Determining Key (Predictor) Course Modules for Early Identification of Students At-Risk

Daqing Chen[1], Geoffrey Elliott[2]
[1]Department of Informatics London South Bank University, London SE1 0AA, UK
[2]Faculty of Business London South Bank University, London SE1 0AA, UK
{chend & elliotgd}@lsbu.ac.uk

Abstract - This paper addresses the problem of early identification of at-risk students, and seeks to determine modules on a given course, referred to as predictor modules, in which a student's performance is implicitly correlated to the end-of-the-first-year performance of the student. Such predictor modules may therefore be used to predict the likelihood of a student's year progression. A data mining project has been conducted for this study, and decision tree-based predictive models have been created using various historical records of students' grades and year progressions. The study reveals that a key predictor module exists, and the pass rate of the key predictor module can be used to predict students' year progression rate. A set of recommendations is given based on the key predictor module identified from the management point of view in relation to improving student retention. The study also suggests that a students' performance in a key predictor module can be directly linked to both key performance indicator and key result indicator in course management and student support.

Index Terms - Educational data mining, Student retention, Decision tree induction, Key performance indicator

1. Introduction

In recent years, education data mining has become an increasingly promising and popular research area. Universities and colleges are embracing data mining techniques as effective tools to gain a profound understanding of possible relationships and patterns among various factors that collectively affect students' academic performance and retention. Such understanding, if well-established, can be further fed as an essential input into student management and support process for better informed management decision-making.

Usually the factors and their possible relationships that are of interest to institutions in educational data mining include, but are not limited to:

- How students' socio-demographic characteristics, such as age, gender, educational background, entry qualifications and certificates, family status, work status, marital status, etc., are related to students' academic performance and progressions;
- Which key modules on a course might serve as predictors such that a student's performance in these modules can be used to indicate the likelihood of the student's progression and persistence;
- What impact students' weekly attendance may have on module grades, and how significant the impact is;
- Which course modules students tend to fail together; and
- How a student's engagement and grades in a given set of phase tests and/or coursework of a module are linked to the student's overall performance of the module.

In this paper, we place our emphasis on identifying whether, on a given course in semester one for first year students, there are any predictor modules, defined as modules in which a student's performance might serve as an indicator to predict the likelihood of the student's end-of-the-first-year academic performance and progression. Knowledge of such predictor modules is vital in identifying at-risk students at an early stage in order to provide timely support to first year students and prevent them from potentially dropping out from their courses. Relevant data sets extracted from the Student Management Unit at London South Bank University (LSBU) is used. A set of decision tree-based predictive models has been constructed using SAS Enterprise Guide and Enterprise Miner. These models reveal that a key predictor module exists, and the pass rate of the key predictor module can be used to predict students' year progression rate.

The rest of this paper is structured as follows. In Section 2 the background of this study is described and a related literature review is given. The methodology and the data set explored in the present data mining project are specified in Section 3, including variable descriptions and the main data pre-processing tasks to be carried out. In Section 4 the analysis process is outlined and its findings are discussed in detail. In section 5, some recommendations are made with regards to how the findings can be used effectively in student management and support. Finally, the main conclusion of the study is summarized in Section 6 along with suggestions for future research.

2. Background and Related Work

Student retention is one of the key issues faced by almost every educational institution. Research on this topic has a long history and some of the early work can even be traced back to the 1930s [1]. In the past decades, various models - qualitative, quantitative or a combination of both - have been proposed from different points of view in an attempt to understand and interpret the factors affecting student retention, and to use such knowledge in student support. Vincent Tinto's model of student retention is one of the widely-accepted such models [2-5]. In addition, [1, 6, 7] give a detailed and historical review of the research area, and highlight some practical guides to academics, administrators, and management

in student support based on the research findings. More recently, [8, 9] provide a useful survey on educational data mining.

In contrast to qualitative models, quantitative models of student retention allow for the use of analytical tools to formulate their predictions and to study the validity and strength of their hypothesized causal relationships precisely. In recent years quantitative models have attracted considerable research attention. This is due to, on the one hand, the rapid development of database management technologies, and on the other hand the emergence of a number of industry-strength data analysis algorithms and their integration into commercial analytical products such as SAS Enterprise Miner, IBM SPSS Clementine, and Oracle Data Miner. All these have made it much easier to deal with very large-scale data sets in data collection, storage, processing, and analysis.

In student management and support, it is a general belief that identifying any students who are struggling with their studies at the earliest possible stage, for example, in the first few weeks after their courses have commenced, or at the end of the first semester in the first year, can noticeably reduce a student's likelihood of dropping-out, if the needs of these students can be timely established and appropriate support can be provided. Research in relation to early identification of at-risk students includes [10-12]. In these studies, quantitative modelling techniques, e.g., classification and decision tree (CART), artificial neural networks (ANNs), logistic regressions and clustering analysis, have played a significant role with a clear emphasis on exploring student enrolment data (mainly socio-demographic data).

In this paper we address the problem of early identification of at-risk students by determining predictor modules on a given course we have offered that may be used to indicate and predict the likelihood of a student's year progression. Mainly, a student's performance in a module is measured by the grades awarded to the student for the coursework and/or examination of the module that the student has taken. In order to verify whether any predictor modules exist on each of the courses offered, a data mining project has been conducted. A set of decision-tree based predictive models has been built by using SAS Enterprise Miner and SAS Enterprise Guide based on records of students' grades and year progressions.

It is interesting to note that, from the course and programme management point of view, students' performance in a predictor module, if it exists, can be considered as a key performance indicator (KPI) for course and programme management, and correspondingly, students' year progression rate can be considered as a key result indictor (KRI). Therefore, identifying predictor modules can potentially contribute sensible inputs into course and programme management process.

In the following sections, detailed discussion about this data mining project is given in a step-by-step way along with the relevant findings.

3. Methodology

A. Major Data Mining Task

In this study, we focus on first year students. The reason for this is that, apart from entry qualifications/certificates, we do not have sufficient first-hand knowledge of the new students with regards to their academic motivations, personalities, and learning skills and capabilities. Moreover, first year students have to experience a transition process in which they need to effectively adapt themselves to and integrate themselves into the new social and academic environment of university. The time taken for each student to complete this process varies diversely with uncertain outcomes in general. As a result, it is usually more difficult to provide individualized support early on in the academic year to first year students compared to those of later years. Therefore, first year students seem to be more "vulnerable" and "unstable" in the new university environment. As such, the major data mining task in this project is to uncover any implicit correlations/relationships between module grades and year progressions of first year students.

In this data mining project, the well-known SEMMA (Sample, Explore, Modify, Model, and Assess) methodology for data mining is adopted. Originally proposed by SAS Institute, Inc, this methodology has been integrated into SAS Enterprise Miner [13]. In addition, prior to model construction, SAS Enterprise Guide has been used in performing essential data pre-processing tasks.

B. Data Requirement and Description

The data set to be explored (i.e., the target data set) in this research includes the performance and year progression records of all the first year students enrolled on Computing and Business Information Technology (BIT) courses at LSBU during the academic years 2004-2009. Usually, each first year student on these two courses had 8 modules to study across two semesters, 4 modules for each semester. Apparently, one would be only interested in how each student's performance in the modules of semester one is correlated to the student's progression at the end of the first year.

The original data set contains more than 30 variables (fields). Some of the variables are not relevant to the present study, for instance, a student's name, age, gender, home address, and therefore are excluded. Finally only 12 variables have been selected for the modelling purpose as shown in Table I. The relationship of them to be identified may be expressed as

$$PROG = f(UNIT\_GRADE\_1,\ldots,UNIT\_GRADE\_11)$$

where *PROG* represents year progression result, and *UNIT_GRADE_i* (*i=1, 2,…,11*) donates the grade that a student has achieved in the *i*th course module. Note that there were totally 11 modules and 6 of them were common to all first year students. Each student took other 2 different modules depending on their courses.

TABLE I Variables in the Target Data Set

Variable	Data Type	Description; Typical Values and Meanings
STUDENT_ID	Nominal	Student ID number; 2123456: the first 2 digits indicate enrolment year
ACAD	Nominal	Academic period; 05/06: academic year 2005/6
CRSE	Nominal	Course code; 353: Computing course
SESS	Nominal	Course year code 1FS00: year one
PROG	Nominal	Progression code; P: Pass to next year F: Fail and re-enrolment not allowed COU: Continue outstanding modules and re-enrolled onto the same year of the course RYA: Repeat the year or just failed modules MLS: Repeat failed modules and may study additional modules at next level POU: Pass with outstanding modules
AOS_CODE	Nominal	Module code
ASS_ID	Nominal	Module assessment type (component) and percentage; CW1_100: 100% coursework EX1_40: 40% examination CW1_60: 60% coursework
ASS_GRADE	Nominal	Module assessment grade; D: distinction P: Pass M: Merit F: Fail R: Referred RF: Failed referred assessment and module DF: Deferred module FD: Fail following deferral FE: Fail and re-attend RP: Passed after referral
ASS_MARK	Numeric	Module assessment mark (%)
ASS_RESIT	Nominal	Module assessment retaken mark (%)
UNIT_GRADE	Nominal	Module overall grade; D: distinction P: Pass M: Merit F: Fail
UNIT_MARK	Numeric	Module overall mark (%)
CAP_MARK	Numeric	Capped assessment mark if retaken (%); 40%
STAGE	Nominal	Student enrollment status; EFE: Education for Employment

C. Related Data Pre-processing

Data pre-processing plays a vital role in delivering quality data mining results. In this project, the relevant data pre-processing tasks include:

- Retrieve original data from the LSBU Students Management database;
- Filter out the target data set with the selected variables from the original data set (Note that this can also be conducted by using SAS Enterprise Miner when assigning each of the variables a model role for model construction);
- Sort out the data set by student ID number and identify any records that contain missing values and/or inconsistent values;
- Resolve missing values and any inconsistencies by consulting the data administrator of the database and further replacing them with appropriate values;
- Transport the cleansed data set into such a data set table that the headers of the table consist of the 12 variables, i.e., *UNIT_GRADE_i* (*i*=1, 2, …, 11) and *PROG* together with *STUDENT_ID*. Each row of the table is distinct and corresponds a complete record of module performance and year progression of one student only;
- Transform the cleansed and transported data set into SAS format so that it can be processed by SAS suite; and
- Partition the whole data set into 5 sub-sets, each corresponding to the students' record in one of the 5 consecutive academic years, respectively.

The original data set retrieved was in CSV format, and was uploaded into SAS Enterprise Guide 4.2 for data pre-processing. A number of appropriate SAS procedures, such as Proc Data, Proc Tran, and Proc SQL, were applied to the data set using SAS Enterprise Guide in order to conduct all the required data pre-processing tasks.

4. Modelling, Analysis and Findings

The pre-processed target data set was uploaded into SAS Enterprise Miner 5.2 for constructing a predictive model using decision tree induction algorithm. The project diagram in SAS Enterprise Miner 5.2 is shown in Fig. 1.

For the decision tree model to be created, the variable *PROG* was used as the target variable of the model and all the 11 variables *{UNIT_GRADE_i}* were used as independent (input) variables of the model. The variable *STUDENT_ID* was dropped. Accordingly each of the variables in the project diagram was set to an appropriate model role for the decision tree model, as illustrated in Fig. 2. Due to the data type of the target variable *PROG*, Entropy was chosen as the criterion in the Property Panel of the Decision Tree node, as depicted in Fig. 3, to determine which variable should be selected to split the data set iteratively in the process of decision tree induction.

Using each of the 5 sub-data sets in turn a decision tree model was created in the Autonomous Manner in SAS Enterpriser Miner first, and then was refined by using the Interactive Manner. Eventually, 5 decision trees have been created, respectively.

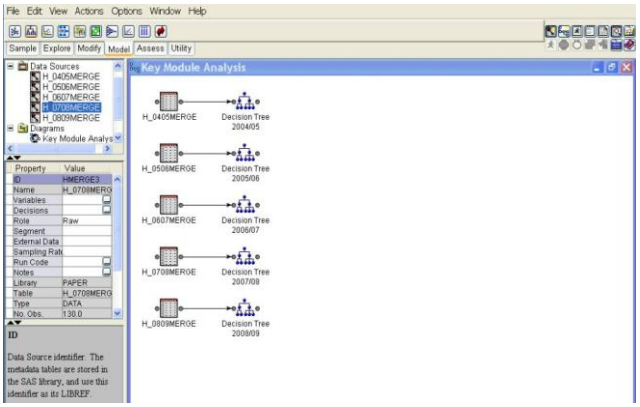

Fig. 1 Project diagram in SAS Enterprise Miner.

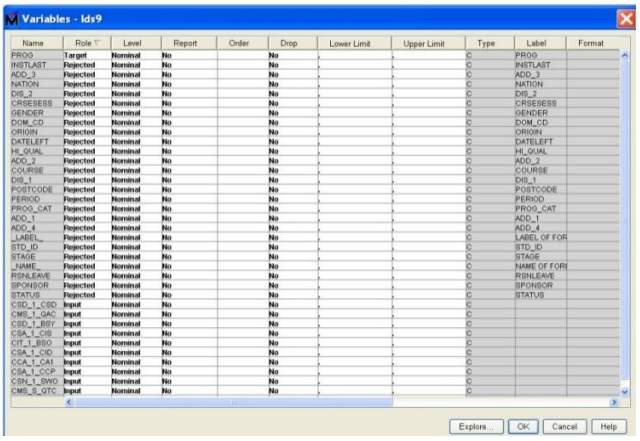

Fig. 2 Model role setting of variables.

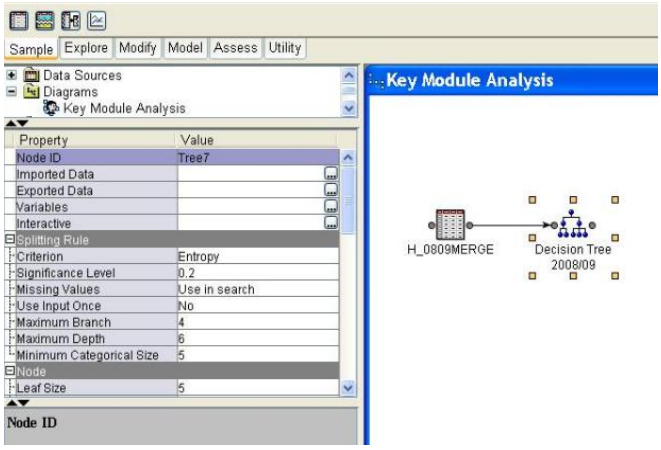

Fig. 3 Decision tree model setting in SAS Enterprise Miner.

From the 5 resultant decision trees, we have found that there was one particular semester one module, CSD-1-CSD, which was always associated with the root node of each of the 5 decision trees. In other words, this correlation pattern remains stable and unambiguous for all the 5 years' records, although, in theory, any of the 11 course module-related variables *{UNIT_GRADE_i}* might be associated with the root node in one or more of the decision trees constructed. As an example, Fig. 4 depicts the decision tree built for the sub-data set of the academic year 2008-2009.

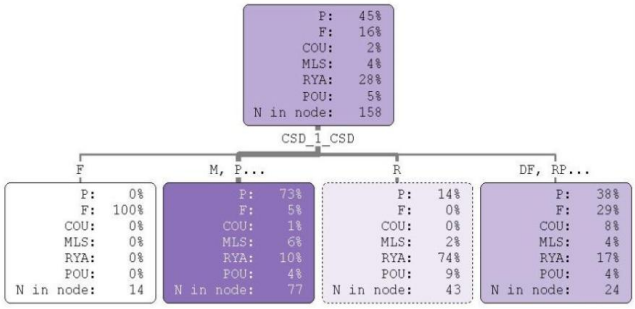

Fig. 4 Resultant decision tree based on the records of academic year 2008/09.

This finding suggests that semester one module CSD-1-CSD seems to be a predictor module. Examining all the 5 decision trees carefully we can find that there was a strong correlation between a student's performance in this particular module and the year progression of the student. Take the decision tree of the academic year 2008-2009 as an example (see Fig. 4). There were totally 158 student samples in that year's cohort, and 77 of them passed this module. Among these 77 students, only less than 4 students (5%) did not progress to the second year of their studies, and 50 of the students succeeded (72.7%) to the second year. On the other hand, there were 14 students who failed this module, and none of them succeeded to the second year! Also, there were 43 students who had re-taken the examination of this module, and only 5 of them (14%) were able to progress to the second year successfully. To make this clearer, Table II gives the set of decision rules represented by the decision tree shown in Fig. 4.

In addition, Table III illustrates the correlations between students' performance in the predictor module and students' year progression based on the results obtained from all the 5 decision trees (For the meanings of F and RYA, refer to Table I).

To summarize the findings from the above analyses, we can conclude that: for any student who has passed the key predictor module the student seems to be more likely to progress to the second year successfully. On the other hand, for any student who has failed the predictor module the student seems to be more likely to repeat some of first year modules, or repeat the whole first year, or, in the worst cases, even fail the course, i.e., drop out from a course. This pattern remains valid for all of the records of the five consecutive academic years.

We next examined whether and how the number of the students who progressed to the second year might be linked to the number of the students who passed the predictor module. A careful examination on the data set has revealed that the number of the students who progressed to the second year is roughly similar to or not more than the number of the students who passed the predictor module, as shown in Table IV. In other words, the pass rate of the predictor module can be used to predict students' year progression rate.

TABLE II Decision Rules Represented by the Decision Tree (2008/2009 Data Set)

IF CSD_1_CSD EQUALS F *Then* *NODE: 1* *N: 14* *P: 0.0%* *F: 100.0* *COU: 0.0%* *MLS: 0.0%* *RYA: 0.0%* *POU: 0.0%*
IF CSD_1_CSD IS ONE OF: M P CP *Then* *NODE: 2* *N: 77* *P: 72.7%* *F: 5.2%* *COU: 1.3%* *MLS: 6.5%* *RYA: 10.4%* *POU: 3.9%*
IF CSD_1_CSD IS ONE OF: R FE *Then* *NODE: 3* *N: 43* *P: 14.0%* *F: 0.0%* *COU: 0.0%* *MLS: 2.3%* *RYA: 74.4%* *POU: 9.3%*
IF CSD_1_CSD IS ONE OF: DF RP D RC RE *THEN* *NODE: 4* *N: 24* *P: 37.5%* *F: 29.2%* *COU: 8.3%* *MLS: 4.2%* *RYA: 16.7%* *POU: 4.1%*

TABLE III Correlation Analysis

Academic year	% of students who passed the predictor module and succeeded to the second year	% of students who failed the predictor module and did not succeed to the second year
2008/2009	72%	100% (F: 100%)
2007/2008	74%	100% (F: 64%, RFA: 36%)
2006/2007	82%	100% (F: 50%, RFA: 50%)
2005/2006	81%	100% (F: 73%, RFA: 27%)
2004/2005	73%	100% (F: 59%, RFA: 41%)

5. Recommendations and Discussions

The findings discussed in the previous section are very encouraging. The existence of stable predictor module(s) on a given course can be potentially used to enhance student management and support. Based on these findings some recommendations are given and discussed in this section in order to explore various possible ways to apply them into the practice of student management and support.

Recommendation 1: Monitoring student performance in all course modules, particularly in the predictor module(s). Course directors need to monitor closely every student's performance in each module in semester one in the first year, in particular, in the predictor module(s). Any failure in a predictor module should be followed up immediately, so that the student involved can be flagged as at-risk at an early stage and can be given timely help. At university level, an effective and workable plan and procedure should be set up to standardize and support such monitoring activities.

Recommendation 2: Making students aware of the predictor modules. At the beginning of semester one in the first year, students should be well-informed the fact that a failure in a predictor module may not only simply mean a failure in a single course module, so that the students can be alerted to seek for effective help proactively if they find themselves struggling to cope with any key predictor modules.

Recommendation 3: Using student performance in predictor modules as a key performance indictor for management purpose. From the viewpoint of course and programme management and student support, students' overall performance in the predictor modules should be considered as a KPI as it directly relates to the students' year progression rate. Monitoring this KPI closely can potentially generate a good prediction at an early stage for students' end-of-first-year progression.

Recommendation 4: Integrating the information of the predictor module(s) into course development and student enrollment process. In course syllabus design and development, it is worthy to consider how each predictor module is logically linked to other course modules in terms of the contents, rationale, prerequisite learning, level of technical challenges, and pedagogical approaches, etc., in order to better understand why the predictor modules are so crucial that it essentially testifies a student's academic capability. Furthermore it is important for the management to explore possible ways to reflect the particular intellectual challenges posed by predictor modules in the process of student enrolment, so as to maximize the enrolment of students academically capable of completing their courses.

TABLE IV Correlation Analysis

Academic year	Number of students who progressed to year 2	Number of students who passed key module
2008/2009	71	79
2007/2008	65	66
2006/2007	74	77
2005/2006	91	93
2004/2005	80	86

6. Conclusion and Future Work

As shown in this project, educational data mining can be use to help better understand our students and enhance student support and course management. Students' performance in the identified key predictor module can be directly linked to both KPIs and KRIs of course management and student support. Monitoring a student's performance in a predictor module can provide definitive and unique information for course management.

Future research involves applying association analysis (market basket analysis), as an alternative technique, into the required modelling. It is also interesting to see if the models created by decision tree induction and association analysis have any similarities and/or differences in terms of identifying stable patterns relating to key predictor modules. A further theoretical analysis on these issues deserves a great deal of research effort as well. Other alternative modelling techniques to be considered include cluster and segmentation analysis, and risk estimation and scoring.

From the long term point of view, the methodology adopted in this research can be extended to explore other courses across our university to identify whether any predictor modules exit, and eventually to investigate and model the whole life-cycle of students at university.

Acknowledgment

This project was supported partially by the Research Opportunity Fund of London South Bank University 2009-10. The authors would like to thank Jennifer Laws for her help in preparing the data set for this study.

References

[1] A. Seidman (Ed.), *College Student Retention – Formula for Student Success*, American Council on Education and Praeger Publisher, USA, 2005.

[2] V. Tinto, "Dropout from higher education: A theoretical synthesis of recent research," *Review of Educational Research*, vol. 45, pp. 89-125, 1975.

[3] V. Tinto, "Limits of theory and practice in student attrition," *Journal of Higher Education*, vol. 53, no. 6, pp. 687-700, 1982.

[4] V. Tinto, "Theory of Student Departure Revisited," In: J. C. Smart (Ed.), *Higher Education: Handbook of Theory and Practice*, vol. 2, pp.359-384, Agathon Press, New York, 1986.

[5] V. Tinto, *Leaving College: Rethinking the Causes and Cures of Student Attrition*, 2nd ed., Chicago: University of Chicago Press, 1993.

[6] A. Cook and B.S. Rushton (Ed.), *How to Recruit and Retain Higher Education Students: A Handbook of Good Practice*, Routledge: New York and London, 2009.

[7] M. Yorke, and B. Longden, *Retention and Student Success in Higher Education*, London: Open University, 2004.

[8] C. Romero, and S. Ventura, "*Educational data mining: A survey from 1995 to 2005*", *Expert Systems with Applications*, vol. 33, pp. 135-146, 2007.

[9] R. S. J. D. Baker and K. Yacef, "*The state of educational data mining in 2009: A review and future visions*", *Journal of Educational Data Mining*, vol. 1, pp. 3-17, 2009.

[10] Z. J. Kovačić, "*Early Prediction of Student Success: Mining Students Enrolment Data," Proceeding of Informing Science & IT Education Conference (InSITE)*, pp. 647-665, Southern Italy, June 19-24, 2010.

[11] L. Horstmanshof, and C. Zimitat, "*Future time orientation predicts academic engagement among first-year university students*," *British Journal of Educational Psychology*, vol. 77, no. 3, pp. 703-718, 2007.

[12] T. L. Strayhorn, "*An examination of the impact of first-year seminars on correlates of college student retention*," *Journal of the First-Year Experience & Students in Transition*, vol. 21, no. 1, pp. 9-27, 2009.

[13] K. S. Sarma, *Predictive Modeling with SAS Enterprise Miner*, NC: SAS Institute Inc., Cary, USA (2007).