



Data Mining Applications in Higher Education

Jing Luan, PhD
Chief Planning and Research Officer, Cabrillo College
Founder, Knowledge Discovery Laboratories

Table of contents

Introduction	2
Data mining overview	2
Data mining models and algorithms	2
Frequently used algorithms	3
Data mining in higher education	3
Supervised and unsupervised modeling	3
Data mining applications in higher education	4
Case study one: Creating meaningful learning outcome typologies	4
Case study two: Academic planning and interventions—transfer prediction	5
Case study three: Predicting alumni pledges	6
Conclusion	7
About SPSS Inc.	7



Introduction

One of the biggest challenges that higher education faces today is predicting the paths of students and alumni. Institutions would like to know, for example, which students will enroll in particular course programs, and which students will need assistance in order to graduate. Are some students more likely to transfer than others? What groups of alumni are most likely to offer pledges? In addition to this challenge, traditional issues such as enrollment management and time-to-degree continue to motivate higher education institutions to search for better solutions.

One way to effectively address these student and alumni challenges is through the analysis and presentation of data, or data mining. Data mining enables organizations to use their current reporting capabilities to uncover and understand hidden patterns in vast databases. These patterns are then built into data mining models and used to predict individual behavior with high accuracy. As a result of this insight, institutions are able to allocate resources and staff more effectively. Data mining may, for example, give an institution the information necessary to take action before a student drops out, or to efficiently allocate resources with an accurate estimate of how many students will take a particular course.

This white paper addresses the capabilities of data mining and its applications in higher education. Three case studies demonstrate how data mining saves resources while maximizing efficiency, and increases productivity without increasing cost. The paper begins with an overview of data mining capabilities.

Data mining overview

Data mining uses a combination of an explicit knowledge base, sophisticated analytical skills, and domain knowledge to uncover hidden trends and patterns. These trends and patterns form the basis of predictive models that enable analysts to produce new observations from existing data.

Gartner Inc.'s definition of data mining is the most comprehensive: "...the process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of data stored in repositories, and by using pattern recognition technologies, as well as statistical and mathematical techniques." Data mining should be performed on very large or raw datasets using either supervised or unsupervised data mining algorithms. Note that data mining cannot occur without direct interaction with unitary data.

The most successful data mining projects comply with the guidelines and steps in the Cross-Industry Standard Process for Data Mining (CRISP-DM). As the demand for data mining increases and more algorithms are created, CRISP-DM ensures good practices that everyone can follow. For more information on CRISP-DM, refer to the white paper, "CRISP-DM 1.0: Step-by-step data mining guide," available from the SPSS Web site at www.spss.com/downloads.

Data mining models and algorithms

Models house the steps, modules, and resources of the data mining process. Some data mining models include the entire process for a particular purpose, be it to cluster or predict. A model is, however, different from an algorithm. An algorithm is a specific, mathematically driven data mining function, such as a neural network, classification and regression tree (C&RT), or K-means.



Frequently used algorithms

Beyond those mentioned in this paper, there are the genetic, market basket analysis, Kohonen network, link analysis, time/sequence, and text mining algorithms, to name just a few. Most of the traditional statistics, such as logistic regression and principal component analysis, are also treated as data mining tools. In addition, university laboratories often produce new algorithms for specific business or scientific research purposes.

Data mining in higher education

Data mining is a powerful tool for academic intervention. Through data mining, a university could, for example, predict with 85 percent accuracy which students will or will not graduate. The university could use this information to concentrate academic assistance on those students most at risk.

In order to understand how and why data mining works, it's important to understand a few fundamental concepts. First, data mining relies on four essential methods: Classification, categorization, estimation, and visualization. Classification identifies associations and clusters, and separates subjects under study. Categorization uses rule induction algorithms to handle categorical outcomes, such as “persist” or “dropout,” and “transfer” or “stay.” Estimation includes predictive functions or likelihood and deals with continuous outcome variables, such as GPA and salary level. Visualization uses interactive graphs to demonstrate mathematically induced rules and scores, and is far more sophisticated than pie or bar charts. Visualization is used primarily to depict three-dimensional geographic locations of mathematical coordinates.

Higher education institutions can use classification, for example, for a comprehensive analysis of student characteristics, or use estimation to predict the likelihood of a variety of outcomes, such as transferability, persistence, retention, and course success.

Supervised and unsupervised modeling

Classification and estimation use either unsupervised or supervised modeling techniques. Unsupervised data mining is used for situations in which particular groupings or patterns are unknown. In student course databases, for example, little is known about which courses are usually taken as a group, or which course types are associated with which student types. Unsupervised data mining is often used first to study patterns and search for previously hidden patterns, in order to understand, classify, typify, and code the objects of study before applying theories.

Supervised data mining, however, is used with records that have a known outcome. A graduation database, for example, contains records of students who completed their studies, as well as of those who dropped out. Supervised data mining is used to study the academic behavior of both groups, with the intention of linking behavior patterns to academic histories and other recorded information.

This so-called “machine learning” uses artificial intelligence to induct rules and delineate patterns that analysts can apply to new data. Once a model performs well, the analyst can feed in another student group, such as new students, and the model applies the learned information to the new group to predict the likelihood of graduation. All of these steps are automated to produce accurate estimations quickly, saving time and resources compared to conventional behavior prediction methods.

Data mining applications in higher education

Data mining is already fundamental to the private sector. Many of the data mining techniques used in the corporate world, however, are transferable to higher education. **Figure 1**, below, shows the higher education equivalents of critical business questions answered by data mining.

Private sector questions	Higher education equivalents
Who are my most profitable customers?	Which students are taking the most credit hours?
Who are my repeat Web site visitors?	Which students are most likely to return for more classes?
Who are my loyal customers?	Who are the “persisters” at my university/college?
Who is likely to increase his/her purchases?	Which alumni are likely to make larger donations?
Which customers are likely to defect to competitors?	What types of courses will attract more students?

Figure 1: Data mining questions in the private sector and their higher education equivalents

The following three case studies illustrate key applications of data mining in higher education.

Case study one: Creating meaningful learning outcome typologies

Challenge

“What do institutions know about their students?” If the answer is a recital of enrollment percentages or other basic counts, institutions do not know their students as well as they could. This case study demonstrates how suburban community colleges can establish learning outcome typologies for students using unsupervised data mining.


A typical suburban community college with an enrollment of 15,000 traditionally identifies its students as “transfer oriented,” “vocational education directed,” or “basic skill upgraders.” These identifications, however, are based on students’ initial declarations of educational goals at enrollment. While these are inclusive classifications, they don’t help to illustrate the differences between each student type.

Solution

To establish appropriate typologies for the 15,000 students, researchers used both TwoStep and K-means, two powerful clustering algorithms. They first applied the algorithms to the general groupings identified above, with mixed results. The boundaries among clusters were unclear and dispersed, and even after repeated testing on holdout datasets, as well as the removal of suspected outliers (cases that do not appear to belong to any group), the results did not improve significantly. It’s possible that the students’ initial declaration of goals did not dictate their academic behavior.

The researchers then used a replacement method that looked at educational outcomes in combination with lengths of study. Defining educational outcomes is easier said than done. Enough time must pass to conclude that a student has reached a certain milestone. Dropping out is also an outcome by itself. Further work was conducted to determine length of study, which required decisions on how to deal with “stopouts,” students who left school and later returned.

All of these situations test the data miner’s domain knowledge. There are no absolutely right or wrong typologies. In essence, a typology is a good one if it serves a particular research objective.



After either removing the outliers or adding them to a particular cluster, the TwoStep algorithm produced the following clusters: “Transfers,” “vocational students,” “basic skills students,” “students with mixed outcomes,” and “dropouts.” K-means validated these clusters. Introducing the length-of-study element gave new dimensions to each cluster. Some transfer students completed their studies quickly; some vocational students took longer; and other students appeared to simply take one or two courses at a time.

Results

Data mining, combined with student demographics and other information, enabled the college to improve its understanding of its student types. Certain older students, for example, tended to take their time, while younger students with more privileged socioeconomic backgrounds often took high credit courses and graduated quickly. One of the most interesting steps in classification is naming the typologies. The college used the term “transfer speeders,” for example, to describe students who quickly accumulated units, while those who took classes for a considerable length of time were “college historians.” Other student clusters were “fence sitters,” “skill upgraders,” etc.

Typologies are important because they go beyond conventional student profiling to identify homogenous groups of students, thus increasing the accuracy of predictive modeling algorithms. Even if a data mining project ends with the discovery of appropriate typologies, the newly discovered patterns and relationships help educators and administrators better meet the needs of varied student groups.

Case study two: Academic planning and interventions—transfer prediction


Challenge

This case study showcases a solution to a vexing higher education problem: How to accurately predict academic outcomes in order to facilitate timely academic intervention. When institutions use data mining to predict which students are most at risk, institutions can prevent a student from failing before the student is even aware that he or she is at risk.

More than half of community college students identify transferring to four-year universities as their goal. Due to academic difficulties, however, many either take a long time to transfer or never transfer at all. While it has traditionally been difficult to discover which students transfer, the National Student Clearing House now allows community colleges and universities to match their data. This means that data miners and decision makers can link the academic behavior of community college students to their transfer outcomes.

Solution

Building an effective data mining model with this data involves a combination of typologies and domain knowledge. Transfer education domain knowledge emphasizes that the most effective means of increasing student transfers is to identify transfer-directed students as early as possible. Grooming those who are most likely to transfer is far more meaningful than counting the number of students who have accumulated enough units to transfer.



Using the transfer outcome data, analysts built a dataset containing students who fell under the general transfer clusters of “speeders” and “laggards.” The dataset was split into a test dataset and a validation dataset, using a proprietary randomization method. The outcome variable was transfer. Other variables, such as demographics, courses taken, units accumulated, and financial aid, were predictors to be analyzed without stepwise testing for significance. Data mining is very tolerant of variable interactions and non-linear relationships in data. Supervised data mining was the obvious and appropriate method; therefore, the analysts ran neural network and rule induction algorithms simultaneously in order to contrast and compare the prediction accuracy.

Results

Data mining enabled the college to accurately identify good transfer candidates. After extensive machine learning, the neural network algorithm, Neural Net, had a prediction accuracy of 72 percent, and the rule induction algorithms, C5.0 and C&RT, had a prediction accuracy of 80 percent. The models then ran against the test dataset and produced similar results, indicating their grasp of the patterns within the data.

Case study three: Predicting alumni pledges

Challenge

For a typical urban university of 25,000, the alumni population can be as much as ten times its enrollment. Most universities send mailings to alumni on a regular basis, even when alumni fail to respond. These mailings typically cost more than \$100,000 a year. This case study shows how data mining helps universities focus on the alumni most likely to make pledges.

Solution

It’s often difficult to determine whether mailings directly affect the volume and value of alumni pledges. Given the same type of mailing, one alumnus may contribute regularly while another may not. Adding to the confusion is the presence of outliers, such as alumni who unexpectedly contribute large sums. How do institutions identify and cultivate relationships with “outlier” alumni?

In **Figure 2**, on the next page, the chart shows the benefit of using data mining to determine alumni mail recipients versus simply mailing to all alumni. The curved line is the optimal return rate (alumni contributions) as predicted by data mining. The straight, 45-degree line is the predicted result if the entire population received the mailing. In this case, the chart indicates that when the mailing reached the 30th percentile of the population predicted by data mining to be responsive, 80 percent would respond with a pledge. If the entire population received the mailing, only 40 percent would respond. If every percentage point = \$2,500, savings = $(70\% * \$2,500) - (30\% * \$2,500) = \$175,000 - \$75,000 = \$100,000$. Without data mining, therefore, it would cost \$100,000 more to reach all 80 percent.

Figure 2, below, shows the benefits that data mining may bring to alumni pledge campaigns.

■ Gain chart

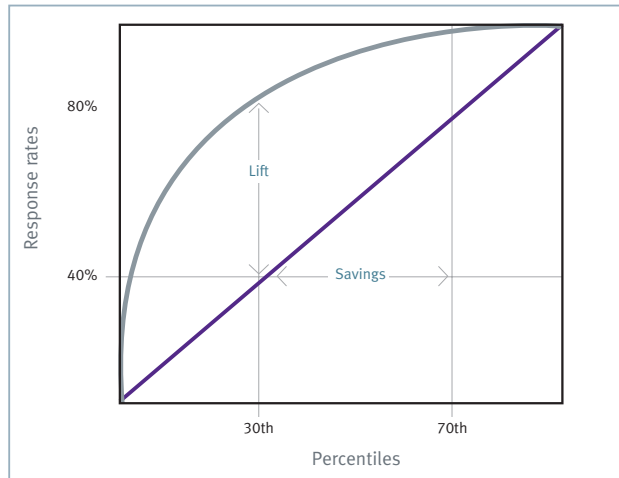


Figure 2: Gain chart for hypothetical data mining of alumni pledges

Results

Using data mining, the college discovered a way to make its mailing more effective and increase alumni pledges, while reducing mailing costs. This is best described using a concept called “lift.” If 20 percent of alumni respond to a pledge request, the college should concentrate on those 20 percent. If data mining can quickly identify potential donors by a ratio of two to four (correctly predicting two out of four who will donate), then the university can achieve results by mailing only to the indicated 40 percent of the alumni population, thus saving considerable time and money.

Conclusion

Data mining is a powerful analytical tool that enables educational institutions to better allocate resources and staff, proactively manage student outcomes, and improve the effectiveness of alumni development. With the ability to uncover hidden patterns in large databases, community colleges and universities can build models that predict—with a high degree of accuracy—the behavior of population clusters. By acting on these predictive models, educational institutions can effectively address issues ranging from transfers and retention, to marketing and alumni relations.

About SPSS Inc.

SPSS Inc. (NASDAQ: SPSS) is the world’s leading provider of predictive analytics software and services. The company’s predictive analytics technology connects data to effective action by drawing reliable conclusions about current conditions and future events. More than 250,000 commercial, academic, and public sector organizations rely on SPSS technology to help increase revenue, reduce costs, improve processes, and detect and prevent fraud. Founded in 1968, SPSS is headquartered in Chicago, Illinois.



To learn more, please visit www.spss.com. For SPSS office locations and telephone numbers, go to www.spss.com/worldwide.

SPSS is a registered trademark and the other SPSS products named are trademarks of SPSS Inc. All other names are trademarks of their respective owners.
© 2004 SPSS Inc. DMHEWP-1004