Educational Data Mining and Retention

Awatif Amin
Johnson C. Smith University

Abstract: Data mining is widely used to mine business, engineering, education and scientific data. Data mining uses pattern based queries, searches, or other analyses of one or more electronic databases/datasets in order to discover or locate a predictive pattern or anomaly indicative of system failure, criminal or terrorist activity, etc. There are various algorithms, techniques and methods used to mine data; including neural networks, genetic algorithms, decision trees. In recent years, throughout industry, academia and government agencies, thousands of data systems have been designed and tailored to serve specific engineering and business needs. Many of these systems use databases with relational algebra and structured query language to categorize and retrieve data. In these systems, data analyses are limited and require prior explicit knowledge of metadata and database relations; lacking exploratory data mining and discoveries of latent information. This paper discusses success stories of data mining to predict student retention.

Keywords

Data mining, Stem, retention, students

Introduction

Data mining is part of knowledge discovery process; Berry and Linoff[3] stated that researchers are using automated or semi-automated ways to find substantial rules and design from huge data. Data mining is a combination of many techniques form different domains such as statistics, machine learning, database system, data warehouse system, pattern recognition, visualization, algorithms, high performance computing and many different application areas. The nature of data mining of involving different techniques and application contributes tremendously to its success in research for different disciplines[6].

Data mining methods allow scientists to create models out of an institution's databases; such models have proven to be extremely valuable in company configurations because they perfectly indicate the current reality of a single company[5]. Apart from statistical analysis mathematical research techniques, data mining techniques are increasing to be well- known and precise in predicting student retention. During the period of 2004–2005, a data mining research was conducted in one of the mid-west universities to model freshmen retention in engineering. Researchers used several models of data mining such as logistic regression, discriminant analysis and neural network[2]

Data mining categories

There are two main data mining categories; descriptive and predictive data mining. Descriptive data mining can provide a description of the data and enough information about the basic structure, associations, correlation, and interconnection of the data. Descriptive models are

unsupervised learning paradigms[4]. Unsupervised learning methods discover statistically prominent features of the input data but do not predict a target value. Predictive data mining, unlike descriptive data mining, is a supervised learning paradigm that predicts a target. The supervised learning term is useful to data in which a specific classification already detected and documented in a training sample data. Then the researcher will build a model to predict those classifications by using another set of testing sample data[2].

The purpose of the classification analysis is to build a model to predict a target by submitting records of independent variables without the value of the dependent variable (target variable) and the classification model will guess the values of the target variable based on the patterns discovered from the training sample data[1]. Classification models in data mining distinguish patterns that define the group to which the dependent variable belongs.

The model examines the existing independent variables that are already classified and hence inferring a set of rules. Examples of classification models are decision trees, neural networks, and Bayesian classification[2].

Methodology and Design

The purpose of this study was to identify, through data mining, factors that can help to increase retention of STEM students at an HBCU by determining influences on STEM students' college degree attainment or withdrawal without a degree.

The research question is : What was the relationship between students' academic background and retention at X University? Students' academic background factors were (a) high school GPA and (b) SAT score/ACT score related to STEM student retention at JCSU.

Retention of students who took the ACT exam. Table1 contains descriptive statistics of the data analyzed in the ACT scores by using the model decision tree. The dependent attribute variable was degree/no degree. The independent attributes were exam scores in four ACT subjects consisting of English, Math, Reading, Science, an ACT Composite score, gender, high school grade point average (HSGPA), and In-state/Out-of-state status.

Table 1
Descriptive Statistics of 28 Students Taking ACT Exam

ID		*Min*	*Max*	*Average*
Dummy Code	Integer	4	97	47.821
Dependent Variable:		No Degree	Degree	
Degree/No Degree	Polynominal	13	15	
Independent Variables:				
		Min	*Max*	*Average*
ACT English Score	Integer	6	22	14.778
ACT Math Score	Integer	13	18	15.556
ACT Reading Score	Integer	10	24	16.000

ACT Science Score	Integer	10	22	17.000
ACT Composite Score	Integer	13	21	15.963
High School Grade Point		*Least*	*Most*	*Values*
Average	Polynominal	C-7	U-12	B-9, C-7, U-12
Gender	Polynominal	Male-14	Female-14	
In/Out of State	Polynominal	Instate-6	Outstate-22	

Note. U = Unknown grade.

The dependent attribute was degree/no-degree representing retention. The independent attributes of the decision tree in Figure 2 and the corresponding narrative description in Table 2 consisted of the four ACT Subject scores, ACT Composite Score, gender, in-state/out of-state status, and high school GPA.

Table. 2
Retention of 28 Students Who Took the ACT Exam: Narrative Description of the decision tree model

Act English > 20.500: No Degree {No Degree = 2, Degree = 0}

Act English ≤ 20.500

| Act Composite > 13.500

| | In/out state = instate: Degree {No Degree = 0, Degree = 4}

| | In/out state = outstate:

| | | Act Science > 18.500: No Degree {No Degree = 5, Degree = 1}

| | | Act Science ≤ 18.500

| | | | HSGPA = B or better: Degree {No Degree = 0, Degree = 3}

| | | | HSGPA = C or better: Degree {No Degree = 0, Degree = 4}`

| | | | HSGPA = Unknown

| | | | | Act Science > 15

| | | | | | Gender = Female: No Degree {No Degree = 1, Degree = 1}

| | | | | | Gender = Male: Degree {No Degree = 0, Degree = 2}

| | | | | Act Science ≤ 15: No Degree {No Degree = 3, Degree = 0}

| Act Composite ≤ 13.500: No Degree {No Degree = 2, Degree = 0}

Note: HSGPA = High school grade point average.

Table 2 shows if a student's ACT Composite score was <= 13.5, the student would not earn a degree. If the ACT Composite score was > 13.5, four out of four (100%) in-state students got a degree and seven out of seven out-of-state students (100%) with a high school GPA of C or higher got a degree. For students with an unknown GPA (U), the tree diverted to ACT Science and gender to show if the ACT Science score was > 15, two out of two (100%) males obtained a

degree compared to one out of two (50%) females.

For the class *degree*, the confusion matrix in Table 3 shows the accuracy rate or effectiveness of the model was .625 or 62.5%, calculated as (3+2)/ (3+2+1+2). The precision rate of .75 for the class *degree* indicates that 75 % of those predicted to receive a degree actually got a degree, calculated as 3/(3+1) = .75. Calculation of the true positive rate, termed *recall* rate for the class *degree* was 3/(3+2) =.6, indicating successful identification of the 60% of students who did not get a degree.

Table 3
*Confusion Matrix for ACT Exam Scores Decision Tree: *Accuracy = 62.5%, **Precision = 75%, ***Recall = 60% (Positive Class = Degree)*

Model Prediction	Actual Degree/No Degree Status from Test Set	
	Degree is True	No Degree is True
Degree	True Positive (TP) = 2	False Positive (FP) = 2
No Degree	False Negative (FN)= 1	True Negative (TN) = 3

Note: TP = correct positive prediction; FP = incorrect positive prediction; FN = incorrect negative prediction. TN = correct negative prediction.
*Accuracy % = (TP + TN)/total number of students (TP+FN+FP+TN). **Precision = TP/total predicted positives (TP+FP). ***Recall or true positive rate = TP/total actual positives (TP+FN).

Table 4
Descriptive Statistics of 68 Students Taking SAT I Exam

ID		*Min*	*Max*	*Average*
Dummy Code	Integer	1	99	49.735
Dependent Variable:		No Degree	Degree	
Degree/No Degree	Polynominal	34	34	
Independent Variables:		*Min*	*Max*	*Average*
SAT Math Score	Integer	300	620	418.515
SAT Verbal Score	Integer	40	700	408.235
SAT Composite Score	Integer	460	1310	826.015
High School Grade Point Average	Polynominal	*Least* D-2	*Most* B-27	*Values*: A-3, B-27, C-22, D-2, U-14
Gender	Polynominal	Male 34	Female 34	
In/Out of State	Polynominal	In-state 20	Out-state 48	

Note. U = Unknown grade.

Retention of students who took the SAT I exam. Table 4 contains descriptive statistics of the data analyzed in the SAT I scores by using the model decision tree for the 68 students who took the exam. The dependent variable was the degree/no degree attribute. Independent attributes

Table 5

Retention of Students Who Took the SAT Exam: Narrative Description of the Decision Tree

SAT Composite > 985
| Gender = F: Degree {Degree=5, No Degree=0}
| Gender = M: Degree {Degree=2, No Degree=1}
SAT Composite ≤ 985
| SAT Math > 485: No Degree {Degree=0, No Degree=6}
| SAT Math ≤ 485
| | SAT Math > 355
| | | SAT Verbal > 345: No Degree {Degree=13, No Degree=14}
| | | SAT Verbal ≤ 345
| | | | instate/outstate = instate: Degree {Degree=2, No Degree=1}
| | | | instate/outstate = outstate: Degree {Degree=8, No Degree=0}
| | SAT Math ≤ 355
| | | SAT Verbal > 455: Degree {Degree=2, No Degree=0}
| | | SAT Verbal ≤ 455
| | | | Gender = F: No Degree {Degree=0, No Degree=6}
| | | | Gender = M
| | | | | HSGPA = B or better: Degree {Degree=1, No Degree=1}
| | | | | HSGPA = C or better: No Degree {Degree=1, No Degree=2}
| | | | | HSGPA = Unknown: No Degree {Degree=0, No Degree=3}

Note: F = Female; M = Male. HSGPA = High school grade point average.

were SAT Math score, SAT Verbal score, SAT Composite score, high school grade point average, gender, and in-state/out-of-state status.

SAT1 exam in table 5 shows that all five female students (100%) who had an SATI composite score of > 980 received a degree whereas two students out of three male students (66%) received a degree. The tree also shows when the SAT Math score was > 355 and SAT Verbal score was > 345, 13 students (49%) received a degree out of 27. When SAT Math score was > 355 and SAT Verbal score was <= 345, two out of three instate students obtained a degree and eight out of eight (100%) out of state students received a degree. Two students received a degree with SAT Math<355 but with an SAT Verbal score of > 455. The tree also shows with SAT Math < 355 and SAT Verbal < 455 no female received a degree out of six, but two male students out of eight (1/4) with high school GPA of B or higher received a degree.

For the class *no degree*, the confusion matrix in Table 6 shows the accuracy rate or effectiveness of the model was .65 or 65%, calculated as (9+4)/ (9+6+1+4). The precision rate of .80 for the class *no degree* indicates that 80 % of those predicted to leave without obtaining a degree actually did not obtain a degree, calculated as 4/(4+1) = .80. Calculation of the true positive rate, termed *recall* rate for the class *no degree* was 4/(4+6) =.4, indicating successful identification of the 40% of students who did not get a degree.

Table 6
*Confusion Matrix for SAT I Exam Scores Decision Tree: *Accuracy = 65%, ** Precision = 80%, ***Recall = 40% (Positive Class = No Degree)*

Model Prediction	Actual Degree/No Degree Status from Test Set	
	Degree is True	No Degree is True
Degree	True Positive (TP) = 9	False Positive (FP) = 6
No Degree	False Negative (FN)= 1	True Negative (TN) = 4

Note: TP = correct positive prediction; FP = incorrect positive prediction; FN = incorrect negative prediction. TN = correct negative prediction.
*Accuracy % = (TP + TN)/total number of students (TP+FN+FP+TN). **Precision = TP/total predicted positives (TP+FP). ***Recall or true positive rate = TP/total actual positives (TP+FN).

Results

- ACT Composite score was widely influential, as all students who scored greater than 13.5 remained enrolled until earning a degree.
- Out-of-state students with a score greater than 15 in the ACT Science and a GPA of C or higher obtained degrees.
- Research also shows that male students who scored higher in ACT Science and were more likely to successfully complete a degree than female students.
- When female students had SAT composite scores is greater than 980, they were certain to obtain a degree.
- When SAT scores for Verbal and Math averaged around 350 each, two-thirds of in-state students obtained degrees while all out-of-state students in this range obtained a degree.
- For the student population with SAT verbal scores above average and SAT math scores below average, the female students did not complete a degree program but the male students with the same SAT scores and with a high school GPA of B or higher did obtain a degree

Summary

This paper presented a brief review of the data mining techniques that is relevant to data mining student retention data. Earlier, many researchers tried to acquire inclusive theoretical models for analysis and prediction to improve students' retention in higher education, for example, the models we discussed earlier and Tinto's[7] experimental models. Few years later, to justify the theoretical models, researchers started to use statistical analysis methods. There is no question today that higher education in colleges and universities in the United States of America is a huge business. If researchers looked at retention from the business point of view, researchers would need to collect huge data and build responsive predictive models; these strategies relate retention models to knowledge discovery and hence data mining methods. From the literature, data mining has proven to support dynamic models with very high percentage of accuracy to Wiley predict student retention.

References

1. Al-Radaideh, Q. A., & Nagi, E. A. (2012). Using data mining techniques to build a classification model for predicting employees' performance. International Journal of Advanced Computer Science and Applications, 3(2), 144.
2. Amin, N. G. (2006). Higher education in Sudan and knowledge management applications. 2nd International Conference on Information & Communication Technologies, 1, 60-65.
3. 5Berry, M., & Linoff, G. (2011). Data mining techniques for marketing, sales, and customer relationship management. New York, NY:.
4. 6Bose, I., & Chen, X. (2009). Hybrid models using unsupervised clustering for prediction of customer churn. Journal of Organizational Computing and Electronic Commerce, 19(2), 133.
5. 9Dekker, G, Pechenizkiy M, & Vleeshouwers J (2009) Predicting students drop out: a case study. In Proceedings of 2nd International Conference on Educational Data Mining, 41–50.
6. 11Han, J., Kamber, M., & Jian, P. (2011). Data Mining: Concepts and Techniques (3rd ed.). Burlington, MA: Kaufmann.
7. 23Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. Review of Educational Research, 45(1), 89-125.

Awatif Amin

Awatif is a professor at Johnson C. Smith University since 2001. She primarily focuses on programing and data analytics. Awatif has a B.S. and M.S. in Computer Science. She is a graduate student earning a Doctoral Degree in Management and Organizational Leadership with concentration in Information System Technology (abd), at University of Phoenix and expect to defend her dissertation by May 2019. She attended Faculty in Residence program at Google's world headquarters in Mountain View, California for six weeks summer 2017, for CS curriculum reform.