Predictive Analytics: Lessons Learned from Retention Studies

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

Office of Institutional Planning, Assessment and Research

Source: Cognitive Insights Phase 2 Retention Model Readout, IBM Global Business Services, Dec 7, 2018.

I. Cognitive Insights Project Overview

- Purpose of the study
- Data sources
- ECU participants
- Project timeline

Purpose of the Study

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

ECU

Partnership with IBM

Diverse Data Sources

- Multiple cohorts of students
- Multiple semesters' data
- Diverse data sources
  - Recruiter
  - Banner
  - Blackboard
  - Academic support services
  - Student Affairs
  - Student surveys
  - 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

Watson Tone Variables (Phase I Study)

- Watson assigned a tone score for each Starfish and course evaluation comment. Then an overall tone scores (mean and standard deviation) were calculated for each student.

Joy	Analytical
Fear	Agreeableness
Anger	Confidence
Sadness	Conscientiousness
Disgust	Openness
Emotional Range	Tentative
Extraversion	

Example of Watson Analyses: Course Evaluation Comments

Tone Analysis

This course evaluation comment has a strong positive tone. Watson assigns tone scores closer to 1 for positive tone.

TONE ANALYSIS

AGREEABLENESS SCORE

0.998

0.000

This course is well taught and very interesting. I greatly enjoyed going. I wouldn't change anything about this class. It was great! Maria is a great professor and I loved that she found ways to show us what we were learning in really life situations.

NOTHING AT ALL [redacted] IT IS THE WORST THING. THE ONLINE GRADING AND TEST SCHEDULING HAS CAUSED ME TO FAIL THE COURSE the [redacted] is terrible. communication issues [redacted] have ruined my grade

This course evaluation comment has a strong negative tone. Watson assigns tone scores closer to 0 for negative tone.

Keyword Analysis

Using word patterns from all comments, Watson extracts keywords from each comment.

KEYWORD ANALYSIS

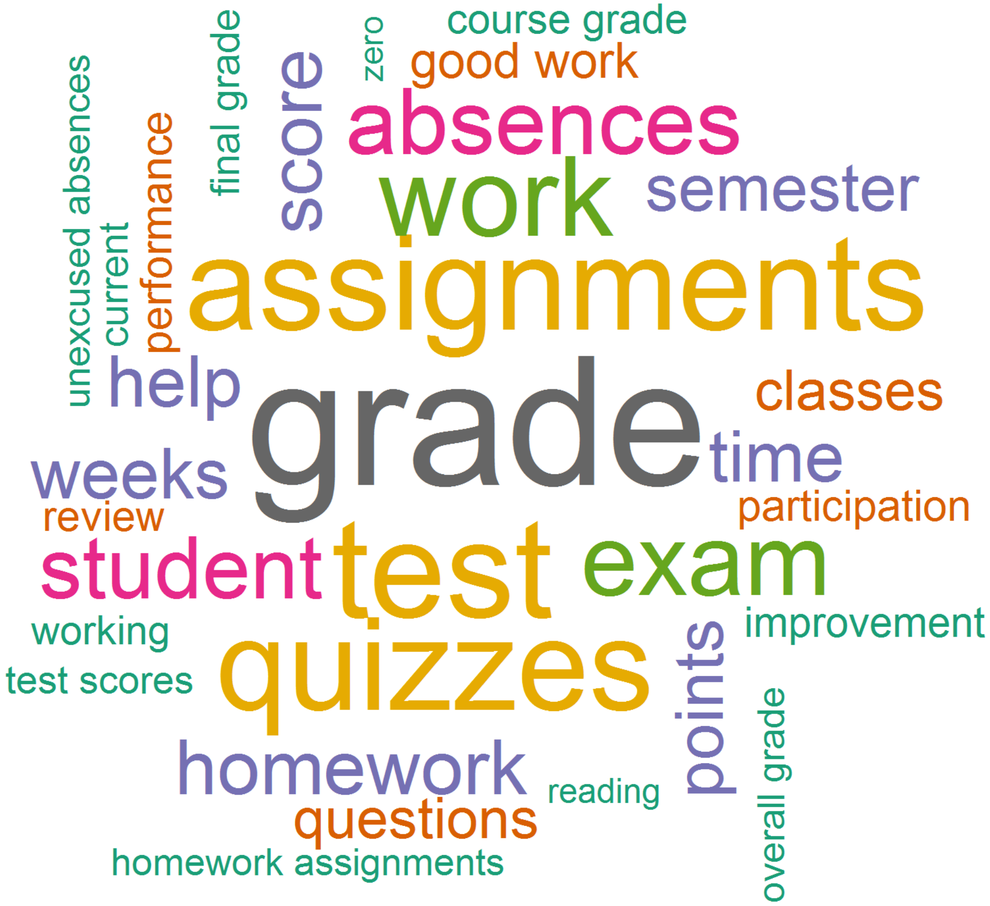
- change
- great professor
- learning
- life

- online
- test
- grade

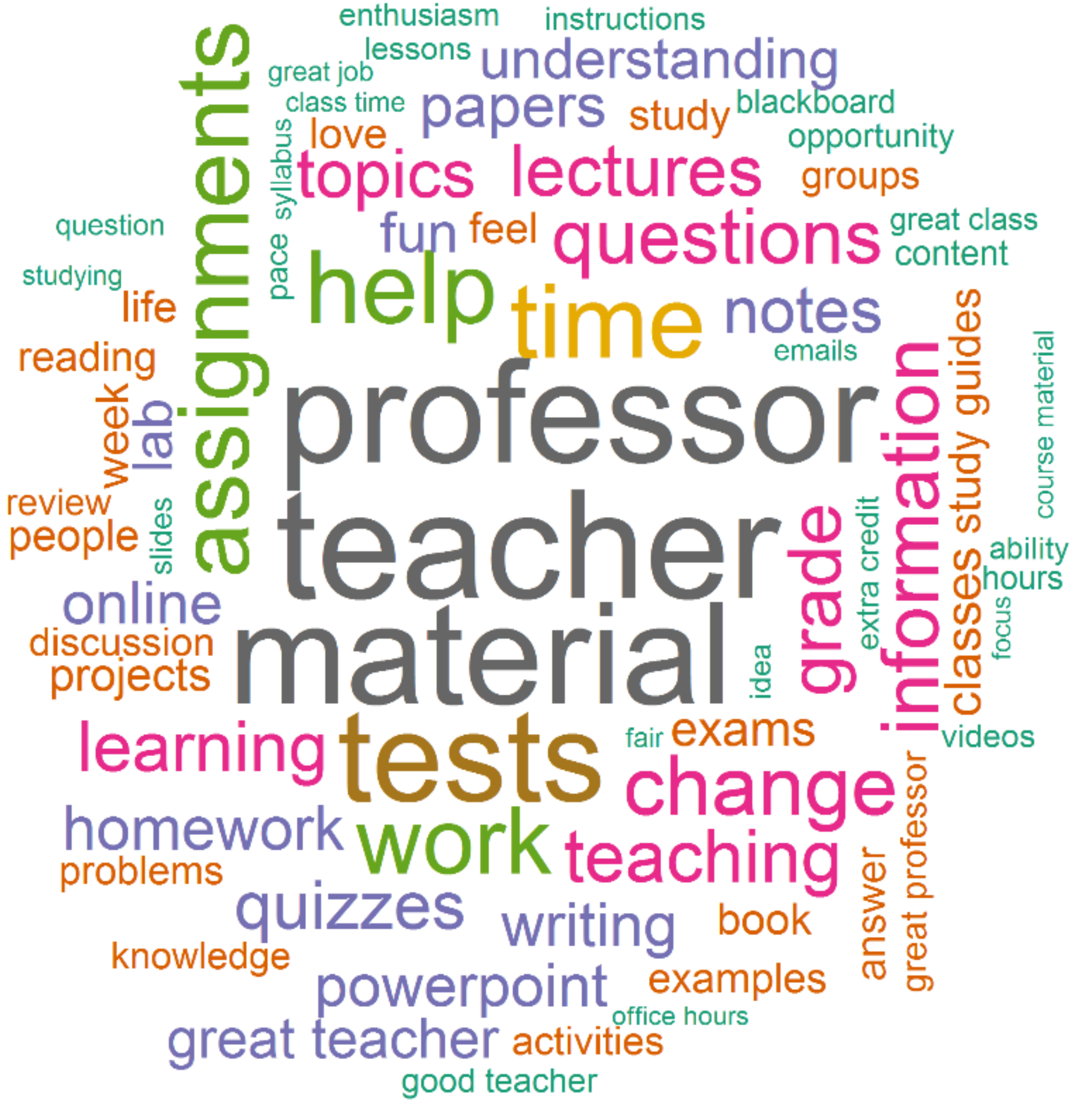
Keywords are converted to new true/false variables to measure use of common words.

Keywords Identified by Watson: Phase I Study

Starfish



Course Evaluations

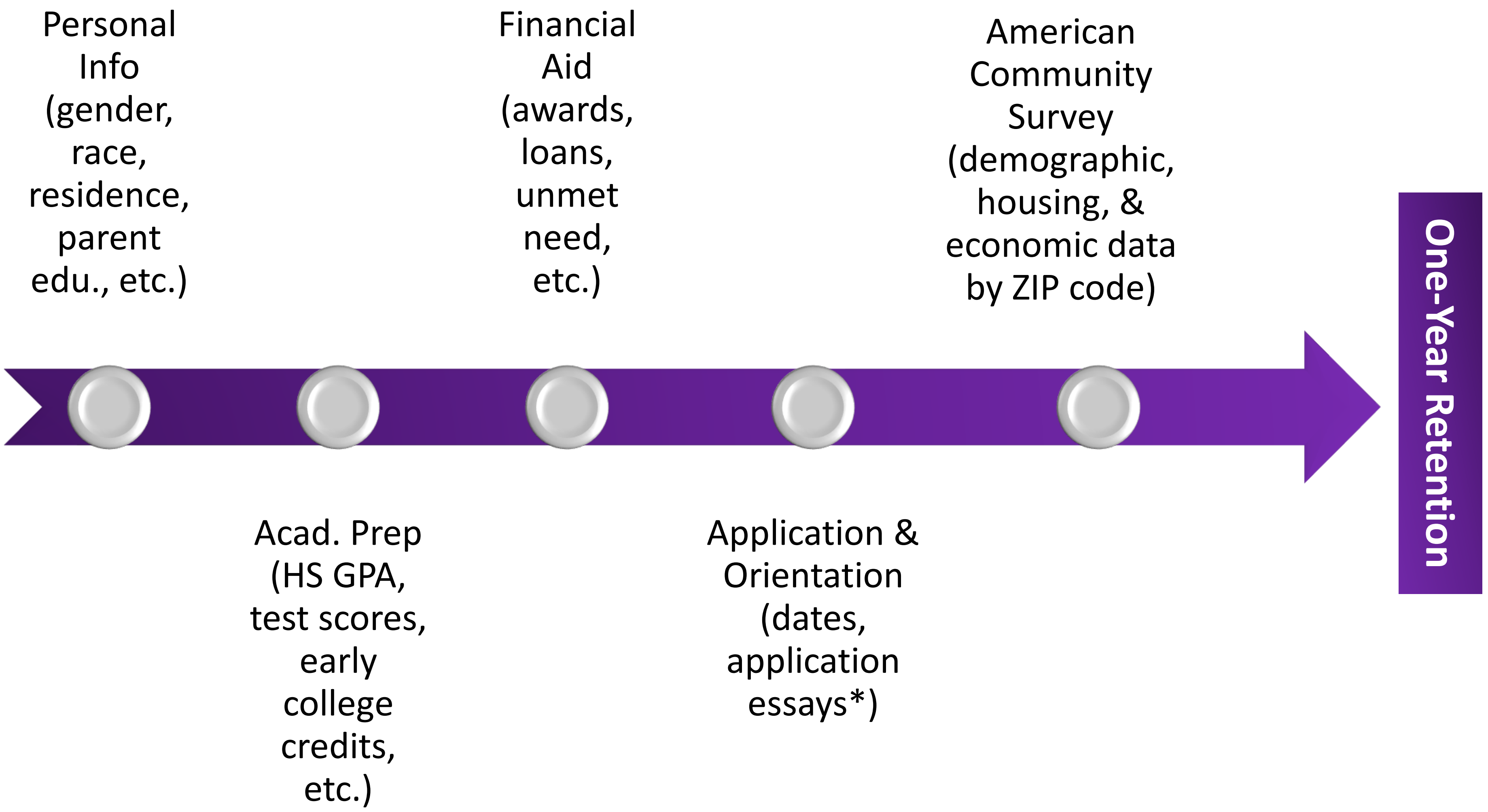


ECU Contributors

- Nicole Caswell
- Kyle Chapman
- Elizabeth Coghill
- James Coker
- Wendy Creasey
- Allison Dannell
- Kristen Dreyfus
- John Fletcher
- Jayne Geissler
- Kathy Hill
- Jerri Hvastkovs
- Beverly King
- Yihui Li
- Chris Locklear
- David Meredith
- Margot Neverett
- Annette Peery
- Julie Poorman
- Amy Shannon
- Scotty Stroup
- John Trifilo
- Jeremy Tuchmayer
- Ruben Villasmil
- Scott Wade
- Hanyan Wang
- Qiang Wu
- Ying Zhou

Phase I and II: One-Year Retention Models

- Phase I: Fall 2012 and Fall 2013 cohorts of **<u>8,416</u>** first-time full-time students
- Phase II: Fall 2015, 2016, and 2017 cohorts of **<u>12,786</u>** first-time full-time students



* Included in Phase II study only.

Phase II: 2nd to 3rd Year Retention Model

Pre-college Data

First Three Reg. Semesters at ECU

Retention Model Variables

Academics: credits, courses, grades, GPAs, bottleneck courses, major, academic standing

Financial Aid: awards, loans, unmet need, etc.

Writing and Tutoring Center Visits

Blackboard Usage: logins, posts, etc.

Student Life: LLC, student conduct, etc.

Comments: Starfish, course evals, & conduct case info

To predict 3rd-year retention

Phase II Timeline: Oct. 2018 – Feb. 2019

Develop Scope of Work → **Select Predictors** → **Build Models** → **Evaluate Results** → **Disseminate Results**

Develop Scope of Work
- Develop research questions
- Assess availability and quality of data
- Engage ECU stakeholders

Select Predictors
- Find strongest predictors of retention
- Remove strongly correlated predictors

Build Models
- Test alternative algorithms
- Assess accuracy & stability of algorithms

Evaluate Results
- Identify drivers of retention
- Evaluate predictive value of the model

Disseminate Results
- Present results
- Discuss implications with stakeholders

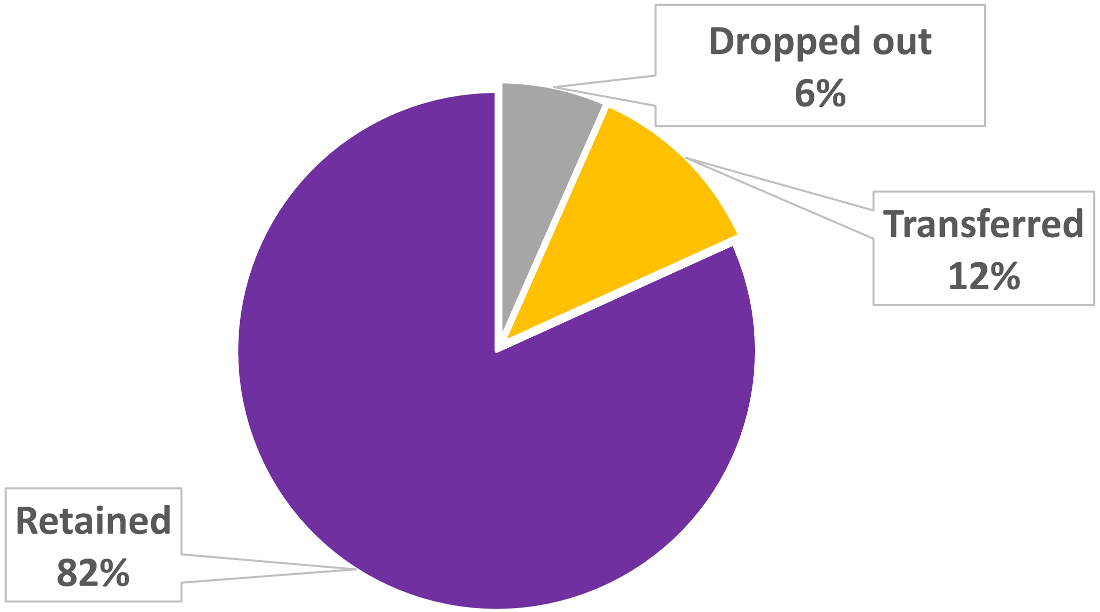
II. Phase II Retention Model Findings

- Student population
- Variables examined
- Selected results
- Caution:
  - The model has very limited power in predicting dropout/transfer outcomes.
  - Due to the complexity of the study, IPAR is still validating the results.
  - End of first semester might be a better checkpoint to predict dropout/transfer outcomes.

One-Year Retention Model

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

Retention Outcomes After One Year



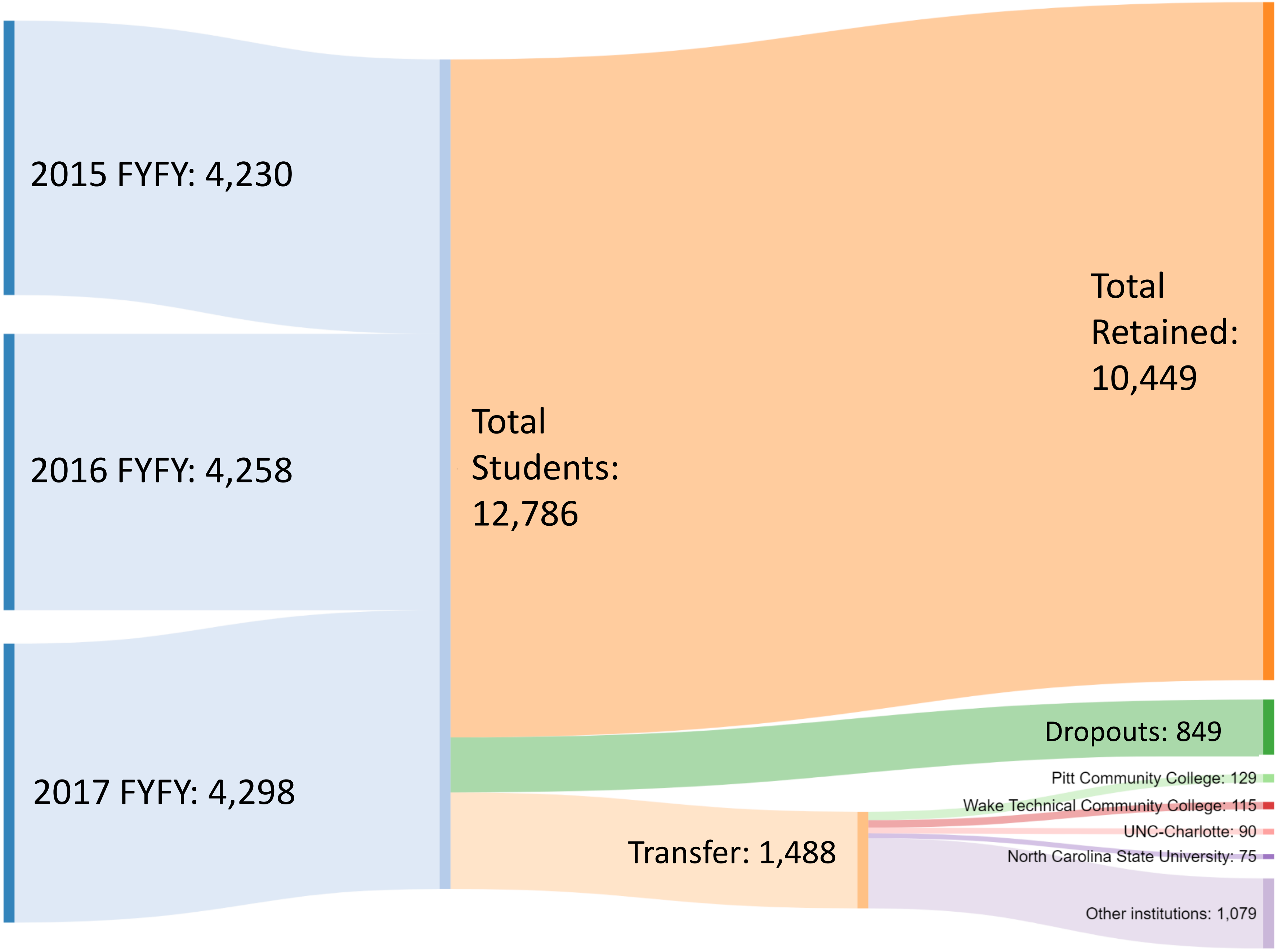
Possible Predictor Variables (170)

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

32 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

ECU

Top Transfer Institutions

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

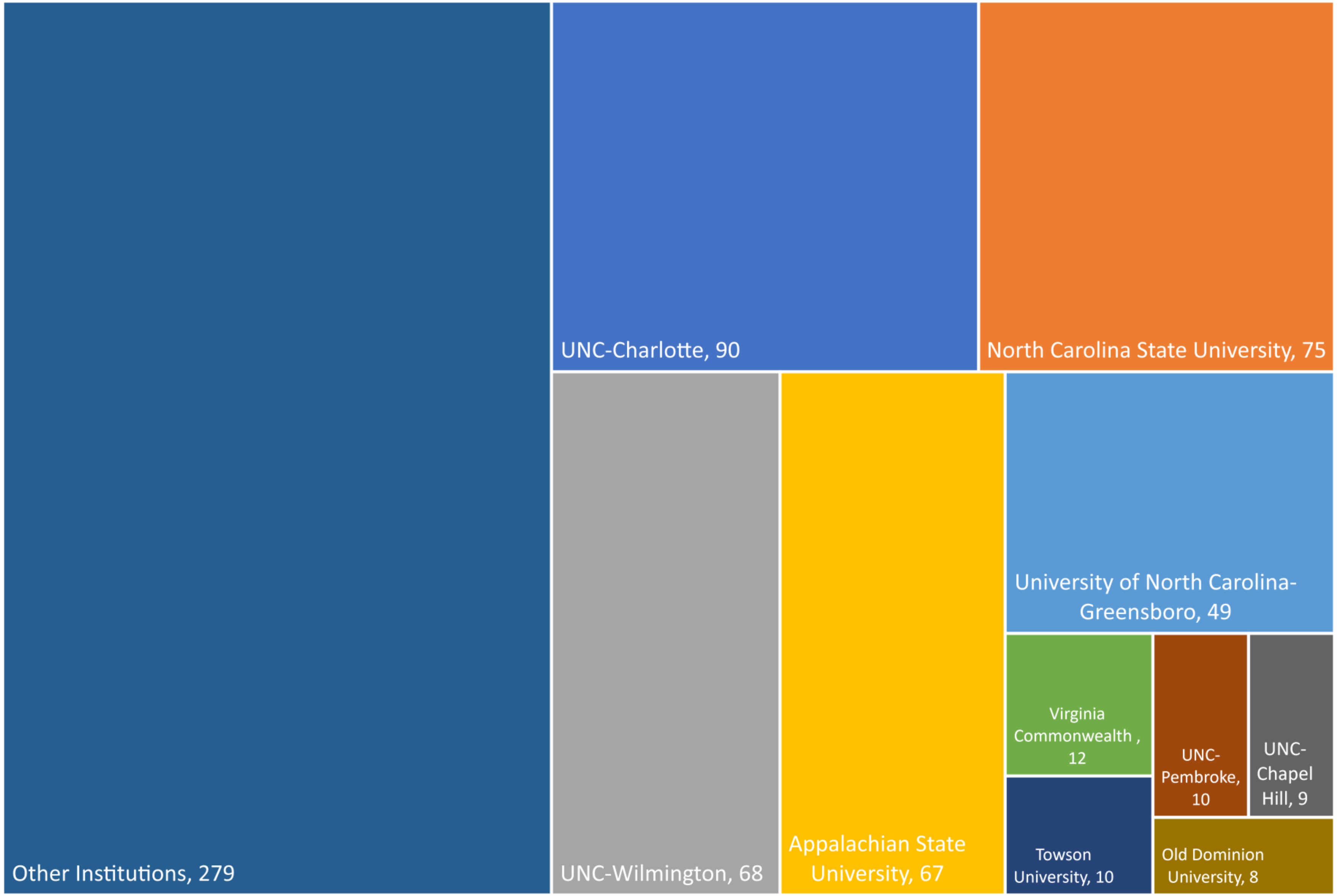
Four Year Institutions

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

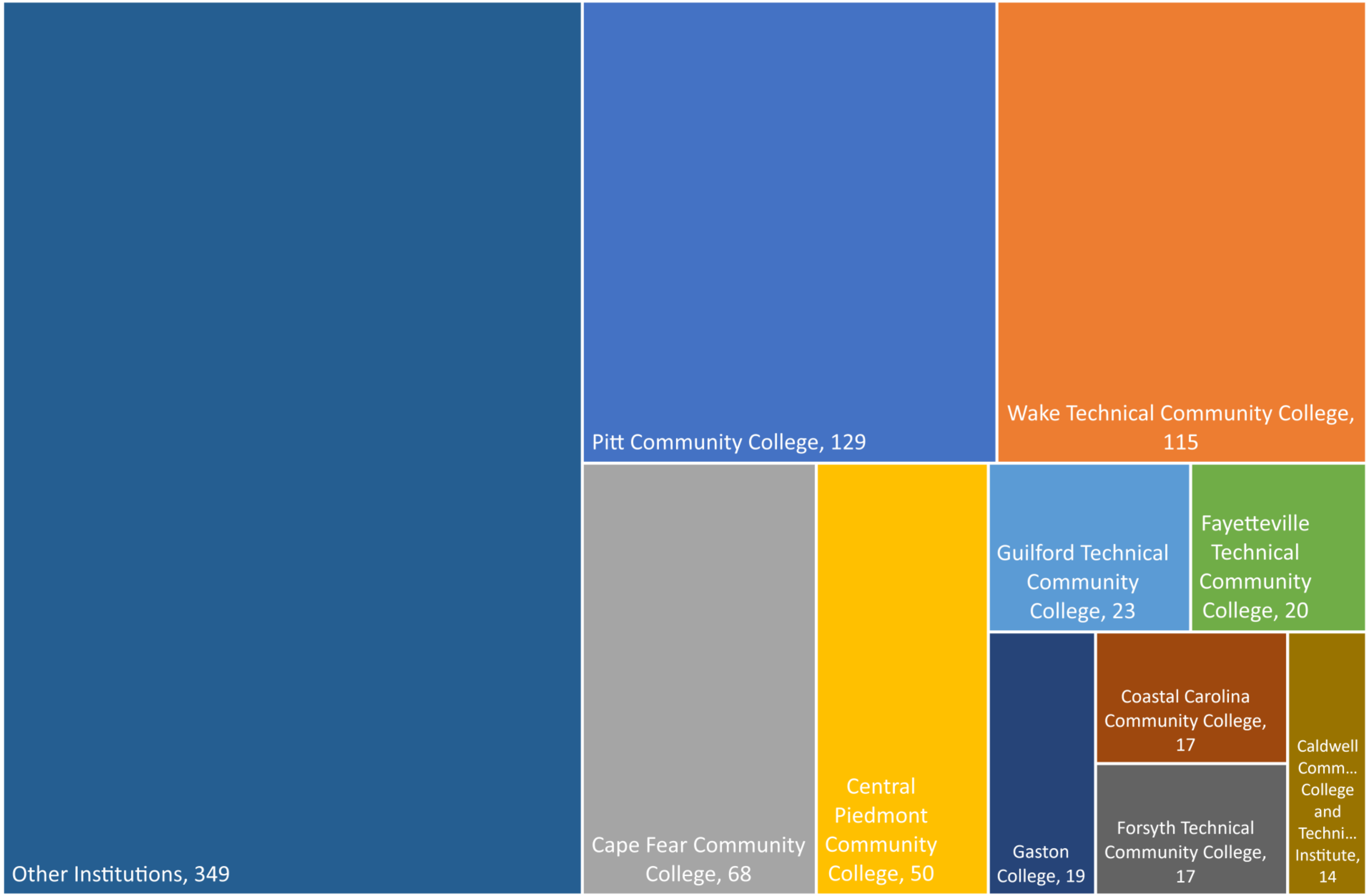
Two Year Institutions

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

Top Transfer Institutions: Four-year Institutions (667 students)



Top Transfer Institutions – Two-Year Institutions (821 students)

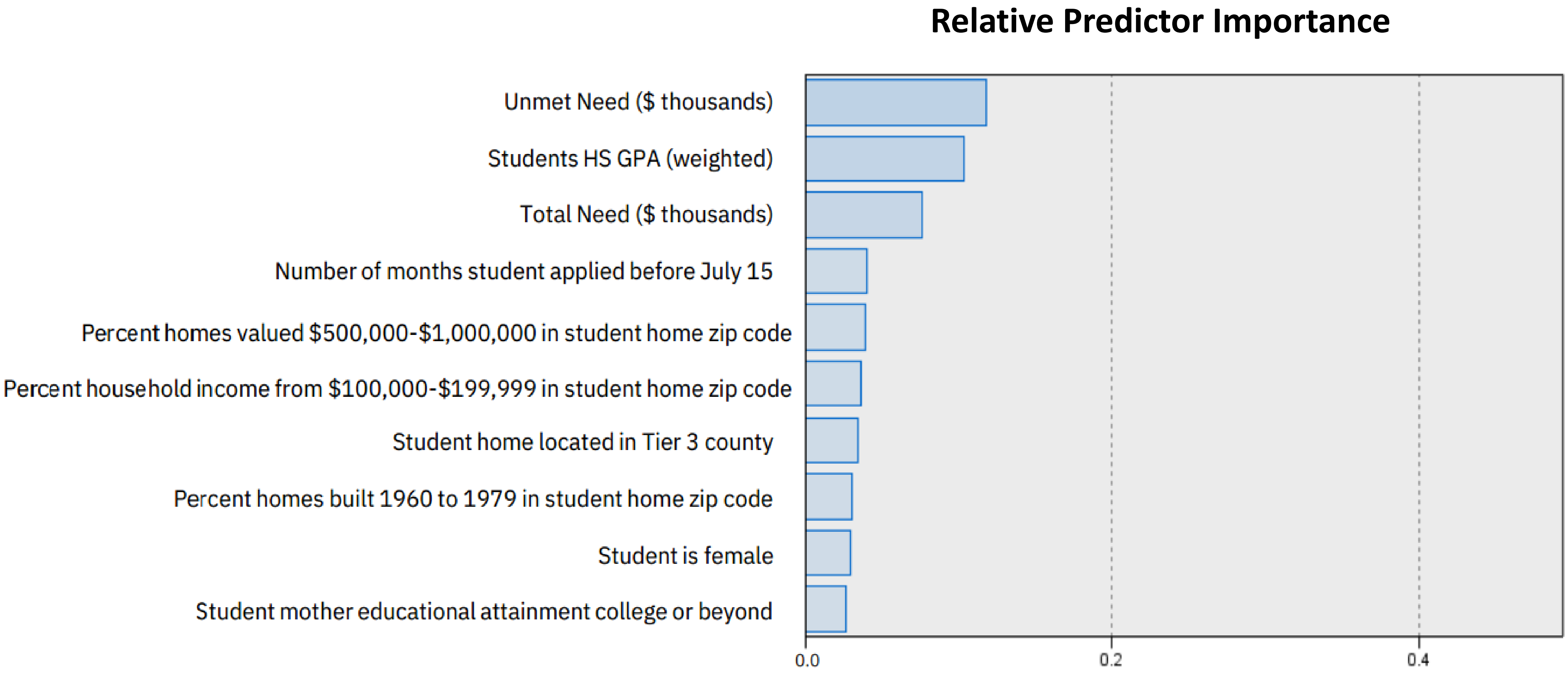


ECU

Comparison: Retained, Dropouts, and Transfers

	Retained	Transferred	Dropped Out
Count	10,449	1488	849
Avg. Weighted HS GPA	3.83	3.62	**3.50**
% rural 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%**

Multinomial Logistic Regression: Strongest Predictors of Dropout and Transfer Risks



Selected Results: Dropout Risk

After controlling for all other variables in the model:

- Every $1,000 increase in unmet **need increases the dropout risk by 12%**.
- Each additional point in weighted HS GPA **reduces the dropout risk by 72%.**
- Students who applied early are less likely to dropout (**every month reduces the dropout risk by 13%**).
- Students from east of I-95 are **49% more likely** to dropout than students from west of I-95 or from another state.
- Male students are **25% more likely** to dropout than female students.
- If the mother's education is college or beyond, the dropout risk **reduces by 20%**.

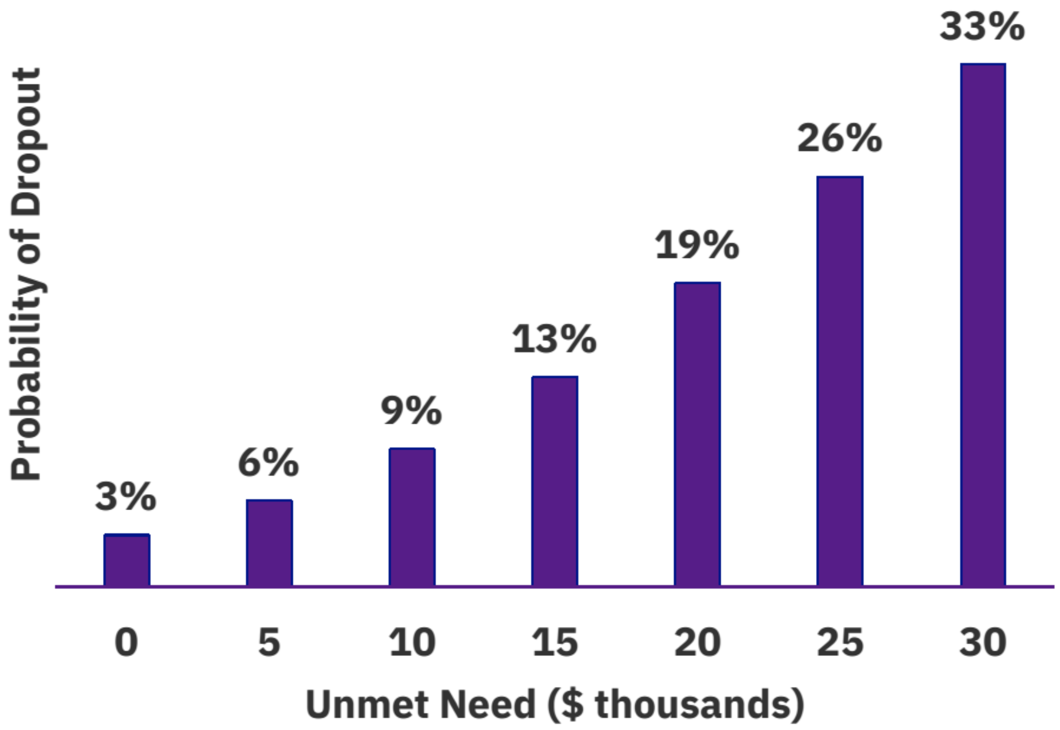
Selected Results: Transfer Risk

After controlling for the other variables in the model:

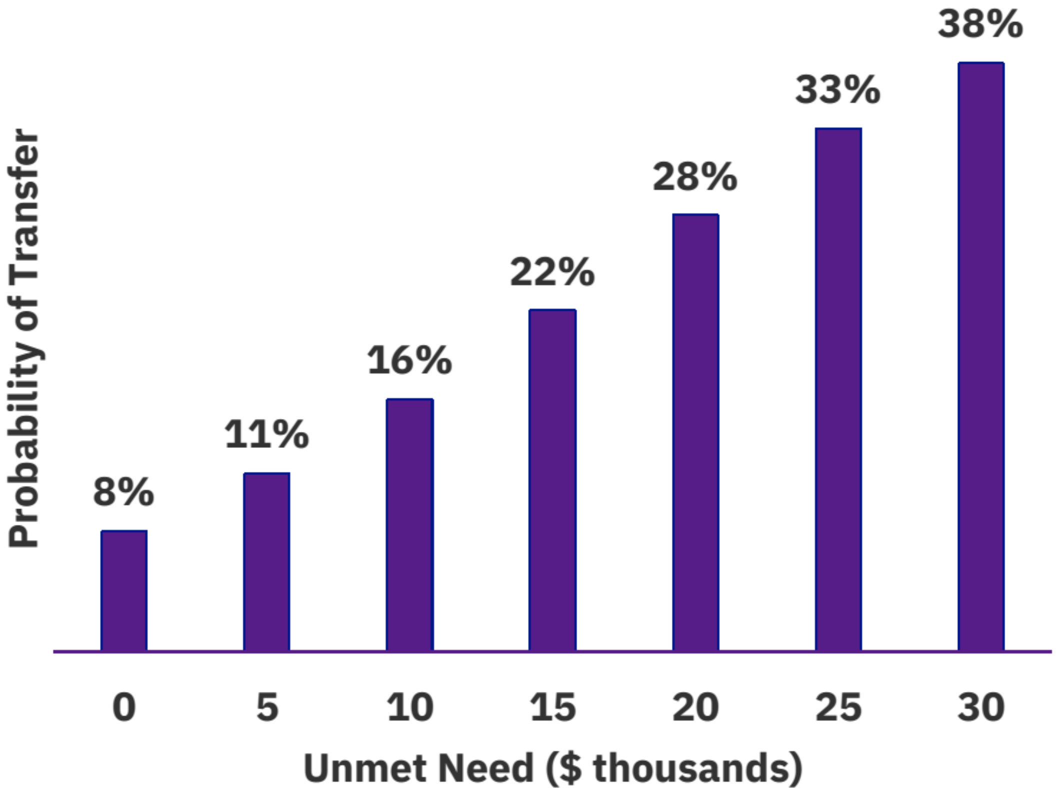
- Every $1,000 increase in unmet **need increases the transfer risk by 9%**.
- Each additional point in weighted HS GPA **reduces transfer risk by 51%.**
- Students who applied early are less likely to transfer (**every month reduces the transfer risk by 6%**).
- Students from Research Triangle are **37% less likely** to transfer.
- Female students are **19% more likely to transfer** than male students.
- White students are **18% more likely to transfer** than non-white students.

Unmet need is a key driver of dropout and transfer risk

Students with the highest unmet need have a probability of <u>dropout</u> that is almost 45% higher than students with no unmet need



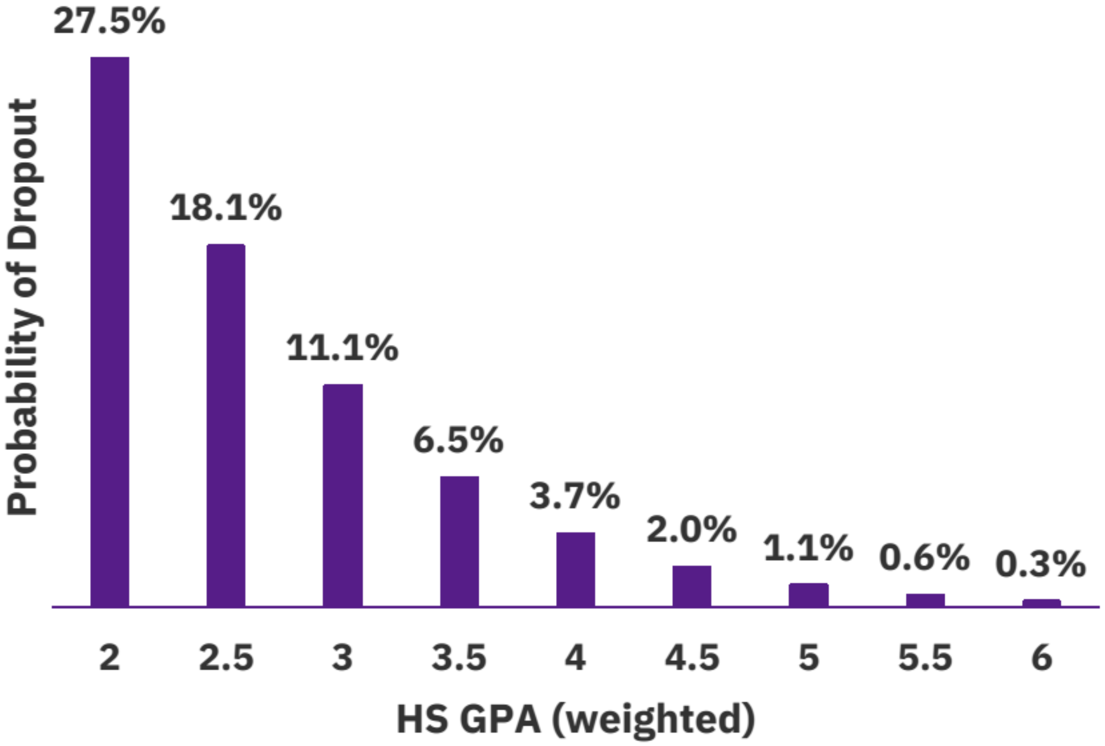
Students with the highest unmet need have a probability of <u>transfer</u> that is over 5x higher than students with no unmet need



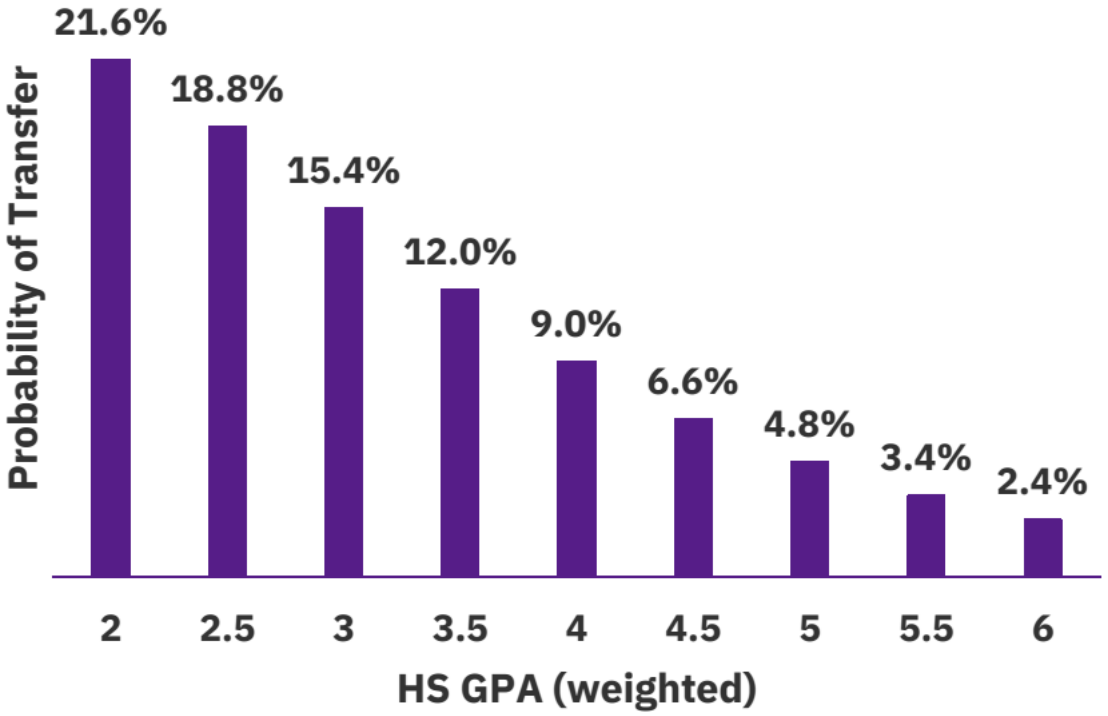
*Range of probabilities shown for both figures assume all other predictors held at the mean value.

Weighted GPA is a key driver of dropout and transfer risk

Students admitted with the lowest weighted HS GPA have a probability of dropout that is over 4x higher than students with the average weighted HS GPA



Students admitted with the lowest weighted HS GPA have a probability of transfer that is over 2x higher than students with the average weighted HS GPA



*Range of probabilities shown for both figures assume all other predictors held at the mean value.

Application Essays

- 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 hedonistic personality (score=1) are almost twice more likely to drop out or transfer than those with the score of 0 (not statically significant).
- Students with application essays that contained the word “work” have a slightly higher transfer probability (statistically significant)

Watson uses natural language understanding algorithms to extract a personality profile for the author of the admission essays

This admission essay snippet has a high hedonistic value. Watson assigns hedonism percentiles closer to 1 for those in the highest percentiles

HEDONISM ANALYSIS

0.520

HEDONSIM VALUE SCORE

0.000

I've always been one to try to impress people and pile on more than I can handle, whether it be extracurricular activities, work hours, promises. I pile on more than I can usually handle. I always end up regretting it too though, because in the moment, I have a lot of stress on my mind. In the end I always realize the struggle was worth it. I always end up happy with the outcome, and always learn from the situation.

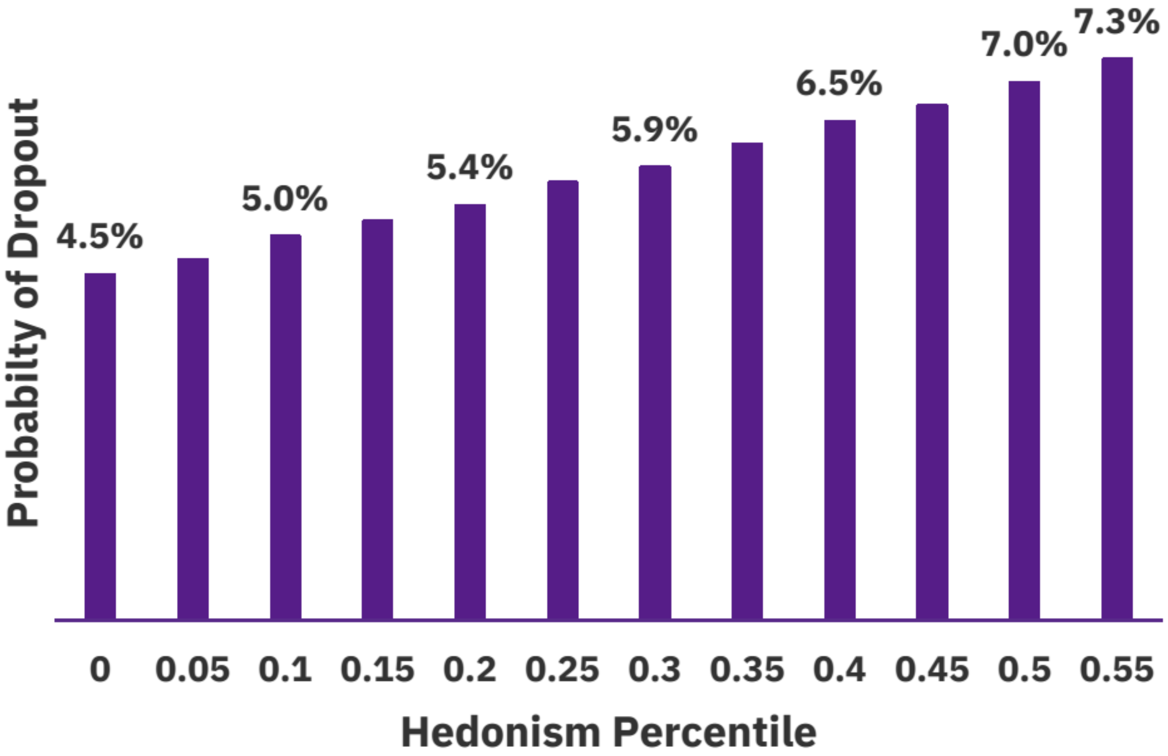
We did an experiment in AP Biology class during my junior year with fruit flies and genetic mutations. As each fly mated, it was fascinating to see how each trait revealed itself in each generation. I have always been intrigued by genetics, but the class broadened my interest. I had never thought about how evolution impacts the entire ecosystem.

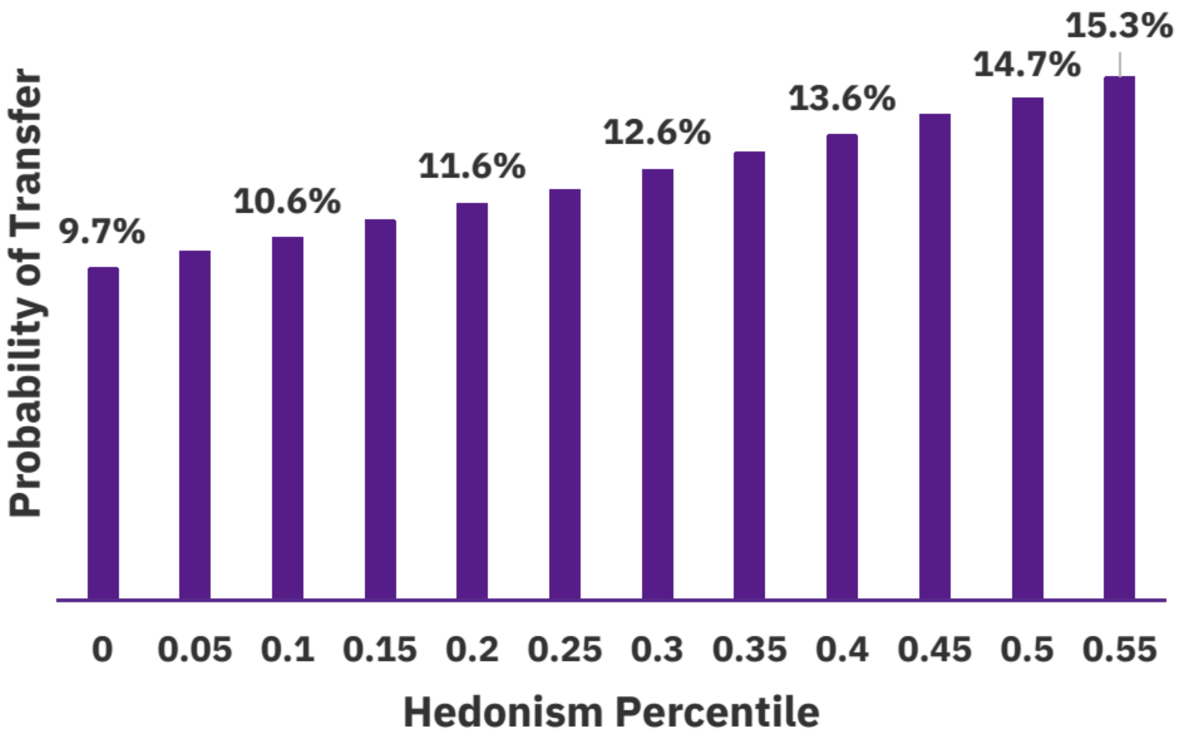
This admission essay snippet has a weak hedonistic value. Watson assigns hedonism percentiles closer to 0 for those in the lowest percentiles

Students who exhibit higher levels of hedonism in their decision making process have higher dropout and transfer risk

A student who is influenced by seeking pleasure for themselves when making decisions has a probability of dropout that is nearly 3 percentage points higher

A student who is influenced by seeking pleasure for themselves when making decisions has a probability of transfer that is over 5 percentage points higher





*Range of probabilities shown for both figures assume all other predictors held at the mean value

III. Predictive Analytics: Lessons Learned

- Challenges of Predictive Analytics
- Potential Use of the Results
- Next Steps

Challenges and Successes

Challenges:

- Multiple data sources used in the study are stored outside of Banner.
- Data integration is labor intensive and variables are defined inconsistently.
- Missing data imputation is a major issue, especially with student comments.
- Because of the complexity of the study, interpretation and communication of the results can be difficult.

Successes:

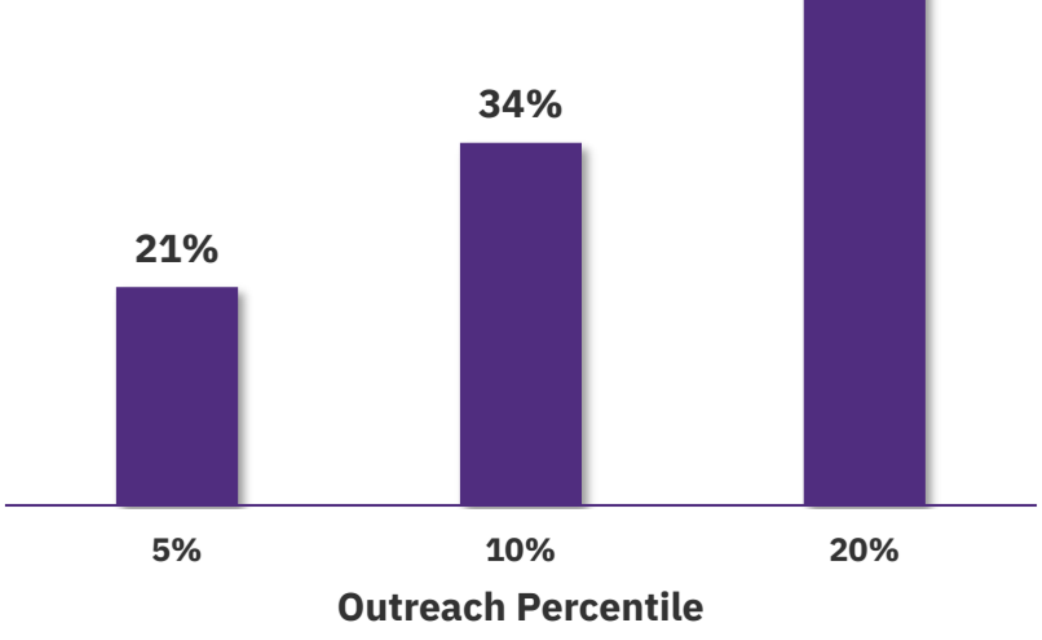
- Key factors identified in the models match previous research.
- IBM has paved a pathway for further research on retention.

Use of Predictive Analytics for Student Outreach

Mitigating Dropout Risk

Outreach to students in the top 10% highest predicted <u>dropout</u> risk will capture almost 35% of students at risk

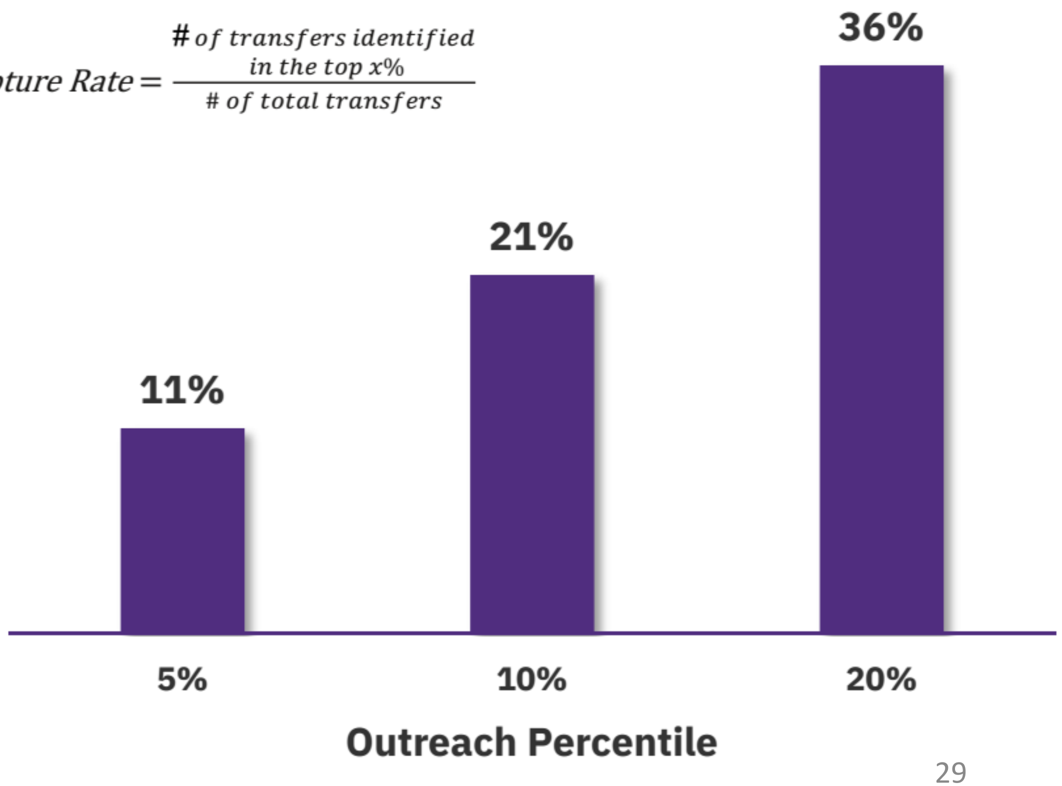
$$Capture\ Rate = \frac{\#\ of\ dropouts\ identified\ in\ the\ top\ x\%}{\#\ of\ total\ dropouts}$$



Mitigating Transfer Risk

Outreach to students in the top 10% highest predicted <u>transfer</u> risk will capture over 20% of the students at risk

$$Capture\ Rate = \frac{\#\ of\ transfers\ identified\ in\ the\ top\ x\%}{\#\ of\ total\ transfers}$$



Potential Use of Predictive Analytics Results: 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
  - Special attention to unmet need
  - Intentional recruitment and marketing efforts: directing at-risk students to academic and student support programs
- Financial literacy program for all students
  - SACSCOC requires a broad-based financial literacy program

Current and Future Effort of IPAR

- Explored data analytics in summer 2018
  - Explored different tools: R, SAS Text Miner, SAS JMP, and Linguistic Inquiry and Word Count (LIWC)
  - Compared variables created by R and Watson
- Develop expertise in predictive analytics through partnership with IBM
- Further improve IBM's predictive models
- Collaborate with other units to make sure critical data elements are stored in Banner, updated timely, and used properly
- Collaborate with ECU faculty and staff in predictive analytics projects
- Promote the awareness and appropriate use of predictive analytics results