Using Analytics to Predict Student Success and Focus Resources

Arkansas Department of Higher Education
Placement Meeting
June 15, 2016

David G. Underwood, Ph.D.
Arkansas Tech University

Data used in this study was provided by the Office of Institutional Research and supported by the Office of Academic Affairs

ARKANSAS TECH UNIVERSITY

Below are a few caveats to go along with this presentation:

1) Anyone wishing to use this type of data must do the analysis using their own student population.

2) Discriminant analysis is not the only technique that can be used to classify students.

3) I intentionally used only variables in the model that would be readily available in most institutional datasets. Other variables that exist on individual campuses may make a better model.

4) This was actually a joining of two different sets of analyses that were done a few years apart. They were put together to illustrate what is possible and specifically for the placement workshop. As a result, it may appear a little disjointed without the talking through of the full process.

Arkansas Tech University

- Public 4-year institution
- 12, 054 total enrollment
- 9, 070 undergraduate enrollment
- 52.8% first generation
- Economic Challenges–59% Pell grant recipients
- 66% first to second year retention rates full-time students
- 47% graduation rate
- **59% graduation rate—national average**
- Math Remediation Rate Fall 2015 - 40%

ARKANSAS TECH UNIVERSITY

Primary Question

Can a statistical model be developed, using variables that most public 4-year institutions in Arkansas will have in their database, that will identify students scoring less than 19 on the ACT Math section who are most likely to be successful in College Algebra if additional assistance is provided, at better than chance selection accuracy?

Stipulations:

- Data must be “ambient” – data that are likely to be readily available to any state institution in Arkansas.
- The statistical methodology should be within the ability of most campuses to perform.
- It should be relatively easy to interpret.

(This analysis was completed using SPSS, which most campuses would have)

Discriminant Analysis Rationale

- DA is used to classify cases into groups and to decide how to assign new cases to those groups.
- The interpretation is similar to multiple regression.
- The Canonical Correlation can be squared and interpreted similarly to R^2 such that squaring it indicates the amount of variance accounted for by the model.
- The Canonical Discriminant Function Coefficients may be interpreted similarly to beta weights in multiple regression and so forth.

Data used were from the fall 2012 student body and included all students who were taking remedial mathematics for the first time during the fall 2012 semester. Success is defined as completing enough modules to enter college algebra with a grade equivalent to an “A”, “B”, or “C”.

Those classified as unsuccessful received a grade lower than “C”, or a “W”.

The grade of “W” was included for two reasons 1) those students did not successfully complete the class, and 2) although Analysis of Variance showed four significant differences between students who received a grade of “W” and those who received a failing grade on the variables in the analysis, in all cases the mean was lower for students receiving a “W” than those with an “F”.

The total number used in the analysis was 640.

The groups were almost evenly split with 318 in the “unsuccessful” group and 322 in the “successful” group.

Variables Included In Analysis

ACT Composite Score
ACT Math Score
ACT Science Score
ACT Reading Score
High School Grade Point Average
High School Class Rank
High School Class Size
High School Class Rank as a Percentile Score

Variables Found to Be Significant Predictors

ACT Comp Score
ACT Math Score
ACT Science Score
High School Grade Point Average
High School Class Rank
High School Class Rank as a Percentile

Decision to Use Stepwise

1) The original analysis using all variables was found to violate the assumption of equality of covariance matrices, although large group sizes decrease the importance of the assumption being met.

2) several of the variables included in the full model, i.e., Class Rank and Class Rank as a Percentile, and ACT Math Score, ACT Science Score and ACT Comp Score, etc., could be highly correlated and therefore responsible for the violation of the assumption of equality of covariance matrices due to multicollinearity.

3) the stepwise function is designed to find the best set of predictors from among a larger number and use only those contributing a significant amount of unique variance to the model.

The stepwise procedure was used with an F to enter of .05 and an F to remove of .1 to identify only those variables adding a significant amount of explained variance to the model.

The stepwise method identified three significant predictors accounting for 22.8% of the explained variance. Box's M was found to be insignificant, indicating the assumption of equality of covariance matrices was met.

The significant predictors identified from the Structure Matrix were High School Grade Point Average (.964), ACT Math Score (.239) and ACT Reading Score (.109).

Based on the Canonical Discriminant Function Coefficients, the discriminant function, used to compute a discriminant score, can be stated as:

D = (.223*ACT_Math) +(.-.056*ACT_Reading)
+(2.534*HSGPA) -9.856

The model exceeds the commonly accepted level of providing at least a 25% improvement over chance assignment. Summing the squared prior probabilities provides a prior chance probability of 50%. Multiplying 50% by 1.25 provides a figure of 62.5%. **An acceptable model should be equal to or greater than 62.5%**. **The cross validated classification model of 71.6% is above the commonly accepted threshold.**

Using this data, a score of +1.5 or greater identified 76 students. Of those, 70 were actually successful for a classification accuracy of 92.1%.

A pilot was attempted as a 5 hour class, if assigned by the model 75% were successful, if not only 50%.

For those assigned directly to College Algebra, the primary predictors of success were **High School GPA and High School Class size**.

Next - Gateway Course Analysis

G2C Project Began in 2012

Courses Chosen, Number of Students and Original DFWI Rates:

• ACCT 2003	433	54.0%
• BIOL 1014	1,112	30.9%
• HIST 1903	1,110	33.5%
• MATH 1113	1,387	38.5%
• PSY 2003	1,426	24.6%

Primary Question

- Can a statistical model be developed, using variables that most institutions will have in their database, that can identify students in gateway courses who are most likely to pass or fail, at better than chance selection accuracy?

Stipulations:

- Data must be “ambient” – data that are likely to be readily available to any institution.
- The statistical methodology should be something within the ability of most campuses to perform.
- It should be relatively easy to interpret.

(This analysis was completed using SPSS, which is readily available to all faculty, staff, and students)

Discriminant Analysis Rationale

- DA is used to classify cases into groups and to decide how to assign new cases to those groups.
- The interpretation is similar to multiple regression.
- The Canonical Correlation can be squared and interpreted similarly to R^2 such that squaring it indicates the amount of variance accounted for by the model.
- The Canonical Discriminant Function Coefficients may be interpreted similarly to beta weights in multiple regression and so forth.

Procedure

A discriminant analysis model was developed using four semesters of G2C data. Fall 2013, Spring 2014, Fall 2014, and Spring 2015

Students were classified into one of two categories “pass” or “fail”

Definitions

- The total number used in the analysis was 12,368.
- **Pass** was defined as completing the G2C course with a grade equivalent to an “A”, “B”, or “C”.
- **Fail** received a grade lower than “C”, or a “W”.
- The grade of “W” was included because those students did not successfully complete the class and they may have been successful with an intervention.

Variables Included In Analysis

ACT Composite Score
ACT Math Score
ACT Science Score
ACT Reading Score
ATU GPA
High School Grade Point Average
High School Class Rank
High School Class Size
High School Class Rank as a Percentile Score
Earned Hours
Transfer Hours

Details of The Discriminant Analysis

- A stepwise model was used.
- The stepwise function is designed to find the best set of predictors from among a larger number and use only those contributing a significant amount of unique variance to the model.
- The stepwise procedure was used with an F to enter of .05 and an F to remove of .1 (the default) to identify only those variables adding a significant amount of explained variance to the model.

Results

- The stepwise method identified four significant predictors accounting for **32.3%** of the variance.
- Box's M was found to be significant, indicating the assumption of equality of covariance matrices was not met, however, Burns & Burns (2012) suggest it is not a problem since the sample size is large.

Results Continued

- Based on the Canonical Discriminant Function Coefficients, the discriminant function, used to compute a discriminant score, can be stated as:

D = (1.176*ATUGPA)+(.472 *HSGPA)+ (.02*ACTSCIENCE)+(.017*ACTMATH) + (-5.625)

- The procedure also allows for individually assigning students to Pass/Fail categories

Acceptability of the Model

- Summing the squared prior probabilities provides a prior chance probability of 57%.
- An acceptable model should improve the chance probability by at least 25%. Multiplying 57% by 1.25 provides a figure of 71.3%.
- An acceptable model should be equal to or greater than **71.3%**.
- The cross validated classification model of **79.6%** is above the commonly accepted threshold.

First Set of Conclusions

- Discriminant Analysis can be used to identify students who are most likely to be successful or unsuccessful depending on which students one needs to identify.
- The classification is better than chance accuracy.

Scoring the Fall 2015 Students

The model developed on the previous four semesters was used to “score” each student enrolled for the fall 2015 semester at the beginning of the term.

Scoring allows one to use the model previously developed to “predict” which students will pass or fail in the new incoming class based on the identified characteristics of the model.

Benefits of Scoring

In scoring, each student is classified into one of the two categories: Pass or Fail

The students enrolled in gateway courses for fall 2015 were scored at the beginning of the fall semester and assigned to either “pass” or “fail”

Once the grades for the semester were recorded, a determination was made to identify the actual accuracy of the prediction.

Fall 2015 Data

The fall 2015 data consisted of 3,544 students in G2C courses who actually received grades.

Students who dropped classes before receiving a grade were not included in the analysis, and, as before, a “W” grade was counted as a “Fail”.

Overall Classification Table (Includes “W”)

	Predicted Outcome Fail	Predicted Outcome Pass	Total	Overall Accuracy
Actual Outcome Fail	**613 (TN)** (76.5%)	471 (FP) (17.2%)	1084	(TN + TP)/ Total
Actual Outcome Pass	188 (FN) (23.5%)	**2272 (TP)** (82.8%)	2460	(613+2272)/3544 = (2885)/3544 =
Total	**801**	**2743**	**3544**	**81.4% Correctly Classified**

Classification Accuracy By G2C Class (Including "W")

	TN	TP	Accuracy
ACCT 2003	88.2%	70.7%	72.2%
BIOL 1014	75.6%	82.9%	81.3%
HIST 1903	83.3%	82.7%	82.8%
MATH 1113	78.6%	81.4%	80.8%
PSY 2003	69.5%	89.4%	83.9%
TOTAL	**76.5%**	**82.8%**	**81.4%**

Second Set of Conclusions

- We can predict which students will pass and which students will fail with 81.4% accuracy
- Out of 3,544 students we can predict with 82.8% accuracy which ones will pass.
- Out of 3,544 we can predict with 74.5% accuracy which ones will fail.
- We can make those predictions at the beginning of the semester, or earlier.

Questions Guiding the Analysis

- If we knew which students are most likely to succeed or fail, could we do something differently?
- Could we tailor services for those students unlikely to be successful?
- Could we require those students to receive additional help? (study labs, corequisite courses, etc.)
- Could we reduce the amount of resources necessary if we know in advance how many students really need the service?

Questions Still to Be Answered

- How accurate would the model need to be for us to act on those questions?
- What would be the best approach to make use of the results of the model?
- How do we use this type of data as an institution to improve student success?

Reference

- Burns, R., & Burns, R. (2008). *Business Research Methods and Statistics using SPSS*. London: Sage Publications Ltd.
- Burns, R. & Burns, R. (2008). Chapter 25: Discriminant Analysis (WWW page). URL http://www.uk.sagepub.com/burns/website%20material/Chapter%2025%20-%20Discriminant%20Analysis.pdf

got questions