Optimal Data Analysis
2013, Vol. 2, Release 1 (September 27, 2013), 48-53

2155-0182/11/$3.00

Manual *vs.* Automated CTA: Predicting Freshman Attrition

Paul R. Yarnold, Ph.D., Fred B. Bryant, Ph.D.,
Optimal Data Analysis, LLC Loyola University Chicago

and

Jennifer Howard Smith, Ph.D.
Applied Research Solutions, Inc.

The enumerated model was 20% more accurate, but 43% less parsimonious and 31% less efficient than the manually-derived model. Granularity afforded by the enumerated model enabled prediction of seven of eight incoming freshmen who left college. Substantive, policy, and methodological implications are considered.

The relationship between student involvement and freshman retention was investigated in a test of Person-Environment (PE) fit theory.[1] Incoming freshmen (N=382) were prospectively followed the summer before beginning college, and again during the spring of their freshman year, in a two-wave panel study. Involvement levels, various summer and spring *preferences* (Ps), and spring *perceptions* (Es) regarding specific aspects of their college environment were assessed, and 12 PE fit indicators were derived. Manually-derived CTA identified nine student clusters (five of returners and four of dropouts), revealing that different subgroups of freshmen chose to return (and leave) for different reasons. Findings suggested that student end-of-the-year preferences are more important predictors of retention than anticipated preferences, college perceptions, or PE fit levels. The model correctly classified 84.8% of the total sample, yielding ESS=68.8 (see Table 1).

An enumerated CTA model was then derived by automated software[2] for the same data used in the original analysis. To be consistent between analyses, attributes were only allowed to enter the model if their associated ESS was stable (did not diminish) in leave-one-out (jackknife) validity analysis.[3] The enumerated model correctly classified 94.6% of the total sample and yielded ESS=82.4, or 11.6% and 19.8% improvement compared to the manually-derived model, respectively (see Table 1).

The enumerated model incorporated 14 attributes rather than eight as used in the manual model, and thus was 42.9% less parsimonious. Compared to the manual model the enumerated model had greater ESS (19.8%), overall PAC (11.6%), specificity (3.5%), sensitivity (31.5%), and both positive (0.2%) and negative (100%) predictive values. In contrast, the manual model was 45.8% more efficient than the enumerated model as a result of its greater parsimony.

2155-0182/11/$3.00

Table 1: Comparing Performance of Manually-Derived *vs*. Enumerated CTA Models

		Manually-Derived Model			Enumerated Model		
		Predicted Class Status			*Predicted Class Status*		
		Dropped	Stayed		Dropped	Stayed	
Actual Class Status	Dropped	26	5	83.9	33	5	86.8
	Stayed	48	269	84.9	14	298	95.5
		35.1	98.2		70.2	98.4	
Total N Classified		348			350		
PAC (%)		84.8			94.6		
Model ESS		68.8			82.4		
Number of Attributes		8			14		
Model Efficiency		8.6			5.9		
CPU Seconds		*na*			400		
Number of Models		1			37		
CPU Seconds / Model		*na*			10.8		

Note: Values given to the right of the Stayed *columns* are the specificity (for dropped) and sensitivity (for stayed), and values given under the Stayed *row*, beneath columns, are the negative (for dropped) and positive (for stayed) predictive values.[3] Total N classified varies as a function of missing data. PAC=percentage accuracy in classification=100% x (sum of correctly classified observations)/(total N classified).[3] ESS=effect strength for sensitivity, a normed index on which 0 is the level of classification accuracy that is expected by chance, and 100 is perfect accuracy.[3] The number of nodes in the CTA model is given, and model efficiency is defined as model ESS divided by number of nodes—expressed in terms of mean ESS-units-per-node, it measures the mean level of explanatory power per node which is used in the model, or "bang-for-the-buck".[3] CPU Seconds=time required to achieve a solution using a 3 GHz Intel Pentium D microcomputer (*na*=not available). Number of Models=number of different CTA models identified during enumeration. CPU Seconds/Model=the mean number of CPU seconds required to solve a single CTA model.

The enumerated model predicted 98.4% accurately that [100% x (5+298)/350], or 86.6% of the total sample would stay, and correctly identified 95.5% of all the students who stayed. And, the enumerated model predicted 70.2% accurately that 13.4% of the total sample would drop, and correctly identified 86.8% of all the students who dropped.

Considered from a *policy perspective* the enumerated model implies that resources should be expended to monitor and assist a total of 47 (33+14) students predicted to drop, from a total class size of 381 (students missing data on attributes employed by the CTA model were not classified). This total of 47 students corresponds to 12.3% of the sample (one in every eight incoming students). Of these 47 students, 33 (70.2%) will drop (seven of every ten incoming students). However, if attributes in the model are actionable on the part of student or counselor, and efforts ultimately are successful, then the model could successfully identify and aid in the circumvention of loss of 33/38, or 86.8% of students (seven of eight) who would otherwise drop, and whom together represent 8.7% (100% x 33/381) of the total freshman sample (one in eleven).

2155-0182/11/$3.00

The size of the sample strata (endpoints) which were identified by the CTA models was somewhat consistent between models, but varied considerably within model. For the manually-derived model the largest strata (N=176, 50.6% of classified sample) is 29-times larger than the smallest strata (N=6, 1.7% of classified sample). For the enumerated model the largest strata (N=125, 35.7% of classified sample) is 42-times larger than the smallest strata (N=3, 0.9% of classified sample).

Table 2: AID Analysis for CTA Example

Attribute	N and % of Sample Evaluated Partly on the Basis of the Attribute	
Desire to join political groups	350	100.0%
Desire for work pressure	231	66.0%
Global activity participation	187	53.4%
Attendance of cultural events	119	34.0%
Pretest desire to attend lectures	66	18.9%
Pretest desire for quality-demand	62	17.7%
Desire for effort-oriented arena	53	15.1%
Desire for email communication	44	12.6%
Desire to foster risk-taking	36	10.3%
Desire to attend a pre-game rally	30	8.6%
Pretest desire to attend mass	25	7.1%
Pretest desire for dorm events	20	5.7%
Pretest desire for college identity	17	4.9%
Pretest-posttest change in image of ideal college environment	15	4.3%

Because of the numerous nodes—that is, as a result of complexity in the model, not every attribute loading in the enumerated CTA model influenced the classification decisions for a substantial portion of the total sample. Table 2 presents an AID (Attribute Importance in Discrimination) analysis, indicating the percent of the total sample (of classified observations) which was classified on the basis of the attribute. Note that three posttest attributes were particularly important in predicting freshman-year matriculation status: frequency of participation in 16 school-related activities predicted outcome for half the incoming freshman class; desire to work under pressure predicted outcome for two of three freshmen; and desire to become active in political groups on campus predicted outcome for all freshmen.

Substantive Implications. Table 3 describes the sample strata of students whom the enumerated model predicted would drop out.

Examining the "drop out" strata groups in the enumerated model, the *first* and largest group consists of students who report a lower desire to join on-campus political groups, stronger desire to attend college-sponsored cultural events, and stronger desire for an environment that is effort-oriented; these students want more cultural stimulation and greater rewards for working hard. From a policy-development perspective, the university might increase retention for this group by providing more cultural activities and finding formal ways to reward hard work academically.

The *second* largest "drop out" group reports a greater desire to join on-campus political groups, seeks greater work pressure, participates very little in extracurricular activities, and has less desire to attend a pre-game pep rally; these students want more opportunity for political involvement and greater academic demands, seemingly in preparation for post-college life. As a strategy for retaining students from this strata group, the university might offer "fast track" honors programs aimed at promoting excellence and encouraging elite students to become political and scholastic leaders.

2155-0182/11/$3.00

Table 3: Description of Sample Strata (Endpoints) the Enumerated CTA Model Predicted would Drop

Strata	N_{Total}	N_{Drop}	N_{Stay}	% Drop	Descriptor
1	17	10	7	58.8	Desire to join political groups<6, Attendance of cultural events>3, Desire for effort-oriented arena>5
2	11	6	5	54.6	Desire to join political groups>6, Desire for work pressure>3, Global activity participation<-5.48, Pretest desire for quality-demand>6, Desire to attend a pre-game rally<6
3	6	5	1	83.3	Desire to join political groups<6, Attendance of cultural events>3, Desire for effort-oriented arena<5, Desire to foster risk-taking<4, Pretest desire for college identity<5
4	5	5	0	100	Desire to join political groups>6, Desire for work pressure<3, Desire for email communication>2, Pretest desire to attend mass>6
5	5	4	1	80	Desire to join political groups>6, Desire for work pressure<3, Desire for email communication>2, Pretest desire to attend mass<6, Pretest desire for dorm events<5
6	3	3	0	100	Desire to join political groups <6, Attendance of cultural events<3, Pretest desire to attend lectures<4, Absolute value of pretest-posttest change in image of ideal college environment<0.53

The *third* largest "drop out" group reports lower desire to join on-campus political groups, greater attendance of cultural events, less desire for an effort-oriented environment, less desire for an environment that fosters risk-taking, and less desire for a college identity before coming to Loyola; these students want less work, fewer challenges, and more opportunity for a life outside of college. Retaining these students might require the university to offer a less demanding degree program for students who want to move at their own pace—an approach that may well be antithetical to the university's goals and ideals.

The *fourth* "drop out" group reports a greater desire to join on-campus political groups, wants less work pressure, seeks more electronic correspondence with faculty and classmates, and had a stronger pretest desire to attend mass at college; these students want stronger social and spiritual connections. As an intervention aimed at retaining students in this strata group, the university might make it easier for students to communicate with each other and with faculty and to become involved in worship services.

Resembling the fourth strata group, the *fifth* "drop out" group also reports a greater desire to join on-campus political groups, wants less work pressure, and seeks more electronic correspondence with faculty and classmates, but before coming to college they had less desire to

2155-0182/11/$3.00

attend mass at college and less interest in attending events in their residence hall; these students want less work pressure but have less interest in group-based social activities. As with the third strata group, retaining students in the fifth strata group might also require the university to provide a less rigorous curriculum for students who want a less challenging environment. Given the university's mission and academic standards, a more viable approach might well involve prescreening students in these two strata groups at the application stage, so as not to admit them in the first place.

Finally, the *sixth* "drop out" group reports lower desire to join on-campus political groups, less attendance of cultural events, had less interest before coming to Loyola in attending faculty lectures outside class, and reports a great change in their image of the ideal college environment during their freshman year; these students have changed their minds about what college should be like. To retain more students in this strata group, the university might try to shape realistic expectations about college at the start, so as to prevent students from changing their minds about what they are seeking.

The automated enumerated CTA model has several important similarities with the manually-derived CTA model. In particular, both models include predictors of drop-out that reflect less desire to identify oneself as a member of the university, develop socially and spiritually, connect with faculty outside class, and work in a challenging and competitive academic arena. These points of convergence demonstrate generalizability between the models and reinforce prior evidence suggesting these attributes are related to attrition.

There were also important differences between the two models. Examining the attributes listed in Table 2—all of which loaded in the enumerated CTA model—five attributes (36%) are pretest variables that reflect what students said they were hoping to find at Loyola, whereas the manually-derived model included no pretest attributes. The enumerated model thus provides a more effective means for policy makers to target students prospectively who are at risk of dropping out.

In summary, the automated model presents a more accurate but less parsimonious way of classifying the sample, which identified many of the same predictive concepts, but included more baseline attributes, compared to the manual CTA model. Many factors emerging in both CTA models are theoretically important, and well supported in college student attrition research.[4,5] The CTA models represent alternative theoretically viable and empirically supported paths to matriculating beyond freshman year.

Policy Implications. For some applications the cost of misclassification is extreme and misclassifications are to be avoided to the fullest extent possible.[6] Considering the lifetime of effort, achievement and sacrifice which usually precedes acceptance into college, and the opportunity loss to individual, family, school and society which is associated with attrition, using available resources to monitor students in strata predicted by the model to fail (Table 3) seems wholly appropriate. Successful monitoring and intervention performed for one in every eight incoming freshmen would retain seven of eight of the students who would otherwise leave college.

Methodological Implications. Clearly, automated—in particular enumerated—CTA is capable of developing models expressing exceptional granularity, which nevertheless satisfy the criteria for a statistically appropriate model. Highly granular solutions may yield nearly perfect classification performance, but model endpoints having small numbers of observations can be criticized on an *a priori* basis in the context of limited statistical power, and on a *post-hoc* basis in the context of the potential cross-generalizability of the finding: small denominators leave little room for inconsistent results before effects found in training analysis vanish in validity analysis.

2155-0182/11/$3.00

The number of attributes required to achieve a highly granular solution can be substantial (Table 2). This is typically troublesome for linear models, for which a large number of attributes, or a high attribute-to-observation ratio, can induce multicollinearity (obviating any solution) or overfitting (in which case the model will not cross-generalize).[7,8] For CTA this is not a problem in the classic sense: not every observation is classified on the basis of every attribute (the maximum depth of the enumerated model presently was five, meaning that at most five attributes were used to classify any given observation). In addition, CTA models, whether algorithmic or enumerated, all employ an integral pruning methodology to control the experimentwise Type I error rate, and employ jackknife validity analysis to assess the potential cross-generalizability of each component of the tree model—and as a criteria for model growth in applications for which avoiding reduced effect strength in validation is key (e.g., financial markets which swing badly if experts overestimate earnings).

AID analysis identified four attributes instrumental in classifying one or more of every three observations in the sample, and five additional attributes important in classification decisions involving between one and two of every ten observations in the sample. In contrast, the final five attributes loading in the model influenced classification decisions for between one in 12 (desire to attend a pre-game rally) and one in 23 (pretest-posttest change in image of ideal college environment) observations. This finding, and the conceptually related finding of wide variability in strata size, suggest the need for sensitivity analysis in both studying and conducting exploratory automated CTA. Specifically, systematic investigation is needed to better understand the interplay between sample size, number of attributes, tree depth, minimum endpoint denominator, granularity, and the training and validity performance of models developed using CTA.

References

[1]Smith JH, Bryant FB, Njus D, Posavac EJ. Here today, gone tomorrow: understanding freshman attrition using person-environment fit theory. *Optimal Data Analysis* 2010, 1:101-124.

[2]Soltysik RC, Yarnold PR. Automated CTA: fundamental concepts and control commands. *Optimal Data Analysis* 2010, 1:144-160.

[3]Yarnold PR, Soltysik RC. *Optimal data analysis: a guidebook with software for Windows*. APA Books, Washington, DC, 2005.

[4]Braxton, JM. *Reworking the student departure puzzle*. Vanderbilt University Press, Nashville, TN, 2000.

[5]Herzog, S. Return vs. dropout/stopout vs. transfer: a first-to-second year analysis of new freshman. *Research in Higher Education*, 1996, 46: 883-928.

[6]Harvey RL, Roth EJ, Yarnold PR, Durham JR, Green D. Deep vein thrombosis in stroke: the use of plasma D-dimer level as a screening test in the rehabilitation setting. *Stroke* 1996, 27: 1516-1520. Abstracted in *American College of Physicians Journal Club* 1997, 126:43.

[7]Grimm LG, Yarnold PR. (Eds.). *Reading and understanding multivariate statistics*. APA Books, Washington, DC, 1995.

[8]Grimm LG, Yarnold PR. (Eds.) *Reading and understanding more multivariate statistics*. APA Books, Washington DC, 2000.

Author Notes

Send eMail to: Journal@OptimalDataAnalysis.com.