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:

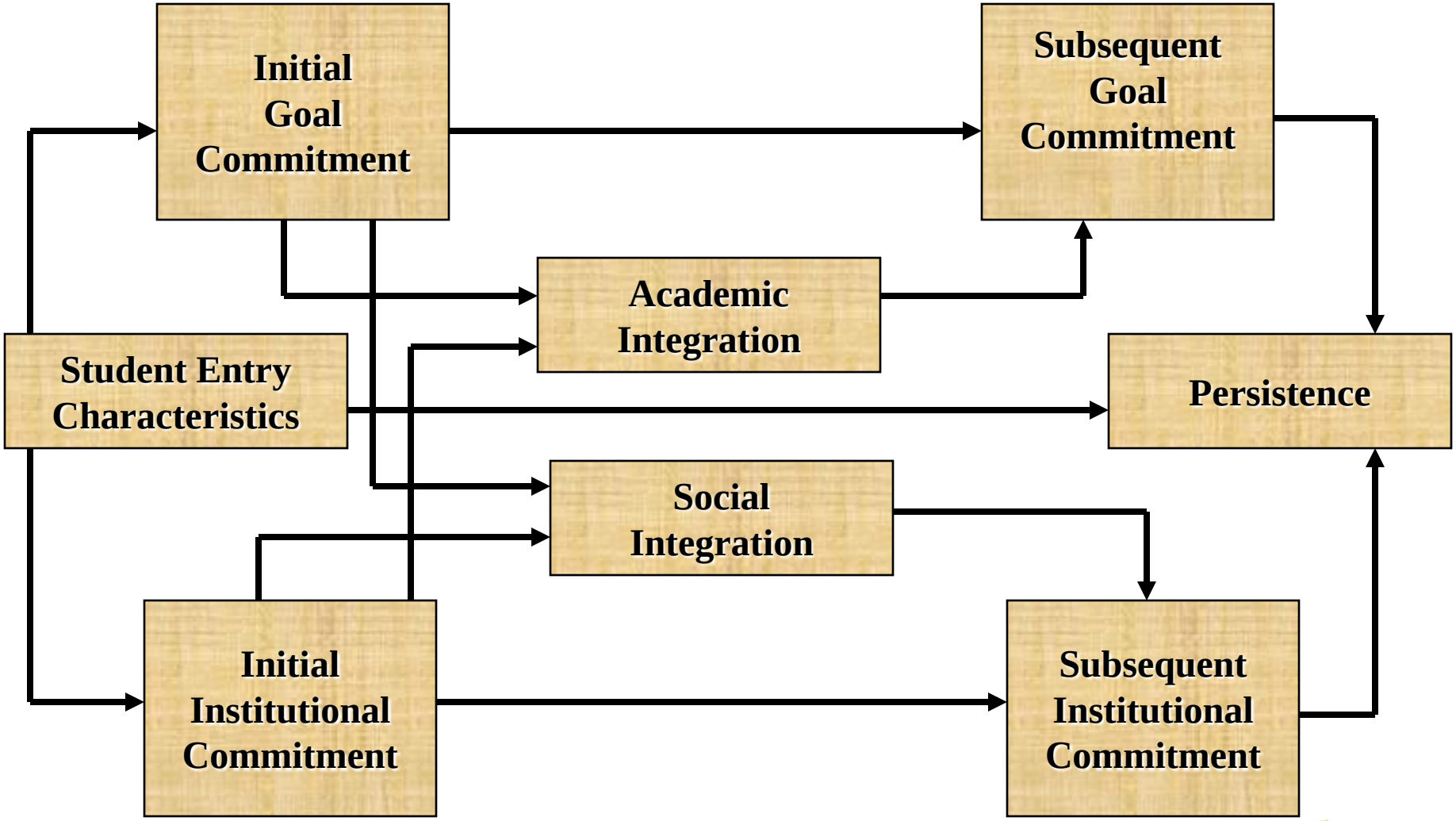
- Are highly structured
- Are interlocked with other programs/services
- Rely on extended, intensive student contact
- Are based on strategy of engagement
- Place special emphasis on staff quality
- Focus on affective as well as cognitive needs
- Track and monitor level of student satisfaction

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



Braxton et al (2004)

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

Year	Total Enrollment	Fall FTIC	HS GPA	SAT	% Residence Halls	Retention
1994	25,363	2,089	3.2	1085	32%	70%
Enrollment and Academic Services						
1998	30,000	3,127	3.5	1129	40%	75%
First Year Advising						
Student Development and Enrollment Services						
Enhanced Funding						
2001	36,013	3,759	3.66	1152	65%	81%
2002	38,795	3,922	3.74	1167	67%	84%
Majors Fair						
2003	42,000	4,134	3.81	1172	68%	84% projected
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
Predicting the Future

Data Mining is NOT a Crystal Ball

It is a Prŏcess (or Prōcess)

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%)

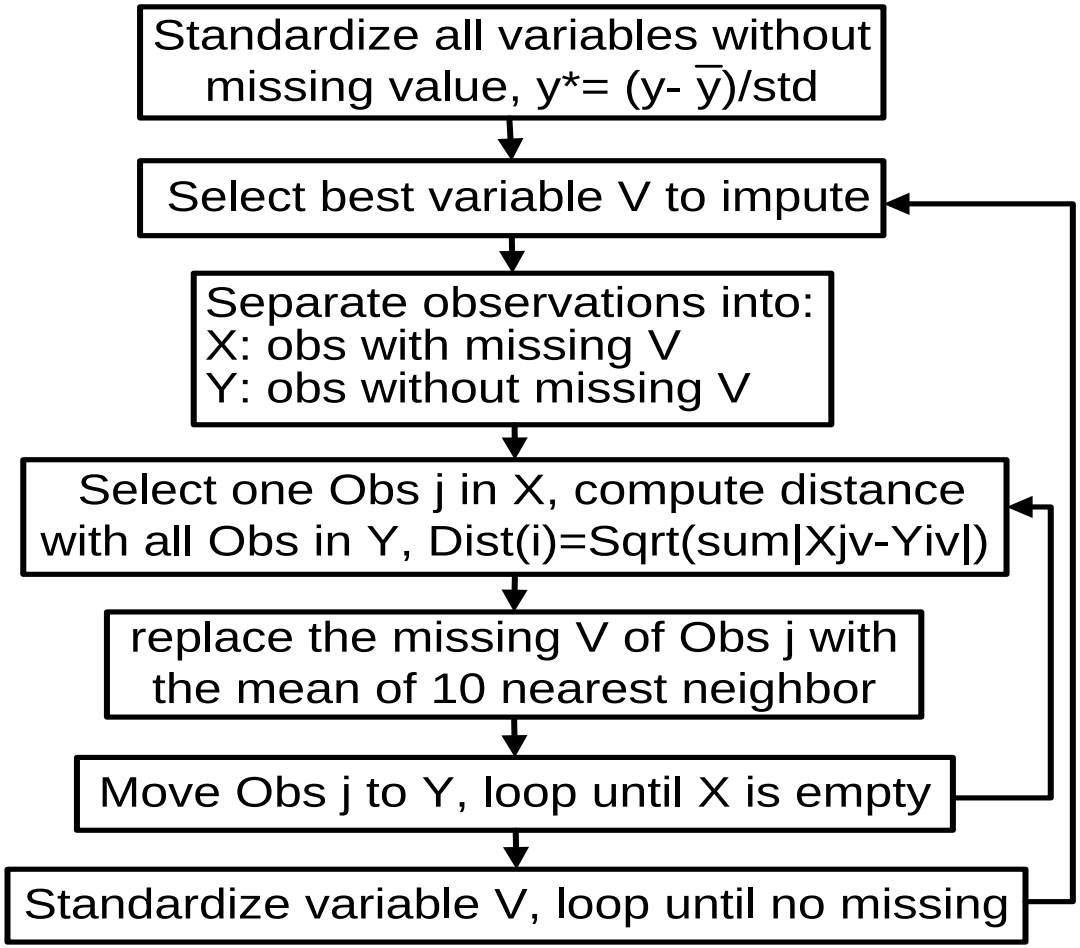
Data Problems

- 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

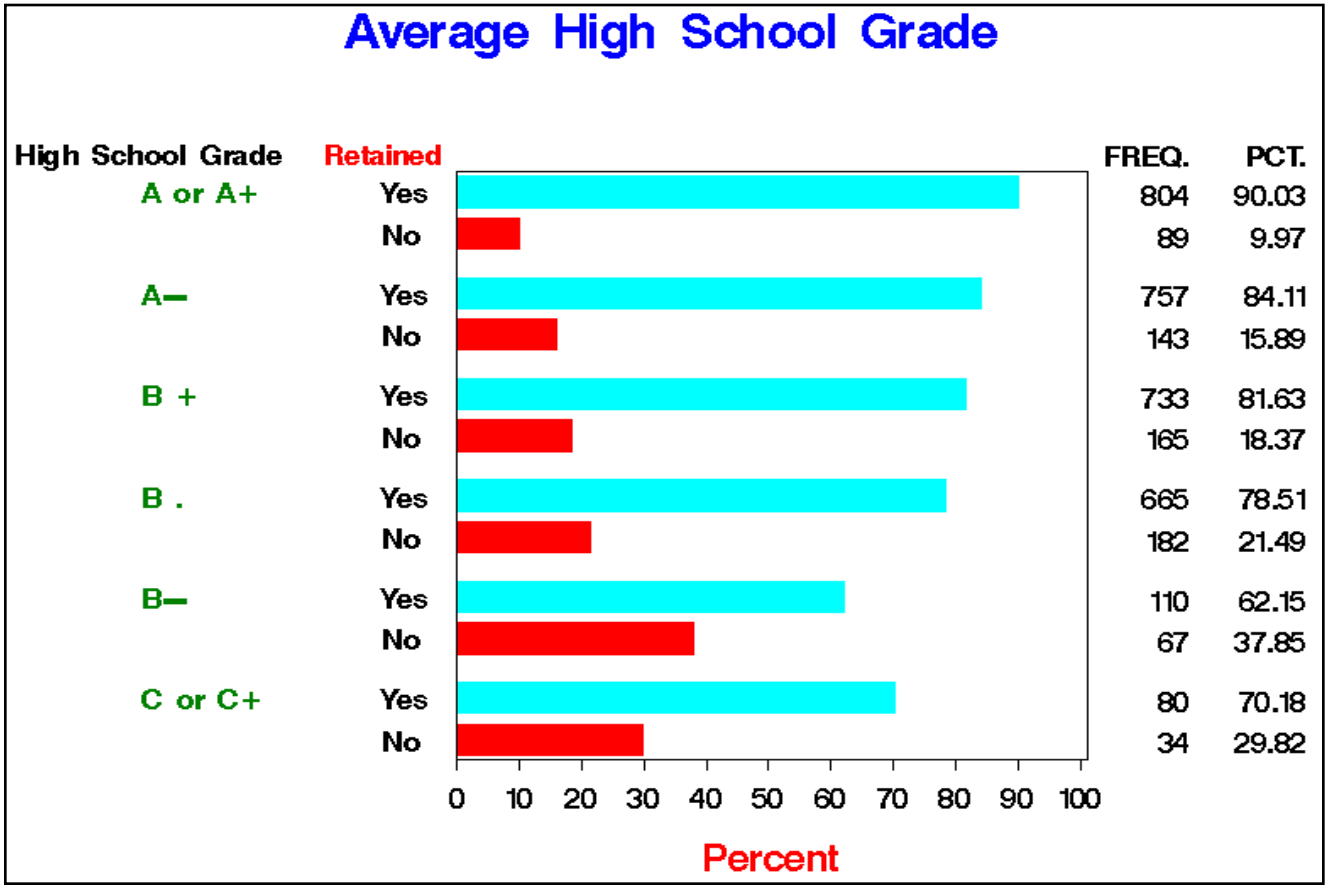
Fix Data Problems

- Missing Value Imputation
- Categorical variables with many categories
- Reduce the number of Variables
- etc.

Continuous variable imputation – Nearest Neighbor Algorithm

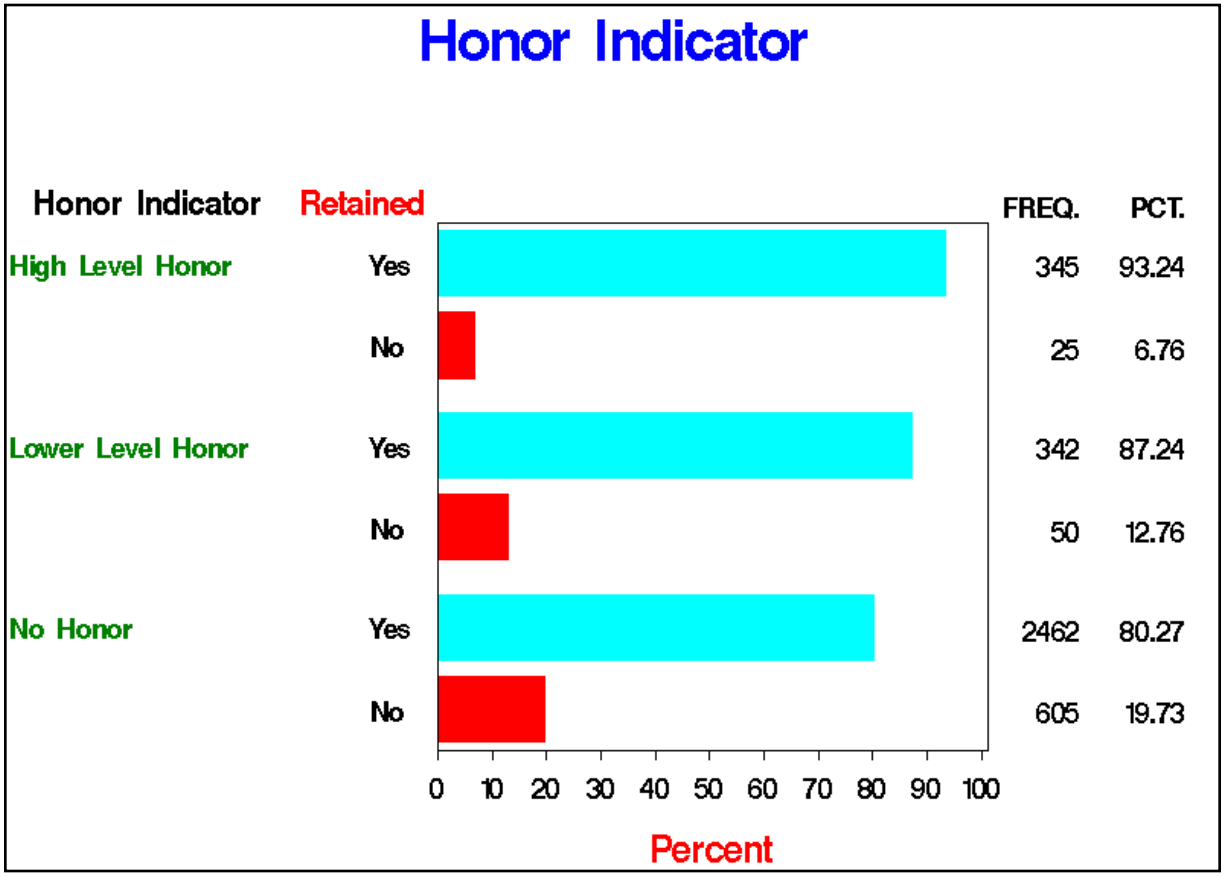


Data Exploration–High School Grade



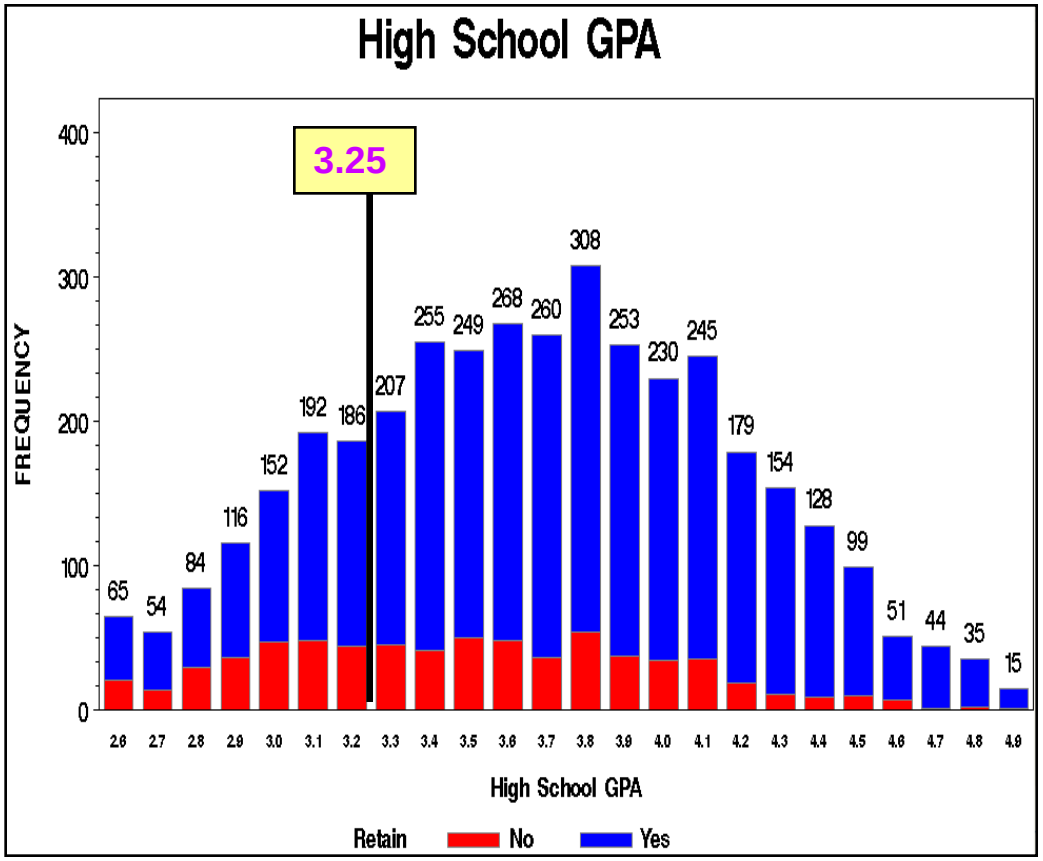
Students with lower high school grade have higher chance of not being retained after their freshmen year.

Data Exploration–Honor Indicator



Entering freshmen with a higher level “Honor” status have higher chance of being retained.

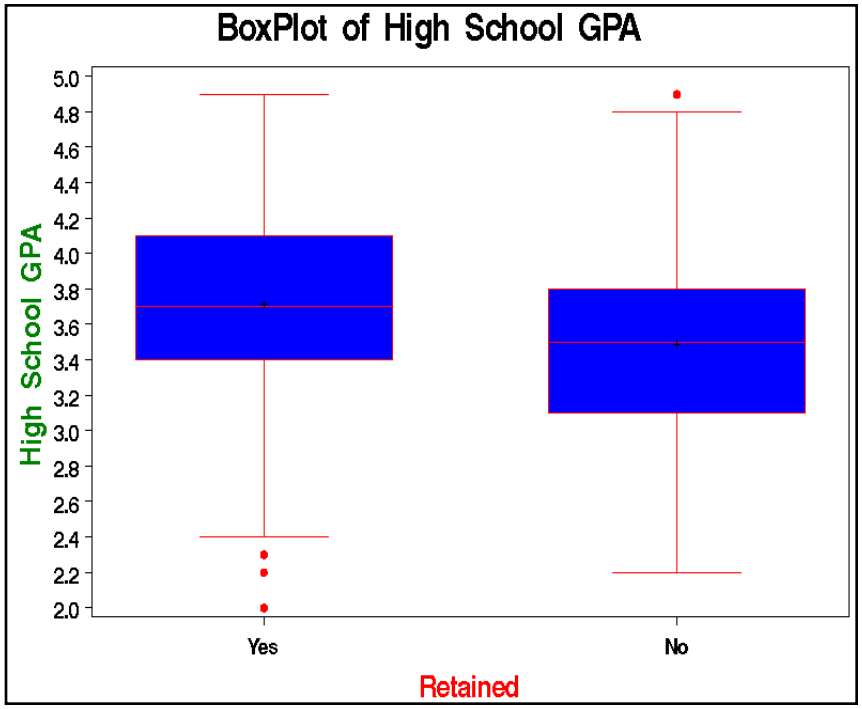
Data Exploration–High School GPA



Indicator of High School GPA	Retained		Total
	Yes	No	
High School GPA < 3.25	80 70.2%	34 29.8%	114
High School GPA >=3.25	3069 82.6%	646 17.4%	3715
Total	3149	680	3829

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

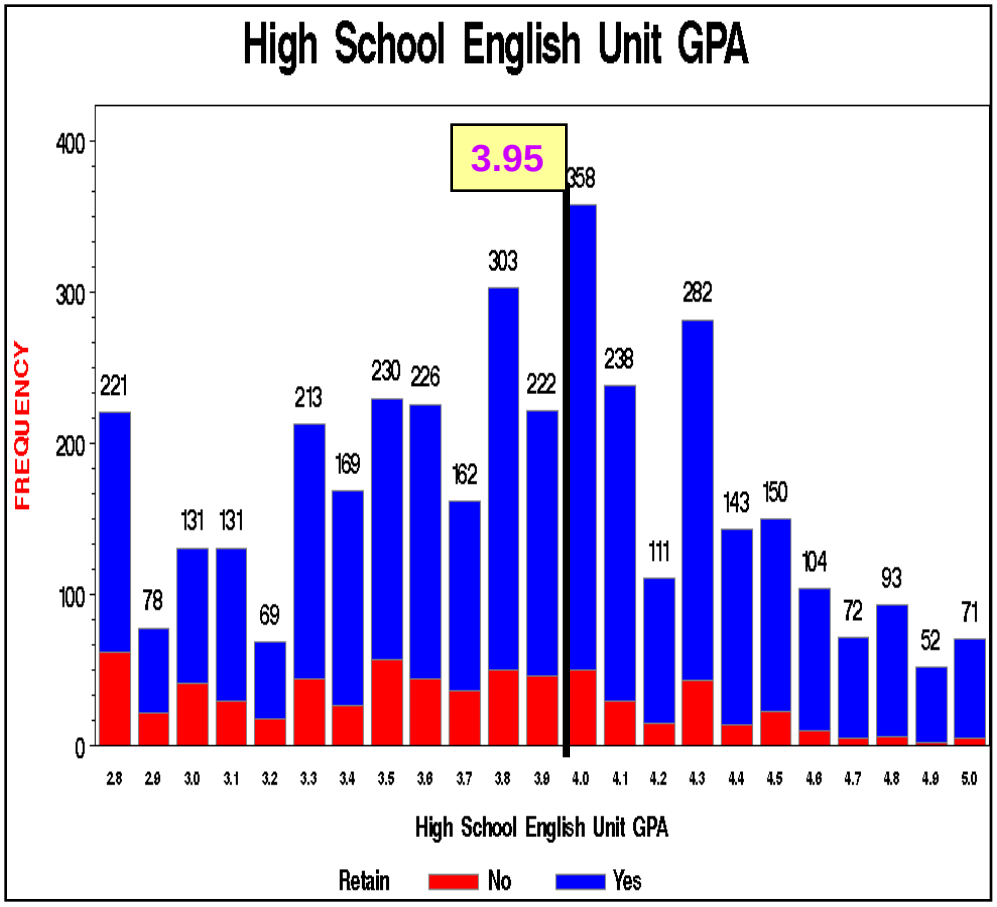


	Retained	Not Retained
Count	3149	680
Mean	3.72	3.48
Std Dev.	0.50	0.50

T test: t value = 10.75
p value < 0.0001 (significant)
Reject null hypothesis

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.

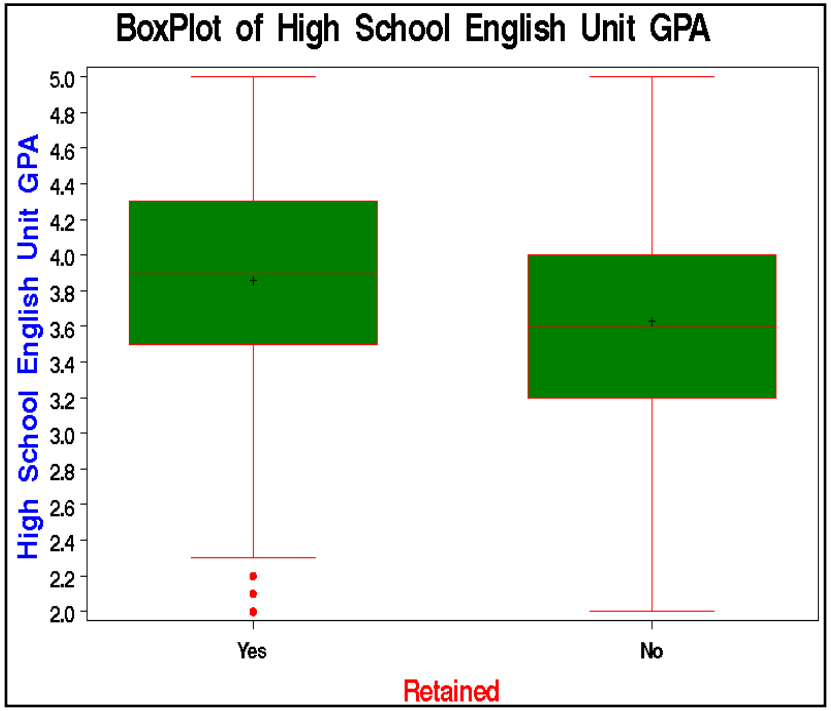
Data Exploration–English Unit GPA



Indicator of English Unit GPA	Retained		Total
	Yes	No	
English Unit GPA < 3.95	1678 77.9%	477 22.1%	2155
English Unit GPA >=3.95	1471 87.9%	646 12.1%	1674
Total	3149	680	3829

High school English is the most important subject for students to succeed in college.

Data Exploration–English Unit GPA

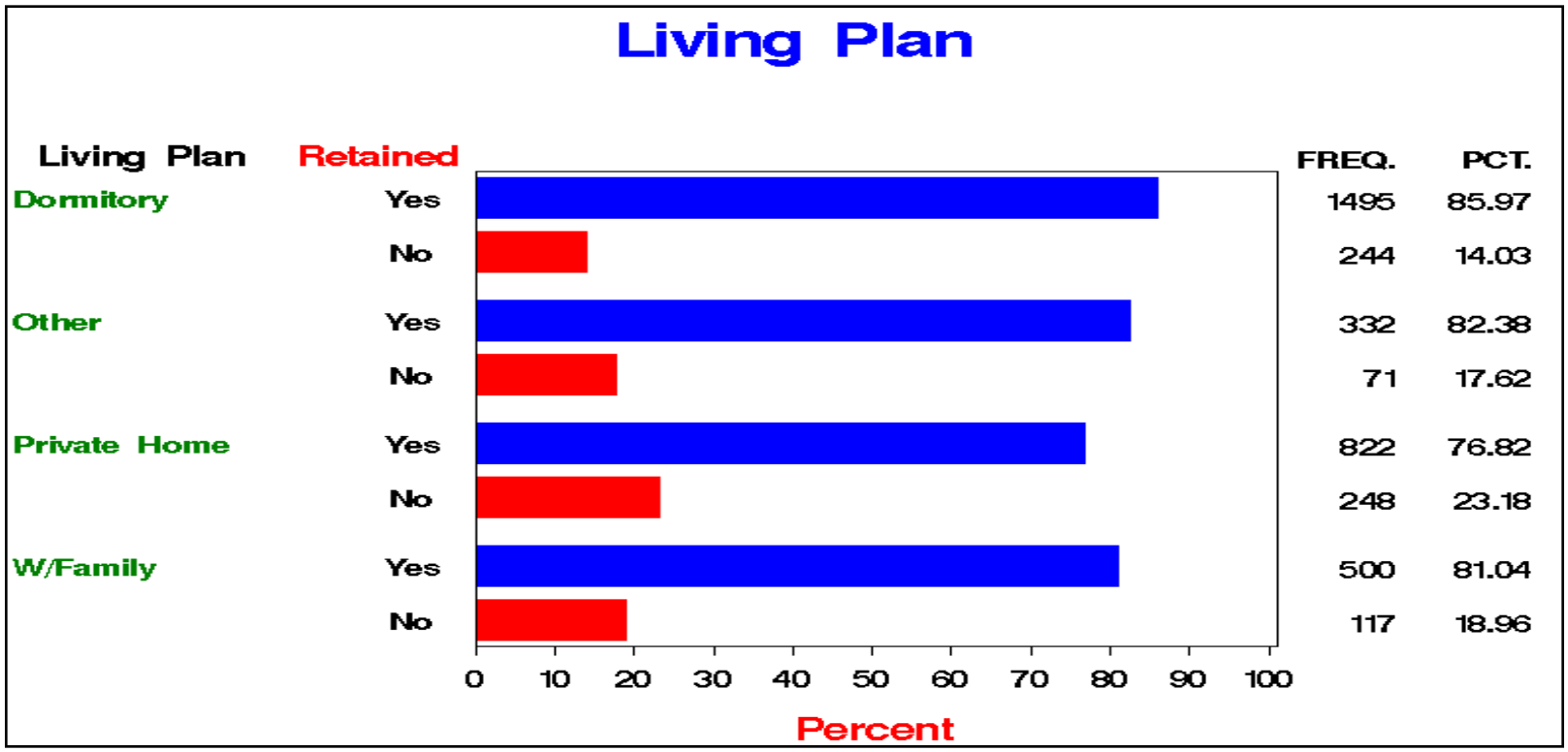


	Retained	Not Retained
Count	3149	680
Mean	3.86	3.63
Std Dev.	0.58	0.57

T test: t value = 9.43
p value < 0.0001 (significant)
Reject null hypothesis

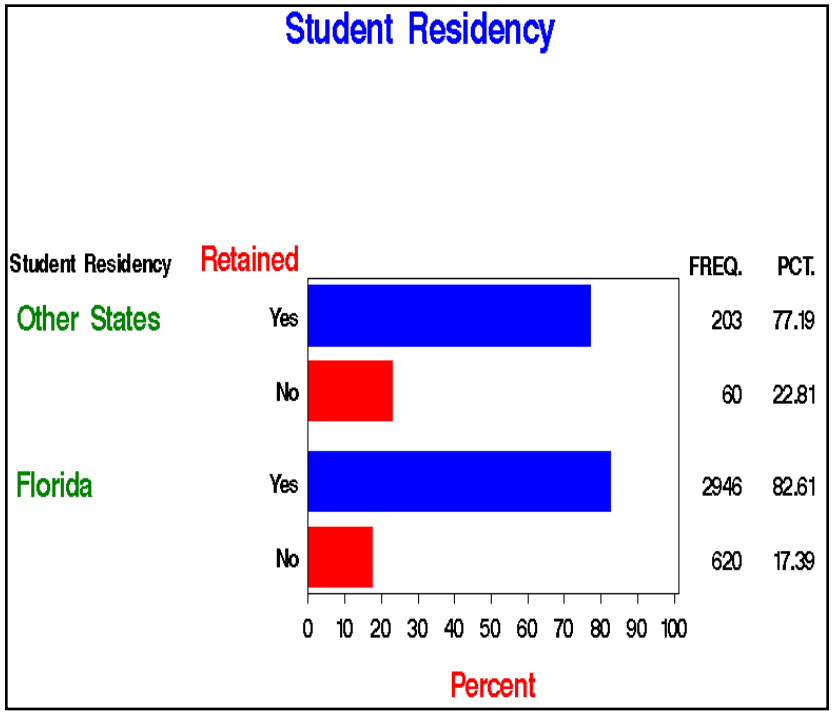
The high school English GPA for students who are not retained after their freshmen year is on the average 0.23 below their counterpart.
Besides, from T test, it shows that comparing retained students to not retained students, the Mean of English Unit GPA is significantly different.

Data Exploration—Living Plan



Students have a higher retention rate if they decide to live in the dormitory.

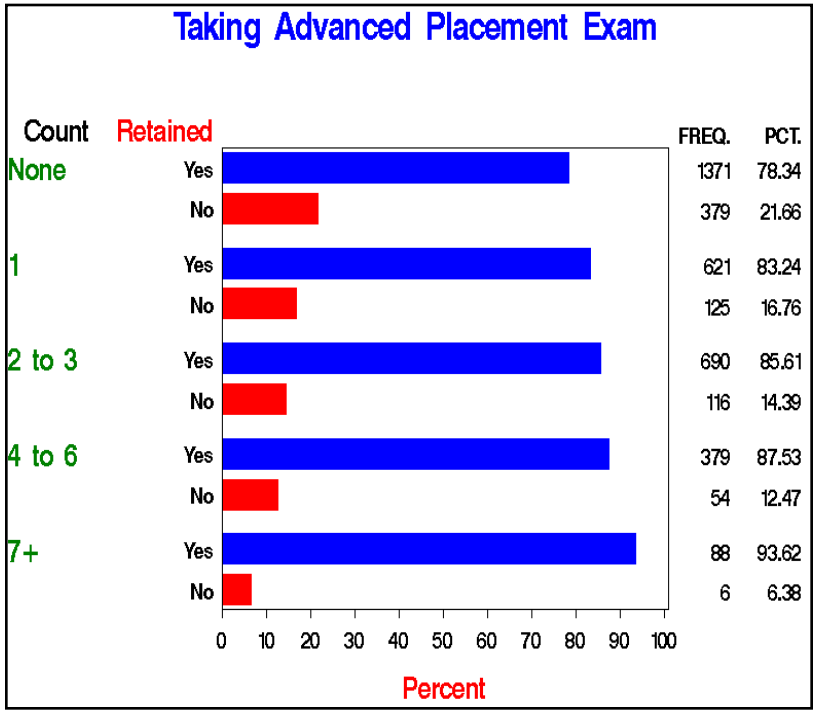
Data Exploration—Student Residency



Indicator of Student Residency	Retained		Total
	Yes	No	
Student comes from Florida	2946 82.6%	620 17.4%	3566
Student comes from other States	203 77.2%	60 22.8%	263
Total	3149	680	3829

Obviously, most students at UCF come from Florida, and they have the higher chance of being retained.

Data Exploration—Taking Advanced Placement Exam



The Number of Taking Advanced Placement Exam	Retained		Total
	Yes	No	
None	1371 78.3%	379 21.7%	1750
1	621 83.2%	125 16.8%	746
2 to 3	690 85.6%	116 14.4%	806
4 to 6	379 87.5%	54 12.5%	433
More than 7	88 93.6%	6 6.4%	94
Total	3149	680	3829

The more Advanced Placement Exams taken, the higher the chance of being retained.

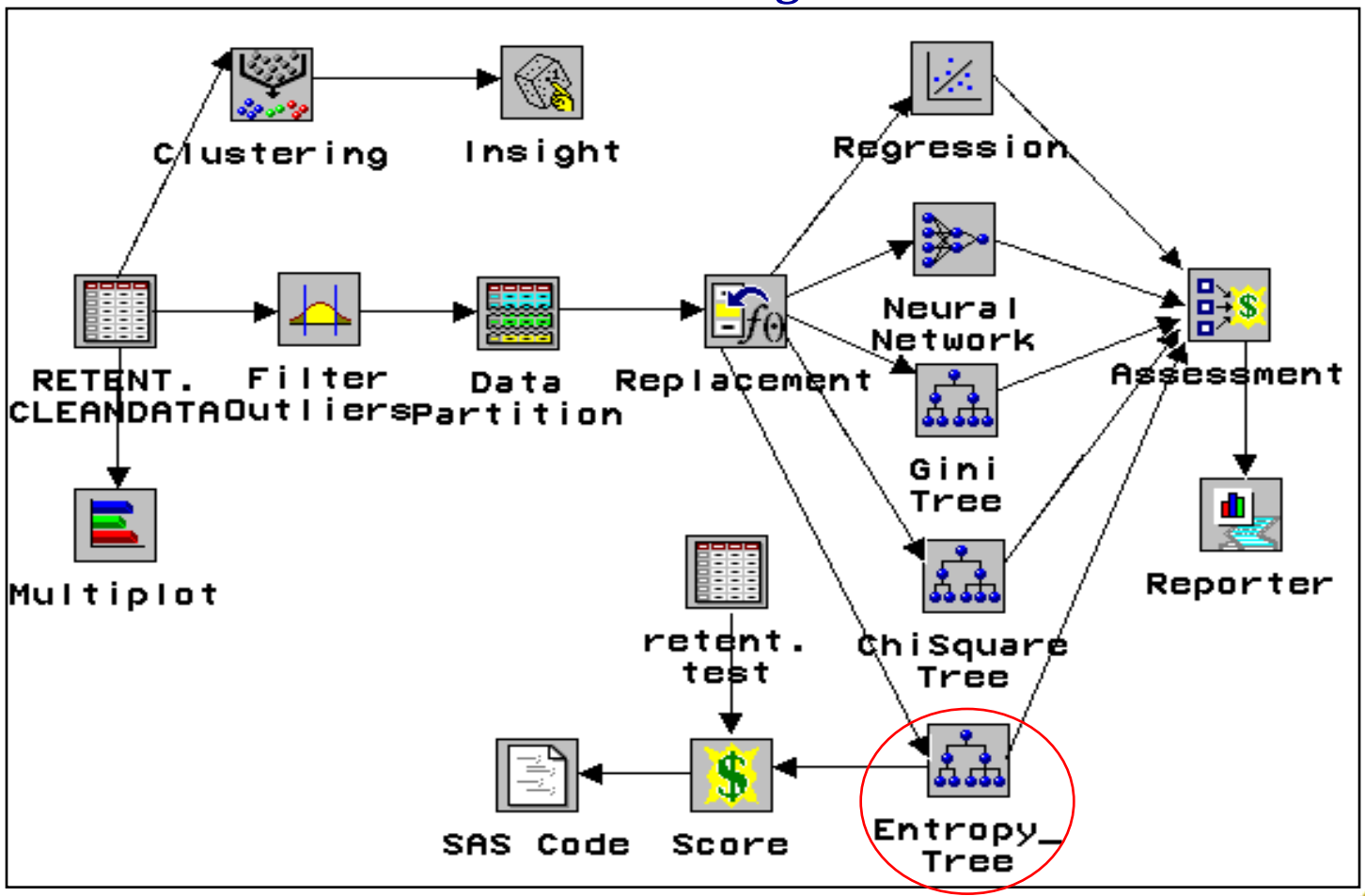
Model Building

- 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)

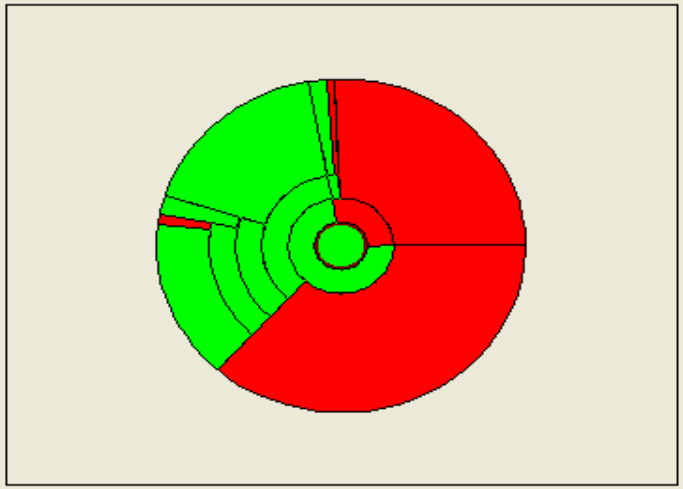
Process Flow Diagram



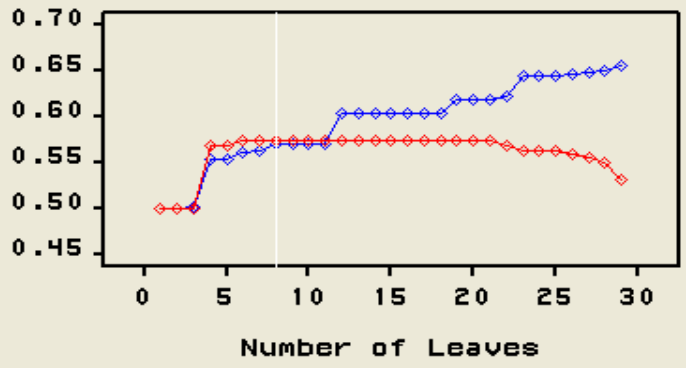
Entropy Decision Tree Summary

Model | All | Summary | Tree Ring | Table | Plot | Score | Log | Notes

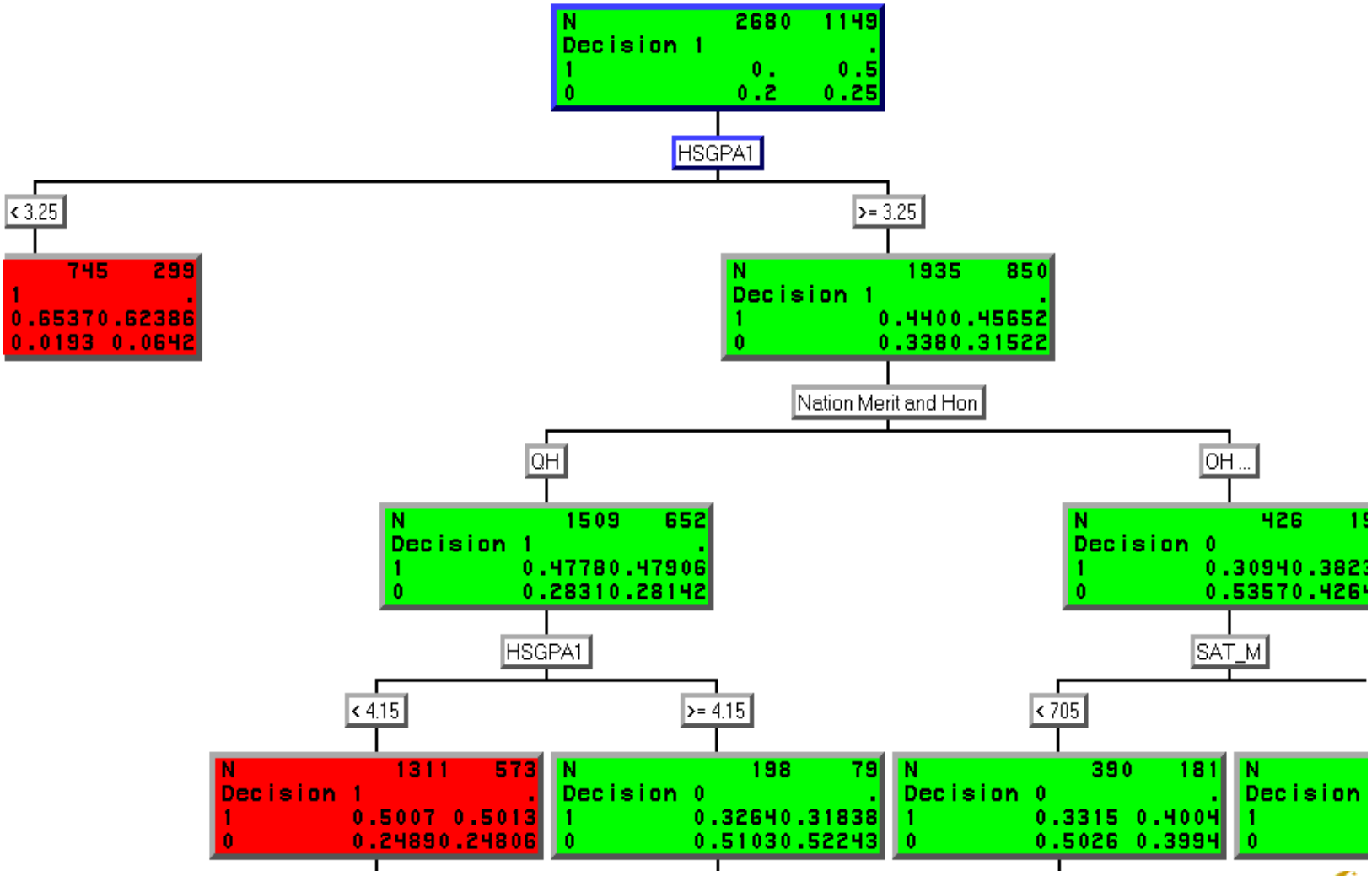
PRIORS	SOURCE	STAT	CHURN	==> 1	==> 0	
NOT APPLIED	TRAIN	N	1	451	47	*
NOT APPLIED	TRAIN	N	0	1568	614	*
NOT APPLIED	TRAIN	N	+	2019	661	*
NOT APPLIED	TRAIN	Row%	1	91	9	*
NOT APPLIED	TRAIN	Row%	0	72	28	*
NOT APPLIED	TRAIN	Row%	+	75	25	*
NOT APPLIED	TRAIN	Col%	1	22	7	*
NOT APPLIED	TRAIN	Col%	0	78	93	*
NOT APPLIED	TRAIN	Col%	+	100	100	*
NOT APPLIED	TRAIN	%	1	17	2	*
NOT APPLIED	TRAIN	%	0	59	23	*
NOT APPLIED	TRAIN	%	+	75	25	*



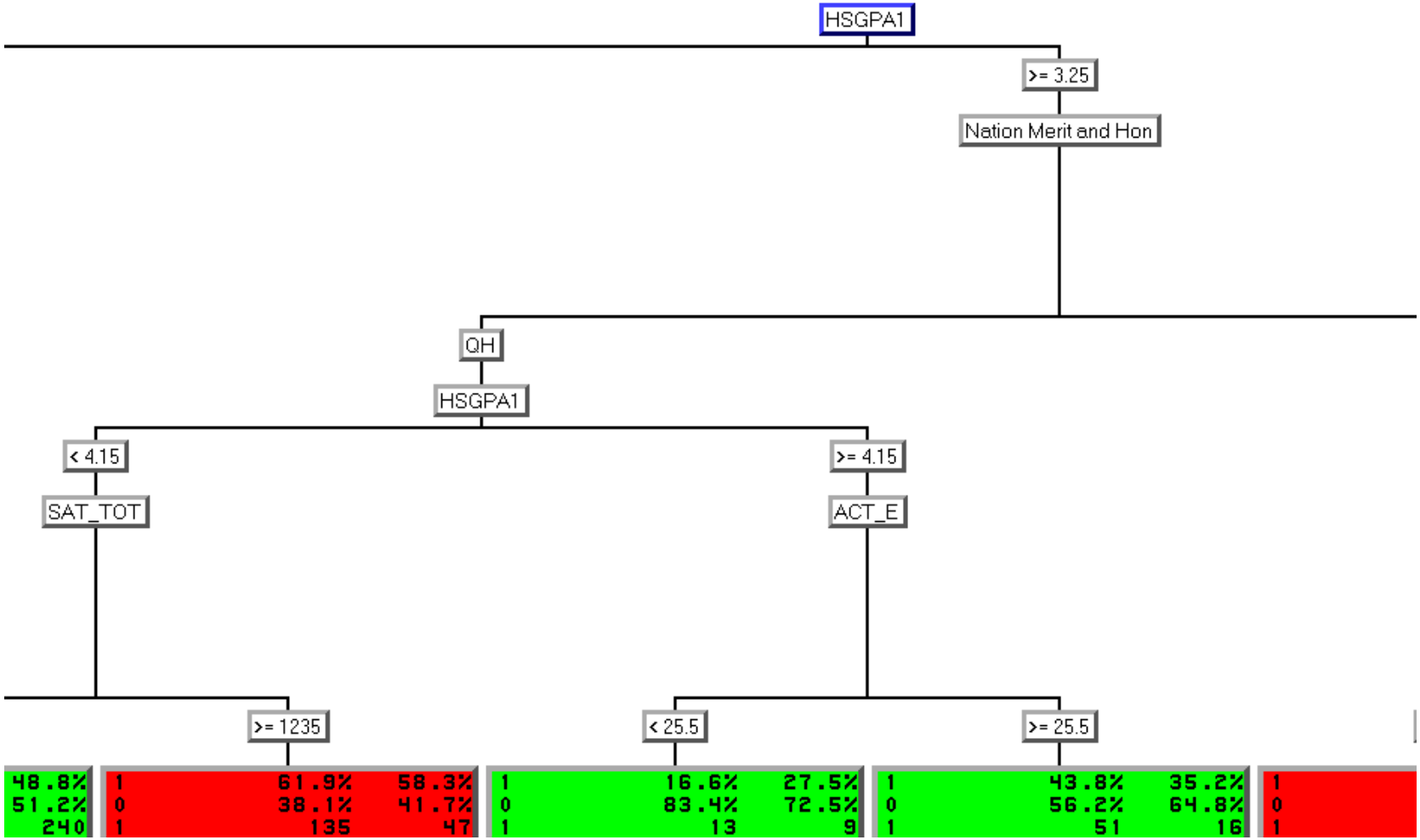
Leaves	Training	Validation
8	0.5699	0.5747
9	0.5699	0.5747
10	0.5699	0.5747
11	0.5699	0.5747
12	0.6031	0.5749
13	0.6031	0.5749
14	0.6031	0.5749
15	0.6031	0.5749
16	0.6031	0.5749
17	0.6031	0.5749



Decision Tree from High School Data



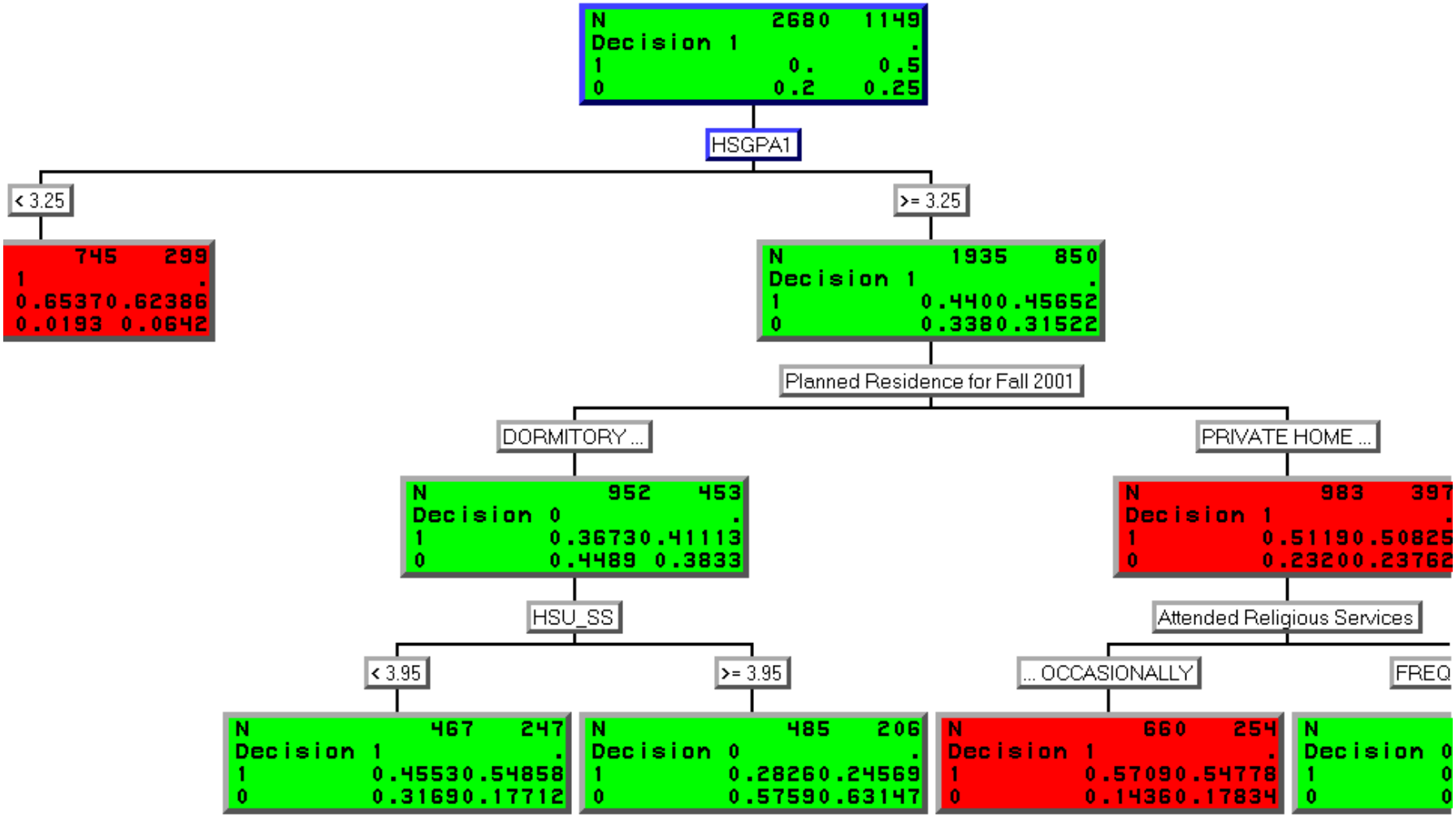
Decision Tree from High School Data cont'd.



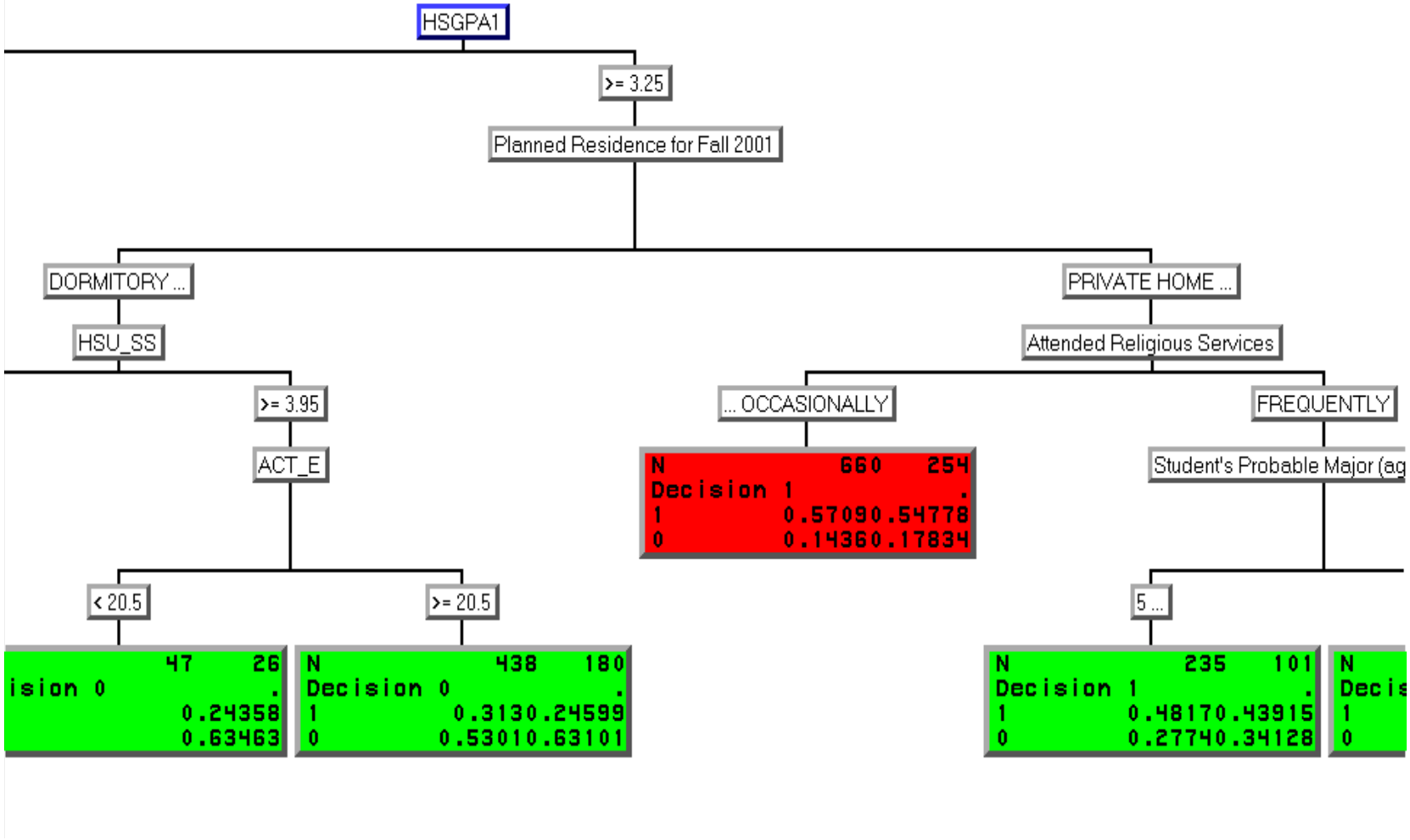
Important variables from High School data

Name	Importance	Role	Rules	Variable Label
HSGPA1	1.0000	input	2	High School GPA
NATION_MERIT_AND_HO	0.5639	input	1	Honors Indicator
ACT_E	0.4504	input	2	ACT English Score
SAT_TOT	0.4191	input	2	SAT Total
HSU_ENG	0.4148	input	2	High School Unit English GPA
HSU_MATH	0.3676	input	2	High School Unit Math GPA
ACT_COMP_TOT	0.3103	input	1	ACT comprehensive Total
SAT_M	0.2735	input	1	SAT Math
SAT_V	0.2587	input	1	SAT Verb
ACTCOMP	0.0000	rejected	0	ACT Composite
FLAG1	0.0000	rejected	0	Indicator From Talented_20 and Honor College
CITIZEN	0.0000	rejected	0	U. S. Citizen
FLAG2	0.0000	rejected	0	Indicator of SAT difference
HSU_NS	0.0000	rejected	0	High School Natural Science Unit GPA
SEX	0.0000	rejected	0	Student's Gender
FULLSTAT	0.0000	rejected	0	Full/Part Time Status
ACT_M	0.0000	rejected	0	ACT Math Score
FT_PT	0.0000	rejected	0	Part time / Full Time
AGE	0.0000	rejected	0	Student's Age

Decision Tree from Overall Data



Decision Tree from Overall Data cont'd.



Important variables from Overall data

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

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

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

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

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

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

Predicted Value

True Value	Not retained	Retained	Total
Not retained	N_{11}	N_{12}	$N_{1.}$
Retained	N_{21}	N_{22}	$N_{2.}$
Total	$N_{.1}$	$N_{.2}$	N

$$\text{Hit Rate} = N_{11} / N_{1.}$$

- 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

Model	Model Description	Hit % in Training data	Hit % in Validation data	Hit % in Testing data
Decision Tree 1	**Entropy** split criterion	**91%**	**90%**	**88%**
Decision Tree 2	Chi-square split criterion	84%	83%	82%
Decision Tree 3	Gini Index split criterion	84%	83%	82%
Logistic Regression	Stepwise regression	78%	77%	73%

What Now??

Conclusion

- Data Mining is a powerful tool for analyzing student retention.
- These model 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