Using Decision Trees to Analyze Students at Risk of Dropping Out in Their First Year of College Based on Data Gathered Prior to Attending Their First Semester

Dawn Broschard, EdD
Senior Research Analyst
Office of Retention and Graduation Success
dbroscha@fiu.edu

Overview

1. First-Year Retention
2. Graduate Success Initiative
3. CRISP-DM/ Decision Trees
4. Findings
5. Use of Findings
6. Lessons Learned
7. What Now?

First Year a Critical Time

The largest percentage of students who drop out are freshmen who leave after their first year. Failing a class one of top reasons for dropout, along with clinical depression, losing financial aid, roommate conflicts, and large increase in tuition (Pleskac, et. al., 2011).

- 20% of FIU FTIC students do not return in the second year
- GPA is too low (< 2.0), not meet Satisfactory Academic Progress
- Not earn 67% or more of attempted credits (including DR, WI, IN)
- 25% of students who lost Bright Futures dropped out, compared to 4% who didn't lose Bright Futures
- HOWEVER, half of the students who drop out have a GPA > 2.0, so not just leaving for academic reasons.

How a student feels after failing a course, not being sure of their major, or losing Bright Futures …

Graduation Success Initiative GSI.FIU.EDU

- Goal: Increase on-time (4 year)graduation rates by 3% per year

- Goal for 2009 FTIC Cohort = <u>56%, on track</u>
  - 51.6% have graduated so far, 1,615
  - 3.0% are potential Spring graduates
  - 9.9% still enrolled, some are potential Summer grads

Graduation Success Initiative GSI.FIU.EDU

Key points:

- Increase availability to professional, trained advisors
  - Suite of technological tools to help advisors and students stay connected.

- Help students identify the right major early on (My FIU, My MajorMatch, Major Maps)
- Identify students who are falling off track in key courses (My eAdvisor, Panther Degree Audit, Alert Systems)
- Identify bottleneck courses

So?

Does pre-college data exist that we can use to identify students who may be at risk of dropping out in their first year?

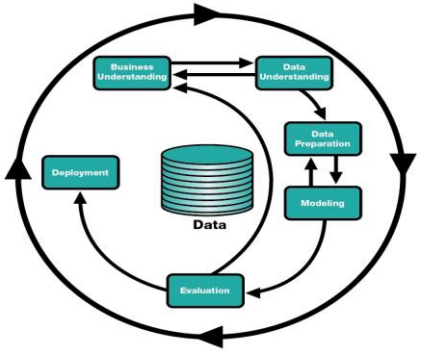
If so how do we analyze it?

Identify

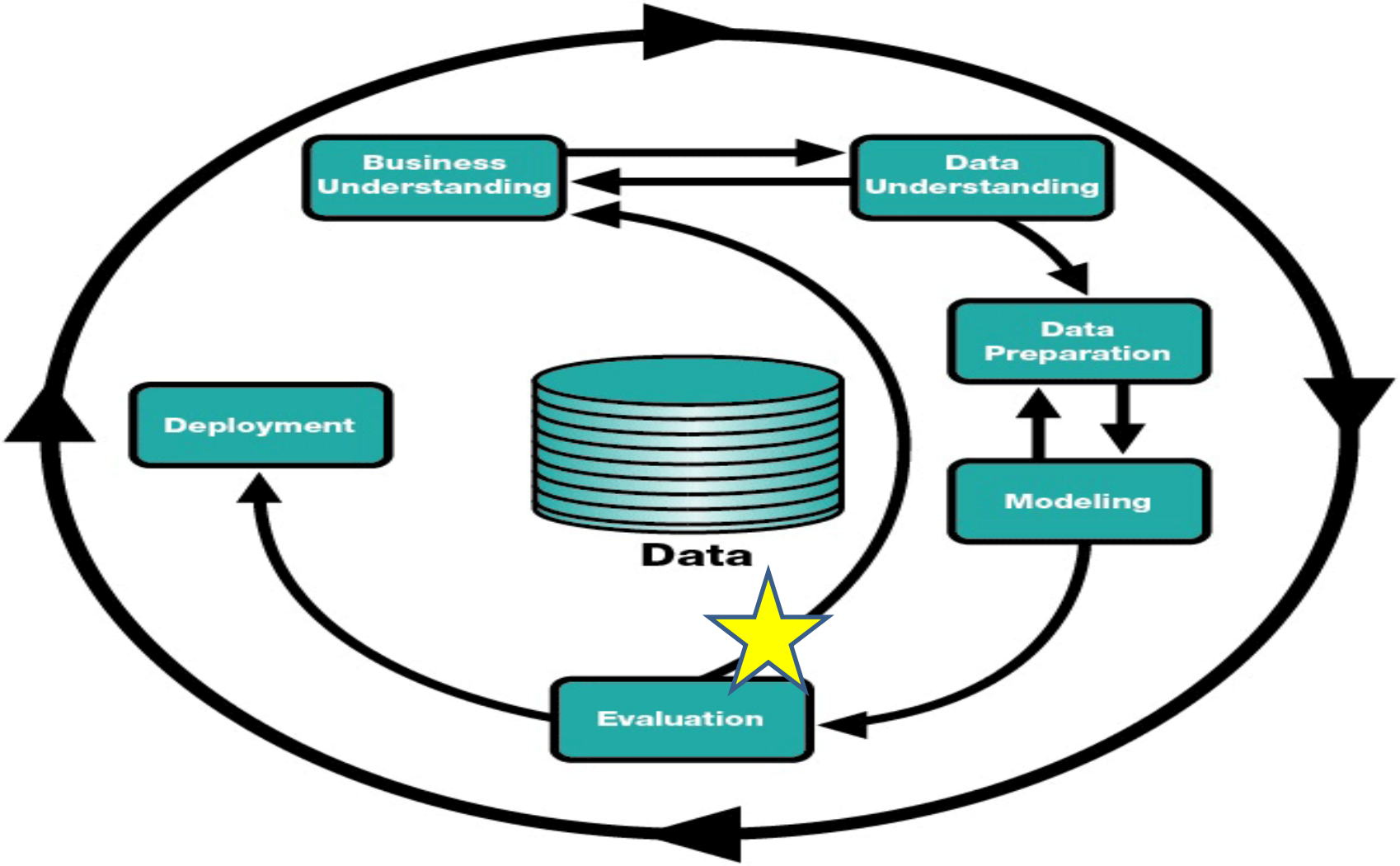
- Goal: Identify incoming freshmen who may be at risk of dropping out *before* they start at FIU

- Who's at risk and why?

- Why? So we can intervene early and help, provide early guidance and support



CRISP-DM



http://www.sv-europe.com/about-us/contact-us/

Decision Trees?

- Decision trees are a data mining tool used to predict an outcome variable based on predictor variables:

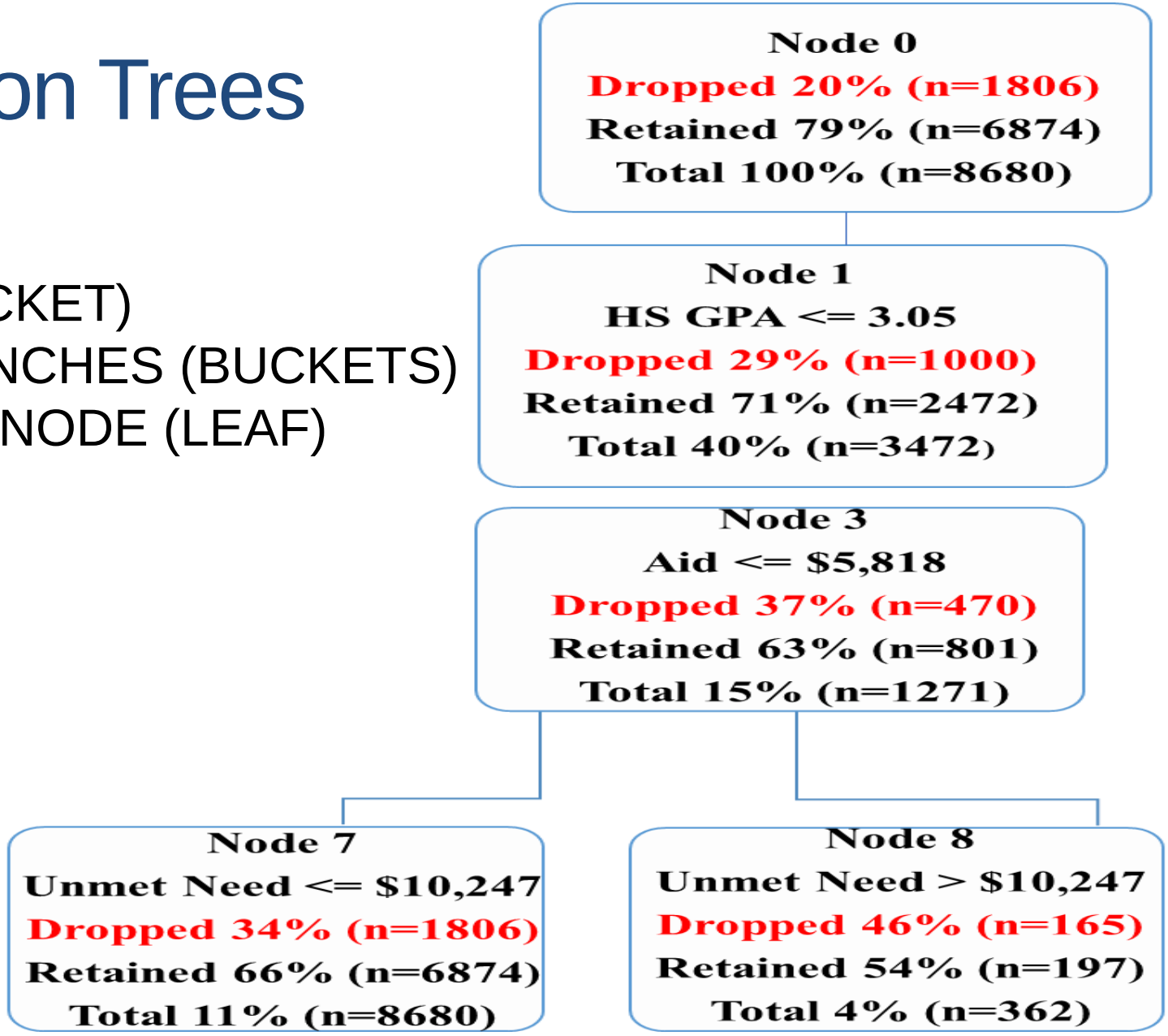
- ***"Multiple variable analysis is particularly important in current problem-solving because almost all critical outcomes that determine success are based on multiple factors."*** (de Ville & Neville, pg. 6).

- Analyzed data from over 17,000 previous freshmen (2010-2013 FTIC students)

- Main variables predicting retention (enrolling for sophomore year): ?

Decision Trees

NODE (BUCKET)
TREE BRANCHES (BUCKETS)
TERMINAL NODE (LEAF)



Advantages

1. You can set your target of interest (in our case not retained).

2. There are several algorithms to choose from (I chose Classification and Regression – CRT/CART)

3. The algorithm excludes variables that do not contribute to the overall accuracy of the model.

Advantages

4. The model is recursive and re-evaluates other variables continuously.

5. You can filter questionable branches (bucket) in the tree and test to see if they are significant and remove ones that are not.

6. You can add pruning techniques where branches with extremely small numbers (which may be insignificant).

Advantages

And my three favorites….

7. There are validation tools available to do a training and test set!

8. You can set up “misclassification costs” for the model!!

9. You can save the model rules and apply it to data where the outcome is still unknown and keep modifying the model as more data exists!!!

   (If this, and this, then this…)

Our Research

- Total Sample was 17,346 (2010-2013 FTIC) students
- This procedure employed the Classification and Regression Tree (CRT/CART) algorithm
- Half of the total sample was used for the training set (n=8,680), and the **other half for the validation set (n= 8,666)** (i.e., a 50/50 training/test set).
- To identify characteristics that may put students at-risk for drop out, the initial node (NODE 0) **"target of interest"** was set at **not** being retained after one year.
- Set **"misclassification cost"** higher for students who leave so they would not be incorrectly classified as retained.

Variables Examined

Demographics:

- Gender
- Ethnicity
- High school country
- Dependent status
- State of Florida reported high school grade
- Distance from FIU
- Parents' highest level of education

FIU Data Available before Enrollment:

- Housing
- First major
- Cost of Attendance
- Financial need
- Financial aid
- Unmet need

High School and Test Data:

- Overall high school GPA
- Math high school GPA
- English high school GPA
- SAT
- SATV
- SATE
- SATM
- ACT
- ACTE
- ACTR
- ACTW
- ALEKS Math Placement Test

Findings

Variables in Order of Importance to the Model

1. High School Overall GPA
2. Financial Aid
3. Unmet Need
4. Ethnicity
5. Major Type (STEM, Nursing, Engineering)
6. FIU Housing

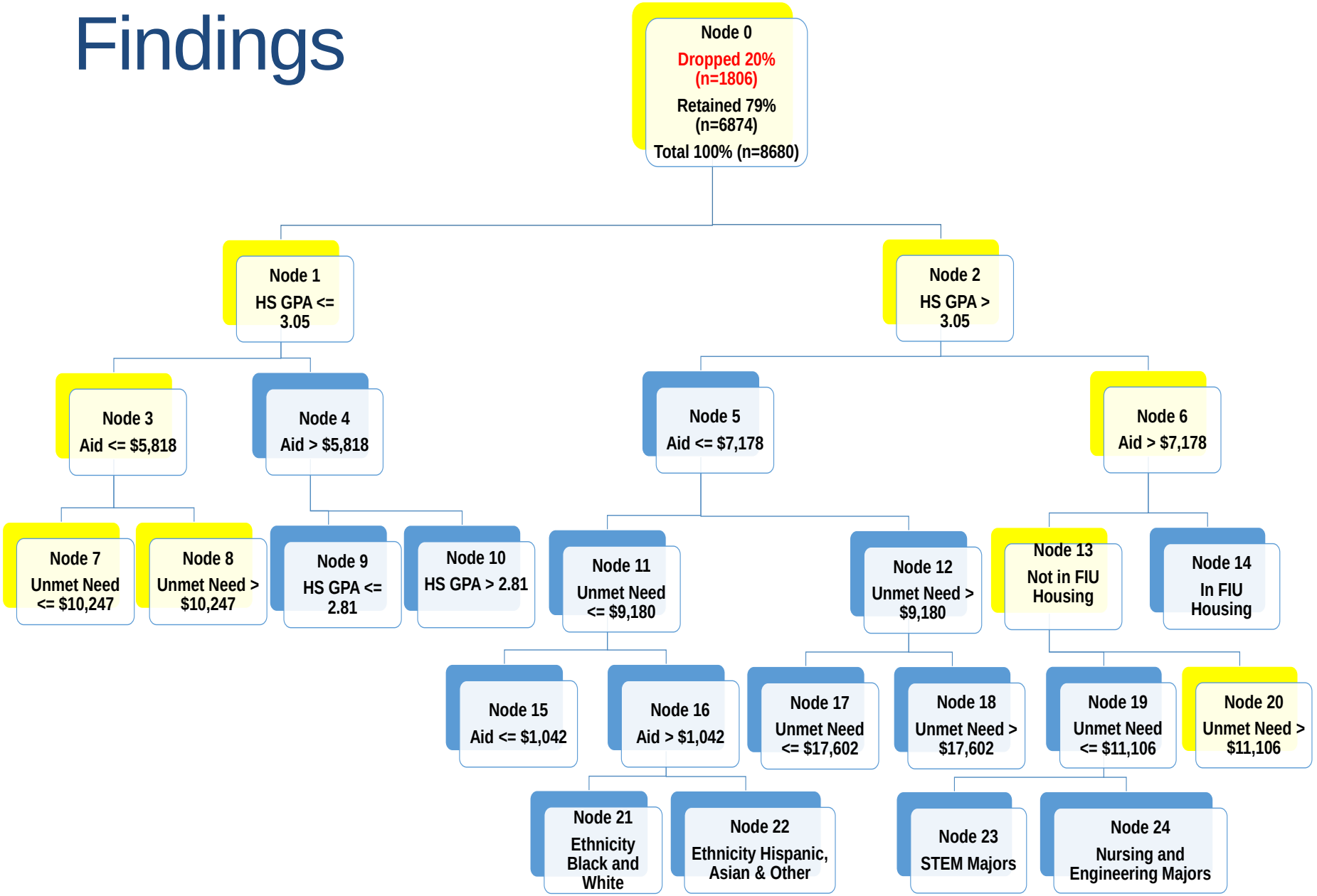
Node 0

Dropped 20% (n=1806)

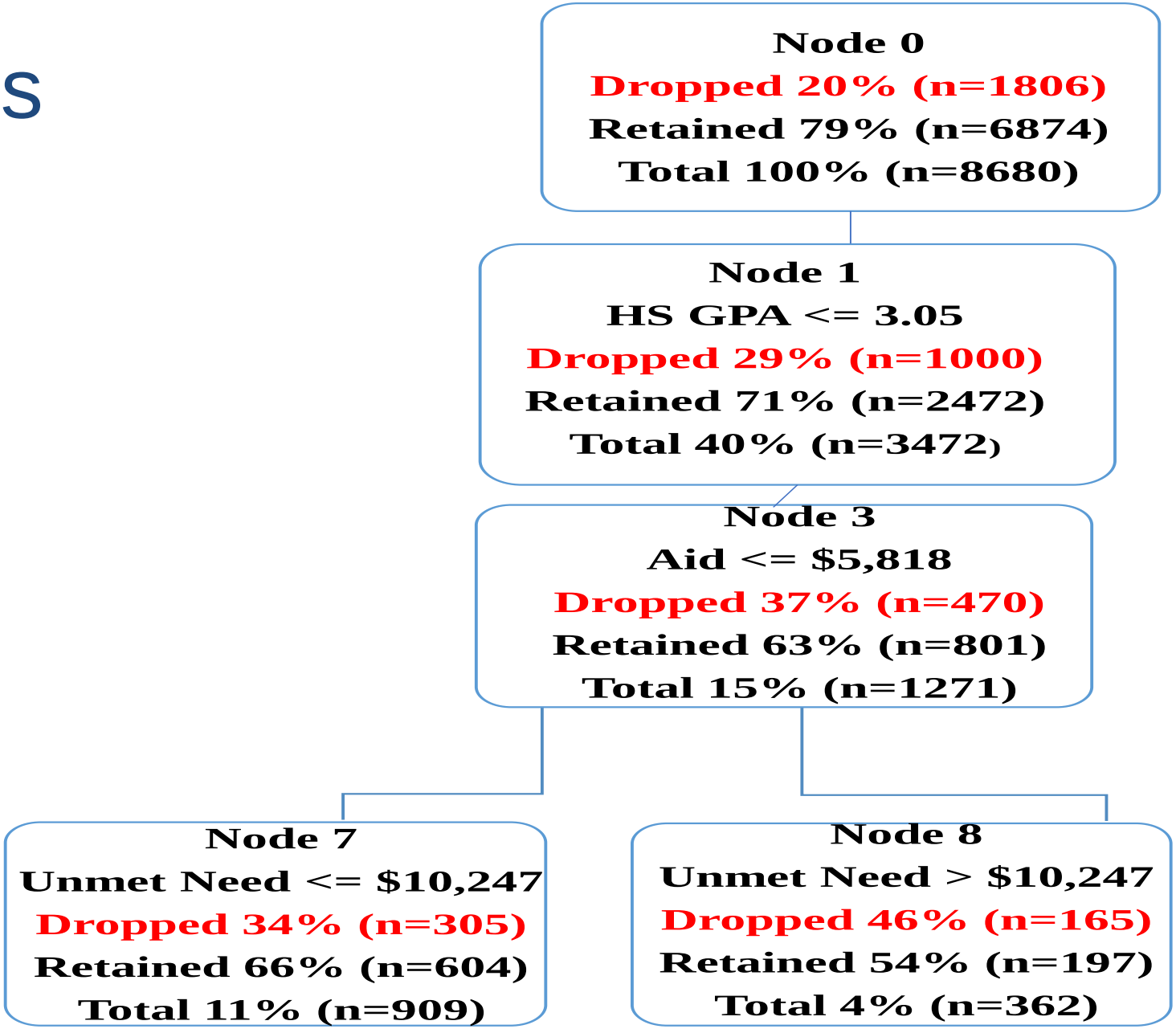
Retained 79% (n=6874)

Total 100% (n=8680)

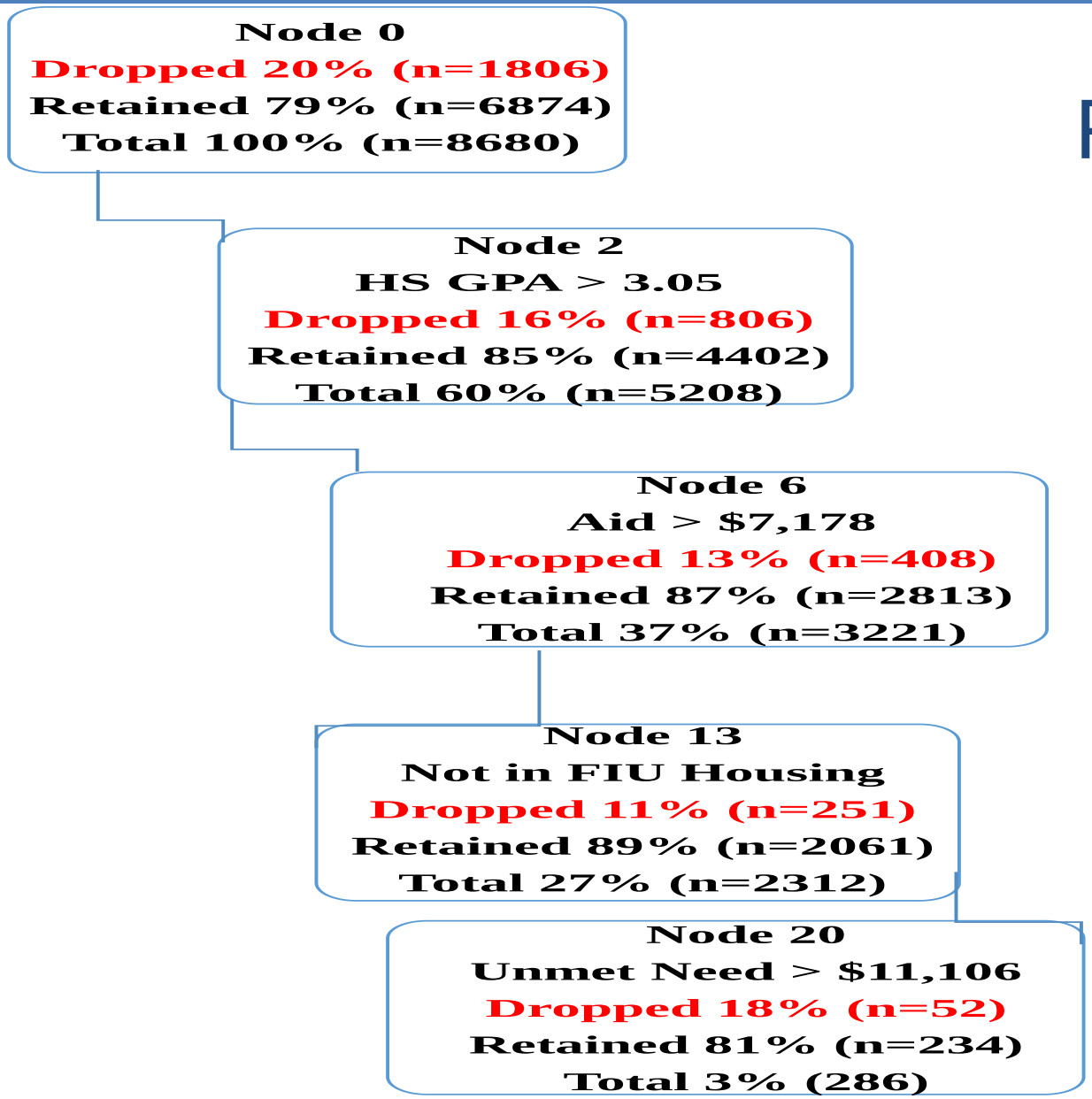
Findings


Node 0
Dropped 20% (n=1806)
Retained 79% (n=6874)
Total 100% (n=8680)
Node 1
HS GPA <= 3.05
Node 2
HS GPA > 3.05
Node 3
Aid <= $5,818
Node 4
Aid > $5,818
Node 5
Aid <= $7,178
Node 6
Aid > $7,178
Node 7
Unmet Need <= $10,247
Node 8
Unmet Need > $10,247
Node 9
HS GPA <= 2.81
Node 10
HS GPA > 2.81
Node 11
Unmet Need <= $9,180
Node 12
Unmet Need > $9,180
Node 13
Not in FIU Housing
Node 14
In FIU Housing
Node 15
Aid <= $1,042
Node 16
Aid > $1,042
Node 17
Unmet Need <= $17,602
Node 18
Unmet Need > $17,602
Node 19
Unmet Need <= $11,106
Node 20
Unmet Need > $11,106
Node 21
Ethnicity Black and White
Node 22
Ethnicity Hispanic, Asian & Other
Node 23
STEM Majors
Node 24
Nursing and Engineering Majors

Findings


Node 0
Dropped 20% (n=1806)
Retained 79% (n=6874)
Total 100% (n=8680)
Node 1
HS GPA <= 3.05
Dropped 29% (n=1000)
Retained 71% (n=2472)
Total 40% (n=3472)
Node 3
Aid <= $5,818
Dropped 37% (n=470)
Retained 63% (n=801)
Total 15% (n=1271)
Node 7
Unmet Need <= $10,247
Dropped 34% (n=305)
Retained 66% (n=604)
Total 11% (n=909)
Node 8
Unmet Need > $10,247
Dropped 46% (n=165)
Retained 54% (n=197)
Total 4% (n=362)

Findings


Node 0
Dropped 20% (n=1806)
Retained 79% (n=6874)
Total 100% (n=8680)
Node 2
HS GPA > 3.05
Dropped 16% (n=806)
Retained 85% (n=4402)
Total 60% (n=5208)
Node 6
Aid > $7,178
Dropped 13% (n=408)
Retained 87% (n=2813)
Total 37% (n=3221)
Node 13
Not in FIU Housing
Dropped 11% (n=251)
Retained 89% (n=2061)
Total 27% (n=2312)
Node 20
Unmet Need > $11,106
Dropped 18% (n=52)
Retained 81% (n=234)
Total 3% (286)

Findings

	Classification Accurracy		
	Overall	Dropped	Retained
Logistic Regression	80%	0%	100%
Decision Tree Training Set	67%	56%	70%
Decision Tree Validation Set	67%	55%	70%

Applying the ‘Rules’

- The model rules (with the same variables and coding) were used to categorize 2014 FTIC students into decision tree nodes

- Since there was not yet Fall 2015 retention data on these students, we used enrollment in Spring 2015 as a proxy measure.

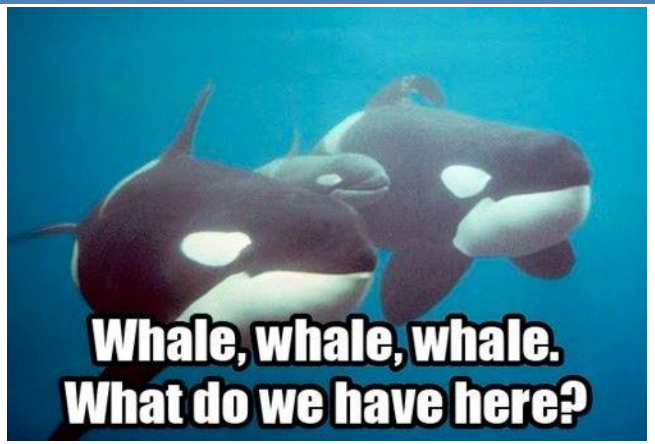
- Students who enrolled in Spring 2015 were counted as “retained,” those who did not enroll in the Spring were counted as “not retained.”

2014 Findings

- Overall, 6% of students of the 2014 FTIC Cohort were not retained into the Spring term

- Of the 227 students who did not register for the Spring, the model had identified 47% of them as at risk of dropout

2014 Results



	Classification Accurracy		
	Overall	Dropped	Retained
Decision Tree Training Set	67%	56%	70%
Decision Tree Validation Set	67%	55%	70%
Rules Predicting 2014	70%	47%	72%

Use of Findings

- Advisors of 2014 freshman students who have enrolled for spring 2015 but are predicted as at high risk of drop out during the first year (do not label “students at-risk”).

- Advisor phone calls, emails, outreach to these students
- Financial aid packages, issues and holds

?? May move them to earlier in the registration cycle

Use of Findings

- Provide **academic and social support** programs for these types of students:
  - including mentoring,
  - smaller classes,
  - summer bridge programs,
  - and major-specific/STEM cohorts.

- The University of Texas at Austin (Tough, 2014) has successfully implemented programs for incoming at-risk students. The programs were so successful that they expanded them to all students.

Lessons Learned

1. We have to start somewhere….use what is available

2. Data preparation is critical (and the most time consuming)! (Thanks Danilo LeSante!)

3. Make sense of your model! Just because the overall classification results show an overall high classification accuracy, doesn't mean the model explains anything.

Lessons Learned

4. Future models should include other variables that have been used in previous research:
   - Date of admission application,
   - whether FIU was a first choice,
   - and by asking students prior to attending FIU if they plan to graduate from FIU

   (Agnihotri & Ott, 2013; Burley, 2007).
   - *Research that added these variables had classification findings that were higher in accuracy (**70% or greater**).*
5. University-wide vs. college specific findings
6. Get in touch with other researchers who are doing this

(I did)!!

Lessons Learned

7. Get ready to revise… revise …revise

Next Steps

- Continue revising pre-college model in 2015 with additional data. FIU's recent purchase and utilization of a Customer Relationship Management (CRM) tool in Admissions will facilitate the collection and use of this type of data.

- Examine how we might use CIRP/NSSE data

- Examine how we might use Institutional variables (IPEDS data)

- Create a model using freshman and sophomore student data and examine 2nd to 3rd year retention patterns (and see if what FIU variables would contribute).

What are Your Questions???

The Team

- Connie Boronat, PhD, Director
- Arlene Garcia, PhD, Assistant Director, Major Maps, Rules
- Dawn Broschard, EdD, Researcher, Statistical analysis
- Tekla Nicholas, PhD, Researcher, Qualitative analysis
- Danilo Le Sante, MA, Researcher, Dashboard, Queries, Analysis

Resources & References

Agnihotri, L. & Ott, A. (2013). ***Who are your at-risk students? How to use data mining target intervention efforts.*** **9th Annual National Symposium, University of Oklahoma, C-IDEA.**

(Author and Date unknown) Decision Trees – What are they? https://support.sas.com

(Author unknown) (2000). ***About CRISP-DM.*** **(Summary of the CRISP-DM Data Mining Guide). Smart Vision Europe http://www.sv-europe.com/crisp-dm-methodology/**

Baizyldayeva, U., Uskenbayeva R., & Amanzholova S., (2013). ***Decision making procedure: Applications of IBM SPSS Cluster Analysis and Decision Tree.***

Bogard, M., Helbig, T., Huff, G. & James, C. (date unknown). ***A comparison of empirical models for predicting student retention.*** **Western Kentucky University. http://www.wku.edu/instres**

Burns, R. & Burns, R. (2009) ***Business Research Methods and Statistics using SPSS*****. SAGE Publishing Inc., Chapter 25 pp. 589-608**

Burley, K. (2007). ***Data mining techniques in higher education research.*** **The example of student retention. Paper presented at 29th EAIR Forum, Austria.**

Delen, D. (2010) ***A comparative analysis of machine learning techniques for student retention management.*** **Decision Support Systems www.elsevier.com/locate/dss**

de Ville, B. & Neville, P. (2013). Decision trees for analytics using SAS Enterprise Miner. Cary, NC: SAS Institute.

IBM Corporation. (2012). IBM SPSS Decision Trees 21 Manual.

Tough, P. (2014, May 15). Who gets to graduate? ***New York Times*****. Retrieved February 6, 2015, from http://www.nytimes.com/2014/05/18/magazine/who-gets-to-graduate.html?_r=0**

Table Version of Tree

Rule	Path (nodes traversed in tree)	Characteristics	% and n Lost	Total Number in Terminal Node
	0	Dropped = 1,860; Retained = 6,874	21%	
1	1/3/7	GPA <=3.05; Aid <=$5,818; Unmet <=$10,247	34% n=305	909
2	1/3/8	GPA <=3.05; Aid <=$5,818; Unmet >$10,247	46% n=165	362
3	1/4/9	GPA <=3.05; Aid >$5,818; HSOV <= 2.81	29% n=304	1,043
4	1/4/10	GPA <=3.05; Aid >$5,818; HSOV > 2.81	20% n=226	1,555
5	2/6/14	GPA > 3.05; Aid >$7,178 (2nd iteration); Housing YES	17% n=157	909

Table Version of Tree

Rule	Path (nodes traversed in tree)	Characteristics	% and n Lost	Total Number in Terminal Node
6	2/5/11/15	GPA > 3.05; Aid <=$7,178; Unmet <=$9,180; Aid (2nd iteration) $1,042	40% n= 40	99
7	2/5/12/17	GPA > 3.05; Aid <=$7,178; Unmet >$9,180 (2nd iteration); Aid <=$17,602 (second iteration)	28% n=100	252
8	2/5/12/18	GPA > 3.05; Aid <=$7,178; Unmet >$9,180	45% n=36	44
9	2/6/13/20	GPA > 3.05; Aid >$7,178; Unmet >$11,106; Housing NO	18% n=52	286

Table Version of Tree

Rule	Path (nodes traversed In tree)	Characteristics	% and n Lost	Total Number in Terminal Node
10	2/5/11/16/21	GPA > 3.05; Aid <=$7,178; Unmet <=$9,180; Aid >$1,042 (2nd iteration); Black and White	27% n=61	228
11	2/5/11/16/22	GPA > 3.05; Aid <=$7,178; Unmet <=$9,180; Aid >$1,042 (2nd iteration); Hispanic and Asian and Other	13% n=161	1,228
12	2/6/13/19/23	GPA > 3.05; Aid >$7,178; Unmet <=$11,106 Housing NO; STEM	9% n=165	1,766
13	2/6/13/19/24	GPA > 3.05; Aid >$7,178; Unmet <=$11,106; Housing NO; Engineering and Nursing	13% n=34	260