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 2016
New Orleans, LA, May 30th – June 3rd
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

New Freshmen Enrollment

<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)
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.*

*Adjusted 2010-$.

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 baseline retention to maximize correct classification
  - Identify statistical outliers to get trimmed dataset
  - Choose model with optimal balance in correct classification
- Dropout risk scoring for new freshmen
  - Transformation of the logit($p$) into probability scores
  - Automated classification and probability score with SPSS
  - Decile grouping of scored students
- 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 2011-14 (training set, N ~ 10,000)
  - Predicted cohort: fall 2015 (holdout set, N ~ 3,300)

- **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, Non-Local, 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, FirstChoice

- **Student social integration** (5 predictors)
  - LLC, CampWork, OnCampus, PlanWorkNo, PlanWorkFT

- **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 R^2, HL sig.)
  – Check for outliers with diagnostic tools (Std residuals, Cook’s)
  – 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 (Constant)</td>
<td>.2167</td>
<td>.027</td>
<td></td>
<td>79.054</td>
</tr>
<tr>
<td>ACT_COMP</td>
<td>.060</td>
<td>.001</td>
<td>.419</td>
<td>51.618</td>
</tr>
</tbody>
</table>

a. Dependent Variable: HS_CORE_GPA
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
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 AdvStanding NoMath NoEngl DistEd Undeclared Age19plus Male Hisp
  Blk OS NonLocal
  WUE OnCampus CampWork Pell Unmet Loans Merit FirstChoice EdGoalGrad
  MoEd4yrColl FathEd4yrColl
  PlanWorkFT PlanWorkNo LLC Crs13to15 Crs16up EFCLow$7725 EFCMid$22995
  EFCHigh HSPrep NoMath NoEngl
  /SAVE=PRED PGROUP COOK ZRESID
  /PRINT=GOODFIT
  /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
SPSS Output File

Goal 3: Estimate dropout risk

- R-square = .11 ; HL test sig. = .004
- Correct classification rate (CCR) for spring dropout is almost nil in both training and holdout data

<table>
<thead>
<tr>
<th>Classification Table^a</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Observed</strong></td>
</tr>
<tr>
<td>SprRetention</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Step 1</td>
</tr>
<tr>
<td>SprRetention</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>Overall Percentage</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. 92.4%)

```
DATASET ACTIVATE DataSet1.
LOGISTIC REGRESSION VARIABLES SprRetention
   /SELECT=Training EQ 1
   /METHOD=ENTER AdvStanding NoMath NoEngl DistEd Undeclared Age19plus Male Hisp Blk OS NonLocal
   WUE OnCampus CampWork Pell Unmet Loans Merit FirstChoice EdGoalGrad MoEd4yrColl FathEd4yrColl
   PlanWorkFT PlanWorkNo LLC Crs13to15 Crs16up EFCLow$7725 EFCMid$22995 EFCHigh HSPrep NoMath NoEngl
   /SAVE=PRED PGROUP COOK ZRESID
   /PRINT=GOODFIT
   /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.924).
```
SPSS Output File

- CCR for spring dropout at 66% for training and 75% for holdout cohort
- Can we raise correct classification rate of dropout students?..............Check for outlier cases

<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>517</td>
<td>175</td>
<td>75.1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3259</td>
<td>1355</td>
<td>56.4</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td>65.5</td>
<td></td>
<td>57.7</td>
</tr>
</tbody>
</table>

a. The cut value is .924
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

Identify Outlier Cases
Goal 3: Estimate dropout risk

SPSS Menu Tasks

• Exclude outliers via ‘select cases if’ function
• Use ‘filter_Outliers (included in data file)’
Results from Trimmed Data

- Cut value adjusted to .926 to reflect trimmed training data
- Dropout CCR at 76.4 % for holdout data
- R-square = .12, but HL reached significance (<.05)

Goal 3: Estimate dropout risk

### Classification Table

<table>
<thead>
<tr>
<th>Observed</th>
<th>Selected Cases</th>
<th>Predicted</th>
<th>Unselected Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SprRetention</td>
<td>Percentage Correct</td>
<td>SprRetention</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Step 1</td>
<td>505</td>
<td>242</td>
<td>67.6</td>
</tr>
<tr>
<td></td>
<td>3223</td>
<td>6191</td>
<td>65.8</td>
</tr>
<tr>
<td>Overall Percentage</td>
<td></td>
<td>65.9</td>
<td></td>
</tr>
</tbody>
</table>

a. The cut value is .926
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.
Results from Trimmed Data

- Lower accuracy rate in middle deciles, i.e. not equal accuracy across all deciles
Results from Second Trimming

- Use `filter_Outliers2` (included in data file), exl Z-res < -2
- Cut value adjusted to .927 to reflect re-trimmed training data
- Dropout CCR at 77.3% for holdout data
- R-square = .123, HL no longer significant (> .05)

### 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>68.4</td>
<td>0</td>
<td>180</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>65.0</td>
<td>1</td>
<td>1383</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td>65.3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- a. The cut value is .927
- 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

Determine Balanced CCR: ROC Charts
Determine Balanced CCR: ROC Charts

- 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: 0.726 (trimmed second time)
- Suggested cutoff point to maximize overall CCR is ~ 0.7
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

<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 Relative (10=highest Spring; 1=lowest)</th>
<th>Relative Spring Retention %tile</th>
</tr>
</thead>
<tbody>
<tr>
<td>18LBA</td>
<td></td>
<td>ART</td>
<td></td>
<td>18</td>
<td>BA</td>
<td>ART</td>
<td>BA-AHI</td>
<td>9</td>
<td>14.92</td>
</tr>
<tr>
<td>18LBA</td>
<td></td>
<td>ANTH</td>
<td></td>
<td>18</td>
<td>BA</td>
<td>ANTH</td>
<td>BA-AN</td>
<td>8</td>
<td>28.52</td>
</tr>
<tr>
<td>18LBA</td>
<td></td>
<td>ANTH</td>
<td></td>
<td>18</td>
<td>BA</td>
<td>ANTH</td>
<td>BA-AN</td>
<td>7</td>
<td>36.80</td>
</tr>
<tr>
<td>18LBA</td>
<td></td>
<td>ANTH</td>
<td></td>
<td>18</td>
<td>BA</td>
<td>ANTH</td>
<td>BA-AN</td>
<td>7</td>
<td>39.18</td>
</tr>
<tr>
<td>18LBA</td>
<td></td>
<td>ANTH</td>
<td></td>
<td>18</td>
<td>BA</td>
<td>ANTH</td>
<td>BA-AN</td>
<td>6</td>
<td>46.87</td>
</tr>
<tr>
<td>18LBA</td>
<td></td>
<td>ANTH</td>
<td></td>
<td>18</td>
<td>BA</td>
<td>ANTH</td>
<td>BA-AN</td>
<td>4</td>
<td>66.48</td>
</tr>
<tr>
<td>19LBA</td>
<td></td>
<td>ANTH</td>
<td></td>
<td>19</td>
<td>BA</td>
<td>ANTH</td>
<td>BA-AN</td>
<td>1</td>
<td>92.42</td>
</tr>
<tr>
<td>18LBA</td>
<td></td>
<td>ANTH</td>
<td></td>
<td>18</td>
<td>BA</td>
<td>ANTH</td>
<td>BA-AN</td>
<td>1</td>
<td>95.57</td>
</tr>
</tbody>
</table>
### Sample Data for Advisors

<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)

Retention Rate

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

Selected Factors and 2nd Fall Retention
Spring-Retained Fall Cohorts 2002-09 (N=15,570)

Retention Rate

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

Decile (Low to High)

Retention Rate
None 1 2 3 4 5 6 7 8 9 10
Data Analysis

MAP-Works Risk Assessment, Fall 2010 Cohort

*Assesses fall 2011 dropout risk of spring-retained students.
### Predicting Student Success

#### Gauging Survey Value

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Baseline Wald</th>
<th>Baseline Sig.</th>
<th>MW Sep Survey Wald</th>
<th>MW Sep Survey Sig.</th>
<th>MW Nov Survey Wald</th>
<th>MW Nov Survey Sig.</th>
</tr>
</thead>
<tbody>
<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 $R^2$</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
Mimic* dataset based on data from:

- ~ 4,300 student enrollment
- Open access
- Large % of under-represented, low income, and first generation students
- 60% male
- Average age is 26 years old
- 66% part-time enrollment
- Over half of academic programs are vocational/career technical
- 18% grad rate (150%)
- 72% fall-to-spring retention first-time freshmen; 50% fall-to-fall retention

*The CC Dataset used in this class has been de-identified, randomized, and altered for instructional and sharing purposes. These “mimic” data do not match actual institutional data, trends, or outcomes.
Community College Data Set Details

- **Data Sources**
  - Matriculation system (Banner, data warehouse)

- **Student cohorts**
  - New first-year students (part-time and full-time)
  - Historical cohorts: fall 2011-13 (training set, N=2,243)
  - Predicted cohort: fall 2014 (holdout set, N=626)
  - Newest cohort: fall 2015 (holdout set #2, N=702)

- **Data elements (predictors) at start of first semester**
  - Student socio-geo-demographics (age, gender, ethnicity, geographic proximity to campus, residency, military)
  - Academic preparation (Compass test scores, high school attended, remediation/developmental courses needed)
  - Financial aid profile (unmet need, pell)
  - Student motivation proxies (degree audit logins, educational goals survey responses)
  - Student academic experience (credit load, math/English enrollment, major type)
• Student socio-demographics (12 predictors)
  – AGE, AGE19PLUS, FEMALE, URM, URMINCAPISINO, WHITE, ISLANDWEST, ISLANDURBAN, ISLANDRURAL, OUTOFSTATE, MILITARY, LOWPERFORMHIGHSCHOOL

• Academic preparation (9 predictors)
  – COMPASS READING, COMPASS WRITING, COMPASSANYMATHHIGHEST, REMEDIAL/DEVELOPMENTAL/COLLEGELEVEL (Math/English) FLAGS,

• Financial aid profile (2 predictors)
  – PERCENTUNMETNEED, PELL

• Student motivation (4 predictors)
  – EDGOAL1, EDGOAL2, STARUSAGE, STARUSAGEAVERAGEFLAG,

• Student academic experience (8 predictors)
  – CREDITSATTEMPTED, CREDITSLESS9, FULLTIME, DISTANCEEDENROLL, ECED MAJOR, APPLIEDTRADESMAJOR, ANYMATHENROLL, ANYENGLISHENROLL
Goal 2: Data File Setup

Step 1: Filter out the 2015 data

Select *Data, Select Cases, If condition…*

*COHORTYEAR ~= 2015*
CC Data: SPSS Menu Tasks

- Select *Analyze, Regression, Binary*
  - *Use same menu options learned in the UNR example.*
  - Click Paste (creates syntax in new window).
- From here on, we will edit syntax as needed to re-specify parameters, re-estimate the dropout risk

```spss
DATASET ACTIVATE DataSet1.
LOGISTIC REGRESSION VARIABLES RETENTIONSPRING
  /SELECT=TRAININGVARIABLE EQ 1
  /METHOD=ENTER CREDITSATTEMPTEDFALL 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).
```
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%)

Goal 3: Estimate dropout risk

Here, we calculated the baseline fall-to-spring retention rate
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
   COMPASSANYMARTHIGHEST REMEDIALMATH REMEDIALENG
 /SAVE=PRED PGROUP COOK ZRESID
 /PRINT=GOODFIT
 /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.721).
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**

```
Observed | Predicted
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Selected Cases</td>
<td>Unselected Cases</td>
</tr>
<tr>
<td></td>
<td>RETENTIONSCELL</td>
<td>RETENTIONSCELL</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Step 1</td>
<td>RETENTIONSCELL</td>
<td>440</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>490</td>
</tr>
<tr>
<td>Overall</td>
<td>Percent</td>
<td>69.9</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 $4/n=\text{number of cases}$
  - Standardized residuals greater |2|
Goal 3: Estimate dropout risk

Identify Outlier Cases

Cook's Values of 0.1 or higher merit outlier exclusion from "eye-balling" the scatterplot.
Goal 3: Estimate dropout risk

SPSS Menu Tasks

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

COHORTYEAR ~= 2015 & (COO_3 < .1 & ZRE_3 < 3 & ZRE_3 > -3)
SPSS Syntax Version of Filter Tasks (fyi)

DATASET ACTIVATE DataSet1.
USE ALL.
COMPUTE filter_$=(COHORTYEAR ~= 2015  &  COO_3 < 1 & ZRE_3 < 3 & ZRE_3 > -3).
VARIABLE LABELS filter_$ 'COHORTYEAR ~= 2015  &  (COO_3 < 1 & ZRE_3 < 3 & ZRE_3 > -3) (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
CC Data: SPSS Menu Tasks

- Run regression syntax again with the 0.721 baseline retention rate

```spss
DATASET ACTIVATE DataSet1.
LOGISTIC REGRESSION VARIABLES RETENTIONSPRING
   /SELECT=TRAININGVARIABLE EQ 1
   /METHOD=ENTER CREDITSATTEMPTEDFALL 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
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>Classification Table&lt;sup&gt;a,b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Observed</strong></td>
</tr>
<tr>
<td>RETENTIONSPRING</td>
</tr>
<tr>
<td>Step 0 RETENTIONSPRING 0</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td><strong>Overall Percentage</strong></td>
</tr>
</tbody>
</table>

- Constant is included in the model.
- The cut value is .721
- Selected cases TRAININGVARIABLE EQ 1
- Unselected cases TRAININGVARIABLE NE 1
CC Data: SPSS Menu Tasks

- Re-run regression syntax AGAIN with the new baseline retention rate = 0.723

```spss
DATASET ACTIVATE DataSet1.
LOGISTIC REGRESSION VARIABLES RETENTIONSPRING
   /SELECT=TRAININGVARIABLE EQ 1
   /METHOD=ENTER CREDITSATTEMPTEDFALL DISTANCEEDENROLLMENT URM FEMALE ISLANDRURAL OUTOFSTATE LOWPERFORMHIGHSCHOOL ECEDMAJOR AGE19PLUS EDGOAL1 PELL PERCENTUNMETNEED STARUSAGE COMPASSREADING COMPASSANYMATH HIGHEST REMEDIALMATH REMEDIALENG
   /SAVE=PRED PGROUP COOK ZRESID
   /PRINT=GOODFIT
   /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.723).
```
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)

Goal 3: Estimate dropout risk

82.7% accuracy in identifying dropped students
Results from Trimmed Data

- Some false positives in Decile 1 for predicting retainers, but overall results suggest stability.
### Results from Trimmed Data

- Parameter estimates results. 10 variables significant at .05 level.

<table>
<thead>
<tr>
<th>Variables in the Equation</th>
<th>B</th>
<th>S.E.</th>
<th>Wald</th>
<th>df</th>
<th>Sig.</th>
<th>Exp(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CREDITSATTEMPTEDFALL</td>
<td>.149</td>
<td>.016</td>
<td>84.554</td>
<td>1</td>
<td>.000</td>
<td>1.160</td>
</tr>
<tr>
<td>DISTANCEEDENROLLMENT</td>
<td>-.375</td>
<td>.188</td>
<td>3.981</td>
<td>1</td>
<td>.046</td>
<td>.687</td>
</tr>
<tr>
<td>URM</td>
<td>-.657</td>
<td>.113</td>
<td>34.002</td>
<td>1</td>
<td>.000</td>
<td>.519</td>
</tr>
<tr>
<td>FEMALE</td>
<td>-.063</td>
<td>.119</td>
<td>.277</td>
<td>1</td>
<td>.599</td>
<td>.939</td>
</tr>
<tr>
<td>ISLANDRURAL</td>
<td>-.576</td>
<td>.200</td>
<td>8.263</td>
<td>1</td>
<td>.004</td>
<td>.562</td>
</tr>
<tr>
<td>OUTOFSTATE</td>
<td>-.580</td>
<td>.267</td>
<td>4.723</td>
<td>1</td>
<td>.030</td>
<td>.560</td>
</tr>
<tr>
<td>LOWPERFORMHIGHSCHOOL</td>
<td>-.494</td>
<td>.142</td>
<td>12.077</td>
<td>1</td>
<td>.001</td>
<td>.610</td>
</tr>
<tr>
<td>ECEDMAJOR</td>
<td>-.031</td>
<td>.269</td>
<td>.013</td>
<td>1</td>
<td>.909</td>
<td>.970</td>
</tr>
<tr>
<td>AGE19PLUS</td>
<td>-.277</td>
<td>.118</td>
<td>5.531</td>
<td>1</td>
<td>.019</td>
<td>.758</td>
</tr>
<tr>
<td>EDGOAL1</td>
<td>1.628</td>
<td>.260</td>
<td>39.276</td>
<td>1</td>
<td>.000</td>
<td>5.095</td>
</tr>
<tr>
<td>PELL</td>
<td>2.951</td>
<td>.269</td>
<td>120.543</td>
<td>1</td>
<td>.000</td>
<td>19.131</td>
</tr>
<tr>
<td>PERCENTUNMETNEED</td>
<td>-4.406</td>
<td>.425</td>
<td>107.412</td>
<td>1</td>
<td>.000</td>
<td>.012</td>
</tr>
<tr>
<td>STARUSAGE</td>
<td>.003</td>
<td>.020</td>
<td>.025</td>
<td>1</td>
<td>.876</td>
<td>1.003</td>
</tr>
<tr>
<td>COMPASSREADING</td>
<td>-.001</td>
<td>.002</td>
<td>.374</td>
<td>1</td>
<td>.541</td>
<td>.999</td>
</tr>
<tr>
<td>COMPASSANYMATHHIGHEST</td>
<td>.005</td>
<td>.003</td>
<td>1.964</td>
<td>1</td>
<td>.161</td>
<td>1.005</td>
</tr>
<tr>
<td>REMEDIALMATH</td>
<td>-.165</td>
<td>.127</td>
<td>1.675</td>
<td>1</td>
<td>.196</td>
<td>.848</td>
</tr>
<tr>
<td>REMEDIALENCPH</td>
<td>-.193</td>
<td>.136</td>
<td>2.013</td>
<td>1</td>
<td>.156</td>
<td>.825</td>
</tr>
<tr>
<td>Constant</td>
<td>-.239</td>
<td>.283</td>
<td>.714</td>
<td>1</td>
<td>.398</td>
<td>.787</td>
</tr>
</tbody>
</table>

a. Variable(s) entered on step 1: CREDITSATTEMPTEDFALL, DISTANCEEDENROLLMENT, URM, FEMALE, ISLANDRURAL, OUTOFSTATE, LOWPERFORMHIGHSCHOOL, ECEDMAJOR, AGE19PLUS, EDGOAL1, PELL, PERCENTUNMETNEED, STARUSAGE, COMPASSREADING, COMPASSANYMATHHIGHEST, REMEDIALMATH, REMEDIALENCPH.
Final Step: “Go Live” and score the incoming cohort

- Update filter in menu: Select Data, Select Cases, If condition... In the syntax, change “~=” to “=” for “COHORTYEAR...”

COHORTYEAR = 2015 | (COO_3 < 1 & ZRE_3 < 3 & ZRE_3 > -3)
Goal 3: Estimate dropout risk

CC Data: SPSS Menu Tasks

- Re-run regression syntax AGAIN with the last baseline retention rate = 0.723
- Change “TRAININGVARIABLE2 EQ 1” to score the 2015 cohort.

```spss
DATASET ACTIVATE DataSet1.
LOGISTIC REGRESSION VARIABLES RETENTIONSPRING
   /SELECT=TRAININGVARIABLE2 EQ 1
   /METHOD=ENTER CREDITSATTEMPTEDFALL 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(0.723).
```
Goal 3: Estimate dropout risk

CC Data: SPSS Output

As you see, we generated predicted probability scores (PRE_2) and Group Membership (PGR_2) for the 2015 cohort.
Last Step: Translate Dropout Risk into deciles for easier interpretation by academic support office

- Convert retention probability to dropout risk deciles (1 = highest, 10 = lowest)
- Filter Data for “2015” cohort only.
- 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)
Goal 4: Assist Student Support

Last Step: Filter 2015 cohort and create new dataset

Copy selected cases to a new dataset; give it a name.

COHORTYEAR = 2015
Last Step: Group into deciles using binning function:

- Transform, Visual Binning, Make 9 cutpoints on “PRE_2”, Label ‘Deciles’, check ‘reverse scale’
Now your new 2015 dataset has 10 deciles with an even distribution of low-to-high risk scores. Decile 10 is the highest risk.
Now that your data is ready, create a spreadsheet for delivery to your advisors/success coaches. Here is an example:

### Goal 4: Assist Student Support

<table>
<thead>
<tr>
<th>ID</th>
<th>LAST NAME</th>
<th>FIRST NAME</th>
<th>EMAIL</th>
<th>CURRENT CREDITS</th>
<th>RESIDENT</th>
<th>AP/CLEP</th>
<th>HS GPA</th>
<th>WORK ON CAMP</th>
<th>1ST YR EXP CLASS</th>
<th>% FIN NEED MET</th>
<th>STAR LOGINS</th>
<th>ADVISOR PREVIOUS CONTACT</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td></td>
<td></td>
<td></td>
<td>15</td>
<td>HI</td>
<td>6</td>
<td>3.80</td>
<td>Y</td>
<td>Y</td>
<td>77%</td>
<td>5</td>
<td>Y</td>
</tr>
<tr>
<td>002</td>
<td></td>
<td></td>
<td></td>
<td>14</td>
<td>HI</td>
<td>0</td>
<td>3.33</td>
<td>N</td>
<td>Y</td>
<td>63%</td>
<td>3</td>
<td>N</td>
</tr>
<tr>
<td>003</td>
<td></td>
<td></td>
<td></td>
<td>12</td>
<td>CA</td>
<td>6</td>
<td>3.00</td>
<td>N</td>
<td>N</td>
<td>45%</td>
<td>0</td>
<td>N</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>AGE</th>
<th>GENDER</th>
<th>ETHNICITY</th>
<th>COLLEGE</th>
<th>MAJOR</th>
<th>DEGREE</th>
<th>Ed Goal Specified</th>
<th>Relative Risk Value</th>
<th>Risk Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>18</td>
<td>F</td>
<td>CH</td>
<td>CA&amp;H</td>
<td>ART</td>
<td>BA</td>
<td>Yes</td>
<td>14.92</td>
<td>LOW</td>
</tr>
<tr>
<td>002</td>
<td>18</td>
<td>F</td>
<td>HW</td>
<td>CSS</td>
<td>SOC</td>
<td>BA</td>
<td>Yes</td>
<td>36.88</td>
<td>MEDIUM</td>
</tr>
<tr>
<td>003</td>
<td>18</td>
<td>M</td>
<td>UNDEC</td>
<td>UNDEC</td>
<td>UNDEC</td>
<td>UNDEC</td>
<td>No</td>
<td>89.18</td>
<td>HIGH</td>
</tr>
</tbody>
</table>
Progress on Implementation at Honolulu Community College (2014)

– Delivering student dropout risk scores to HCC’s Academic Success Center (via an Excel file).
– Training staff members on using the data.
– Academic advisors 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 models.
- IR and Advising staff from U. of Nevada-Reno travelled to UH in 2012 to share insights on implementing predictive analytics.
- Successful is using models for explanatory purposes initially. Now actively delivering the prediction scores to academic support personnel.
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
- Engagement with senior management
- 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