The Use of Data Mining in Higher Education Strategic Enrollment Management

Sharon O'Boyle

George Mason University

IT 103, Section B02

June 20, 2013

GMU Honor Code

By placing this statement on my webpage, I certify that I have read and understand the GMU Honor Code on http://oai.gmu.edu/honor-code/ . I am fully aware of the following sections of the Honor Code: Extent of the Honor Code, Responsibility of the Student and Penalty. In addition, I have received permission from the copyright holder for any copyrighted material that is displayed on my site. This includes quoting extensive amounts of text, any material copied directly from a web page and graphics/pictures that are copyrighted. This project or subject material has not been used in another class by me or any other student. Finally, I certify that this site is not for commercial purposes, which is a violation of the George Mason Responsible Use of Computing (RUC) Policy posted on http://universitypolicy.gmu.edu/policies/responsible-use-of-computing/ web site.

Introduction

Data mining is a process for extracting useful information from large amounts of data. Data mining has been used for many years in various industries, and recently colleges and universities have begun to use this process. Colleges and universities have much data available. Just as in other industries, these institutions must develop the ability to determine meaningful and predictive insights from the data. This paper will describe how applications of data mining in strategic enrollment management are being developed.

Background on Data Mining and Strategic Enrollment Management in Colleges and Universities

According to SAS (an organization that produces analytical software), data mining is "an iterative process of selecting, exploring and modeling large amounts of data to identify meaningful, logical patterns and relationships among key variables" (Patel et al., 2010). One of the common uses for data mining is customer segmentation and targeting. For colleges and universities, the customer would be potential enrolled students. Strategic Enrollment Management (SEM) is being used at colleges and universities to meet their enrollment goals. According to Lingrell (2012), data and information are what are needed for the "strategic" aspect of SEM.

In traditional data analysis, summary statistics of a group are commonly the result: averages, percentages, etc. In data mining, information about an individual is sought. Chang (2006) says that data mining can identify hidden patterns in data and allow predictions to be made at the individual level, which is what higher education institutions desire to do. Recruiters can then use this information during the recruitment cycle to identify which individuals are more likely to apply and eventually enroll at their institution.

Current Use and Benefits

Two data mining processes that are currently being used are SEMMA and CRISP-DM. A description of these two processes follows, along with an example of how each has been used in a university setting.

The SEMMA data mining process was developed by SAS. The steps in this process are as follows: **S**ample -> **E**xplore -> **M**odify -> **M**odel -> **A**ssess. The SAS technology that utilizes this approach is SAS Enterprise Miner. In the Sample phase, the sample must be large enough so that hidden relationships and patterns can be detected, but small enough to be manageable. In the Explore phase techniques like clustering, classification and regression look for relationships to study during the process. Anomalies and outliers are also examined. The Modify phase selects and transforms variables for the next phases. The Model phase uses various analytical tools to determine the best model for predicting outcomes. Finally, the Assess phase studies the reliability and usefulness of the results. Modifications might be necessary and some of the steps might need to be repeated.

An example of a university using the SAS Enterprise Miner approach is The University of Central Florida, Division of Graduate Studies which has used data mining as a tool in graduate admissions. Data were collected from several sources and consolidated into a single data mart. They used the SAS Enterprise Miner software as their data mining tool. After initial analysis, they selected 23 predictor variables (specific graduate program, academic level, gender, ethnic group, etc.) and one response variable (whether or not the student enrolled). They used a logistic regression model which is an appropriate model when predicting a binary response (enroll/not enroll). Half of the data was used to build the model. The remaining half was used to test the fit of the model. Each predictor was given a weight depending on the strength of its relation to the

response variable. The findings indicated a valid model and they used the resulting model to predict enrollment for the fall 2007 semester.

The second widely used data mining approach is CRISP-DM (**CR**oss-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining). The CRISP-DM data mining process is as follows: (Business Understanding -> Data Understanding -> Data Preparation -> Modeling ->Evaluation -> Deployment). Chang (2006) describes the use of the CRISP-DM data mining process in the freshman admissions process at a large unnamed state university. The data mining/modeling software Clementine was used in this study. Clementine (now known as SPSS Modeler) is a product of SPSS (a company, now a part of IBM, which produces analytical software).

In the Business Understanding phase, the basic questions to be answered are defined. For example, "Which types of students are likely to enroll?" "Can demographic features be used as predictors?" In the Data Understanding phase a study is done to examine what data is available and how it can be mined. Chang examined demographic, academic and communications activity data. Fifteen predictor variables (high school GPA, gender, ethnicity, etc.) and one outcome variable (enrollment status) were chosen. Data Preparation includes cleansing, combining and transforming the data.

In the Modeling phase, various models are implemented, and in the Evaluate phase the models are tested for validity. Three different modeling techniques were used: classification and regression tree (C&RT), neural networks, and logistic regression. Similar to the SEMMA approach, part of the data is used to build the model while the remaining part is used for the validity test. Results indicated that enrollment could be predicted to some degree. Finally, validated models are put into practice in the Deployment phase. The models were then used to predict enrollment in future years.

Security Concerns

There are a number of security concerns surrounding the use of data mining in the college enrollment process. Data mining, by its very nature, involves tremendous amounts of data. In the college admissions process, the data is often obtained for and from children under 18. The types of data that are collected and stored are often personal and sensitive such as birthday and ethnicity. The data can are obtained from multiple sources. Security breaches are possible at any stage (ex. collection, transmission, storage) and could result in serious consequences such as identity theft.

Legal, Ethical and Social Issues

There are legal and ethical issues related to the collecting, storing, analyzing and reporting of potential enrolled students. Two examples of these issues are described below.

In 2003 the Federal Trade Commission settled a lawsuit against Education Research Center of America, Inc. (ERCA) and Student Marketing Group, Inc. (SMG). The FTC claimed that these companies deceptively collected detailed personal information about K-12 students saying that it was for educational purposes (i.e. for use by colleges). However, the data was rarely used for the stated purpose. Instead, the companies sold it to marketers for commercial purposes. As a result of the suit, the companies were restricted from using the already collected data and were instructed to provide full disclosure about how they will use any data to be collected in the future.

In 2011 U.S. Representatives Edward Markey (D-Mass.) and Joe Barton (R-Texas), co-chairs of the Congressional Privacy Caucus, requested information from the College Board and ACT, Inc. Each year these two companies collect information from millions of students when they take the SAT and ACT tests for college admission. The Representatives were concerned

about how these companies collect, store and sell the names and personal information of students who take their tests. The two representatives were working on legislation that would protect the privacy of children under the age of 18.

Since data mining involves customer segmentation and categorizing an individual based on his or her attributes (including ethnicity, gender, etc.), there are possible social risks associated with this practice. For example, the College Board collects data on ethnicity and religion. Colleges and universities can buy this data and then use this to target specific populations.

Further Required Research and Potential Future Use

The use of data mining in enrollment management is a fairly new development. Current data mining is done primarily on simple numeric and categorical data. In the future, data mining will include more complex data types. In addition, for any model that has been designed, further refinement is possible by examining other variables and their relationships. Research in data mining will result in new methods to determine the most interesting characteristics in the data. As models are developed and implemented, they can be used as a tool in enrollment management.

Conclusion

Data mining, along with traditional data analysis, is a valuable tool that that is being used in Strategic Enrollment Management to achieve desired enrollment targets in colleges and universities. In situations where it has been applied, it has been proven to successfully predict enrollment, at least to a degree. More research is needed to fully take advantage of the data mining processes and technologies.

References

Chang, L. (2006). Applying data mining to predict college admissions yield: A case study. *New Directions for Institutional Research,* (131), 53-68. Retrieved June 7, 2013, from http://ehis.ebscohost.com/ehost/results?sid=3d6f79b6-122e-4adb-917b-77ad6a5f4c92%40sessionmgr14&vid=1&hid=4&bquery=SO+(New+directions+for+institutional+research)+and+(%22Applying+Data+Mining+to+Predict+College+Admissions+Yield+A+Case+Study%22)+and+DT+%222006%22&bdata=JmRiPWE5aCZ0eXBlPTEmc2l0ZT1laG9zdC1saXZlJnNjb3BlPXNpdGU%3d

This is a journal article that describes the application of a data mining approach to freshman admissions at a large public university. The author is Director of Institutional Research and Analysis at Colorado State University-Pueblo.

Federal Trade Commission. (2003). Student survey companies settle FTC charges [Press release]. Retrieved June 7, 2013 from http://www.ftc.gov/opa/2003/01/ecra.shtm

This is a press release that summarizes the settlement of a lawsuit between the FTC and two companies that were collecting and selling student data.

Jones, T., & Vaiciulis, A. (2007). Data mining, predictive modeling, and recruiting targets. *College and University, 82*(2), 47-49. Retrieved June 7, 2013, from http://search.proquest.com/docview/225611369?accountid=14541

This is a journal article that describes a data mining approach used in the graduate admissions process at the University of Central Florida. Jones is Executive Director for Graduate Studies at the University of Central Florida; Vaiciulis is a Master's student in the Data Mining Program at the University of Central Florida.

Lingrell, S. (2012). Getting it right: Data and good decisions. In B. Bontrager, D. Ingersoll & R. Ingersoll (Eds.), *Strategic enrollment management: Transforming higher education* (pp.155-

171). Washington, DC: American Association of Collegiate Registrars and Admissions Officers.

This is a chapter from a book on Strategic Enrollment Management. It describes the importance of good data. The author is Vice President for Student Affairs and Enrollment Management at the University of West Georgia.

Loren, J. (2011, May 27). SAT test demanding teen information prompts regulator query. Retrieved June 7, 2013, from http://www.bloomberg.com/news/2011-05-26/sat-test-owner-to-face-query-on-teen-privacy-from-lawmakers.html

This is an article that describes an inquiry by two U.S. Representatives into how ACT Inc. and the College Board collect, store and sell the data from their test-takers.

Patel, P., Thompson, W. & Stephens, C. (2010, June). *Data mining 101: How to reveal new insights in existing data to improve performance* [White paper]. Cary, NC: SAS Institute Inc.

This is a white paper published by the SAS Institute Inc. that describes the methodology used in their data mining product Enterprise Miner. Patel is a Global Marketing Manager; Thompson is an Analytics Product Manager; Stephens is a Product Manager.