# Leveraging Data Mining Techniques to Facilitate Transcript Analysis

Using association rule mining to explore the impact of change in policy on student behavior

Randall Hickman 2015 MIAIR Conference

# Research Objective

- Explore potential impacts of policy implemented in fall 2013 that mandated the taking and passing of the college success skills course
  - Policy requires students placing into all three remedial areas to enroll in the college success skills course (CSSK-1200) in their first semester. Students must pass this course within two attempts in order to avoid being prohibited from enrolling in other courses. Students who do not pass the course after three attempts are not allowed further enrollment.

# Research Questions

- Are there changes in the demand for certain courses?
- Are there changes in the demand for certain *groups* of courses?
- Are there changes in the extent to which certain courses are associated?
- *Which changes might be an impact of the policy mandating the passing of the college success skills course?*

# Methodology

- Compare the first semester enrollment behavior of the fall 2012 incoming cohort of at-risk students—students placing into all three remedial areas—with the enrollment behavior of the fall 2013 incoming cohort of at-risk students, using association rule mining in a form of "market basket analysis"
- Dataset
  - Fall 2012 (N = 436) and fall 2013 (N = 496) incoming cohorts of at risk students (students placing into all three remedial areas)
  - Courses taken in first semester, where enrolling in a course defined as remaining in a course long enough to write data to the transcript

# Methodology: Association Rule Mining

- The kind of analysis that Amazon or Netflix employs when it makes recommendations concerning books or movies you might be interested in, based on your previous purchases and the purchases that others have made—a form of ARM known as “market basket analysis”
- Although ARM is probably most often used for ”market basket analysis”, it is a general form of analysis applicable to any research concerned to identify patterns of association among behaviors, transactions, or properties
- A natural form of analysis for exploring student decision making in a number of areas, such as the taking (and sequencing) of courses, student services utilized, and pursuit of awards

# Methodology: Association Rule Mining

- Produces models that are sets of rules that describe associations in the data, at large or small scales (i.e., applicable to large or small portions of the dataset), and that are readily interpretable
- Two types of models
  - Models that generate rules that ignore the sequencing of items over time
  - Models that generate rules concerning the sequencing of items over time
- Although ARM is an unsupervised machine learning technique, the rules can be viewed as predictions

# Methodology: Association Rule Mining

- Rule structure
  - If {A:Itemset members} then {B:Itemset members}, where A is the "antecedent" or condition and B is the "consequent" or prediction
  - E.g., if {one or more courses} are taken, then {one or more other courses} are taken
- Rule generation
  - A single itemset with *k* items can generate $3^k - 2^{k+1} + 1$ rules, when all subsets of the itemset are included
  - For example, a single "pattern" (itemset) of five courses can generate 180 rules
  - Generating rules is controlled by imposing constraints related to rule properties deemed "interesting"

# Methodology: Association Rule Mining

- "Rule interestingness" measures—key statistics generated that are relevant for evaluating patterns
  - Rule support: proportion satisfying the rule
  - Rule confidence (a conditional probability): proportion of records that satisfy the antecedent that also satisfy the consequent
  - Antecedent support: proportion satisfying the antecedent
  - Lift: ratio of joint probability of {Antecedent $\wedge$ Consequent} to expected probability based on marginal probabilities of {Antecedent} and {Consequent}, i.e. under the assumption of statistical independence—a measure of the strength and direction of the association

# Methodology: Association Rule Mining

- What makes a rule “interesting”—Piatetsky-Shapiro criteria
  - Interestingness should be zero if rule support = (antecedent support x consequent support), i.e. if antecedent and consequent are statistically independent
  - Interestingness should increase monotonically with increase in rule support, other things remaining the same
  - Interestingness should decrease monotonically with increase in either antecedent support or consequent support, other things remaining the same

# Methodology: Modeling and Rule Interestingness Measures

- Controlling modeling through rule interestingness parameters
  - Minimum rule support level of 5% (related to prevalence of association)—desirable to avoid "overlearning", mistaking random variation for "signal"
  - Minimum confidence level of 10% (related to strength of association)

# Results: Rule Interestingness Measures

- Fall 2012 cohort
  - Number of different courses taken: 156
  - Number of rules: 178
  - Maximum rule support: 16.1%; mean: 5.9%
  - Maximum antecedent support: 56.1%; mean: 11.1%
- Fall 2013 cohort
  - Number of different courses taken: 144
  - Number of rules: 61
  - Maximum rule support: 42.4%; mean: 10.9%
  - Maximum antecedent support: 71.8%; mean: 30.2%

# Results: Differences in Course Taking Behavior

- Enrollment in CSSK-1200
  - Fall 2012: 13.9%
  - Fall 2013: 71.8%
- Enrollment in remedial English
  - Fall 2012: 56.1%
  - Fall 2013: 50.6%
- Enrollment in remedial reading
  - Fall 2012: 16.9%
  - Fall 2013: 18.7%

# Results: Differences in Course Taking Behavior

- Enrollment in remedial (lower-level) math
  - Fall 2012: 22.1%
  - Fall 2013: 27.5%
- Enrollment in remedial (upper-level) math
  - Fall 2012: 13.2%
  - Fall 2013: 8.1%
- Enrollment in remedial English and remedial reading
  - Fall 2012: 14.6% (expected: 9.5%)
  - Fall 2013: 14.7% (expected: 9.5%)

# Results: Differences in Course Taking Behavior

- Enrollment in all three remedial areas (lower-level math)
  - Fall 2012: 3.2% (expected: 2.1%)
  - Fall 2013: 7.0% (expected: 2.6%)
- Enrollment in all three remedial areas (lower-level math) and CSSK-1200
  - Fall 2012: 0.5% (expected: 0.3%)
  - Fall 2013: 6.8% (expected: 1.9%)

# Results: Differences in Course Taking Behavior

- Stronger association between enrolling in courses in all three remedial areas and enrolling in CSSK-1200
  - Conditional probability of enrolling in CSSK-1200, given enrolling in remedial English, reading, and math (lower-level)
    - Fall 2012: 15.4% (Rule support: 0.5%) Lift: 1.1
    - Fall 2013: 96.8% (Rule support: 6.8%) Lift: 1.4

# Results: Differences in Course Taking Behavior

- Stronger association among the remedial courses in the 2013 cohort
  - Conditional probability of enrolling in remedial English and remedial reading, given enrolling in lower-level remedial math
    - Fall 2012: 14.6% (Rule support: 3.2%) Lift: 1.0
    - Fall 2013: 25.4% (Rule support: 7.0%) Lift: 1.7
  - Conditional probability of enrolling in remedial English and lower-level remedial math, given enrolling in remedial reading
    - Fall 2012: 19.1% Lift: 1.2
    - Fall 2013: 37.4% Lift: 2.1

# Results: Differences in Course Taking Behavior

- Stronger association among the remedial courses in the 2013 cohort
  - Conditional probability of enrolling in lower-level remedial math and remedial reading, given enrolling in remedial English
    - Fall 2012: 5.8% Lift: 1.8
    - Fall 2013: 13.8% Lift: 1.7
  - Average conditional probability of enrolling in the other two remedial areas, given enrolling in one of the remedial areas
    - Fall 2012: 13.2% Lift: 1.3
    - Fall 2013: 25.5% Lift: 1.8

# Discussion

- Fall 2013 cohort had reduced variability/greater homogeneity in course taking by comparison with the fall 2012 cohort; this was manifested in the remedial courses and the college success skills course
- The large increase in the number of students enrolling in the college success skills course is clearly an impact of the policy mandating the taking of that course for at-risk students
- The stronger association in the 2013 cohort between enrolling in remedial courses in all three areas (including lower-level math) and enrolling in the college success skills course may also be an effect of the policy

# Discussion

- The stronger association among the remedial courses in the 2013 cohort would seem on the face of it to be an unlikely consequence of the college success skills policy, since that policy does not address the taking of remedial courses
- The stronger association among the remedial courses in the 2013 cohort may be a reflection of the implementation of the “Easy Start” intake process and associated changes in counseling practices

# Concluding Remarks

- Substantial changes in enrollment behavior occurred in the at risk population after implementation of the CSSK policy, and some of these changes are too large to be the result of random sampling variability
- Some of these changes are clearly the result of the CSSK policy; the extent to which other changes reflect the policy, rather than the new student intake process remains unclear
- ARM—a data mining technique—facilitates the analysis of transcript data; ARM easily/quickly generates key information related to enrollment behavior