Using Data Mining Techniques to Predict Student Development and Retention

Morgan C. Wang
Department of Statistics and Actuarial Science
University of Central Florida
Presenters

- University of Central Florida – Department of Statistics
  - Morgan C. Wang, Professor of Statistics
Agenda

- Background
- UCF History and Approach
- Project Description
  - Data
  - Model Building
  - Findings
- Conclusions
- Further Research
Retention

- Institution’s capacity to engage faculty and administrators in a collaborative effort to construct educational settings that engage all students in learning.

  Tinto
Retention

- Establishing a meaningful early connection and commitment to the institution that positively influences continued progress towards the degree from one year to the next.

Ehasz
The Most Successful Retention Programs:

- Are highly structured
- Are interlocked with other programs/services
- Rely on extended, intensive student contact
- Are based on strategy of engagement
- Place special emphasis on staff quality
- Focus on affective as well as cognitive needs
- Track and monitor level of student satisfaction

Noel-Levitz
Retention Is Negatively Affected By:

- Unclear career goals
- Uncertainty about major
- Lack of academic challenge
- Transition/adjustment problems
- Limited/unrealistic expectations
- Lack of engagement
- Low level of integration
Tinto Model

Initial Goal Commitment → Academic Integration → Persistence

Student Entry Characteristics → Social Integration

Initial Institutional Commitment

Subsequent Goal Commitment

Subsequent Institutional Commitment

Academic Challenges

- Low High School GPA
- Low High School senior grades
- High School senior courses
- Test scores and sub-groups
- Key courses
- Key majors
- Probation
- Rigor
- Uncertainty
Integration Challenges

- Ethnicity
- Residency
- Institution preference
- Family background
- Emotional support
- Attitude toward education
- Self reliance
- Run-around
- Negativity
- Weak campus community
- Unwelcome environment
Involvement Challenges

- Off-campus residence
- Off-campus job
- Limited co-curricular program
- Self-responsibility
- Freedom
University of Central Florida Fast Facts

- **LOCATION:** 13 miles east of downtown Orlando
- **CONSTRUCTION BEGAN:** January, 1967
- **DATE OF FIRST CLASSES:** October, 1968
- **ORIGINAL ENROLLMENT:** 1,948 students
- **FALL 2004 ENROLLMENT:** 42,837
- **Fall 2004 FTICs Enrolled:** 4,092
- **Summer 2004 FTICs Enrolled Fall 2004:** 1866
- **Average SAT Total:** 1186
- **Average H.S. GPA:** 3.84
### University of Central Florida
First Year Retention Rates and Key Events

<table>
<thead>
<tr>
<th>Year</th>
<th>Total Enrollment</th>
<th>Fall FTIC</th>
<th>HS GPA</th>
<th>SAT</th>
<th>% Residence Halls</th>
<th>Retention</th>
</tr>
</thead>
<tbody>
<tr>
<td>1994</td>
<td>25,363</td>
<td>2,089</td>
<td>3.2</td>
<td>1085</td>
<td>32%</td>
<td>70%</td>
</tr>
</tbody>
</table>

**Enrollment and Academic Services**

<table>
<thead>
<tr>
<th>Year</th>
<th>Total Enrollment</th>
<th>Fall FTIC</th>
<th>HS GPA</th>
<th>SAT</th>
<th>% Residence Halls</th>
<th>Retention</th>
</tr>
</thead>
<tbody>
<tr>
<td>1998</td>
<td>30,000</td>
<td>3,127</td>
<td>3.5</td>
<td>1129</td>
<td>40%</td>
<td>75%</td>
</tr>
</tbody>
</table>

**First Year Advising**

**Student Development and Enrollment Services**

**Enhanced Funding**

<table>
<thead>
<tr>
<th>Year</th>
<th>Total Enrollment</th>
<th>Fall FTIC</th>
<th>HS GPA</th>
<th>SAT</th>
<th>% Residence Halls</th>
<th>Retention</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>36,013</td>
<td>3,759</td>
<td>3.66</td>
<td>1152</td>
<td>65%</td>
<td>81%</td>
</tr>
<tr>
<td>2002</td>
<td>38,795</td>
<td>3,922</td>
<td>3.74</td>
<td>1167</td>
<td>67%</td>
<td>84%</td>
</tr>
</tbody>
</table>

**Majors Fair**

<table>
<thead>
<tr>
<th>Year</th>
<th>Total Enrollment</th>
<th>Fall FTIC</th>
<th>HS GPA</th>
<th>SAT</th>
<th>% Residence Halls</th>
<th>Retention</th>
</tr>
</thead>
<tbody>
<tr>
<td>2003</td>
<td>42,000</td>
<td>4,134</td>
<td>3.81</td>
<td>1172</td>
<td>68%</td>
<td>84% projected</td>
</tr>
</tbody>
</table>

**LINK**

**Bus Stop Advising**

**Golden Opportunities**
FTIC Retention Success – 2001

- National Merit finalists
- Burnett Honors College
- LEAD Scholars Program
- Greek membership
- On-campus housing
  - Sumter Hall
  - Academic Village
- Bright Future recipients
FTIC Retention Challenges – 2001

- Out-of-State residents
- Ethnicity
- Off-campus residents
- Selected housing unit residents
- Program of study
Current Retention Efforts

- At the present time, UCF retention studies have been limited to simple year-by-year demographic summaries which do not fully explain student progression patterns or trends.

- Student Development and Enrollment Services has been gathering data on program attendance, attitudes, and opinions from various sources: Housing, Financial Assistance, Recreation and Wellness Center, Greek Organizations, Academic Advising, and Assessment.

- We believe that student behavior can be explained with a more sophisticated method of data analysis.
Proposed Approach – Data Mining

- No additional data collection needed
- Treat each student as an individual
- Prevent student from dropping out instead of documenting student who already dropped out
- Rules found must be very easy to guide the administration to develop prevention programs to target the at-risk students
Data Mining is NOT a Crystal Ball
It is a Prŏcess (or Prŏcess)
Data

Data Sources:
- CIRP (Cooperative Institutional Research Project) Survey in 2002
- High School data from Academic Year 2001-2002

Number of Students: 3829

Number of Variables: 285
- 23 numerical variables: SAT_Verb, SAT_Math, Income
- 175 nominal variables: Ethnic, Student_status, Goal
- 36 ordinal variables: HSGPA, Age, …
- 47 binary variables: Gender, Full_status, Non_retain…
- 4 derive variables: Flag1 – Flag4

Study Target: Student who has lower chance to be retained
- Retained after freshmen year: 3149 (82.24%)
- Not Retained after freshmen year: 680 (17.76%)
Data Problems

- Many variable with missing values: More than 60% observations have one or more variables that have missing values
  - ACT_Composite_Score: 50%
  - Highest_Degree_Plan: 39%
  - Finance_AID_From_Other: 53%
  - Finance_AID.Must_Repay: 31%

- Variables with different scales:
  - “Text” Format
  - “Numerical” Format

- Nominal variable with many levels
Fix Data Problems

- Missing Value Imputation
- Categorical variables with many categories
- Reduce the number of Variables
- etc.
Continuous variable imputation – Nearest Neighbor Algorithm

1. Standardize all variables without missing value, \( y^* = (y - \bar{y}) / \text{std} \)

2. Select best variable \( V \) to impute

3. Separate observations into:
   - \( X \): obs with missing \( V \)
   - \( Y \): obs without missing \( V \)

4. Select one Obs \( j \) in \( X \), compute distance with all Obs in \( Y \), \( \text{Dist}(i) = \sqrt{\text{sum}|X_{jv} - Y_{iv}|} \)

5. Replace the missing \( V \) of Obs \( j \) with the mean of 10 nearest neighbor

6. Move Obs \( j \) to \( Y \), loop until \( X \) is empty

7. Standardize variable \( V \), loop until no missing
Students with lower high school grade have higher chance of not being retained after their freshmen year.
Entering freshmen with a higher level “Honor” status have higher chance of being retained.
Students whose High School GPA is below 3.25 have higher risk of not being retained after their freshmen year.
Data Exploration—High School GPA

The high school GPA for students who are not retained after their freshmen year is on the average 0.24 below their counterpart. Besides, from T test, it shows that comparing retained students to not retained students, the Mean of High School GPA is significantly different.
High school English is the most important subject for students to succeed in college.
The high school English GPA for students who are not retained after their freshmen year is on the average 0.23 below their counterpart. Besides, from T test, it shows that comparing retained students to not retained students, the Mean of English Unit GPA is significantly different.
Students have a higher retention rate if they decide to live in the dormitory.
Data Exploration—Student Residency

<table>
<thead>
<tr>
<th>Student Residency</th>
<th>Retained</th>
<th>FREQ.</th>
<th>PCT.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Florida</td>
<td>Yes</td>
<td>2946</td>
<td>82.6%</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>620</td>
<td>17.4%</td>
</tr>
<tr>
<td>Other States</td>
<td>Yes</td>
<td>203</td>
<td>77.2%</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>60</td>
<td>22.8%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Indicator of Student Residency</th>
<th>Retained</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student comes from Florida</td>
<td>Yes</td>
<td>2946</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>620</td>
</tr>
<tr>
<td>Student comes from other States</td>
<td>Yes</td>
<td>203</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>60</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>3149</td>
</tr>
</tbody>
</table>

Obviously, most students at UCF come from Florida, and they have the higher chance of being retained.
The more Advanced Placement Exams taken, the higher the chance of being retained.
Model Building

- **Data Partition:**
  - 70% Training
  - 30% Validation

- Models are constructed using training data sets and evaluate model performance using validation data sets, and using other data sources as testing data sets.

- Several modeling techniques are used, e.g., **logistic regression**, **neural network**, **decision trees**, and **clustering**
Predictive Model
– Decision tree models (Enterprise Miner)

Process Flow Diagram
## Entropy Decision Tree Summary

### Model Summary

<table>
<thead>
<tr>
<th>Priors</th>
<th>Source</th>
<th>Stat</th>
<th>Churn</th>
<th>== &gt; 1</th>
<th>== &gt; 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Applied</td>
<td>Train</td>
<td>N</td>
<td>1</td>
<td>451</td>
<td>47</td>
</tr>
<tr>
<td>Not Applied</td>
<td>Train</td>
<td>N</td>
<td>0</td>
<td>1568</td>
<td>614</td>
</tr>
<tr>
<td>Not Applied</td>
<td>Train</td>
<td>N</td>
<td>+</td>
<td>2019</td>
<td>661</td>
</tr>
<tr>
<td>Not Applied</td>
<td>Train</td>
<td>Row%</td>
<td>1</td>
<td>91</td>
<td>9</td>
</tr>
<tr>
<td>Not Applied</td>
<td>Train</td>
<td>Row%</td>
<td>0</td>
<td>72</td>
<td>28</td>
</tr>
<tr>
<td>Not Applied</td>
<td>Train</td>
<td>Row%</td>
<td>+</td>
<td>75</td>
<td>25</td>
</tr>
<tr>
<td>Not Applied</td>
<td>Train</td>
<td>Col%</td>
<td>1</td>
<td>22</td>
<td>7</td>
</tr>
<tr>
<td>Not Applied</td>
<td>Train</td>
<td>Col%</td>
<td>0</td>
<td>78</td>
<td>93</td>
</tr>
<tr>
<td>Not Applied</td>
<td>Train</td>
<td>Col%</td>
<td>+</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Not Applied</td>
<td>Train</td>
<td>%</td>
<td>1</td>
<td>17</td>
<td>2</td>
</tr>
<tr>
<td>Not Applied</td>
<td>Train</td>
<td>%</td>
<td>0</td>
<td>59</td>
<td>23</td>
</tr>
<tr>
<td>Not Applied</td>
<td>Train</td>
<td>%</td>
<td>+</td>
<td>75</td>
<td>25</td>
</tr>
<tr>
<td>Not Applied</td>
<td>Train</td>
<td>%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Leaves

<table>
<thead>
<tr>
<th>Leaves</th>
<th>Training</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>0.5699</td>
<td>0.5747</td>
</tr>
<tr>
<td>9</td>
<td>0.5699</td>
<td>0.5747</td>
</tr>
<tr>
<td>10</td>
<td>0.5699</td>
<td>0.5747</td>
</tr>
<tr>
<td>11</td>
<td>0.5699</td>
<td>0.5747</td>
</tr>
<tr>
<td>12</td>
<td>0.6031</td>
<td>0.5749</td>
</tr>
<tr>
<td>13</td>
<td>0.6031</td>
<td>0.5749</td>
</tr>
<tr>
<td>14</td>
<td>0.6031</td>
<td>0.5749</td>
</tr>
<tr>
<td>15</td>
<td>0.6031</td>
<td>0.5749</td>
</tr>
<tr>
<td>16</td>
<td>0.6031</td>
<td>0.5749</td>
</tr>
<tr>
<td>17</td>
<td>0.6031</td>
<td>0.5749</td>
</tr>
</tbody>
</table>

### Number of Leaves

- Number of leaves: 0 to 30
- Training accuracy: 0.45 to 0.70

---

MWSUG

UCF
Decision Tree from High School Data cont’d.

- **HSGPA1**
  - >= 3.25
    - Nation Merit and Hon
  - < 4.15
    - SAT_TOT
      - >= 1235
        - Green: 48.8% 1, 61.9% 1, 58.3% 1
        - Red: 51.2% 0, 38.1% 0, 41.7% 0
      - < 25.5
        - Green: 1, 16.6% 1, 27.5% 1
        - Red: 0, 83.4% 0, 72.5% 0
    - HSGPA1
      - >= 4.15
        - Green: 1, 43.8% 1, 35.2% 1
        - Red: 0, 56.2% 0, 64.8% 0
      - < 25.5
        - Green: 1, 51% 1, 16% 1
        - Red: 1, 13% 1, 9% 1

MWSUG
### Important variables from High School data

<table>
<thead>
<tr>
<th>Name</th>
<th>Importance</th>
<th>Role</th>
<th>Rules</th>
<th>Variable Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>HSGPA1</td>
<td>1.0000</td>
<td>input</td>
<td>2</td>
<td>High School GPA</td>
</tr>
<tr>
<td>NATIONMerit_AND_HO</td>
<td>0.5639</td>
<td>input</td>
<td>1</td>
<td>Honors Indicator</td>
</tr>
<tr>
<td>ACT_E</td>
<td>0.4504</td>
<td>input</td>
<td>2</td>
<td>ACT English Score</td>
</tr>
<tr>
<td>SAT_TOT</td>
<td>0.4191</td>
<td>input</td>
<td>2</td>
<td>SAT Total</td>
</tr>
<tr>
<td>HSU_ENG</td>
<td>0.4148</td>
<td>input</td>
<td>2</td>
<td>High School Unit English GPA</td>
</tr>
<tr>
<td>HSU_MATH</td>
<td>0.3676</td>
<td>input</td>
<td>2</td>
<td>High School Unit Math GPA</td>
</tr>
<tr>
<td>ACT_COMP_TOT</td>
<td>0.3103</td>
<td>input</td>
<td>1</td>
<td>ACT comprehensive Total</td>
</tr>
<tr>
<td>SAT_M</td>
<td>0.2735</td>
<td>input</td>
<td>1</td>
<td>SAT Math</td>
</tr>
<tr>
<td>SAT_V</td>
<td>0.2587</td>
<td>input</td>
<td>1</td>
<td>SAT Verb</td>
</tr>
<tr>
<td>ACTCOMP</td>
<td>0.0000</td>
<td>rejected</td>
<td>0</td>
<td>ACT Composite</td>
</tr>
<tr>
<td>FLAG1</td>
<td>0.0000</td>
<td>rejected</td>
<td>0</td>
<td>Indicator From Talented_20_and Honor College</td>
</tr>
<tr>
<td>CITIZEN</td>
<td>0.0000</td>
<td>rejected</td>
<td>0</td>
<td>U. S. Citizen</td>
</tr>
<tr>
<td>FLAG2</td>
<td>0.0000</td>
<td>rejected</td>
<td>0</td>
<td>Indicator of SAT difference</td>
</tr>
<tr>
<td>HSU_NS</td>
<td>0.0000</td>
<td>rejected</td>
<td>0</td>
<td>High School Natural Science Unit GPA</td>
</tr>
<tr>
<td>SEX</td>
<td>0.0000</td>
<td>rejected</td>
<td>0</td>
<td>Student’s Gender</td>
</tr>
<tr>
<td>FULLSTAT</td>
<td>0.0000</td>
<td>rejected</td>
<td>0</td>
<td>Full/Part Time Status</td>
</tr>
<tr>
<td>ACT_M</td>
<td>0.0000</td>
<td>rejected</td>
<td>0</td>
<td>ACT Math Score</td>
</tr>
<tr>
<td>FT_PT</td>
<td>0.0000</td>
<td>rejected</td>
<td>0</td>
<td>Part time / Full Time</td>
</tr>
<tr>
<td>AGE</td>
<td>0.0000</td>
<td>rejected</td>
<td>0</td>
<td>Student’s Age</td>
</tr>
</tbody>
</table>
Decision Tree from Overall Data

HSGPA1

< 3.25

N 745 299
Decision 0
1 0.65370 0.62386
0 0.0193 0.0642

>= 3.25

N 1935 850
Decision 1
1 0.4400 0.45552
0 0.3380 0.31522

Planned Residence for Fall 2001

DORMITORY...

N 952 453
Decision 0
1 0.36730 0.41113
0 0.4489 0.3833

PRIVATE HOME...

N 983 397
Decision 1
1 0.51190 0.50825
0 0.23200 0.23762

HSU_SS

< 395

N 467 247
Decision 1
1 0.45530 0.54858
0 0.31690 0.17712

>= 3.95

N 485 206
Decision 0
1 0.28260 0.24569
0 0.57590 0.63147

Attended Religious Services

... OCCASIONALLY

N 660 254
Decision 1
1 0.57090 0.54778
0 0.14360 0.17834

FREQ

N 0 0
Decision 0
1 0 0
0 0 0
Decision Tree from Overall Data cont’d.

HSGPA1

>= 3.25

Planned Residence for Fall 2001

DORMITORY...

HSU_SS

>= 3.95

ACT_E

<20.5

N: 47 26

Decision 0: 0.24358

1: 0.31302

>= 20.5

N: 438 180

Decision 0: 0.63463

1: 0.53010

PRIVATE HOME...

Attended Religious Services

.. OCCASIONALLY

N: 660 254

Decision 1: 0.57090

1: 0.54778

0: 0.14360

FREQUENTLY

Student’s Probable Major (ac

N: 235 101

Decision 1: 0.48170

1: 0.43915

0: 0.27740

N: 0

Decision 1: 0

MWSUG

UCF
## Important variables from Overall data

<table>
<thead>
<tr>
<th>Name</th>
<th>Importance</th>
<th>Role</th>
<th>Rules</th>
<th>Variable Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>HSGPA1</td>
<td>1.0000</td>
<td>input</td>
<td>1</td>
<td>High School GPA</td>
</tr>
<tr>
<td>LIVEPLAN</td>
<td>0.6440</td>
<td>input</td>
<td>1</td>
<td>Planned Residence for Fall 2002</td>
</tr>
<tr>
<td>HSU_SS</td>
<td>0.5393</td>
<td>input</td>
<td>1</td>
<td>High School Social Science Unit GPA</td>
</tr>
<tr>
<td>ACT0201</td>
<td>0.5348</td>
<td>input</td>
<td>1</td>
<td>Attended Religious Services</td>
</tr>
<tr>
<td>MAJOR02A</td>
<td>0.5345</td>
<td>input</td>
<td>1</td>
<td>Student’s Probable Major</td>
</tr>
<tr>
<td>ACT0223</td>
<td>0.4594</td>
<td>input</td>
<td>1</td>
<td>Communicated Via E-mail</td>
</tr>
<tr>
<td>AID1</td>
<td>0.4547</td>
<td>input</td>
<td>1</td>
<td>Family Resource</td>
</tr>
<tr>
<td>GOAL0221</td>
<td>0.4383</td>
<td>input</td>
<td>1</td>
<td>Understand of Other Countries / Cults</td>
</tr>
<tr>
<td>ACT_E</td>
<td>0.4143</td>
<td>input</td>
<td>1</td>
<td>ACT English Test Score</td>
</tr>
<tr>
<td>CRED2</td>
<td>0.0000</td>
<td>rejected</td>
<td>0</td>
<td>4 Year College or University</td>
</tr>
<tr>
<td>DISAB2</td>
<td>0.0000</td>
<td>rejected</td>
<td>0</td>
<td>Disability Hearing</td>
</tr>
<tr>
<td>CRED1</td>
<td>0.0000</td>
<td>rejected</td>
<td>0</td>
<td>2 Year College</td>
</tr>
<tr>
<td>DISAB6</td>
<td>0.0000</td>
<td>rejected</td>
<td>0</td>
<td>Disability Health-related</td>
</tr>
<tr>
<td>FUTACT05</td>
<td>0.0000</td>
<td>rejected</td>
<td>0</td>
<td>Work Full-time</td>
</tr>
<tr>
<td>CITIZEN</td>
<td>0.0000</td>
<td>rejected</td>
<td>0</td>
<td>U. S. Citizen</td>
</tr>
<tr>
<td>FATHEDUC</td>
<td>0.0000</td>
<td>rejected</td>
<td>0</td>
<td>Father’s Education</td>
</tr>
<tr>
<td>FUTACT04</td>
<td>0.0000</td>
<td>rejected</td>
<td>0</td>
<td>Get Job to Pan Expenses</td>
</tr>
<tr>
<td>FLAG2</td>
<td>0.0000</td>
<td>rejected</td>
<td>0</td>
<td>Indicator of SAT Difference</td>
</tr>
<tr>
<td>FLAG3</td>
<td>0.0000</td>
<td>rejected</td>
<td>0</td>
<td>ACT is missing value</td>
</tr>
<tr>
<td>FLAG4</td>
<td>0.0000</td>
<td>rejected</td>
<td>0</td>
<td>High School units English or SS or NS or Math is 0</td>
</tr>
</tbody>
</table>
Rule #1 : If...

High School GPA is less than 3.25

Then...

The probability of student retained is 71.53%

And

The probability of student not retained is 28.47%
Rule #2: If...

SAT Total score is greater than 1235

And

High School GPA is between 3.25 and 4.15

And

National Merit and Honor Indicator equals “QH”

Then...

The probability of student retained = 74.71%

And

The probability of student not retained = 25.29%
Rule #3: If...

SAT Total score is greater than 995
And
High School Unit SS GPA is greater than 4.05
And
SAT Math score is greater than 455
And
High School Unit English GPA is greater than 4.75
Then...

The probability of student being retained is 82.92%
And
The probability of student not retained is 17.08%
## Summary of Rules

<table>
<thead>
<tr>
<th>Rule</th>
<th>Students Not retained</th>
<th>Total # of Students in this rule</th>
<th>Not retained Hit Rate % in this rule</th>
<th>Not retained Hit Rate % in all data</th>
<th>Odds Ratio</th>
<th>95% Confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule 1</td>
<td>240</td>
<td>849</td>
<td>28.27%</td>
<td>35.29%</td>
<td>2.54</td>
<td>(1.26, 24)</td>
</tr>
<tr>
<td>Rule 2</td>
<td>65</td>
<td>257</td>
<td>25.3%</td>
<td>9.56%</td>
<td>2.95</td>
<td>(1.3, 29)</td>
</tr>
<tr>
<td>Rule 3</td>
<td>267</td>
<td>1563</td>
<td>17.08%</td>
<td>39.26%</td>
<td>4.85</td>
<td>(1.5, 46)</td>
</tr>
</tbody>
</table>

**Notes:** Rule 1 – Rule 3 are derived from High School data alone.
Rule #4: If...

High School GPA is less than 3.25

Then...

The probability of student retained is 71.53%

And

The probability of student not retained is 28.47%
Rule #5: If...

High School GPA is greater than 3.25
And
High School Social Science GPA is less than 3.95
And
Planned Residence for Fall 2002 is “Dormitory”, “Other Campus Housing”, or “Undecided”
Then...
The probability of student being retained is 83.15%
And
The probability of student not retained is 16.85%
Rule #6:  If...

Attended Religious Services is “Not at All” or “Occasionally “
And
High School GPA is greater than 3.25
And
Planned Residence for Fall 2002 is “Private Home”, “W/Family”, or “Frat/Sorority”

Then...
The probability of student being retained is 78.04%
And
The probability of student not retained is 21.96%
## Summary of Rules

<table>
<thead>
<tr>
<th>Rule</th>
<th>Students Not retained</th>
<th>Total # of Students in this rule</th>
<th>Not retained Hit Rate % in this rule</th>
<th>Not retained Hit Rate % in all data</th>
<th>Odds Ratio</th>
<th>95% Confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule 4</td>
<td>240</td>
<td>849</td>
<td>28.27%</td>
<td>35.29%</td>
<td>2.54</td>
<td>(1.26,24)</td>
</tr>
<tr>
<td>Rule 5</td>
<td>122</td>
<td>724</td>
<td>16.85%</td>
<td>17.94%</td>
<td>4.93</td>
<td>(1.5,47.8)</td>
</tr>
<tr>
<td>Rule 6</td>
<td>184</td>
<td>838</td>
<td>21.96%</td>
<td>27.06%</td>
<td>3.55</td>
<td>(1.37,34)</td>
</tr>
</tbody>
</table>

**Note:** Rule 4 – Rule 6 are derived from both High School and Survey data.
Rule #7: If...

High School GPA is between 3.25 and 4.15
And
Student comes from Florida equals “Yes”
Then...
The probability of student retained = 84.03%
And
The probability of student not retained = 15.97%
Rule #8: If...

High School GPA is greater than 3.25
And
High School English Unit GPA is less than 3.95
Then...
The probability of student retained = 81.371%
And
The probability of student not retained = 18.63%
What is Hit Rate?

Definition: Not retained Hit Rate

<table>
<thead>
<tr>
<th>True Value</th>
<th>Not retained</th>
<th>Retained</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not retained</td>
<td>N_{11}</td>
<td>N_{12}</td>
<td>N_{1.}</td>
</tr>
<tr>
<td>Retained</td>
<td>N_{21}</td>
<td>N_{22}</td>
<td>N_{2.}</td>
</tr>
<tr>
<td>Total</td>
<td>N_{1.}</td>
<td>N_{2.}</td>
<td>N</td>
</tr>
</tbody>
</table>

Hit Rate = \frac{N_{11}}{N_{1.}}.

Hit Rate is a powerful measurement in model fitting.

Hit Rate represents the prediction accuracy in our retention model.
Testing Data

- **Data Sources:**
  - High School data from Academic Year 2002

- **Number of Students:** 5579

- **Number of Variables:** 26
  - 13 numerical variables: HSGPA, SAT_Verb, SAT_Math…
  - 2 nominal variables: Nation_Merit_and_Hon, Ethnic_Origin
  - 7 binary variables: Gender, Full_status, Non_retain…
  - 4 derive variables: Flag1 – Flag4

- **Study Target:** Student who has lower chance to be retained
  - Retained after freshmen year: 4609 (82.61%)
  - Not Retained after freshmen year: 970 (17.39%)
## Model Comparison by Hit Rate

<table>
<thead>
<tr>
<th>Model</th>
<th>Model Description</th>
<th>Hit % in Training data</th>
<th>Hit % in Validation data</th>
<th>Hit % in Testing data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Tree 1</td>
<td><strong>Entropy</strong> split criterion</td>
<td>91%</td>
<td>90%</td>
<td>88%</td>
</tr>
<tr>
<td>Decision Tree 2</td>
<td>Chi-square split criterion</td>
<td>84%</td>
<td>83%</td>
<td>82%</td>
</tr>
<tr>
<td>Decision Tree 3</td>
<td>Gini Index split criterion</td>
<td>84%</td>
<td>83%</td>
<td>82%</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>Stepwise regression</td>
<td>78%</td>
<td>77%</td>
<td>73%</td>
</tr>
</tbody>
</table>
What Now??
Conclusion

- Data Mining is a powerful tool for analyzing student retention.
- These models can identify more than 88% of the students who dropped out in the test data.
- These models can be used to predict students retention before the start of the freshman year.
- First semester information can be added to further predict risk factors.
- Data Mining provides objective statistical data to support changes to retention efforts.
- Data Mining provides an assessment tool to measure the success of interventions.
Conclusion – Decision Tree model

- The quality of student learning experience (such as High School GPA, SAT) is the most significant factor in retention rate.
- The number of advance placement exams taken plays an important role in predicting retention.
- Student retention is also affected by student’s intended living arrangement.
- Career motivation also affects retention rate.
Strategies for Early Interventions

- Develop a focused retention program:
  - Current interventions focused on approximately 3500 freshmen
  - Using data mining, can focus retention efforts on approximately 850 students
- Provide a higher level of learning support (especially Science and Math) to minimize drop-out rate.
- Enhance the communication between the students and faculty.
- Keep student study interest and motivation alive.
Further Research

- Our approach is not a solution to all the problems that exist with retention.
- Enlarge the data source to look for other significant factors.
- Determine the most appropriate threshold for Decision tree model.
- Check accuracy of predictions on new data source.
- Develop integrated student retention programs.
- Continue to refine the models.
Questions?

UCF Student Development & Enrollment Services Website: www.sdes.ucf.edu
E-mail addresses:
Ron Atwell: ratwell@mail.ucf.edu
Steve Johnson: skjohnso@mail.ucf.edu
Morgan Wang: cwang@mail.ucf.edu