA was calculated for 2nd semester classes and the overall cumulative GPA.

Class attendance was also calculated using a similar scheme, in which the weighted average of absences was calculated for the 2nd semester classes and overall attendance average, using the number of class credits as weights. The average absences was also converted to a scale represented by Low, Medium and High, with an average of less than 8, between 8 and 16 and 16 or more absences, respectively. Attendance records of withdrawn or canceled classes were not considered as well.

The final measure of academic performance is a count of the number of failed classes for the first and second semesters. Three groups were created for this indicator, these being: No failed classes, up to 2 failed classes and 3 or more.

Student records were divided into two groups. The first group, with 3,058 records, was used for training and testing the model, and included records of students who enrolled in 2005 and 2006, who completed the socioeconomic enrollment survey and graduated (within the average graduation time of 5 years for the institution), or dropped out prematurely. The second group with the remaining 8,437 records, contained students regularly enrolled in classes in the second semester of 2010, that a) would not graduate at the end of the semester, b) were are at the minimum in their second year, and c) completed their socioeconomic enrollment survey. These records were used to generate predictions.

4 Results

The training data was processed through WEKA using four classification algorithms. These were ZeroR, a simple classifier, Average One-Dependency Estimator (AODE), a Bayesian probabilistic classifier, J48, an implementation of the C4.5 decision tree, and Multilayer Perceptron, a neural network.

ZeroR classifier was used for drawing a prediction baseline to compare against other classifiers [13]. To establish the upper limit of accuracy, training and testing was done using the complete dataset, and to build the model later used for prediction,

the data was split using ⅔ for training and ⅓ for testing. Prediction accuracy is summarized in Figure 1.

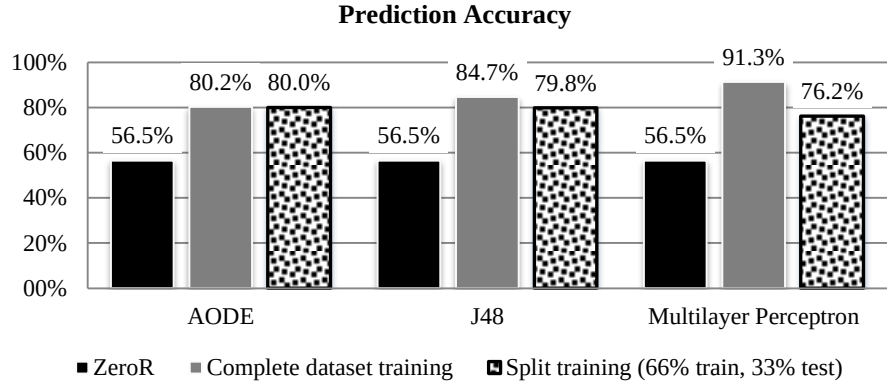

Figure 1: Algorithm prediction accuracy in training

When trained using the entire training dataset, all prediction algorithms showed significant improvement over the baseline results. As would be expected, these high rates drop when trained using a smaller sample, 66% of the set, and tested using the remaining data. Accuracy reduction for AODE was negligible, while J48 and Multilayer Perceptron sustained greater accuracy reductions.

In order to boost overall prediction accuracy across all algorithms, a committee machine was set up using all three prediction outputs (AODE, J48 and Multilayer Perceptron). After testing various committee schemes, simple arithmetic mean was selected because it maintained an overall high prediction result accuracy of 80.6% (slightly superior as achieved for AODE and J48), while being extremely simple to implement. Table 1 presents the summarized training results using the ⅔ ⅓ split.

	AODE	J48	Multilayer Perceptron	Committee Machine
Overall Accuracy	80.0%	79.8%	76.2%	80.6%
Graduate Prediction	80.5%	80.8%	80.9%	81.9%
Evade Prediction	79.2%	78.4%	70.7%	78.7%

Table 1: Summary of the prediction accuracy for each algorithm and the committee machine

The confusion matrix produced by each training algorithm and the committee machine using their results is presented in table 2.

	AODE		J48		Perceptron		Committee	
	Grad.	*Evade*	*Grad.*	*Evade*	*Grad.*	*Evade*	*Grad.*	*Evade*
Graduate	505	86	500	91	449	142	499	92
Evade	122	327	119	330	106	343	110	339

Table 2: Confusion matrices for each training algorithm

The accuracy of the algorithms was also evaluated using receiver operating characteristic (ROC) curves, which indicate the performance of a given classifier [14]. The curves show the relation between true positives (Y axis) and false positives (X axis), expressed as a percentage of the sample. Higher arcs (closer to the x, y point of 0%, 100%) are evidence of more accurate classifiers. The ROC curves plotted for each prediction algorithm are shown in Figure 2.

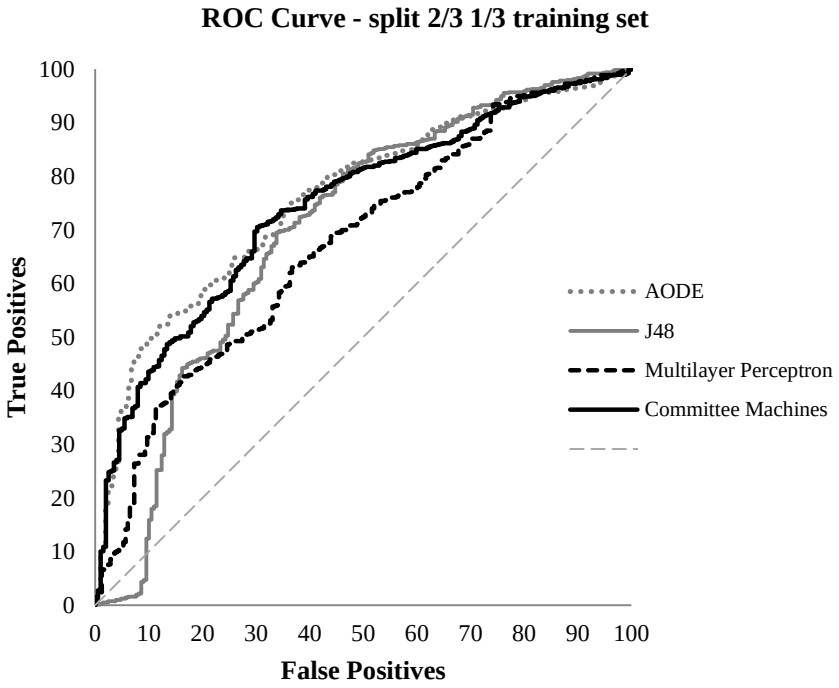

Figure 2: ROC Curves expressed for the split training sets

After training, the 8,437 prediction records were processed using the models created using the ⅔ ⅓ split. The output for all three predictions was combined by averaging each individual student probabilities for each outcome (graduate or evade) using simple arithmetic mean. The results were then converted to a committee prediction representing the most likely outcome expected for the student. Students for whom the prediction was ambiguous (equal probabilities) were predicted to graduate. A sample of the committee results report is presented in table 3 below.

Student ID	Course	Committee Predictions		
		Graduate	Evade	Predicted status
185	Journalism	2.6%	97.4%	Evade
186	Psychology	1.6%	98.4%	Evade
189	Law	57.4%	42.6%	Graduate

Table 3: Committee machine prediction for regularly enrolled students

The complete report indicates the predicted future status for every student enrolled, with the probability for each possible outcome (graduate or evade). This information allows school officials to decide whether to pursue a policy of attempting to retain students with high predicted probability of dropping out or to address borderline cases, those that would, in theory, be easier to address.

The model predicted that 3,250 of the 8,437 students enrolled would evade, which corresponds to 38.5% of students. Although lower than the estimated national average of 44.7% according to INEP [2], the figure is compatible with historic evasion rates for the HEI. The remaining 5,187 students were predicted to graduate normally.

Preliminary analysis by college Deans and Professors indicate that predictions in fact indicate many known at-risk students. Additionally, records of enrollment renewals for 2011 also confirm several of the evasion predictions contained in the full prediction report.

5 Conclusion

Many industries have experienced deep transformation through the use of techniques such as data mining and machine learning. In highly competitive business environments, those organizations that correctly implemented these techniques acquired great competitive advantage, and consequently, great success.

To remain economically viable, higher education institutions have sought to gain competitive advantage through investments in information technologies and recruiting campaigns. But these efforts ignore one very important problem faced by HEIs, which is student retention. Perhaps because historically early detection has been very difficult, HEIs have attempted to deal with the problem at the exit point, after the student has decided to leave, a point at which it is often too late to revert the decision.

This article presents a successful experiment which employed data mining tools in early evasion-risk detection for students of a preeminent higher education institution in Brasília, the capital of Brazil. The results of this experiment indicate that it is possible to identify students with elevated evasion risk, even those students who may not present overt signs of academic difficulty or financial strain, and this information, associated with appropriate pedagogical and financial grievance mitigation strategies, may prove to be exceptionally valuable tools for colleges and universities who seek to reduce student evasion, and provide the critical competitive advantage needed to thrive in the educational services market.

The experiment resulted in a report containing predictions for each undergraduate student that completed their second term, and were regularly enrolled in the institution studied. The report indicated, with accuracy of 80.6%, whether the student would graduate or evade, expressed as a percent of likelihood of one outcome or the other. Of the 8,437 students analyzed, the model indicated that 3,250 students (38.5%) were predicted to evade.

By narrowing the student population to those at risk, the results of this experiment allow department Chairs, college Deans and Bursars officials to engage each student early in their undergraduate career, and in so doing, to identify his or hers individual

risk factors and to devise strategies to combat these factors, both at the level of individual solutions and institutional changes. These may include pedagogical support for basic education deficiencies, academic orientation, vocational orientation or financial aid. Some of these strategies may, in fact, uncover opportunities to create additional educational products that could be offered to students, creating new, and potentially profitable, revenue streams. Without this filter of the student population, such engagement would be far less efficient, perhaps to the point of becoming unfeasible.

Regardless of the strategy adopted, Bergamo et al [8] indicate that developing a closer contact with the student by engaging him or her personally to discuss their needs foster loyalty, and by so doing may be sufficient to overcome evasion intent. This study indicates an objective method to select students to receive such attention.

References

1. Hand, D.; Mannila, H.; Smyth, P.: Principles of Data Mining. The MIT Press, Cambridge (2001)
2. National Institute for Educational Research Anisio Teixeira (Inep): Higher Education Statistical Synopsis. http://portal.inep.gov.br, (2008)
3. National Institute for Educational Research Anisio Teixeira (Inep): Technical Summary – Higher Education Census 2008 (Preliminary Data). http://portal.inep.gov.br, (2009)
4. UNESCO Institute for Statistics: International Standard Classification of Education Key Statistics. http://www.uis.unesco.org, (2011)
5. Cardoso, J.: Brazilian GDP expected to grow 7,5% in 2010 and 4,3% in 2011, forecasts OCDE. Jornal Valor Online, São Paulo, 18 Nov. (2010)
6. Lima, M. C.: The WTO and the "Educational Market". Reasons Behind the Interest and Possible Consequences. VI International Colloquium on Higher Education Management in South America. Blumenau (2006)
7. Campos, E.: Dollar Closes at R$ 1,740 and Negates Losses for the Year. Jornal Valor Online, São Paulo, 16 Nov. (2010)
8. Bergamo, F.; Farah, O. E.; Giuliani, A. C.: Loyalty and Higher Education: Stategic Tool in Client Retention. Revista Gerenciais, vol. 6 n. 1 pp 55-62 , São Paulo (2007)
9. Gaioso, N. P. L.: Student Evasion in Higher Education: Student and Management Perspectives. Universidade Católica de Brasília, Brasília (2005).
10. Wolpert, D.: The Lack of a Priori Distinctions between Learning Algorithms, Neural Computation, vol. 8 n. 7 pp. 1341-1390, MIT Press (1996)
11. Tresp, V.: Committee Machines. In: Hu, Y. H.; Hwang, J.-N. (eds.). Handbook for Neural Network Signal Processing. CRC Press (2001)
12. Chapman, P.; Clinton, J.; Kerber, R.; Khabaza, T.; Reinartz, T.; Shearer, C.; Wirth, R..: CRISP-DM 1.0: Step-By-Step Data Mining Guide. http://www.crisp-dm.org (2000).
13. Bouckaert, R. R.; Frank, E.; Hall, M.; Kirkby, R.; Reutemann, P.; Seewald, A.; Scuse, D.: Weka Manual for Version 3-6-2. University of Waikato, Hamilton

(2010).
14. Witten, I. H.; Frank, E.: Data mining: practical machine learning tools and techniques. Elsevier, San Francisco (2005).