Cognitive Insights: Structured & Unstructured Data in Predictive Analytics

Ying Zhou, zhouy14@ecu.edu
Margot Neverett, neverettm@ecu.edu
Hanyan Wang, wangh17@ecu.edu

Office of Institutional Planning, Assessment and Research
East Carolina University

I. Cognitive Insights Project Overview

- **Purpose of the study**
- **Data sources**

Project Overview

Phase I Model

FTFT cohort

2012 and 2013

One-year retention

Four-year graduation

Phase II Model

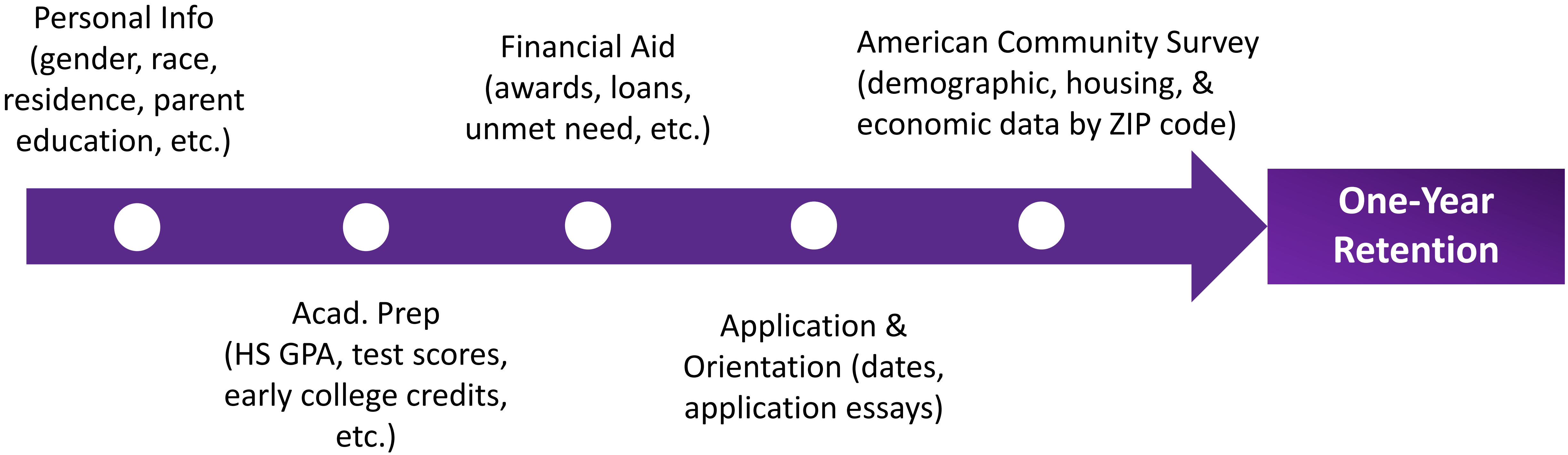
FTFT cohort

2015, 2016, and 2017

One-year retention

2^{nd} – 3^{rd} year graduation

Phase II: One-Year Retention Multinomial Logistic Regression Model



Partnership with IBM

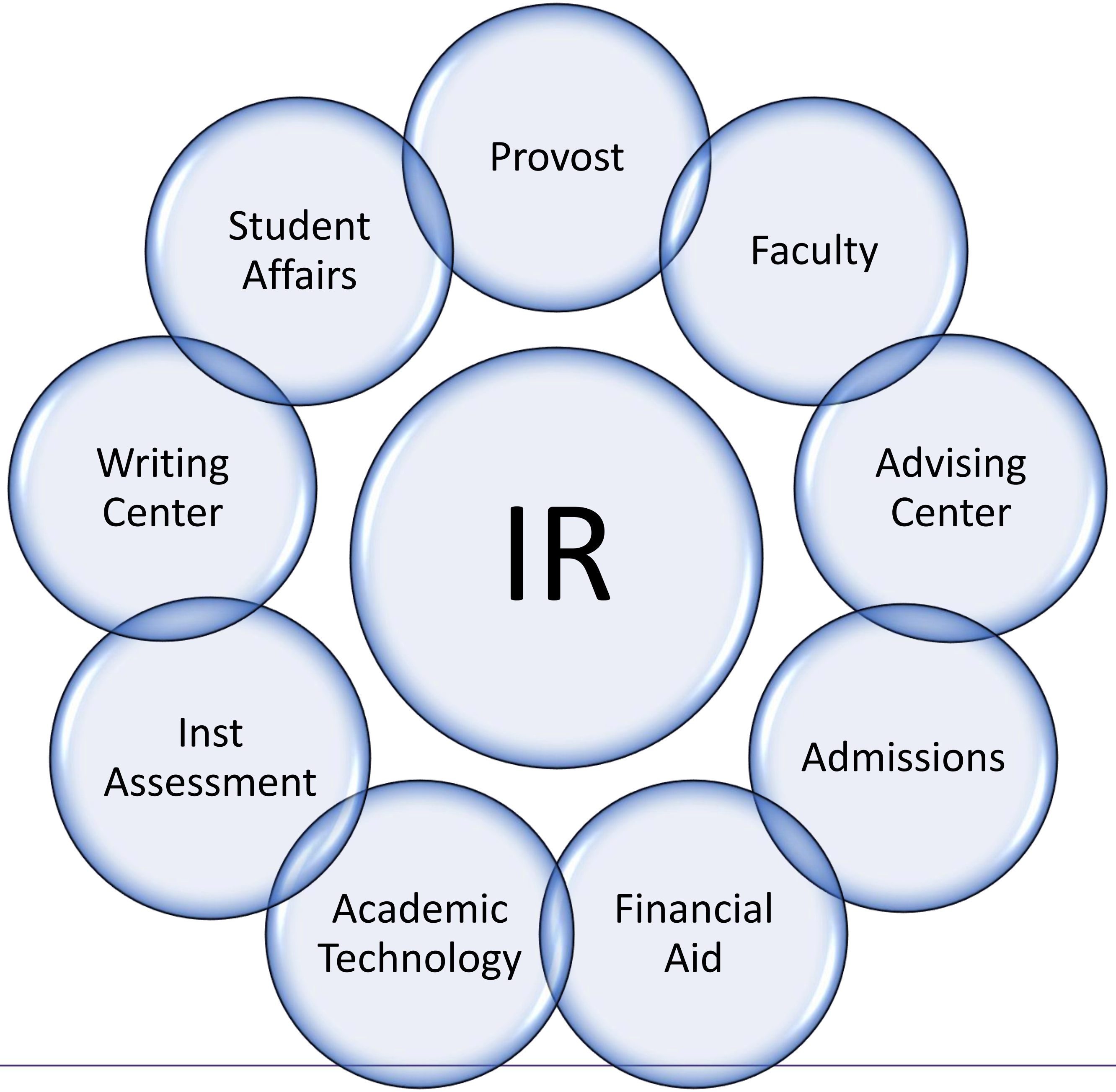
Diverse Data Sources

- Multiple cohorts of students
- Multiple semesters' data
- Diverse Structured Data
  - Banner SIS
  - Blackboard
  - Academic support services
  - Student Affairs
  - American Community Survey

Watson Technology

- Unstructured Data
  - Application essays (Phase II only)
  - Starfish faculty comments
  - Student comments from course evaluations
- Watson Natural Language Understanding
  - Key words
  - Sentiments and Tones
  - Personality

ECU Collaborators



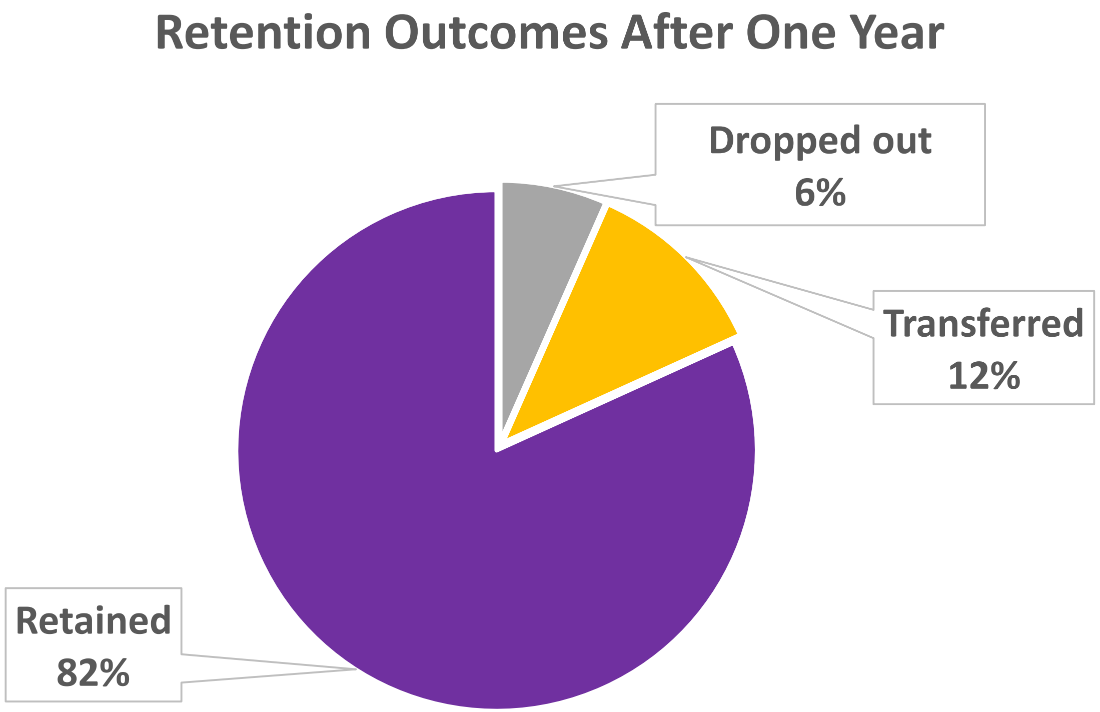
II. Enhanced Phase II One-Year Retention Model Findings

- Student population
- Variables examined
- Selected results from structured data
- Selected results from unstructured data

ECU

Phase II One-Year Retention Model

Total Population: 12,786 First-time Full-time Students (Fall 2015, 16 and 17)



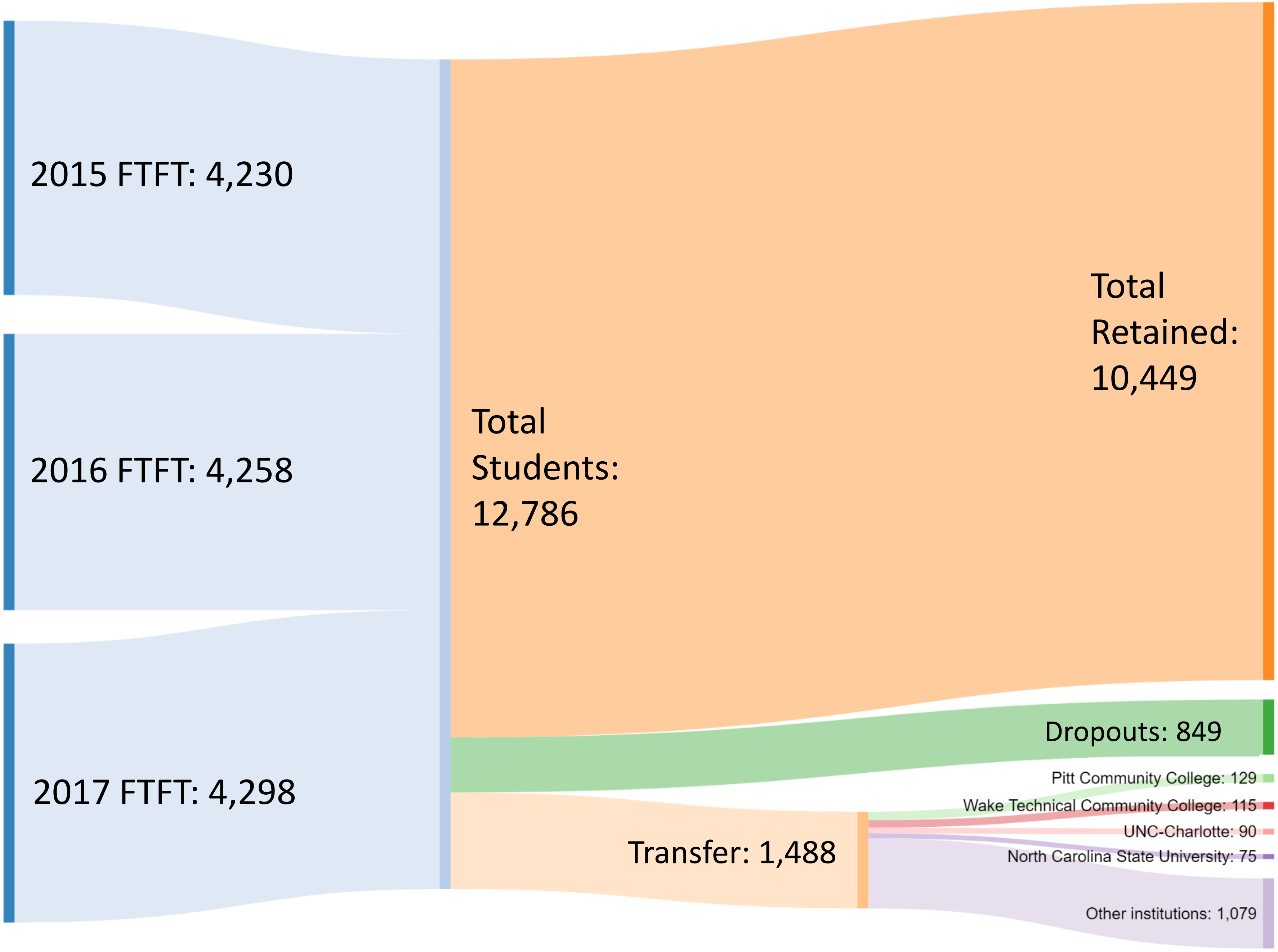
Possible Predictor Variables (170)

- ECU (46 predictors)
  - Admissions
  - FAFSA
  - Orientation
- American Community Survey (77 predictors)
  - Demographic
  - Housing
  - Economic
- Watson (47 predictors)
  - Tone and Personality
  - Keywords

10 included in the final model

4230 from 2015 + 4258 from 2016 + 4298 from 2017 → 12,786 total students

Sankey diagram of model population* and retention outcomes after one year



* First-time full-time students entered ECU in summer and fall semesters

Top Transfer Institutions

(Note: 1,488 of 12,786 students transferred out after one year)

Four Year Institutions

- UNC - Charlotte, 90
- North Carolina State University, 75
- UNC - Wilmington, 68
- Appalachian State University, 67
- UNC - Greensboro, 49

Two Year Institutions

- Pitt Community College, 129
- Wake Technical Community College, 115
- Cape Fear Community College, 68
- Central Piedmont Community College, 50
- Guilford Technical Community College, 23

Comparison: Retained, Dropouts, and Transfers

	Retained	Transferred	Dropped Out
Count	10,449	1488	849
Avg. Weighted HS GPA	3.83	3.62	**3.50**
% NC Rural counties (Tiers 1 and 2)	28%	25%	**35%**
Avg. Unmet Need ($)	3,211	5,159	**6,368**
Avg. distance between home and ECU (miles)	131	**164**	144
% from East of I-95	36%	31%	**42%**
% female	60%	58%	**45%**

Variable Selection

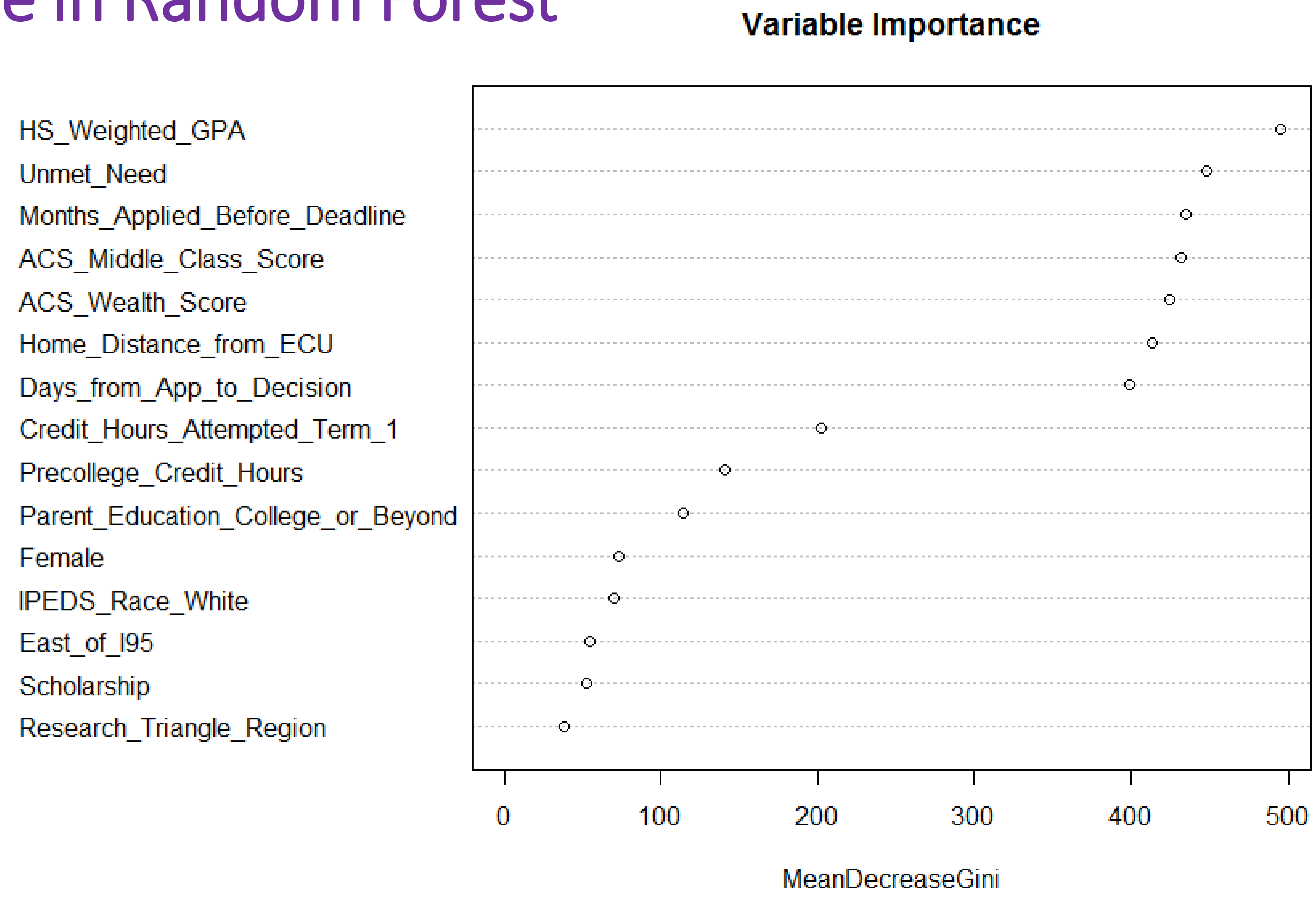
IBM

- Method
  - Cross validation
  - Correlations
  - Backward selection
- Using automated approach

ECU

- Method
  - Careful examination of all variables in IBM's model
  - Factor analysis (American Community Survey)
- Trying to create a simpler model while keeping the same level of accuracy

Strongest Predictors of Dropout and Transfer Risks by Variable Importance in Random Forest



Results from Structured Data

Weighted GPA is a key driver of dropout and transfer risk

Each additional point in weighted high school GPA

- 69%

Dropout risk

- 45%

Transfer risk

Unmet need is a key driver of dropout and transfer risk

Every $1,000 increase in unmet need

+ 8%	+ 6%
Dropout risk	Transfer risk

ACS Variable – “Middle Class” Score

Computed at home ZIP code level from:

- % total household income between $100-200k
- % house value (owner-occupied units) between $200-500k

Findings

After controlling for all other variables in the model:

- Students with a higher “middle class” score are **less likely** to drop out
- “Middle class” score does not have a significant impact on transfer risk

ACS Variable – “Wealth” Score

Computed at home ZIP code level from

- % total household income >$200k
- % house value (owner-occupied units) between $500k – 1 million
- % house value (owner-occupied units) > $1 million

Findings

After controlling for all other variables in the model:

- Students with a higher “wealth” score are **less likely to drop out or transfer**

Selected Results: Dropout Risk

After controlling for all other variables in the model:

- Students who applied early are less likely to dropout (**every month reduces the dropout risk by 13%**).
- For each extra day ECU took to process an application, the dropout risk **increases by 0.2%.**
- For each extra credit hour a student attempted in the first semester, the dropout risk **reduces by 16%.**
- Each college-educated parent (possible categories: **both** parents with college degree, **one** parent with college degree, and **neither** parent with college degree) reduces the dropout risk by **16%**.

Selected Results: Transfer Risk

After controlling for all other variables in the model:

- Students who applied earlier are less likely to transfer (**every month reduces the transfer risk by 5%**).
- Students whose home is far away from ECU are more likely to transfer **(every 100 miles increases the transfer risk by 6%)**.
- For each extra credit hour a student attempted in the first semester, the transfer risk **reduces by 11%.**
- For each extra pre-college credit hour a student earned, the transfer risk **reduces by 0.7%.**

ECU

Results from Unstructured Data

Application Essays

- 72% Submitted
- Variables Tested in the Model
  - Submission (Y/N)
  - Count of all words
  - Count of words after removing stop words
  - Average number of letters in a word
  - Key Words
  - IBM Watson variables

IBM Watson Personality Insights

Big 5 Personality Types	Needs	Values
• Openness	• Harmony	• Helping others
• Conscientiousness	• Curiosity	• Tradition
• Extraversion	• Love	• Hedonism
• Agreeableness	• Challenge	• Achieving Success
• Neuroticism	• Liberty	• Open to change

ECU

Results

- Four of the 47 variables computed by Watson were included in the final model
  - Hedonistic personality
  - key words: East Carolina University, school, and work
- Students with the strongest hedonism score (score=1) are almost three times as likely to drop out or transfer than those with the score of 0 (not statistically significant).
- Hedonism scores range 0.00028 to 0.52, with the mean=0.06 and 90th percentile=0.11.
- Students with application essays that contained the word “work” have a slightly higher transfer probability (statistically significant)

More Quotes of Hedonism

High Score

- … throbbing with nerves' screaming my name…
- … uncontrollable tears, frustrated, embarrassed
- I miss the warm, glowing…sun .. (a)s well as the joyous memories conjured
- Wide eyed and excited I found myself at the front of the crowd to watch. This is too awesome.

Low Score

- …This class has really taught me to work hard in everything I do, and to stay be humble.
- …me into a productive student who cares about (my) grades and the campus around (me).
- ...help those who are in need of it by using my knowledge of the Spanish language anytime I can….I imagine…being a successful and hard-working student who contributes positively to a new community.

III. Predictive Analytics: Lessons Learned

- **Challenges of Predictive Analytics**
- **Potential Use of the Results**

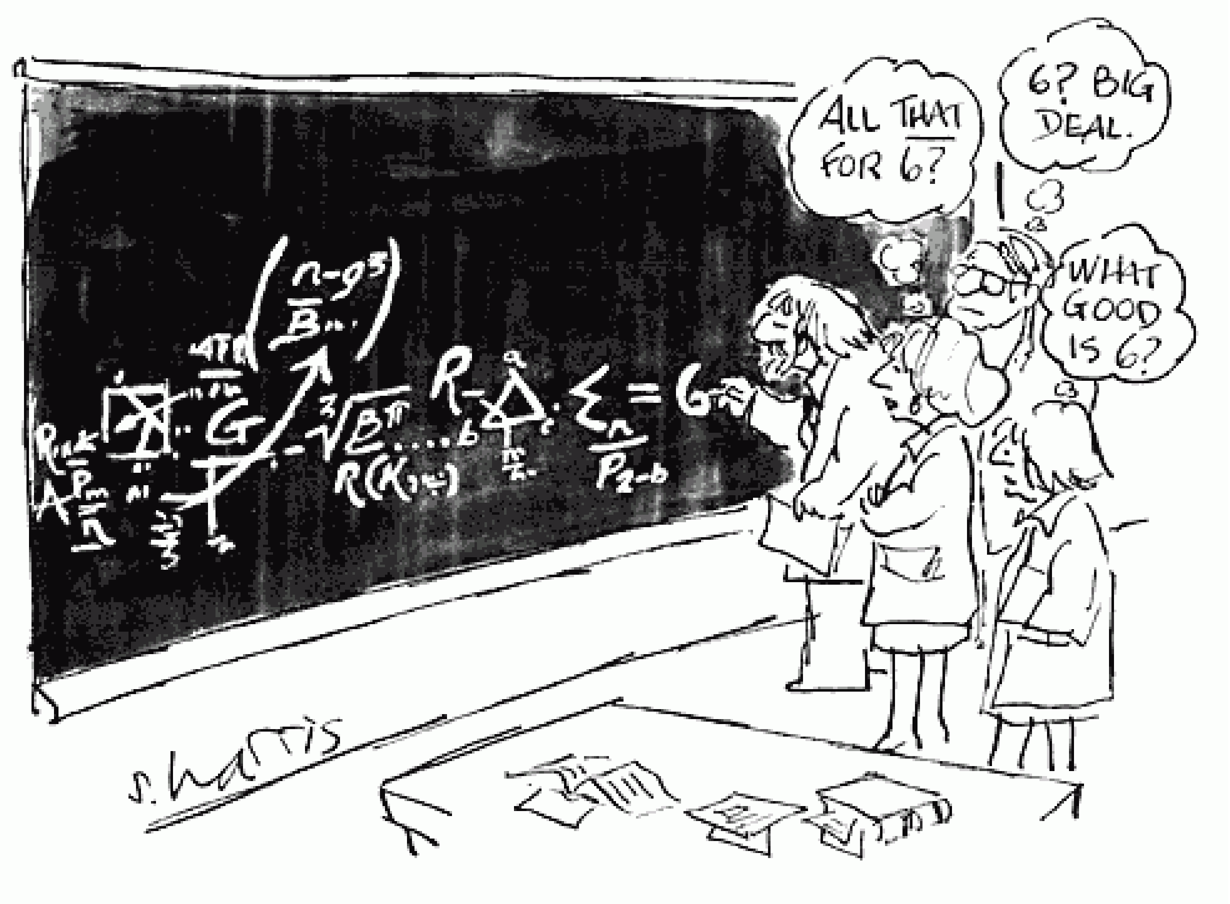
Challenges

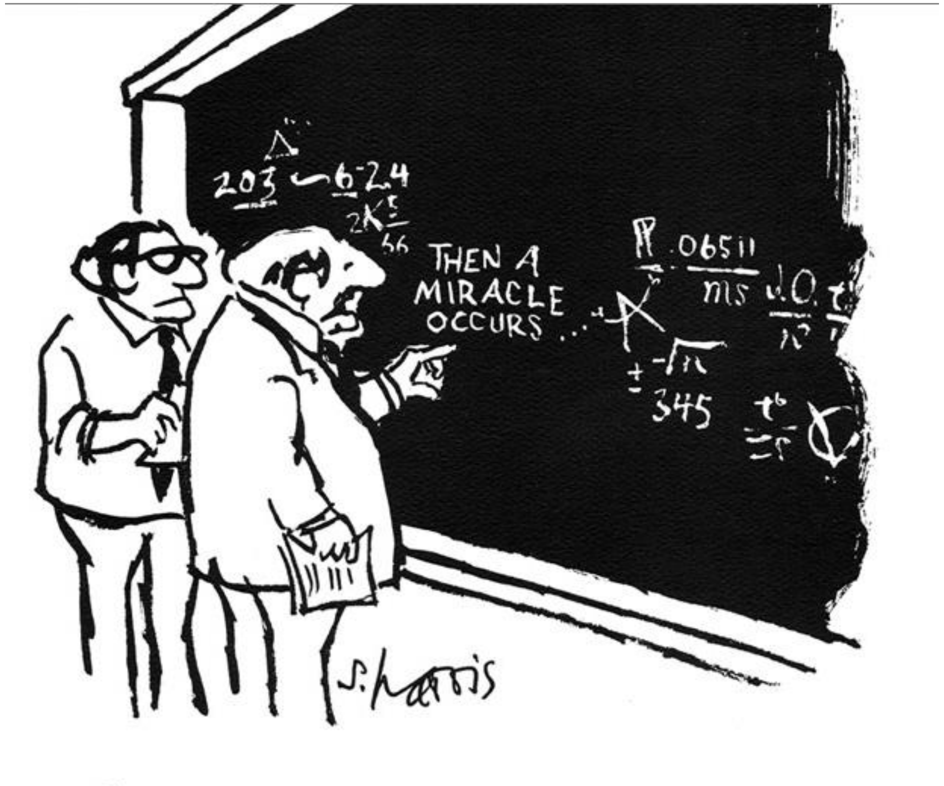
- Multiple data sources used in the study are stored outside of Banner.
- Data integration is labor intensive and variables are defined inconsistently.
- Difficulties in distinguishing transfer vs. dropout risks.
- Because of the complexity of the study, interpretation and communication of the results can be difficult.

Challenges with Unstructured Data

- Watson variables didn't improve any model developed in the study in both Phase I and Phase II.
- Practical Question: what application essays reflect about the applicants?
- Missing value imputation (e.g., no comment in course evaluation, no application essay) is a major issue.
- Text analytics:
  - Could further develop the dictionary for admission essays and course evaluations
  - Different text analytics packages are different

Challenges with Communicating Results



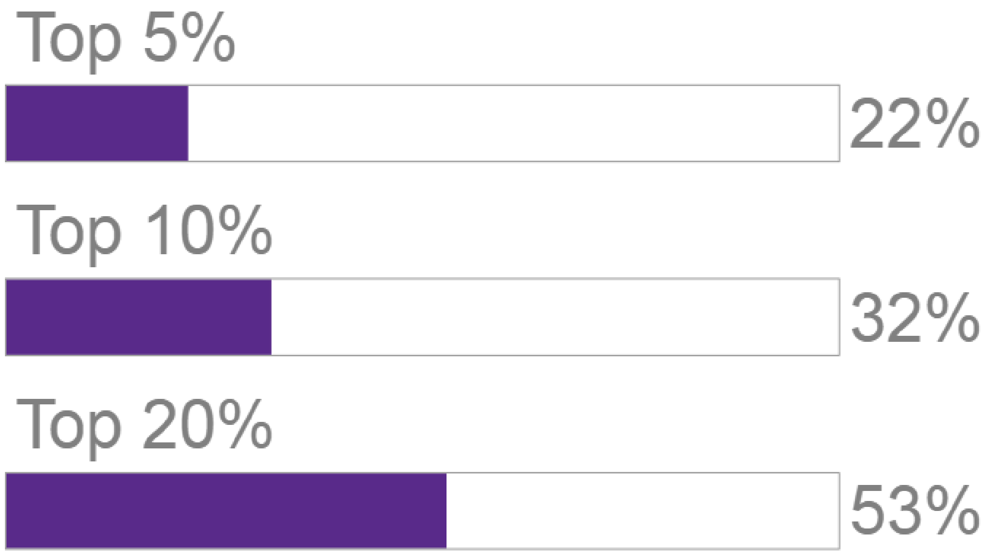


"I THINK YOU SHOULD BE MORE EXPLICIT HERE IN STEP TWO."

Use of Predictive Analytics for Student Outreach

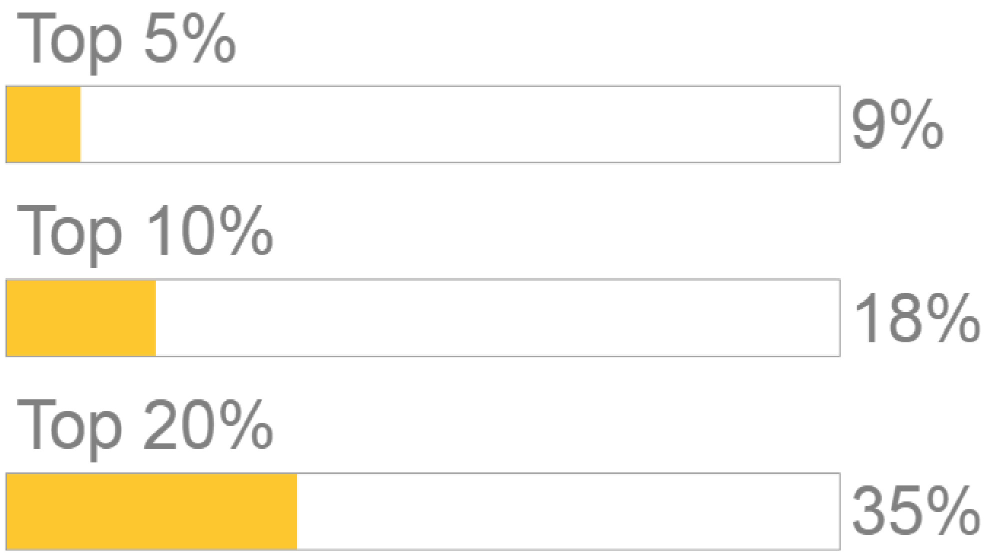
Mitigating Dropout Risk

Outreach to students in the top 20% highest dropout risk will capture more than half of the dropout students.



Mitigating Transfer Risk

Outreach to students in the top 20% highest transfer risk will capture more than one third of the dropout students.



Use of Predictive Analytics Results: Based on Feedback from Stakeholders

- Student outreach before signs of disengagement
  - Designated staff (e.g., advisors) for at-risk student populations
  - Different approaches to mitigating transfer and drop-out risks

- Actions to address unmet need
  - Scholarship: ECU created one thousand $1,000 scholarships for four years for incoming freshmen
  - Financial literacy program (College of Business)
  - Increase on-campus student employment opportunities

Appendix

Multinomial Logistic Regression Results

```
Call:
multinom(formula = Retained_ECU ~ Days_from_App_to_Decision + 
    Months_Applied_before_Cutoff + UNMET_NEED + ORIG_WEIGHTED_GPA + 
    Student_Distance_from_ECU_100mile + MiddleClass + Wealth1 + 
    Hours_attempted_term1 + FD_hours + PARENT_College.or.beyond, 
    data = c1)

Coefficients:
  (Intercept) Days_from_App_to_Decision Months_Applied_before_Cutoff UNMET_NEED ORIG_WEIGHTED_GPA Student_Distance_from_ECU_100mile
O    5.161117              0.0018632196                  -0.12334063 0.08058654        -1.1847407                       0.008263265
T    2.079062              0.0001549986                  -0.04962247 0.05850022        -0.5952626                       0.060813255
  MiddleClass    Wealth1 Hours_attempted_term1     FD_hours PARENT_College.or.beyond
O -0.28227618 -0.1593743            -0.1793481 -0.005821627              -0.17477120
T  0.01819601 -0.1100769            -0.1170162 -0.007436156               0.02546239

Std. Errors:
  (Intercept) Days_from_App_to_Decision Months_Applied_before_Cutoff  UNMET_NEED ORIG_WEIGHTED_GPA Student_Distance_from_ECU_100mile
O   0.4958092              0.0009172535                   0.02377462 0.006060997        0.08552207                        0.02624198
T   0.3918767              0.0007569697                   0.01868733 0.004972967        0.06273586                        0.01591870
  MiddleClass    Wealth1 Hours_attempted_term1    FD_hours PARENT_College.or.beyond
O  0.06529550 0.07779242            0.02702071 0.003543900               0.04838727
T  0.04699038 0.05418676            0.02168642 0.002831437               0.03818858

Residual Deviance: 13735.98 
AIC: 13779.98
```

Multinomial Logistic Regression Results

```
> p
   (Intercept) Days_from_App_to_Decision Months_Applied_before_Cutoff UNMET_NEED ORIG_WEIGHTED_GPA Student_Distance_from_ECU_100mile
O 0.000000e+00                 0.0422243                 2.126650e-07          0                 0                      0.7528472416
T 1.124271e-07                 0.8377581                 7.921275e-03          0                 0                      0.0001333222
   MiddleClass    Wealth1 Hours_attempted_term1    FD_hours PARENT_College.or.beyond
O 0.0000153882 0.04049030          3.192024e-11 0.100441356             0.0003039324
T 0.6985871538 0.04221078          6.820852e-08 0.008632283             0.5049291887
```

```
> exp(coef(m1))
  (Intercept) Days_from_App_to_Decision Months_Applied_before_Cutoff UNMET_NEED ORIG_WEIGHTED_GPA Student_Distance_from_ECU_100mile
O  174.359118                  1.001865                    0.8839625   1.083923         0.3058255                          1.008298
T    7.996963                  1.000155                    0.9515886   1.060245         0.5514178                          1.062700
  MiddleClass   Wealth1 Hours_attempted_term1  FD_hours PARENT_College.or.beyond
O   0.7540654 0.8526772             0.8358149 0.9941953                0.8396491
T   1.0183626 0.8957652             0.8895708 0.9925914                1.0257893
```

Project Overview

Phase I

Phase II

One-Year Retention

Use pre-college data to identify students most at risk **before matriculation** or before typical signs of disengagement appear

Four-Year Graduation

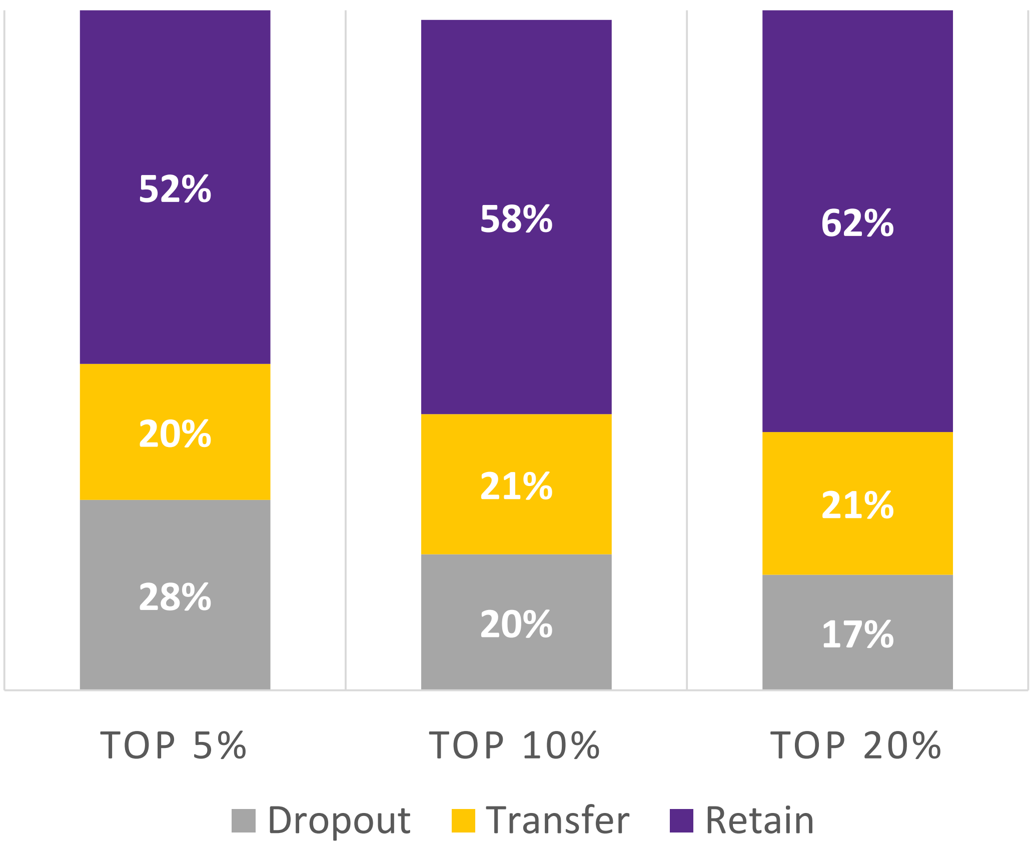
Identify characteristics of students **at the end of the second spring semester** who are the least likely to graduate in four years

2nd – 3rd Year Retention

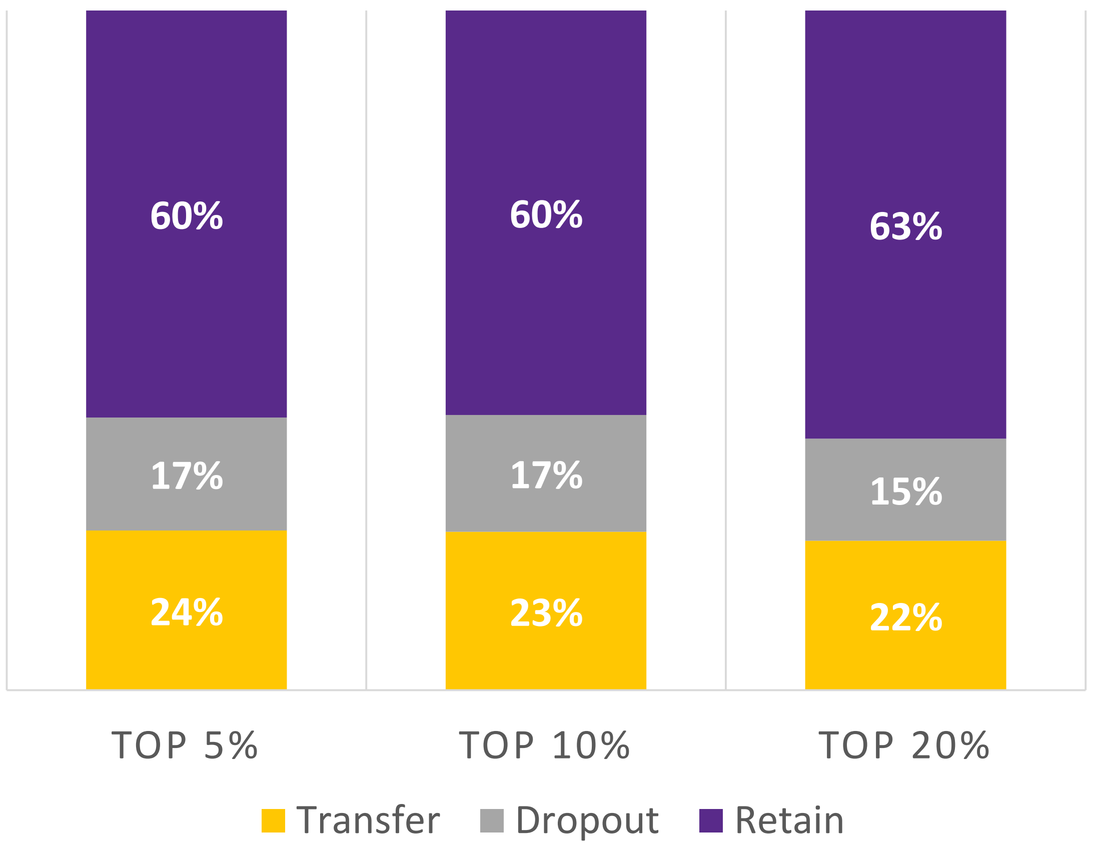
Identify characteristics of students **at the end of the second fall semester** who are most likely to be retained to the third year

Accuracy of Predictive Analytics Results

Among the students with top 10% highest dropout risk scores, 20% of them dropped out and 21% of them transferred.

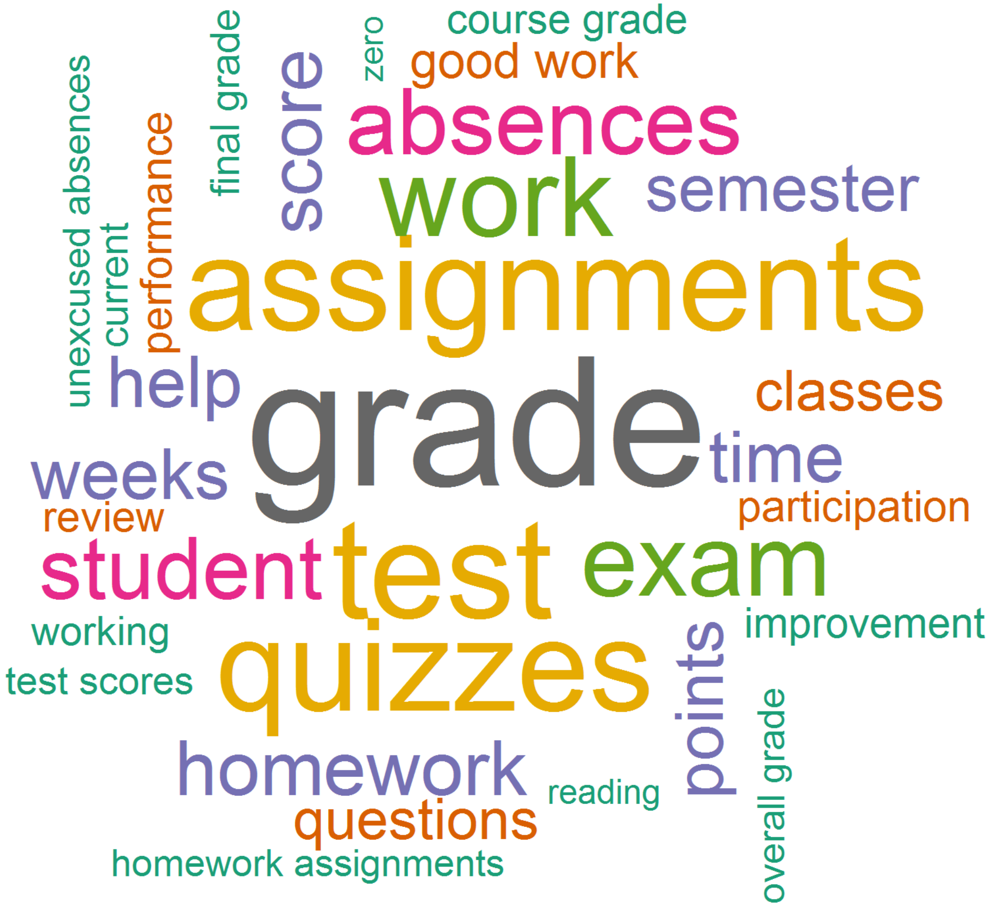


Among the students with top 10% highest transfer risk scores, 23% of them transferred and 17% of them dropped out.



Keywords Identified by Watson: Phase I Study

Starfish



Course Evaluations