What is Predictive Analytics?

• Firstly, “Analytics” is the use of data, statistical analysis, and explanatory and predictive models to gain insights and act on complex issues.”

  -EDUCAUSE Center for Applied Research

• “Predictive Analytics” is simply predictive modeling, or creating a statistical model that best predicts the probability of an outcome.

• Examples include
  • Enrollment forecasting
  • Classifying/scoring at-risk students
  • Predicting Gainful Employment
Predictive Analytics...

...helps answer questions like:

• Which student variables are most useful for predicting outcomes like retention/degree completion/gainful employment?

• What is the “best” combination of variables to optimize predictions in the sample?

• How useful is this combination for identifying at-risk students?
Examples of Higher Ed Institutions Using Predictive Analytics

- University of Nevada at Reno
- Florida State University
- Missouri State University
- San Jose State University
- Virginia Tech University
- Arizona State University
- University of Texas at Austin
- California State University System

.....
Manoa Example

• Build and implement a predictive model for estimating freshmen retention outcomes.

• Campus context:
  • Approximately 2,000 first-time freshmen entering every fall at Manoa.
  • 79% retention rate (fall-to-fall)
    • versus 79% predicted retention rate
    • versus 82% peer average
4 Steps to Modeling Retention

1. Get Freshmen Data.
   (i.e. I used fall 2009, 2010, 2011 data to build a “training” data set.)

2. Build Model.
   - Hawaii
   - HS GPA
   - AP Credit
   - SAT R
   - SAT M
   - Ed Goals
   - FYE Class

   strongest relationships:
   - 222.426***
   - 62.285***
   - 13.730***
   - 11.714***
   - 7.881**
   - 5.068**
   - 3.797

3. Apply model’s parameter estimates to new data.
   (i.e. model validation, scoring)

4. Check the actual 2012 retention outcomes to see how well the model performed.
Relevant Previous Research


Methodology

• Binary Logistic Regression – model a student’s binary choice to return a second fall semester while controlling for a student’s characteristics.

• Types of covariates included in model:
  • Demographic
  • Pre-Collegiate Experience
  • Academic Experience
  • Campus Experience
  • Financial Aid
  • Achievement Milestones
  • Interaction Variables
Examples of Student Variables Analyzed

- **DEMOGRAPHICS**
  - Gender
  - Age
  - Ethnicity
  - Residency
  - Geographic Origin

- **PRE-COLLEGE**
  - High School GPA & Rank
  - SAT
  - AP CLEP
  - Educational Goals
  - Transfer GPA
  - # Transfer Credits

- **ACADEMIC**
  - Major
  - Credit Load
  - Credits Earned
  - First Term GPA
  - Distance Education
  - Dual Enrollment
  - High Failure Rate Courses
  - Courses Taken (including Math & English)

- **RETENTION / GRADUATION**

- **CAMPUS EXPERIENCE**
  - On Campus Employment
  - Housing
  - Student Life Activities
  - Athletics
  - STAR Usage
  - Average Class Size

- **FINANCIAL NEED**
  - Need Based Aid
  - Non-need Based Aid
  - Pell Grant
  - Work Study
  - % of Aid Met

- **MILESTONES**
  - Credits earned
  - Credits attempted
  - Credit Completion Ratio
  - Math/English Enrollment/Completion
  - Continuous Enrollment
  - Milestone metrics

- **INTERACTIONS**
  - Ethnicity by Geographic Origin
  - Employment by Housing
  - High School GPA by First Term GPA
  - Residency by Need Based Aid
  - Ratio of Successful Adds to Drops
Data Description

• Data sources
  • Matriculation system (Banner/ Banner ODS)

• Student cohorts
  • New full-time freshmen
  • Fall entry ’09, 10’, 11 for model dev. (training set, N=2,470)
  • Fall entry 2012 for model validation (holdout set, N=1,912)

• Data elements at start of first semester
  • Student demographics (residency)
  • Academic preparation (high school GPA/test score index)
  • Financial aid profile (unmet need)
  • Credits enrolled, campus housing (y/n), first year experience class (y/n), educational goals questionnaire (y/n)

• Data elements after start of first semester
  • Credits completed
  • End-of-semester GPA
  • 2nd educational goals questionnaire
Data Management Tasks

• Exploratory data analysis
  – Variable selection (bivariate regression on outcome variable)
  – Variable coding (continuous/categorical/dummy in logit model)
  – Missing data imputation, constant-$ conversion (fin. aid data)
  – Composite variable(s) – *not used in today’s model*
    • Acad prep index = (HSGPA*12.5)+(ACTM*.69)+(ACTE*.69)
  – Variables excluded: age, gender, ethnicity, college remediation, ACT/SAT test date

• Logistic regression model
  – Maximize model fit (-2LL test/score, pseudo R², HL sig.)
  – Create balanced sample in training dataset to optimize correct classification rate (CCR) for enrollees vs. non-enrollees (i.e. model sensitivity vs. specificity): all non-enrollees plus random sample of enrollees of ~ equal N)
Data Management Tasks

• Scoring of relative dropout/retention risk

\[ p = \frac{\exp(a + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4 \ldots)}{1 + \exp(a + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4 \ldots)} \]

Where:
- \( p \) = probability of enrollment/non-enrollment
- \( \exp \) = base of natural logarithms (~ 2.72)
- \( a \) = constant/intercept of the equation
- \( b \) = coefficient of predictors (parameter estimates)
## Example: Early Warning Model

### Balanced Model Parameter Estimates

<table>
<thead>
<tr>
<th>Step 1&lt;sup&gt;a&lt;/sup&gt;</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>hawaii</td>
<td>1.223</td>
<td>.106</td>
<td>133.474</td>
<td>1</td>
<td>.000</td>
<td>3.398</td>
</tr>
<tr>
<td>fye_class</td>
<td>.280</td>
<td>.116</td>
<td>5.793</td>
<td>1</td>
<td>.016</td>
<td>1.323</td>
</tr>
<tr>
<td>housing</td>
<td>.071</td>
<td>.103</td>
<td>.478</td>
<td>1</td>
<td>.489</td>
<td>1.073</td>
</tr>
<tr>
<td>ed_goals</td>
<td>.325</td>
<td>.133</td>
<td>5.988</td>
<td>1</td>
<td>.014</td>
<td>1.384</td>
</tr>
<tr>
<td>credits_attempted</td>
<td>.054</td>
<td>.026</td>
<td>4.322</td>
<td>1</td>
<td>.038</td>
<td>1.056</td>
</tr>
<tr>
<td>hs_gpa</td>
<td>.905</td>
<td>.125</td>
<td>52.395</td>
<td>1</td>
<td>.000</td>
<td>2.471</td>
</tr>
<tr>
<td>sat_m</td>
<td>.002</td>
<td>.001</td>
<td>7.881</td>
<td>1</td>
<td>.005</td>
<td>1.002</td>
</tr>
<tr>
<td>sat_r</td>
<td>-.002</td>
<td>.001</td>
<td>11.714</td>
<td>1</td>
<td>.001</td>
<td>.998</td>
</tr>
<tr>
<td>ap</td>
<td>.475</td>
<td>.128</td>
<td>13.739</td>
<td>1</td>
<td>.000</td>
<td>1.608</td>
</tr>
<tr>
<td>unmet_need</td>
<td>-.221</td>
<td>.119</td>
<td>3.425</td>
<td>1</td>
<td>.064</td>
<td>.802</td>
</tr>
<tr>
<td>Constant</td>
<td>-4.546</td>
<td>.632</td>
<td>51.784</td>
<td>1</td>
<td>.000</td>
<td>.011</td>
</tr>
</tbody>
</table>

<sup>a</sup> Single-step variable entry

Pseudo Rsquare = .17
The Wald test statistic was used to indicate strength of the variable instead of the coefficient, standardized beta. Because of the nature of the logistic regression, the coefficient is not easily interpretable to indicate strength.
## CCR Out-of-Sample

### Classification Table Comparison (Training vs Validation Dataset)\(^a\)

<table>
<thead>
<tr>
<th>Observed</th>
<th>RETAINED</th>
<th>Percentage Correct</th>
<th>RETAINED</th>
<th>Percentage Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>RETAINED</td>
<td>B</td>
<td></td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>849</td>
<td>317</td>
<td>72.8</td>
<td>266</td>
</tr>
<tr>
<td>Yes</td>
<td>492</td>
<td>677</td>
<td>57.9</td>
<td>483</td>
</tr>
<tr>
<td>Overall Percentage</td>
<td></td>
<td></td>
<td>65.4</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) The cut value is .550; Nagelkerke R-sq = .17
## Example: End-of-Semester Updated Model

### Balanced Model Parameter Estimates

<table>
<thead>
<tr>
<th></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><strong>Step 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hawaii</td>
<td>1.376</td>
<td>.119</td>
<td>133.951</td>
<td>1</td>
<td>.000</td>
<td>3.960</td>
</tr>
<tr>
<td>fye_class</td>
<td>.055</td>
<td>.128</td>
<td>0.184</td>
<td>1</td>
<td>.668</td>
<td>1.056</td>
</tr>
<tr>
<td>housing</td>
<td>.030</td>
<td>.116</td>
<td>.067</td>
<td>1</td>
<td>.796</td>
<td>1.030</td>
</tr>
<tr>
<td><strong>ed_goals2</strong></td>
<td>.627</td>
<td>.100</td>
<td>38.994</td>
<td>1</td>
<td>.000</td>
<td>1.872</td>
</tr>
<tr>
<td>credits_attempted</td>
<td>-.075</td>
<td>.035</td>
<td>4.647</td>
<td>1</td>
<td>.031</td>
<td>0.928</td>
</tr>
<tr>
<td>hs_gpa</td>
<td>-.069</td>
<td>.146</td>
<td>0.220</td>
<td>1</td>
<td>.639</td>
<td>0.934</td>
</tr>
<tr>
<td>sat_m</td>
<td>.002</td>
<td>.001</td>
<td>7.512</td>
<td>1</td>
<td>.006</td>
<td>1.002</td>
</tr>
<tr>
<td>sat_r</td>
<td>-.003</td>
<td>.001</td>
<td>15.361</td>
<td>1</td>
<td>.000</td>
<td>0.997</td>
</tr>
<tr>
<td>ap</td>
<td>.300</td>
<td>.142</td>
<td>4.440</td>
<td>1</td>
<td>.035</td>
<td>1.350</td>
</tr>
<tr>
<td>unmet_need</td>
<td>-.186</td>
<td>.132</td>
<td>1.988</td>
<td>1</td>
<td>.159</td>
<td>.830</td>
</tr>
<tr>
<td><strong>CREDITS EARNED</strong></td>
<td>.121</td>
<td>.026</td>
<td>22.342</td>
<td>1</td>
<td>.000</td>
<td>1.128</td>
</tr>
<tr>
<td><strong>EOS GPA</strong></td>
<td>.835</td>
<td>.087</td>
<td>92.213</td>
<td>1</td>
<td>.000</td>
<td>2.304</td>
</tr>
<tr>
<td>Constant</td>
<td>-3.072</td>
<td>.696</td>
<td>19.459</td>
<td>1</td>
<td>.000</td>
<td>.046</td>
</tr>
</tbody>
</table>

*a. Single-step variable entry*

Pseudo Rsquare = .37
# Gauging Value of EOS Data

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Early Warning</th>
<th></th>
<th>End-of-Semester</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Wald</td>
<td>Sig.</td>
<td>Wald</td>
<td>Sig.</td>
</tr>
<tr>
<td>hawaii</td>
<td>133.474</td>
<td>***</td>
<td>133.951</td>
<td>***</td>
</tr>
<tr>
<td>fye_class</td>
<td>5.793</td>
<td>*</td>
<td>0.184</td>
<td></td>
</tr>
<tr>
<td>housing</td>
<td>.478</td>
<td></td>
<td>.067</td>
<td></td>
</tr>
<tr>
<td>ed_goals</td>
<td>5.988</td>
<td>*</td>
<td>38.994</td>
<td>***</td>
</tr>
<tr>
<td>credits_attempted</td>
<td>4.322</td>
<td>*</td>
<td>4.647</td>
<td>*</td>
</tr>
<tr>
<td>hs_gpa</td>
<td>52.395</td>
<td>***</td>
<td>0.220</td>
<td></td>
</tr>
<tr>
<td>sat_m</td>
<td>7.881</td>
<td>**</td>
<td>7.512</td>
<td>**</td>
</tr>
<tr>
<td>sat_r</td>
<td>11.714</td>
<td>**</td>
<td>15.361</td>
<td>***</td>
</tr>
<tr>
<td>ap</td>
<td>13.739</td>
<td>**</td>
<td>4.440</td>
<td>*</td>
</tr>
<tr>
<td>unmet_need</td>
<td>3.425</td>
<td></td>
<td>1.988</td>
<td></td>
</tr>
<tr>
<td>CREDITS EARNED</td>
<td>22.342</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EOS GPA</td>
<td>92.213</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nagelkerke R-sq</td>
<td>0.17</td>
<td></td>
<td>0.37</td>
<td></td>
</tr>
<tr>
<td>CCR of At-Risk</td>
<td>71%</td>
<td></td>
<td>73%</td>
<td></td>
</tr>
</tbody>
</table>
Scoring Students

- Scoring of relative dropout/retention risk

\[ p = \frac{\exp(a + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4 \ldots)}{1 + \exp(a + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4 \ldots)} \]

Where:
- \( p \) = probability of enrollment/non-enrollment
- \( \exp \) = base of natural logarithms (\( \sim 2.72 \))
- \( a \) = constant/intercept of the equation
- \( b \) = coefficient of predictors (parameter estimates)
Example: John is at risk of dropping

- John:
  - is from the continental U.S. (0)
  - has a below average high school GPA (3.0)
  - is enrolled in 12 credits (12)
  - has a low % of financial need met (.65)
  - isn’t not working on campus (0)
  - isn’t enrolled in CAS 110 (0)
  - didn’t specify any educational goals in survey (0)

- Probability of Dropping: 0.75
### Sample Data for Advisors/Success Coaches

<table>
<thead>
<tr>
<th>UH 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>

<table>
<thead>
<tr>
<th>UH 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>15</td>
<td>HI</td>
<td>6</td>
<td>3.80</td>
<td>Y</td>
<td>Y</td>
<td>77%</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>Y</td>
</tr>
<tr>
<td>002</td>
<td>14</td>
<td>HI</td>
<td>0</td>
<td>3.33</td>
<td>N</td>
<td>Y</td>
<td>63%</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>N</td>
</tr>
<tr>
<td>003</td>
<td>12</td>
<td>CA</td>
<td>6</td>
<td>3.00</td>
<td>N</td>
<td>N</td>
<td>45%</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>N</td>
</tr>
</tbody>
</table>
Classification Accuracy: Unbalanced vs Balanced Model

<table>
<thead>
<tr>
<th>Observed</th>
<th>Predicted</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Unbalanced Dataset</td>
<td>Balanced Dataset</td>
</tr>
<tr>
<td></td>
<td>RETAINED</td>
<td>Percentage Correct</td>
<td>RETAINED</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>RETAINED</td>
<td>No</td>
<td>47</td>
<td>329</td>
</tr>
<tr>
<td>RETAINED</td>
<td>Yes</td>
<td>42</td>
<td>1267</td>
</tr>
<tr>
<td>Overall Percentage</td>
<td>78.0</td>
<td></td>
<td>64.8</td>
</tr>
</tbody>
</table>

a. The cut value is .550
Model Degeneracy

• Predictive model classifies all samples in dominant category
  • The more skewed the data set is, the higher the overall classification rate is...
  • BUT MODEL DOESN’T HELP!
    • Performs poorly at classifying dropped students which are the very students we want to identify.
Model Performance

Distribution of Estimates by Decile Group
Fall 2012 Freshmen

Average Retention Rate

Risk Values Low to High
(≈170 students in each group)
Model Performance

Distribution of Estimates by Decile Group
Fall 2012 Freshmen (End-of-Semester Model)

- Risk Values Low to High
  - Approximately 170 students in each group

Average Retention Rate
- 25%
- 63%
- 70%
- 74%
- 82%
- 90%
- 91%
- 94%
- 95%
- 96%
# Classification Accuracy: Early Warning vs End-of-Semester Model

## Classification Table Comparison (Validation Datasets)\(^{a}\)

<table>
<thead>
<tr>
<th>Observed</th>
<th>Predicted</th>
<th>Balanced Early Warning Model</th>
<th>Balanced End-of-Semester Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RETAINED</td>
<td>Percentage Correct</td>
<td>RETAINED</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>RETAINED</td>
<td>No</td>
<td>266</td>
<td>110</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>483</td>
<td>826</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td></td>
<td>64.8</td>
</tr>
</tbody>
</table>

\(^{a}\) The cut value is .550
Model Performance

Distribution of Estimates by Decile Group
Fall 2012 Freshmen (Early Warning vs EOS)

- Early Warning Model
- EOS Model

Average Retention Rate

Risk Values Low to High
(≈170 students in each group)
Impact on Campus

- **422** (of 1,912) freshmen from 2012 dropped out in first year.
- Retaining **40** students would have improved Mānoa’s freshmen retention rate from 78% to **80%**.
- Additional Revenue from Tuition and Fees ≈ **$450,224** (for 28 HI, 12 WUE, excludes out-of-state.).
Takeaways on Implementation

• Early-alert data key
• Identify results that are actionable.
• Support for student advising and first-year experience staff
  • Involve colleges and departments.
• Use prediction data as component part of student dropout-risk assessment plan.
• Ways to increase awareness of retention and graduation rates:
  • Campaigns
  • Showing impact on the bottom line
Challenges to Implementation

- Culture change
- Wary of misuse of data
- Questions about data used in model to generate risk scores
- Students’ rights to access risk scores
- 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 predictive analytics and keep abreast of changes in analytical and data reporting tools

• Using prediction data
  • 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:
  • E.g. “Analytics in Higher Education” by J. Bichsel, Educause, 2012
Mahalo

John Stanley
Institutional Analyst
University of Hawaii at Manoa

Questions: jstanley@hawaii.edu

Link to this presentation: