A Step-by-Step Introduction to Building a Student-at-Risk Prediction Model Using SPSS

http://www.unr.edu/ia/research

Serge Herzog, PhD
Director, Institutional Analysis
Consultant, CRDA StatLab
University of Nevada, Reno
Reno, NV, serge@unr.edu

John Stanley, MEd
Associate Director, Institutional Effectiveness
University of Hawaii – West Oahu
Kapolei, HI, jstanley@hawaii.edu

AIR Forum
Denver, CO, May 25-29, 2015
Workshop Objectives

1. Develop a conceptual understanding of how predictive models developed by an IR office can improve institutional effectiveness;

2. Learn how to set up a matriculation system (or census warehouse) data file in SPSS that can be used to develop a predictive statistical model to identify students at risk;

3. Learn how to use historical data to ‘automatically’ develop predictor coefficients to estimate (score) the dropout risk for students in future cohorts; and

4. Learn how to translate the student dropout risk into a relative percentile risk score to assist student support services with ‘actionable’ information.
Two Institutions, One Mission
Challenges for Institutional Research

- Compliance vs. Self-Improvement
- Developing a culture of evidence
- From reporting to analysis
- Converting results into ‘actionable’ statements
- From ‘data silos’ to integrated warehouse
- Leverage technology, stay abreast of tech
- Follow highest standards, best practices
- Know your customers, mission
- Empower staff, continuous honing of skills
The Institutional Context

• **Student success: a strategic imperative**
• Performance-based state funding impending
• Dwindling state support for higher education
• *Tuition-revenue maximization*
• Reputation and marketing
• Effective senior-management support by IR
• K-16 Education Collaborative
  – High school transcript study
  – High school gateway curriculum
  – Reversing the tide of college remediation
## The Institutional Context

### Setting the Stage

<table>
<thead>
<tr>
<th>Total Enrollment</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>Undergraduate</td>
<td>13,660</td>
<td>14,415</td>
<td>14,675</td>
</tr>
<tr>
<td>Graduate</td>
<td>3,248</td>
<td>2,935</td>
<td>2,894</td>
</tr>
<tr>
<td>First-Professional (Medical School)</td>
<td>246</td>
<td>249</td>
<td>251</td>
</tr>
<tr>
<td>Non-Degree</td>
<td>525</td>
<td>405</td>
<td>407</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>American Indian/Alaskan</td>
<td>173</td>
<td>156</td>
<td>152</td>
</tr>
<tr>
<td>Asian American</td>
<td>1,142</td>
<td>1,053</td>
<td>1,148</td>
</tr>
<tr>
<td>Black, Non-Hispanic</td>
<td>469</td>
<td>557</td>
<td>610</td>
</tr>
<tr>
<td>Hispanic</td>
<td>1,911</td>
<td>2,032</td>
<td>2,419</td>
</tr>
<tr>
<td>Multi-Ethnic</td>
<td>534</td>
<td>872</td>
<td>942</td>
</tr>
<tr>
<td>Pacific Islander</td>
<td>127</td>
<td>41</td>
<td>68</td>
</tr>
<tr>
<td>White, Non-Hispanic</td>
<td>12,583</td>
<td>12,329</td>
<td>12,150</td>
</tr>
<tr>
<td>Non-Resident Alien</td>
<td>554</td>
<td>594</td>
<td>512</td>
</tr>
<tr>
<td>Unknown/Unspecified</td>
<td>186</td>
<td>370</td>
<td>226</td>
</tr>
</tbody>
</table>
The Institutional Context
New Freshmen Enrollment

Setting the Stage

<table>
<thead>
<tr>
<th>Year</th>
<th>Total</th>
<th>Out of State</th>
<th>Nevada Residents</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008 Fall</td>
<td>2,296</td>
<td>366</td>
<td>1,930</td>
</tr>
<tr>
<td>2009 Fall</td>
<td>2,172</td>
<td>354</td>
<td>1,818</td>
</tr>
<tr>
<td>2010 Fall</td>
<td>2,764</td>
<td>599</td>
<td>2,165</td>
</tr>
<tr>
<td>2011 Fall</td>
<td>2,880</td>
<td>835</td>
<td>2,045</td>
</tr>
<tr>
<td>2012 Fall</td>
<td>2,780</td>
<td>724</td>
<td>2,056</td>
</tr>
</tbody>
</table>
Examples of Actionable Findings

- Study abroad enhances academic performance

- Impact of classroom facilities/schedule on learning
  - Smaller rooms are preferable
  - After-2pm courses associated with lower performance

- **Student financial aid to maximize retention**
  - Tuition discounts for middle-income students
  - More academic support for low-income students
  - [http://www.uark.edu/ua/der/EWPA/Research/School_Finance/1802.html](http://www.uark.edu/ua/der/EWPA/Research/School_Finance/1802.html)

- Effect of high school environment on freshmen success
  - [http://www.uark.edu/ua/der/EWPA/Research/Achievement/1808.html](http://www.uark.edu/ua/der/EWPA/Research/Achievement/1808.html)

- **IR Data to drive institutional strategic planning**
Raising Graduation Rates
Comparing 4-year and 6-year-plus Graduates

**Opportunity cost** of staying one more year in college = **$32,000** in foregone earnings plus annual increase in tuition cost.*

- HS GPA: 3.5 vs 3.2
- ACT: 24.5 vs 22.2
- First-Y GPA: 3.35 vs 2.71
- CoreHum 201 Grade: 3.3 vs 2.6
- MathGPA: 3.12 vs 2.4
- Honors Courses: 14% vs 5%
- Change in Major: 25% vs 55%
- Capstone GPA: 3.5 vs 3.2
- Avg annual remaining need: $2,610 vs $3,270
- Final GPA: 3.4 vs. 2.9
- Internship: 31% vs 24%
- Difference in avg semester load: 3 credits

Improving the Bottom Line

• Rise in freshmen retention by 4 percentage points due to better at-risk forecasting
  – AY 2010-11 additional net tuition revenues = $215,119 (for 94 NV,19 WUE, excl OS students) for one cohort in one year, without OS $!
  – Downstream cumulative additional net tuition revenues result in $ millions!

• Incentive for student to speed up graduation
  – Opportunity cost per year in foregone earnings = $32,000 per year (published constant 2010-$)
Relevant Previous Research

Impact of this At-Risk Forecasting Model

• *University Retention Rates Hold Steady As States Balance Access with Success.* Scripps Howard Foundation Wire, April 15, 2011.


• Consulting services to IR offices at institutions in Arizona, California, Hawaii, and Texas.
At-Risk Forecasting Model

- Identify at-risk freshmen students after initial matriculation for early intervention program
- Develop regression model to predict dropout risk of future cohort
  - Determine cut value to maximize correct classification rate
  - Identify statistical outliers to get trimmed dataset
  - Chose model with optimal balance in correct classification
- Dropout risk scoring for new freshmen
  - Transformation of the logit(\(p\)) into probability scores
  - Decile grouping of scored students
  - Compare deciles with actual enrollment and other predicted enrollment (MAP-Works: [http://www.unr.edu/mapworks](http://www.unr.edu/mapworks))
- Reporting of dropout risk via secure online access
Goal 2: Data file setup

• Data sources
  – Matriculation system (Peoplesoft, data warehouse)
  – New student survey (in PS starting fall 2011)

• Student cohorts
  – New full-time first-year students (incl. advanced standing)
  – Historical cohorts: fall 2011-13  (training set, N=7,085)
  – Predicted cohort:  fall 2014       (holdout set, N=2,945)

• Data elements (predictors) at start of first semester
  – Student socio-demographics (personal, parent attributes)
  – Academic preparation (high school GPA, test scores)
  – Financial aid profile (unmet need, aid type received, EFC)
  – Student motivation (proxy variables)
  – Student social integration (on-campus experiences)
  – Student academic experience (credit load, math/English)
Goal 2: Data file setup

- **Student socio-demographics** (10 predictors)
  - Age19Plus, Male, Hisp, Blk, OS, WUE, Clark, MotherEd, FatherEd, Pell

- **Academic preparation** (2 predictors)
  - HSPrep (HS Core GPA/Test Score Index), AdvStanding

- **Financial aid profile** (6 predictors)
  - Unmet, Loans, Merit, EFCLow$7725, EFCMid$22995, EFCHigh

- **Student motivation** (2 predictors)
  - EdGoal, UNRFirst

- **Student social integration** (5 predictors)
  - LLCFlag, CampWork, OnCampus, WorkPlanNo, WorkPlanFT

- **Student academic experience** (6 predictors)
  - Crs13to15, Crs16up, NoEngl, NoMath, DistEd, Undeclared
Data Management Tasks

• Exploratory data analysis
  – Variable selection (bivariate regression on outcome variable)
  – Variable coding (continuous vs. dummy/binary)
  – Missing data imputation
  – Derived variable(s)
    • HSPrep = (HSGPA*12.5)+(ACTM*.69)+(ACTE*.69)

• Logistic regression model
  – Preliminary model fit (-2LL test/score, pseudo R², HL sig.)
  – Check for outliers with diagnostic tools
  – Check correct classification rate (CCR) for enrollees vs. non-enrollees (i.e. model sensitivity vs. specificity) using baseline probability and Receiver Operating Characteristics (ROC) curve
Data Management Tasks

• Imputation example: HS Preparation index score for cases with missing core GPA or test score
  – Regress core GPA and test score on each other
  – Use regression coefficients to estimate GPA/test score, respectively
  – Run HSPrep index equation for new cases

Goal 2: Data file setup

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Constant)</td>
<td>2.167</td>
<td>.027</td>
<td>.001</td>
<td>79.054</td>
</tr>
<tr>
<td>ACT_COMP</td>
<td>.060</td>
<td>.419</td>
<td>.000</td>
<td></td>
</tr>
</tbody>
</table>

a. Dependent Variable: HS_CORE_GPA
SPSS Menu Tasks

- Select full-time students only

Goal 3: Estimate dropout risk
Goal 3: Estimate dropout risk

SPSS Menu Tasks

- Select Analyze, Regression, Binary
Goal 3: Estimate dropout risk

SPSS Menu Tasks

- Select Analyze, Regression, Binary, Save
Goal 3: Estimate dropout risk

SPSS Menu Tasks

- Select Analyze, Regression, Binary
  - Under Options, select HL goodness-of-fit
Goal 3: Estimate dropout risk

SPSS Menu Tasks

• Select Analyze, Regression, Binary
  – Under Selection Variable, select Training variable, click Rule, insert 1
  – Click Paste (inserts syntax in syntax window)
Goal 3: Estimate dropout risk

SPSS Menu Tasks

• Select Analyze, Regression, Binary
  – Click Paste (creates syntax in new window)
• Edit syntax as needed to re-specify parameters, re-estimate the dropout risk

DATASET ACTIVATE DataSet1.
LOGISTIC REGRESSION VARIABLES SprRetention
  /SELECT=Training EQ 1
  /METHOD=ENTER UNRFFirst EdGoal MotherEd FatherEd NoEnglish WorkPlanFT WorkPlanNo Unmet Pell Loans
  • Merit NoMath NoEngl DistEd AdvStanding Age19plus Male Hisp Blk Clark WUE OS
OnCampus LLCFlag
  • CampWork HSPrep Crs13to15 Crs16up EFCLow$7725 EFCMid$22995 EFCHigh Undeclared
  /SAVE=PRED PGROUP COOK ZRESID
  /PRINT=GOODFIT
  /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
Goal 3: Estimate dropout risk

SPSS Output File

- R-square = .095 ; HL test sig. = .021
- Correct classification rate (CCR) for spring dropout is nil in both training and holdout data (0.0%)

### Classification Table

<table>
<thead>
<tr>
<th>Observed</th>
<th>SprRetention</th>
<th>Percentage Correct</th>
<th>SprRetention</th>
<th>Percentage Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>229</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>6596</td>
<td>100.0</td>
<td>2722</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td>92.8</td>
<td></td>
<td>92.2</td>
</tr>
</tbody>
</table>

a. The cut value is .500
b. Selected cases Training EQ 1
c. Unselected cases Training NE 1
d. Some of the unselected cases are not classified due to either missing values in the independent variables or categorical variables with values out of the range of the selected cases.
Goal 3: Estimate dropout risk

SPSS Menu Tasks

- Select Analyze, Regression, Binary
  - Click Paste (creates syntax in new window)
- Edit cut value in syntax to reflect baseline probability of spring retention (i.e. 92.8%)

```
DATASET ACTIVATE DataSet1.
LOGISTIC REGRESSION VARIABLES SprRetention
  /SELECT=Training EQ 1
  /METHOD=ENTER UNRFist EdGoal MotherEd FatherEd NoEnglish WorkPlanFT WorkPlanNo Unmet Pell Loans
     Merit NoMath NoEngl DistEd AdvStanding Age19plus Male Hisp Blk Clark WUE OS
OnCampus LLCFLag
   CampWork HSprep Crs13to15 Crs16up EFCLow$7725 EFCMid$22995 EFCHigh Undeclared
/SAVE=PRED PGROUP COOK ZRESID
/PRINT=GOODFIT
/CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.928).
```
Goal 3: Estimate dropout risk

SPSS Output File

- CCR for spring dropout at 65% for training and 67% for holdout cohort, respectively
- Still low correct classification rate of dropout students
  - Check for outliers

### Classification Table

<table>
<thead>
<tr>
<th>Observed</th>
<th>SprRetention</th>
<th>Percentage Correct</th>
<th>Unselected Cases</th>
<th>Percentage Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Step 1</td>
<td>335</td>
<td>177</td>
<td>65.4</td>
<td>154</td>
</tr>
<tr>
<td></td>
<td>2303</td>
<td>4293</td>
<td>65.1</td>
<td>957</td>
</tr>
<tr>
<td>Overall Percentage</td>
<td></td>
<td></td>
<td>65.1</td>
<td></td>
</tr>
</tbody>
</table>

a. The cut value is .928
b. Selected cases Training EQ 1
c. Unselected cases Training NE 1
d. Some of the unselected cases are not classified due to either missing values in the independent variables or categorical variables with values out of the range of the selected cases.
Identify Outlier Cases

- Examine Cook’s distance (COO_) and standardized residuals (ZRE_)
- Exclude cases with
  - Cook’s distance greater than 1, or visual separation
  - Standardized residuals greater |3|
- More stringent exclusion rules
  - Cook’s distance greater than 4/n=number of cases
  - Standardized residuals greater |2|

Goal 3: Estimate dropout risk
Goal 3: Estimate dropout risk

Identify Outlier Cases

Graph showing the distribution of analogs of Cook's influence statistics against CaseNum.
Goal 3: Estimate dropout risk

SPSS Menu Tasks

- Exclude outliers via ‘select cases if’ function
- Use ‘filter_Trim (already included)’
Results from Trimmed Data

- Cut value adjusted to .964 to reflect trimmed training data
- Dropout CCR at 96% for training, 89.2% for holdout data
- Overall CCR at 81% for both training and holdout data
- R-square = .333, but HL reached significance (<.05)

Goal 3: Estimate dropout risk

Classification Table

<table>
<thead>
<tr>
<th>Observed</th>
<th>SprRetention</th>
<th>Percentage Correct</th>
<th>SprRetention</th>
<th>Percentage Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Step 1</td>
<td>235</td>
<td>10</td>
<td>95.9</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>1301</td>
<td>5274</td>
<td>80.2</td>
<td>534</td>
</tr>
<tr>
<td>Overall</td>
<td>1536</td>
<td>5374</td>
<td>80.8</td>
<td>533</td>
</tr>
</tbody>
</table>

a. The cut value is .964
b. Selected cases Training EQ 1
c. Unselected cases Training NE 1
Results from Trimmed Data

- Lower accuracy rate in deciles 2 and 4, i.e. less accurate in higher-risk deciles, not equal accuracy across all deciles

<table>
<thead>
<tr>
<th>Step</th>
<th>SprRetention = 0</th>
<th>SprRetention = 1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Observed</td>
<td>Expected</td>
</tr>
<tr>
<td>1</td>
<td>154</td>
<td>153.870</td>
</tr>
<tr>
<td>2</td>
<td>66</td>
<td>41.547</td>
</tr>
<tr>
<td>3</td>
<td>23</td>
<td>21.589</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>12.075</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>7.177</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>4.288</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>2.404</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>1.294</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0.605</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0.150</td>
</tr>
</tbody>
</table>
Measure of Goodness of Fit

- Simultaneous measure of sensitivity (true positive) and specificity (true negative) for all possible cutoff values
- Calculate area under the ROC curve (exercise)
- Area under the ROC: 0.709 (full) vs. 0.921 (trimmed)
- Suggested cutoff point to maximize overall CCR is ~ 0.7 for full data, and ~ 0.81 for trimmed data
Goal 3: Estimate dropout risk

Measure of Goodness of Fit

Test Variable:
- Predicted probability [FRE_3]

State Variable:
- SprRetention
  Value of State Variable: 1

Display:
- ROC Curve
  - With diagonal reference line
  - Standard error and confidence interval
  - Coordinate points of the ROC Curve
Assess Prediction Accuracy

• Compare results from full-data model with results from trimmed-data model
• Determine the best cut value (classification) based on re-adjusted baseline probability versus ROC-curve derived probability level
• Evaluate relative cost of (in-)accurate prediction of retained students (sensitivity) versus dropout students (specificity)
Unbalanced Data

- Proportion of dropouts is usually much smaller than proportion of retained students.
- Number of cases in rare event (dropout) should be sufficient to yield *minimum* 10:1 ratio with number of predictors.
- Check standard errors in coefficient results table ("Variables in the Equation") for inflated values.
- Check variance inflation factor (VIF) in collinearity diagnostics (must run linear regression) to determine which predictor(s) to remove if ratio well below 10:1 or run *Exact Logistic Regression* (see example at http://www.ats.ucla.edu/stat/stata/dae/exlogit.htm).
Translate Dropout Risk

- Convert retention probability to dropout risk deciles (1 = highest, 10 = lowest)
- Copy retention probability for fall 2014 cohort to HoldoutPredProb
- Group into deciles using binning function:
  - Transform, Visual Binning, Make 9 cutpoints, Label ‘Deciles’, check ‘reverse scale’
- Note bottom high-risk deciles with far lower retention probability (run decile average)
- Create new data file for scored (fall 2014) cohort, including risk score, decile grouping, and other useful data elements to support student assistance personnel
Sample Data for Advisors

- [http://www.unr.edu/ia](http://www.unr.edu/ia)

<table>
<thead>
<tr>
<th>R Number</th>
<th>Last Name</th>
<th>First Name</th>
<th>Email Addr</th>
<th>Age</th>
<th>College</th>
<th>Dept</th>
<th>Major</th>
<th>Dropout Risk Decile</th>
<th>Relative Spring %tile</th>
</tr>
</thead>
<tbody>
<tr>
<td>18 LBA</td>
<td>ART</td>
<td>BA-AHI</td>
<td></td>
<td>9</td>
<td>ART</td>
<td></td>
<td></td>
<td>14.92</td>
<td></td>
</tr>
<tr>
<td>18 LBA</td>
<td>ANTH</td>
<td>BA-AN</td>
<td></td>
<td>8</td>
<td>ANTH</td>
<td></td>
<td></td>
<td>28.52</td>
<td></td>
</tr>
<tr>
<td>18 LBA</td>
<td>ANTH</td>
<td>BA-AN</td>
<td></td>
<td>7</td>
<td>ANTH</td>
<td></td>
<td></td>
<td>36.80</td>
<td></td>
</tr>
<tr>
<td>18 LBA</td>
<td>ANTH</td>
<td>BA-AN</td>
<td></td>
<td>7</td>
<td>ANTH</td>
<td></td>
<td></td>
<td>39.18</td>
<td></td>
</tr>
<tr>
<td>18 LBA</td>
<td>ANTH</td>
<td>BA-AN</td>
<td></td>
<td>6</td>
<td>ANTH</td>
<td></td>
<td></td>
<td>46.87</td>
<td></td>
</tr>
<tr>
<td>18 LBA</td>
<td>ANTH</td>
<td>BA-AN</td>
<td></td>
<td>4</td>
<td>ANTH</td>
<td></td>
<td></td>
<td>66.48</td>
<td></td>
</tr>
<tr>
<td>19 LBA</td>
<td>ANTH</td>
<td>BA-AN</td>
<td></td>
<td>1</td>
<td>ANTH</td>
<td></td>
<td></td>
<td>92.42</td>
<td></td>
</tr>
<tr>
<td>18 LBA</td>
<td>ANTH</td>
<td>BA-AN</td>
<td></td>
<td>1</td>
<td>ANTH</td>
<td></td>
<td></td>
<td>95.57</td>
<td></td>
</tr>
</tbody>
</table>
### Sample Data for Advisors

- **http://www.unr.edu/ia**

<table>
<thead>
<tr>
<th>Gender</th>
<th>Ethnicity</th>
<th>Credits</th>
<th>Resident</th>
<th>State/Cnty</th>
<th>HS GPA</th>
<th>ACTE</th>
<th>ACTM</th>
<th>Has Pell$ (1=yes)</th>
<th>Has Loan$ (1=yes)</th>
<th>Clark Cnty Resi (1=yes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>AS</td>
<td>12 NV</td>
<td>NWA</td>
<td>3.10</td>
<td>24</td>
<td>18</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>WH</td>
<td>15 NV</td>
<td>NCL</td>
<td>3.23</td>
<td>21</td>
<td>18</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>M</td>
<td>WH</td>
<td>16 WU</td>
<td>CA</td>
<td>3.19</td>
<td>23</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>M</td>
<td>WH</td>
<td>17 WU</td>
<td>OR</td>
<td>3.23</td>
<td>24</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>WH</td>
<td>16 NV</td>
<td>NWA</td>
<td>3.18</td>
<td>17</td>
<td>17</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>WH</td>
<td>15 NV</td>
<td>NDO</td>
<td>3.47</td>
<td>30</td>
<td>21</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>M</td>
<td>WH</td>
<td>15 NV</td>
<td>NWA</td>
<td>3.65</td>
<td>26</td>
<td>25</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>AS</td>
<td>16 NV</td>
<td>NCL</td>
<td>3.90</td>
<td>30</td>
<td>28</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Exercise

• Estimate fall-fall dropout risk
Selected Factors and Spring Retention
Fall Cohorts 2002-09 (N=17,311)

Predicting Student Success

Retention Rate

Decile (Low to High)

- Rem Need
- Pell $
- Loan $
- % Credits w/F,W
- AcadIndex
Selected Factors and 2nd Fall Retention
Spring-Retained Fall Cohorts 2002-09 (N=15,570)

Retention Rate

Decile (Low to High)

- Rem Need
- Pell $
- Loan $
- % Credits w/F,W
- AcadIndex

Predicting Student Success
Data Analysis

Predicting Student Success

Predicted Retention Decile
Spring Status of Fall 2010 Cohort

Actual  Retained  Departed

1  71  48  33
2  48  20  18
3  20  18  18
4  18  13  13
5  13  15  15
6  15  10  10
7  10  10  10
8  10  10  10
9  10  10  10
10  10  10  10

42
Data Analysis

Predicting Student Success

MAP-Works Risk Assessment, Fall 2010 Cohort

*Assesses fall 2011 dropout risk of spring-retained
## Gauging Survey Value

### Predicting Student Success

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Baseline</th>
<th></th>
<th>Baseline</th>
<th></th>
<th>Baseline</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Wald</td>
<td>Sig.</td>
<td>Wald</td>
<td>Sig.</td>
<td>Wald</td>
<td>Sig.</td>
</tr>
<tr>
<td>Age</td>
<td>2.1</td>
<td>*</td>
<td>2.4</td>
<td>*</td>
<td>2.2</td>
<td>*</td>
</tr>
<tr>
<td>Asian</td>
<td>0.5</td>
<td></td>
<td>0.1</td>
<td></td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>Credits Enrolled</td>
<td>5.3</td>
<td>***</td>
<td>5.5</td>
<td>***</td>
<td>6.1</td>
<td>***</td>
</tr>
<tr>
<td>ClarkRural</td>
<td>13.6</td>
<td>***</td>
<td>13.6</td>
<td>***</td>
<td>14.4</td>
<td>***</td>
</tr>
<tr>
<td>LoanFlag</td>
<td>1.4</td>
<td>*</td>
<td>0.7</td>
<td></td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>PellFlag</td>
<td>1.3</td>
<td>*</td>
<td>1.7</td>
<td>*</td>
<td>1.9</td>
<td>*</td>
</tr>
<tr>
<td>MillFlag</td>
<td>13.9</td>
<td>***</td>
<td>16.6</td>
<td>***</td>
<td>17.1</td>
<td>***</td>
</tr>
<tr>
<td>AthleteFlag</td>
<td>0.0</td>
<td></td>
<td>0.2</td>
<td></td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>HSGFlag</td>
<td>15.9</td>
<td>***</td>
<td>12.2</td>
<td>***</td>
<td>13.5</td>
<td>***</td>
</tr>
<tr>
<td>AcadIndex</td>
<td>7.1</td>
<td>***</td>
<td>1.5</td>
<td>*</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>RemNeedFlag</td>
<td>2.4</td>
<td>*</td>
<td>2.6</td>
<td>*</td>
<td>2.8</td>
<td>**</td>
</tr>
<tr>
<td>MWR HI</td>
<td></td>
<td></td>
<td>9.9</td>
<td>***</td>
<td>23.5</td>
<td>***</td>
</tr>
<tr>
<td>MWR MO</td>
<td></td>
<td></td>
<td>4.1</td>
<td>***</td>
<td>11.1</td>
<td>***</td>
</tr>
<tr>
<td>LR test pass</td>
<td></td>
<td></td>
<td>yes</td>
<td></td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Nagelkerke R²</td>
<td>0.19</td>
<td></td>
<td>0.21</td>
<td></td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>CCR of At-Risk</td>
<td>76.0%</td>
<td></td>
<td>75.6%</td>
<td></td>
<td>78.0%</td>
<td></td>
</tr>
</tbody>
</table>
Gauging Survey Value

• A sustained 2% point rise in prediction accuracy over 5 years due to MAP-Works may translate into:
  – $237,500 in additional net revenue (5x1900x5x5) per cohort
  – Assuming no freshmen enrollment growth

  *But…*

• Five-year cost of survey implementation
  – Product cost/fee, on-campus HR/IT investment

• Data not available until late in the semester!

• Balanced model (2002-10 data) yields 79% CCR for at-risk students, i.e. better than survey prediction

• Survey prediction furnishes no at-risk deciles
Value of Student Self-Reported Data for At-Risk Prediction

**Sources:**
- On-campus surveys
- ACT Student Profile Q
- SAT Student Descriptive Q
- NSSE, CIRP (HERI-UCLA)

**Limitations:**
- Validity of acad exp questions
- Convergent validity of construct
- Cognitive vs. affective questions
- Interpretive ambiguity
- Mental recall
- Vague quantifiers
Community College Data Set Details

• Data Sources
  – Matriculation system (Banner, data warehouse)
  – Educational Goals student survey (in Banner starting fall 2010)

• Student cohorts
  – New first-year students (part-time and full-time)
  – Historical cohorts: fall 2010-12 (training set, N=2,243)
  – Predicted cohort: fall 2013 (holdout set, N=626)

• Data elements (predictors) at start of first semester
  – Student socio-geo-demographics (personal)
  – Academic preparation (high school GPA, Compass test scores, remediation)
  – Financial aid profile (unmet need, aid type received, EFC)
  – Student motivation (proxy variables, degree audit logins)
    – Student social integration (proxy variables)
  – Student academic experience (credit load, math/English)
• Student socio-demographics (10 predictors)
  – AGE, AGE19PLUS, FEMALE, URM, URMINCFILIPINO, WHITE, ISLANDWEST, ISLANDURBAN, ISLANDRURAL, OUTOFSTATE, MILITARY, LOWPERFORMHIGHHSCHOOL

• Academic preparation (2 predictors)
  – COMPASS, REMEDIAL/DEVELOPMENTAL/COLLEGELEVEL (Math/English),

• Financial aid profile (6 predictors)
  – PERCENTUNMETNEED, PELL, EFC, COA

• Student motivation (2 predictors)
  – EDGOAL1, EDGOAL2, STARUSAGE

• Student academic experience (6 predictors)
  – CREDITSATTEMPTED, CREDITSLESS9, FULLTIME, DISTANCEEDENROLL, MATHENROLL, ENGLISHENROLL, ECEDMAJOR, APPLIEDTRADESMajor
Goal 3: Estimate dropout risk

CC Data: SPSS Menu Tasks

- Select Analyze, Regression, Binary
  - Click Paste (creates syntax in new window)
- Edit syntax as needed to re-specify parameters, re-estimate the dropout risk

```plaintext
DATASET ACTIVATE DataSet2.
LOGISTIC REGRESSION VARIABLES RETENTIONSPRING
/SELECT=TRAININGVARIABLE EQ 1
/METHOD=ENTER CREDITSATTEMPTED DISTANCEEDENROLLMENT URM FEMALE ISLANDRURAL OUTOFSTATE LOWPERFORMHIGHSCHOOL ECEDMAJOR AGE19PLUS EDGOAL1 PELL PERCENTUNMETNEED STARUSAGE COMPASSREADING COMPASSANYMATHHIGHEST REMEDIALMATH REMEDIALENG
/SAVE=PRED PGROUP COOK ZRESID
/PRINT=GOODFIT
/CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```
Goal 3: Estimate dropout risk

SPSS Output File

- R-square = .255; HL test sig. = .103
- Null model correct classification rate (CCR) for spring dropout is nil in both training and holdout data (0.0%)

### Classification Table

<table>
<thead>
<tr>
<th>Observed</th>
<th>RETENTIONSPRING</th>
<th>Percentage Correct</th>
<th>RETENTIONSPRING</th>
<th>Percentage Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Selected Cases</td>
<td></td>
<td>Unselected Cases</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Step 0</td>
<td>RETENTIONSPRING</td>
<td>0 0</td>
<td>626</td>
<td>0 177</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0 0</td>
<td>1617</td>
<td>0 449</td>
</tr>
<tr>
<td>Overall</td>
<td>Percentage</td>
<td>72.1</td>
<td></td>
<td>71.7</td>
</tr>
</tbody>
</table>

a. Constant is included in the model.
b. The cut value is .500
c. Selected cases TRAININGVARIABLE EQ 1
d. Unselected cases TRAININGVARIABLE NE 1
SPSS Menu Tasks

- Select Analyze, Regression, Binary
  - Click Paste (creates syntax in new window)
- Edit cut value in syntax to reflect baseline probability of spring retention (i.e. 72.1%)

```
DATASET ACTIVATE DataSet2.
LOGISTIC REGRESSION VARIABLES RETENTIONSPRING
/SELECT=TRAININGVARIABLE EQ 1
/METHOD=ENTER CREDITSATTEMPTED DISTANCEEDENROLLMENT URM FEMALE ISLANDRURAL OUTOFSTATE
  LOWPERFORMHIGHSCHOOL ECEDMAJOR AGE19PLUS EDGOAL1 PELL PERCENTUNMETNEED STARUSAGE COMPASSREADING
  COMPASSANYMATHHIGHEST REMEDIALMATH REMEDIALENG
/SAVE=PRED PGROUP COOK ZRESID
/PRINT=GOODFIT
/CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.721).
```
Goal 3: Estimate dropout risk

SPSS Output File

- R-square = .255 ; HL test sig. = .103
- CCR for spring dropout at 70% for training and 80% for holdout cohorts
- Good correct classification rate of dropout students
  - Check for outliers to seek further improvement

### Classification Table

<table>
<thead>
<tr>
<th>Observed</th>
<th>RETENTIONSPRING</th>
<th>Percentage Correct</th>
<th>Unselected Cases</th>
<th>Percentage Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RETENTIONSPRING</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>70.3</td>
<td>142</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>490</td>
<td>1127</td>
<td>69.7</td>
</tr>
<tr>
<td>Overall Percentage</td>
<td>69.9</td>
<td>68.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. The cut value is .721
b. Selected cases TRAININGVARIABLE EQ 1
c. Unselected cases TRAININGVARIABLE NE 1
Identify Outlier Cases

- Examine Cook’s distance (COO_) and standardized residuals (ZRE_)
- Exclude cases with
  - Cook’s distance greater than 1, or visual separation
  - Standardized residuals greater |3|
- More stringent exclusion rules
  - Cook’s distance greater than 4/n=number of cases
  - Standardized residuals greater |2|
Goal 3: Estimate dropout risk

SPSS Menu Tasks

- Exclude outliers via ‘select cases if’ function
- Use ‘filter_Trim (already included)’
Goal 3: Estimate dropout risk

Calculate new baseline from trimmed data

- New baseline retention rate is .723 based on trimmed training data

<table>
<thead>
<tr>
<th>Observed</th>
<th>RETENTIONSPRING</th>
<th>Percentage Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>.0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>100.0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>173</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>443</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td>72.3</td>
</tr>
</tbody>
</table>

---

- a. Constant is included in the model.
- b. The cut value is .721
- c. Selected cases TRAININGVARIABLE EQ 1
- d. Unselected cases TRAININGVARIABLE NE 1
Updated Results from Trimmed Data

- Cut value adjusted to .723 to reflect trimmed training data
- Dropout CCR at 72% for training, 83% for holdout data
- Overall CCR at ~70% for both training and holdout data
- R-square = .292, but HL reached significance (<.05)

<table>
<thead>
<tr>
<th></th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Selected Cases</td>
</tr>
<tr>
<td></td>
<td>RETENTIONSPRING</td>
</tr>
<tr>
<td>Observed</td>
<td>0</td>
</tr>
<tr>
<td>Step 1</td>
<td>RETENTIONSPRING</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Overall Percentage</td>
<td></td>
</tr>
</tbody>
</table>

a. The cut value is .723
b. Selected cases TRAININGVARIABLE EQ 1
c. Unselected cases TRAININGVARIABLE NE 1
Progress on Implementation at Honolulu Community College

• Currently doing:
  – Delivering student dropout risk scores to HCC’s Academic Success Center (via an Excel file).
  – Training staff members on using the data.
  – Academic support services moving towards a proactive, targeted approach.
Predictive Analytics at U. of Hawaii

• Relevant previous research has provided a suitable starting point for developing at-risk student forecasting model.
• IR and Advising staff from U. of Nevada-Reno travelled to UH to share insights on implementing predictive analytics.
• Successful is using models for explanatory purposes, but need to evolve towards prediction/prescription.
Takeaway from Collaboration

- Early-alert data key
- Identify results that are actionable.
- Support for academic advising, including training on how to use data.
Barriers to Implementation at the University of Hawaii

- Culture change
- Wary of misuse of data
- More accountability
- Faculty buy-in
Summary

• Predicting students at-risk
  – Keep prediction model parsimonious
  – Keep prediction data for student advising intuitive and simple (actionable)
  – Triangulate prediction data with multiple sources of information
  – Use prediction data as component part of student dropout-risk assessment
  – Follow ‘best practices’ in IR and keep abreast of changes in analytical and data reporting tools

• Using prediction data for student advising
  – Embrace the use of available data
  – Ensure users conceptually understand what’s behind the data
  – Use data as a complementary piece of information when advising students
  – Timing can be critical in terms of student intervention as well as maximizing advising resources

• Stay abreast of new research on predictive analytics:

----------------------------------------

Link to presentation:
http://www.unr.edu/ia/research