Prediction of Student Success Using Enrolment Data

Nihat Cengiz
Epoka University
Department of Computer Engineering
Rr. Tiranë-Rinas,Km. 12
1039 Tirana, Albania
ncengiz@epoka.edu.al

Arban Uka
Epoka University
Department of Computer Engineering
Rr. Tiranë-Rinas,Km. 12
1039 Tirana, Albania
auka@epoka.edu.al

ABSTRACT

Predicting the success of students as a function of different predictors has been a topic that has been investigated over the years. This paper explores the socio-demographic variables like gender, region lived and studied, nationality and high school degree that may influence success of students. We examine to what extent these factors help us to predict students' academic achievement and will help to identify the vulnerable students and their need for extra tutoring or similar supportive services at an early time.

We analyzed the data of the Epoka University students that have been enrolled from 2007 to 2013. The sample includes 1211 undergraduate students where 716 did and were supposed to complete the three-year bachelor studies in the past six semesters.

Based on the data mining techniques the most important predictors for student success were the students' high school GPA and gender. For students with high school grades below average, females were found to have a higher percentage of success than boys. No significant correlation was found between the students' success and the demographic information.

Keywords

Academic achievement, influence, classification tree, outcome

1. INTRODUCTION

Increasing the student graduation and decreasing the dropout rates is a long term goal of the higher education institutions. From the students' perspective, a timely and successful graduation is vital as these two factors would strongly affect their employability rate. Employability rate has become an indicator in determining the ranking of higher education institution (HEI), thus HEIs are focusing more on increasing this rate [2].

Many of the students studying at the university face several difficulties during the first year and thus the performance of the first year has been identified as an important predictor of timely graduation rate. In terms of keeping the students in the university, the retention rate is a factor that has been studied extensively. Mallincrodt and Sedlacek (1987) found that freshman class attrition rate were greater than the other academic years with numbers running up to 30%.[3] Therefore most researchers targeted the first year students. An early identification of the students at high risk of failing will enable a timely intervention with the necessary measures by the educators that would increase the graduation rate. Preventing students' failure depends on the identification of the factors affecting success.

Here in this work we will analyze whether the background information has any effect on the success rate of regular students. The only data we collected during the registration period of Epoka University based on the registration form. The content of this form determined by the local authorities and University Administration. In this study we tried to get answers if we can use this data to predict student success. The main objective of our study is to determine the factors that may affect the study outcomes in Epoka University.

2. DATA AND METHODOLOGY

Epoka University student management system does not provide data in the format ready for a direct statistical analysis and modeling. Therefore a data preparation and cleaning were undertaken to prepare database for modeling.

Table Descriptive statistics – Study outcome (716 students)

Descriptive						
		count		%		
	Domain	FAIL	PASS	FAIL	PASS	Total
GENDER	M	221	189	53.9	46.1	57.3
	F	78	228	25.5	74.5	42.7
COUNTRY	ALB	238	372	39.0	61.0	85.2
	TUR	35	14	71.4	28.6	6.8
	KOS	14	17	45.2	54.8	4.3
	OTH	12	14	46.2	53.8	3.6
NATIONALITY	ALB	256	382	40.1	59.9	89.1
	OTH	43	35	55.1	44.9	10.9
REGION	CITY	262	372	41.3	58.7	88.5
	VILL.	37	44	45.7	54.3	11.3
HS_GPA	UPPER	48	224	17.6	82.4	38.0
	INTER.	89	113	44.1	55.9	28.2
	LOWER	160	77	67.5	32.5	33.1

2.1. Data and Methodology

Outcome that we used in our analysis is for the outcome of the student at the end of three-year study. We measured only outcomes, labeled as: Pass and Fail. Students labeled 'Pass' successfully completed the program at the end of three years. Students labeled as 'Fail' include the withdrawn students from the program voluntarily or by the academic registry for not fulfilling the regulations. Those students who stayed on the program until the end of the study but scored less than the graduation grade (2.00) were also allocated into this category.

The data set with numeric continuous variable such as secondary school grade (HS GPA) was converted into a categorical variable with only three levels A (UPPER), B (INTERMEDIATE) or C (LOWER) denoting grades above 9 out of 10, grades between 8 and 9 and grades less than 8 respectively. Other variables (nationality, citizenship, and region) were classified upon major groups.

In this study we conducted three main types of data mining approaches. Descriptive approach which concerns the nature of the dataset such as the frequency table and the relationship between the attributes obtained using cross tabulation analysis. Predictive approach which is conducted by using four different classification trees and a comparison between these and Logistic regression to confirm the accuracy of the predictors.

Classification tree models can handle a large number of predictor variables, are non-parametric, can capture nonlinear relationships and complex interactions between predictors and dependent variable.[1]

Before generating the classification trees we classified the variables according to the study outcome, i.e. whether students are eligible to be graduated or not. We used attribute selection to rank the variables by their importance for further analysis. Then we generated the classification trees in four different growing methods.

2.2. Summary Data Description

We carried out a cross-tabulation for each variable and the study outcome after cleaning the data as shown in the table above. Table shows that the majority of the successful students are female (over 57%) which is the result of the fact that 74.5% of the female students successfully completed the study. This suggests that female students are more likely to succeed than their male classmates. In terms of country and nationality it is clearly seen that Albanian population is leading the group.

An expected result has been observed in secondary school degrees. We can say that high school degree graduation ratio is directly proportional to the university graduation ratio. While 82% of upper students were able to complete the study on time 56% of intermediate and 32% of lower group students were able to complete.

2.3. Decision Trees

Although the results of the attribute selection suggests continuing analysis with only the subset of predictors, we included all available predictors in our classification trees but only 2 variables were used in the diagrams: HS_GPA and GENDER. Even though some variables may have little significance to the overall prediction outcome, they can be essential to a specific record [1].

Almost all growing methods, (CHAID, exhaustive CHAID, CRT and QUEST) generated exactly the same trees. The largest successful group consists of 272 (38%) students. HS_GPA of this group is over 90%. The largest unsuccessful group contains 237 students (33% of all participants). They have a HS_GPA less than 80%. The next largest group considered also as unsuccessful students are male students having lower HS_GPA.

As the cross-validation estimate of the risk (0.309) indicates that the successful or unsuccessful students are predicted with an error of 30.9% of the cases which means the risk of misclassifying a student is approximately 31%. This result is consistent with the results in the CHAID classification matrix. The Overall percentage shows that the model only classified correctly 70% of students. The classification tables, however, reveal one potential problem with this model: for unsuccessful students, it predicts as successful for only 65.9% of them, which means that 34% of failing students are inaccurately classified with the passing students.

2.4. Logistic regression

The Variables not in the Equation table in block 0 shows that four of the five variables are individually significant predictors of whether a student is successful or not. Region is not a significant predictor. The variables not in the Equation table in block 1 shows that only high school grade point average and gender are significant predictors, but not the other variables. This result also confirms why these two were the only variables used in decision trees

3. CONCLUSIONS

This study examines the background information from enrolment data that impacts upon the study outcome programs at the Epoka University. Based on results, the classification accuracy from the classification trees was significantly high 71% in all tree methods. Although all the variables except the region individually significant predictors as described in attribute selection trees displayed only two variables Gender and secondary school degree. This outcome is also confirmed by the logistic regression. Block 0 classification implied that all except region were good predictors (p<,001) but block 1 classification highlighted that only gender and secondary school degree were significant.

4. REFERENCES

[1]. Kovačić, Z.J. 2010, Early Prediction of Student Success: Mining Students Enrolment Data, proceedings of Informing Science & IT Education Conference (InSITE) 2010, Open Polytechnic, Wellington, New Zealand

[2]. Bratti, M., McKnight, A., Naylor, R., & Smith, J. (2004): Higher Education Out-comes, Graduate Employment and University Performance Indicators. In: Journal of the Royal Statistical Society, 167(3), pp 475-496.

[3]. Mallinckrodt, B., & Sedlacek, W. E. (1987). Student retention and the use of campus facilities by race. NASPA Journal, 24, 28-32.