Predicting Student Persistence Using Data Mining and Statistical Analysis Methods

Koji Fujiwara
Office of Institutional Research and Effectiveness
Bemidji State University & Northwest Technical College

December 18, 2012

* The original version was presented at the AIRUM 2012 conference with Doug Olney.
NTC Enrollment

- Northwest Technical College enrollment grew very quickly between 2004 and 2009, but now it has been falling off.
  - Fall 2004 Headcount = 794
  - Fall 2009 Headcount = 1,604
  - Fall 2012 Headcount = 1,168
Fall Semester Headcount and FTE

FTE = Undergraduate Semester Hours / 15.
## NTC Enrollment Report

### Headcount and FTE for Term by Campus - Northwest Technical College

**Spring 2013**

<table>
<thead>
<tr>
<th>Campus</th>
<th>Unduplicated Headcount</th>
<th>Duplicated Headcount</th>
<th>FTE</th>
<th>Credits</th>
<th>FYE</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTC</td>
<td>576</td>
<td>576</td>
<td>366.13</td>
<td>5492</td>
<td>183.07</td>
</tr>
<tr>
<td>NTC Distance</td>
<td>696</td>
<td>895</td>
<td>309.07</td>
<td>4636</td>
<td>154.53</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1272</strong></td>
<td><strong>1471</strong></td>
<td><strong>675.19</strong></td>
<td><strong>10128</strong></td>
<td><strong>337.60</strong></td>
</tr>
</tbody>
</table>

*Unadjusted numbers*

Date: 12/17/2012  Days to beginning of term: 27.5693
Retention & Grad Rates – IPEDS definition

- Fall 2007 (166): 78%
- Fall 2008 (114): 74%
- Fall 2009 (170): 81%
- Fall 2010 (141): 71%
- Fall 2011 (104): 76%

- Fall to Spring:
  - Fall 2007: 40%
  - Fall 2008: 37%
  - Fall 2009: 48%
  - Fall 2010: 35%
  - Fall 2011: 39%

- 150% Grad Rates:
  - Fall 2007: 32%
  - Fall 2008: 29%

Cohort (Size)
Retention/Success Study at NTC (& other colleges)

- A study was conducted last year to identify the factors that affected student persistence (success, to be defined later).
  - Factors: Credit Completion Rate & GPA

- The first semester credit completion rate was analyzed using a
  - 2 (success status: successful vs. unsuccessful) ×
  - 2 (college ready status: ready vs. underprepared) ×
  - 2 (late registration status: regular vs. late) ANOVA.
What did we learn? (AIRUM 2011 / AIR 2012)

We need to focus on these students.

We need to focus on these students as well.
More Proactive Retention/Success Efforts

- How can an IR office contribute to this project?

- We looked at eight data mining and two statistical models to find the best method to develop a predictive model for student persistence (success).
Proposed Retention/Success Plan

[Summer]
Retention/Success Study & Develop a Model

[Spring]
Interventions

[Fall]
Run the model
Data and Variables
Target Variable

- Target Variable: Second Fall Success
  - Define “Second Fall Success = Success” if

At the beginning of second Fall
- Students came back to NTC (retained) or
- Students transferred out to another institution (transferred) or
- Students completed the program (graduated)

- Why did we use “success” and not retention?
Models

- Decision Tree
- Naive Bayes Classifier
- K-Nearest Neighbor Classifier
- Neural Network
- Support Vector Machine
- Bagging
- Boosting
- Random Forest
- Probit Regression
- Logistic Regression
DATA

- Combined Fall 2008 – 2010 first-year, full-time, degree seeking students data
  - There was no difference in the success status and cohort year (Chi-square test, \( p = 0.72 \)).

- In this study, we included first-time regular students and transfer students whose attempted transferred credits \( \leq 6 \).
Predictor Variables

• Age

• **Gender** (Female/Male)

• **College Ready Status** (Underprepared/Ready)
  • if a student needed to take at least one college readiness course, then the student was categorized as “Underprepared”.

• **Pell Status** (No/Yes)

• **Late Registration Status** (before or on/after **August 1st**)

• **Underrepresented Status** (No/Yes)

• **Admission Status** (Regular/Transfer)
Predictor Variables

• # Attempted Credit Hours
• Enrolled Semester GPA
• Enrolled Semester Credit Completion Rate

  ▪ # complete observations = 622
How to Compare the Models?
How to Estimate a Model Performance?

- Simple Cross-Validation

Original Data
N = 622

Training (for modeling)
N1 = 311

Testing (for validation)
N2 = 311
How to Compare the Results?

- We propose to use these three indicators:
  - **Overall Classification Rate**
  - **Sensitivity**
    - Prob. that the model classified as “Successful” when given to a group of students who were actually “Successful”
  - **Specificity**
    - Prob. that the model classified as “Unsuccessful” when given to a group of students who were actually “Unsuccessful”
Results and Conclusion
## Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Classification</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>81.7%</td>
<td>89.8%</td>
<td>67.8%</td>
</tr>
<tr>
<td>Probit Regression</td>
<td>81.7%</td>
<td>90.8%</td>
<td>66.1%</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>81.4%</td>
<td>89.3%</td>
<td>67.8%</td>
</tr>
<tr>
<td>Naive Bayes Classifier</td>
<td>82.3%</td>
<td>90.3%</td>
<td>68.7%</td>
</tr>
<tr>
<td>K-Nearest Neighbor (5)</td>
<td>82.3%</td>
<td>90.3%</td>
<td>68.7%</td>
</tr>
<tr>
<td><strong>Neural Network</strong></td>
<td><strong>81.4%</strong></td>
<td><strong>86.7%</strong></td>
<td><strong>72.2%</strong></td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>82.6%</td>
<td>92.9%</td>
<td>65.2%</td>
</tr>
<tr>
<td><strong>Bagging</strong></td>
<td><strong>82.3%</strong></td>
<td><strong>89.8%</strong></td>
<td><strong>69.6%</strong></td>
</tr>
<tr>
<td>Boosting</td>
<td>78.9%</td>
<td>87.8%</td>
<td>63.5%</td>
</tr>
<tr>
<td>Random Forest</td>
<td>83.0%</td>
<td>92.9%</td>
<td>65.2%</td>
</tr>
</tbody>
</table>
Conclusion and Future Research

- We are going to use a **Neural Network** model and a **bagging** model.

- **December**
  - These models will be applied to the actual data.
  - The list will be sent to the Student Services.

- **January**
  - The Student Services will start contacting to the students.
  - Interventions (what types?)

- **Next Summer and Fall**
  - Summer – Conduct Retention Study
  - Fall – Evaluate the models
Conclusion and Future Research (contd.)

- **Current Model**
  - Need Students First Semester Academic Performance Data
  - Interventions: Spring Semester

- **Ideal Model**
  - Can run before the Fall semester begins
    - Need other predictors: i.e.) FA data?
  - Interventions: Fall Semester