PURSUIT

CHURN RATE ANALYSIS

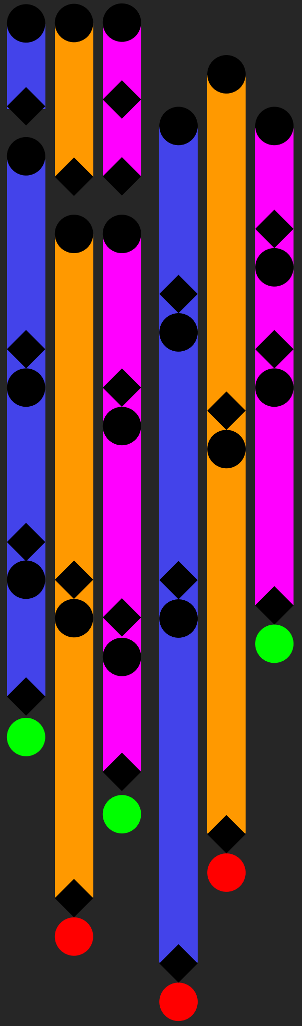
2023 - By Youssef Agour
Edited for Public Sharing

ANSWERING THREE QUESTIONS:

→ What are the chances of looping?

→ Who is looping?

→ Why are they looping?



Project Overview

Section 01 – Scope

- Problem
- Data
- Procedure

Section 02 – Analysis

- Descriptive Statistics
- Predictive Modeling
- Causation

Section 03 – Findings & Aftermath

- Summary of Outcomes
- Applications
- Financials
- Recommendations

What is a ‘Looper’?

Looper - - [Looh-Purr]

A looper is what we define at Pursuit as someone who exits the 1 year long fellowship program at any point in time, **with intention to return.**

For this reason they are not marked as withdrawn upon separation, although some fellows do eventually fully withdraw.

Some of these fellows return to complete on their second, sometimes even third tries, others have yet to even re-enroll.

The Scope

The Problem at a Glance

Pursuit relies on the success of its fellows in Core to ensure their readiness to enter the workforce. **For fellows who cannot complete core in one try, the chances of EVER completing become drastically lower.**

We know Our fellow population

100% From Low-income Populations	**50%** Women & Non-binary	**76%** Black or Latinx
40% Immigrants	**61%** Do not have a Bachelor's Degree	**+50%** Public Assistance Recipients

The Historical chance of looper completion based on latest enrollments

30%

Our Data

Sourced from records across enrollments, and salesforce, our data is an aggregation of records containing 101 historical loopers, and 39 loopers who are currently enrolled in classes XXXXXXXX XXXXX XXX XXXXX XXXX

The Full Set

Fellows currently looped out, regardless of whether it's their first enrollment

Fellows looped out, that went on to eventually withdraw

Fellows who looped, completing the program in 2 or more enrollments

The Prediction Set

Fellows who looped, and are currently enrolled in core

Things to Note

- All fellows gathered based on latest enrollment, this pulls data from their most up to date record
- Fellows looped out on first enrollment were included to understand the chances fellows currently ‘Looped Out’ have to complete core
- There are limited records pre 2020, prior to the 8.0 cohort due to the inability to reference older data as no documentation was available

The Procedure

Method of how we got from A to B

Descriptive Statistics

Identifying if our loopers represent our greater population or come from a certain demographic, and identifying the prototypical looper

Predictive Modeling

Finding the likelihood of looping and making foresight reports through 6 different machine learning algorithms

Causal Regression

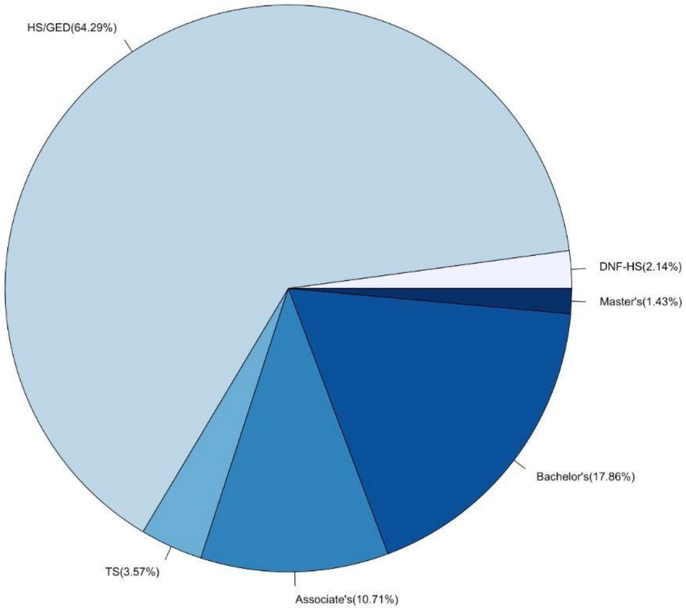
Discovering what factors are most responsible, contributing to the probability a fellow loops out or not

The Analysis

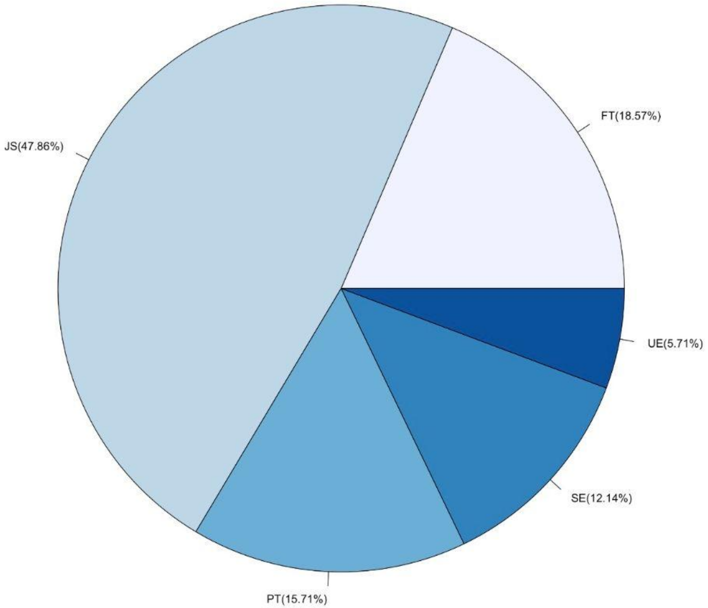
Descriptive Statistics

Describing our Looper Demographics

Education Distribution



Employment Status Distribution



About

1 in 5

Looped fellows hold Bachelor's Degrees

Nearly

50%

Are unemployed and in active search for employment

34 years

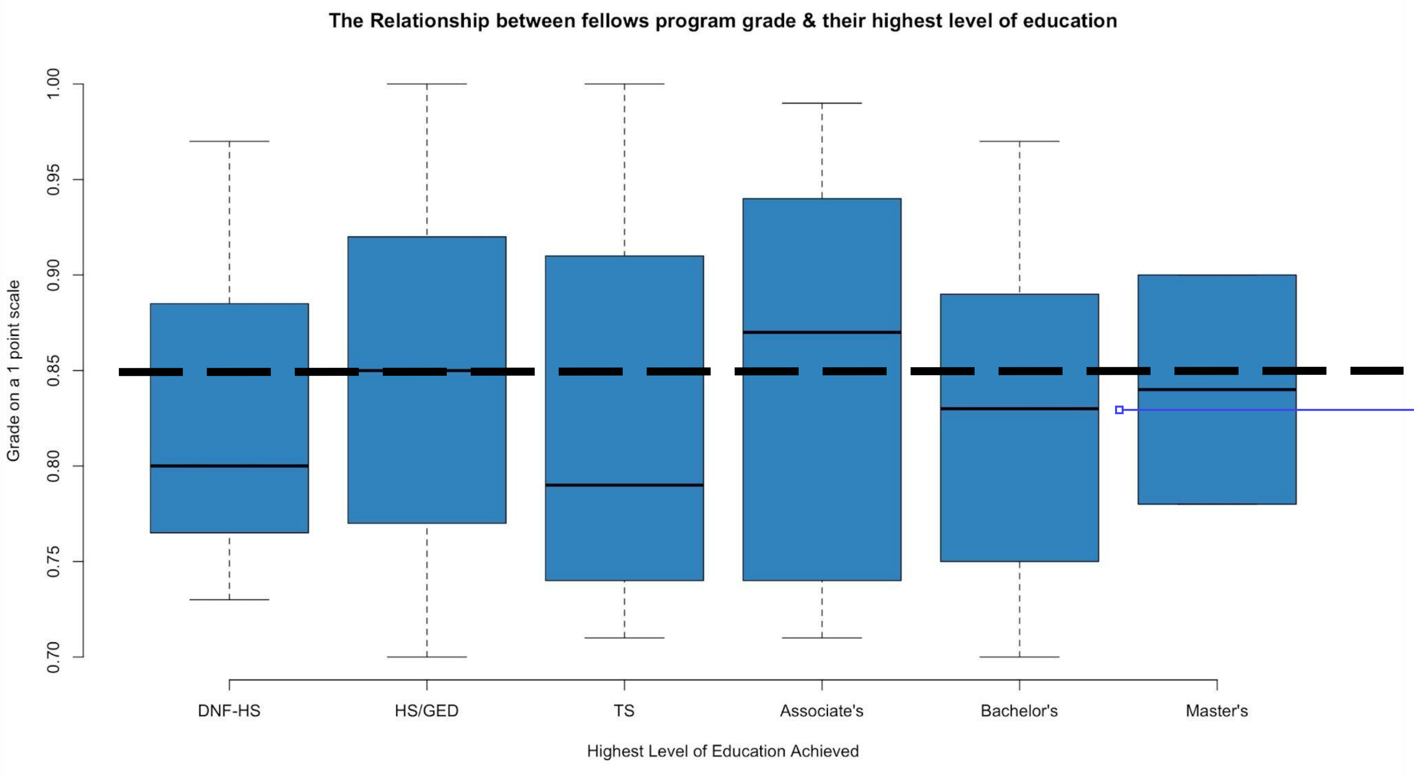
Average Age of Looped Fellows

Nearly

40%

Entered the program with No Income

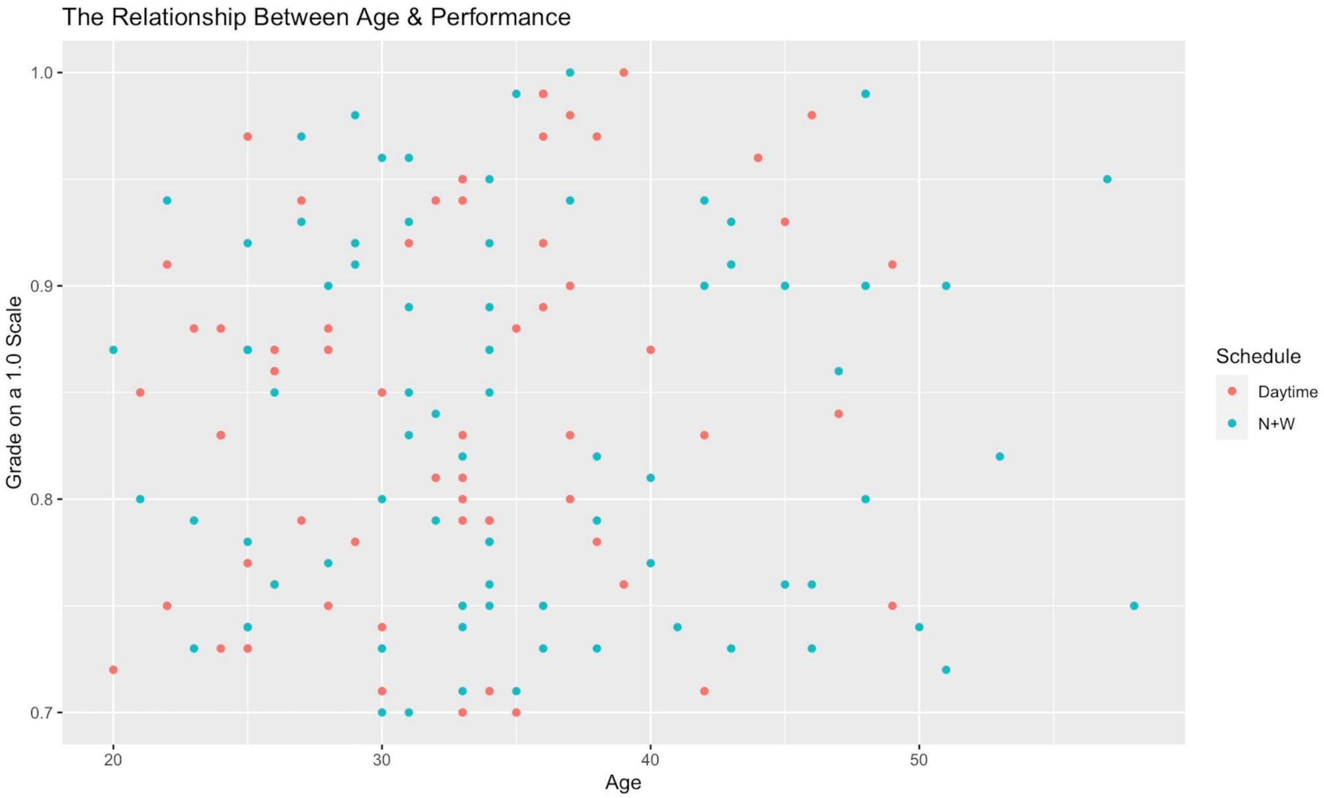
Education **does** impact performance



Fellows with Bachelor's degrees on a median scale **performed worse** than those with a HS diploma

Performance is unaffected by Schedule &/or Age

Although not directly related to performance, shown by both the range of grade and the range of ages, the lack of correlation between these two attributes does not make them any less significant to the purposes of our analysis.



Predictive Modeling

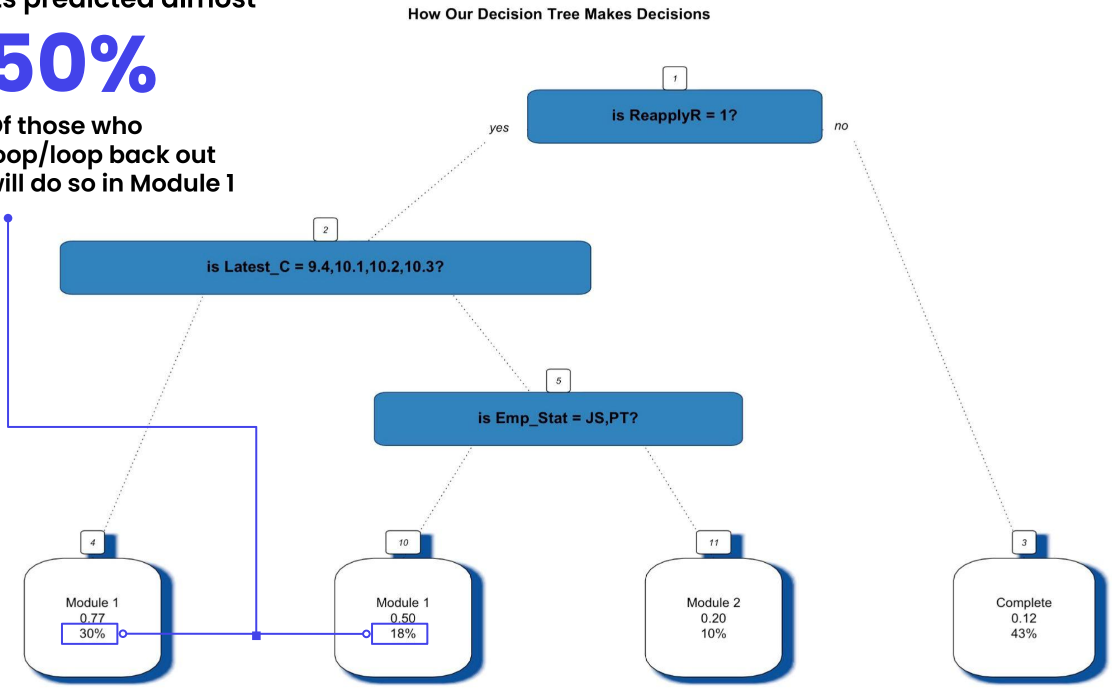
When are people Looping?

Knowing when people are looping is just as important as understand why they do it

Its predicted almost

50%

Of those who loop/loop back out will do so in Module 1



Decision Tree & Random Forest Modeling

Decision trees involve dividing the data into subsets at each split based on the most significant feature until a stopping criterion is met (Completion or Looping Out). A Random Forest does this as the name suggests, with *a lot* of trees; **the more trees, the more accuracy in our predictions**

Although it was hard to reinforce statistical significance, both models helped inform:

- The decision to keep our analysis binary rather than trying to predict separation timing of specific modules as it was originally
- The identification of possible factors that are skewing our results eg. **Latest Class, and Recommendation to Reapply**

Module 1: Our De-Facto Filter Module

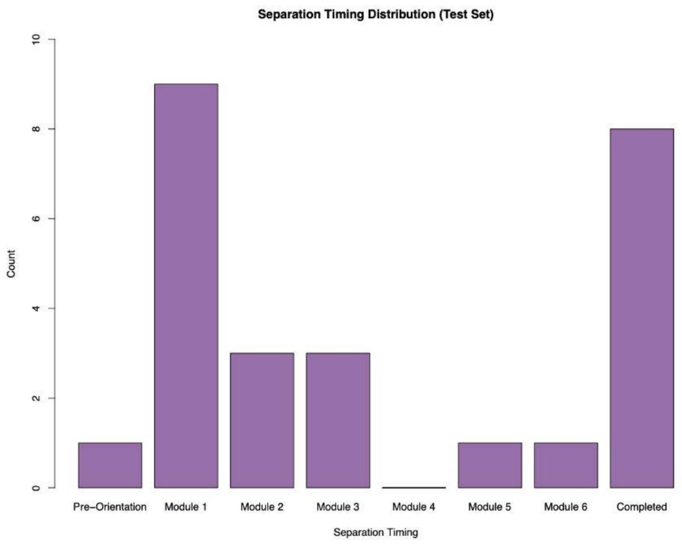
Module 1 without a doubt both historically and predictively is where the majority of our fellows are looping out of the program. Our actual results produce a slightly lower number than the one made by our decision tree. Whether it is their first time looping or not, if someone is going to loop, it seems to happen most frequently at the very start. Are we admitting fellows who are not as interested in careers in tech as we thought, do we need to re-evaluate our behavioral and technical pre assessments? Does the module content need to be revised? These are some of the questions we must ask about module 1 going forward.

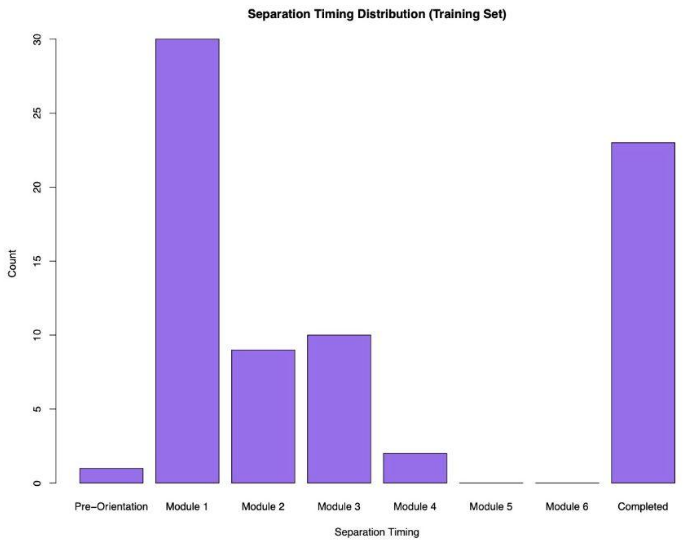
Historical Looper Data

In reality,

40%

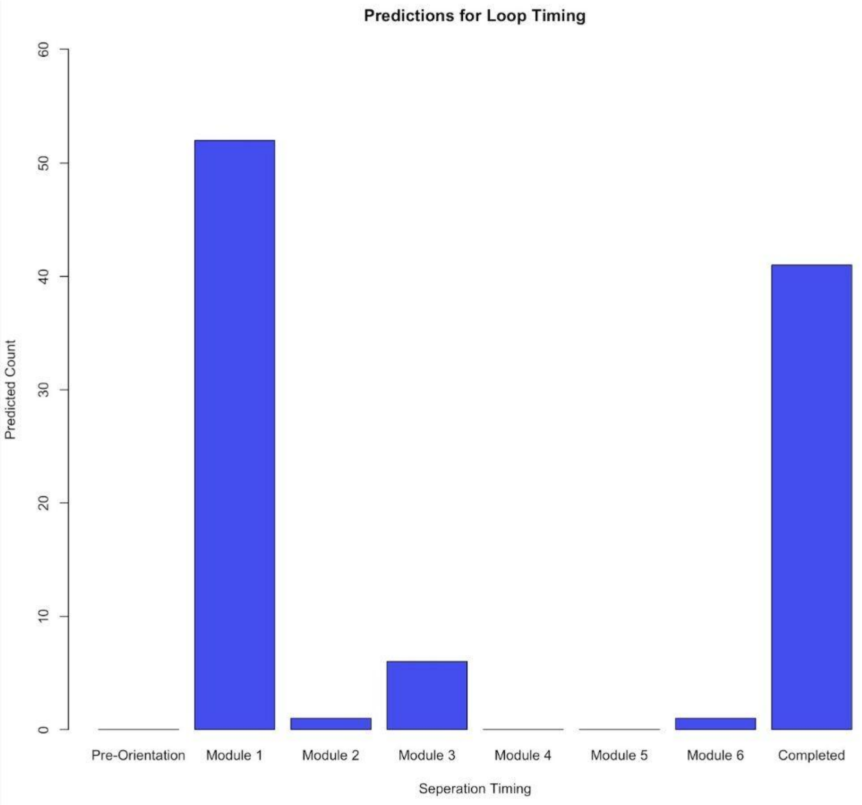
Of those who loop/loop back out will do so in Module 1





Module 1: Our De-Facto Filter Module

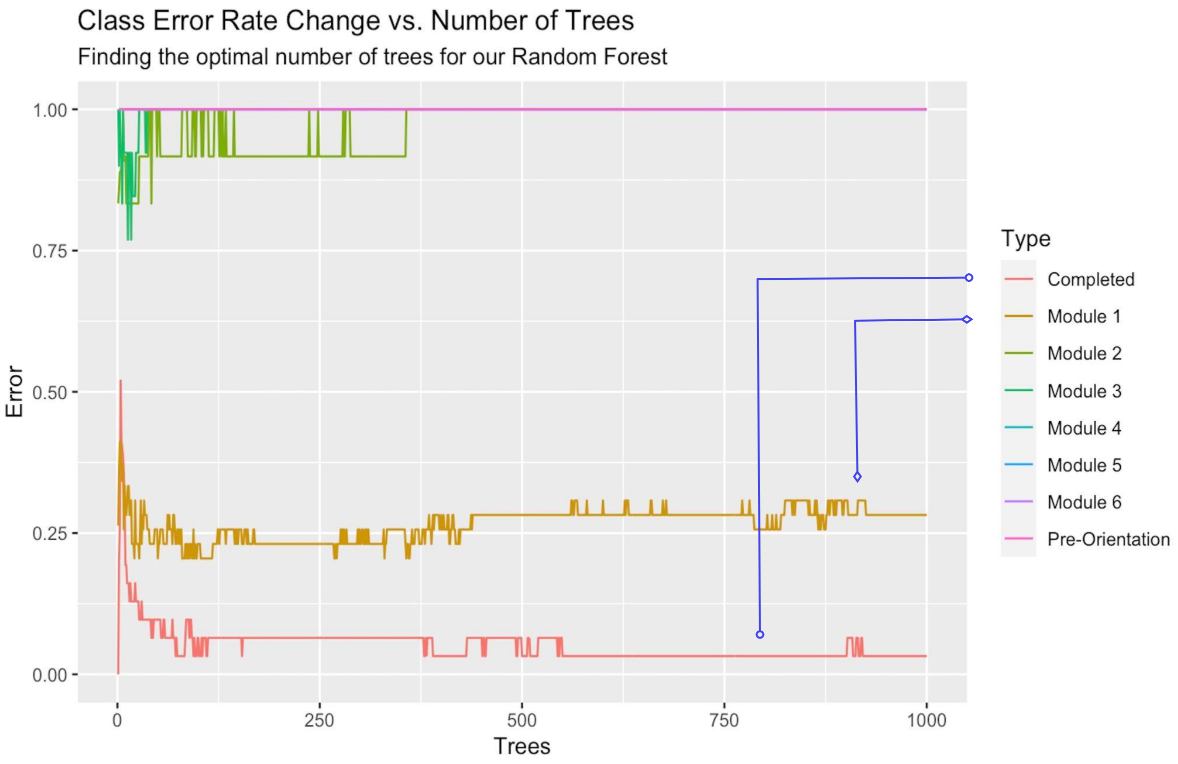
Reinforced by Predictions



Predicted Looper Data

Although we have shown that the Random Forest predictions were not statistically sound enough to draw concrete conclusions, **the overwhelming skew of our data reinforces the fact that module 1 suffers from the greatest number of loopers**

Starting to Identify Patterns



What we gathered from the error rates shown by our Random Forest was a **model overfit,** and an average error rate of just over 40% – which was later optimized to 30%. This was done with 1000 trees, and only reduced the error rate found in our decision tree (just over 60%) by a slight margin. To say a model predicts wrongly 1 out of 3 times, even considering the size of our data set, is not good enough. We were however, able to **identify key attributes which increased the accuracy of predictions.**

Latest Class → Barrier to effective analysis

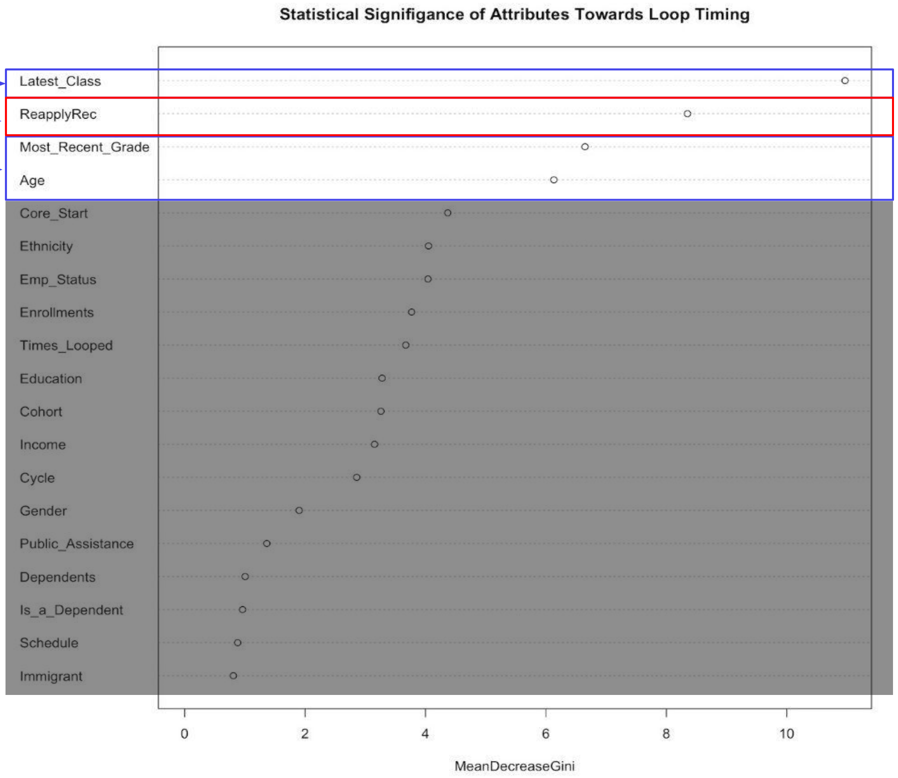
Latest Class Enrollment was deemed our most statistically significant attribute by our Random Forest

However barriers to deeming this an actionable insight are:

- Instructor Changes
- Change in curriculum
- Varied Peer Interaction → COVID

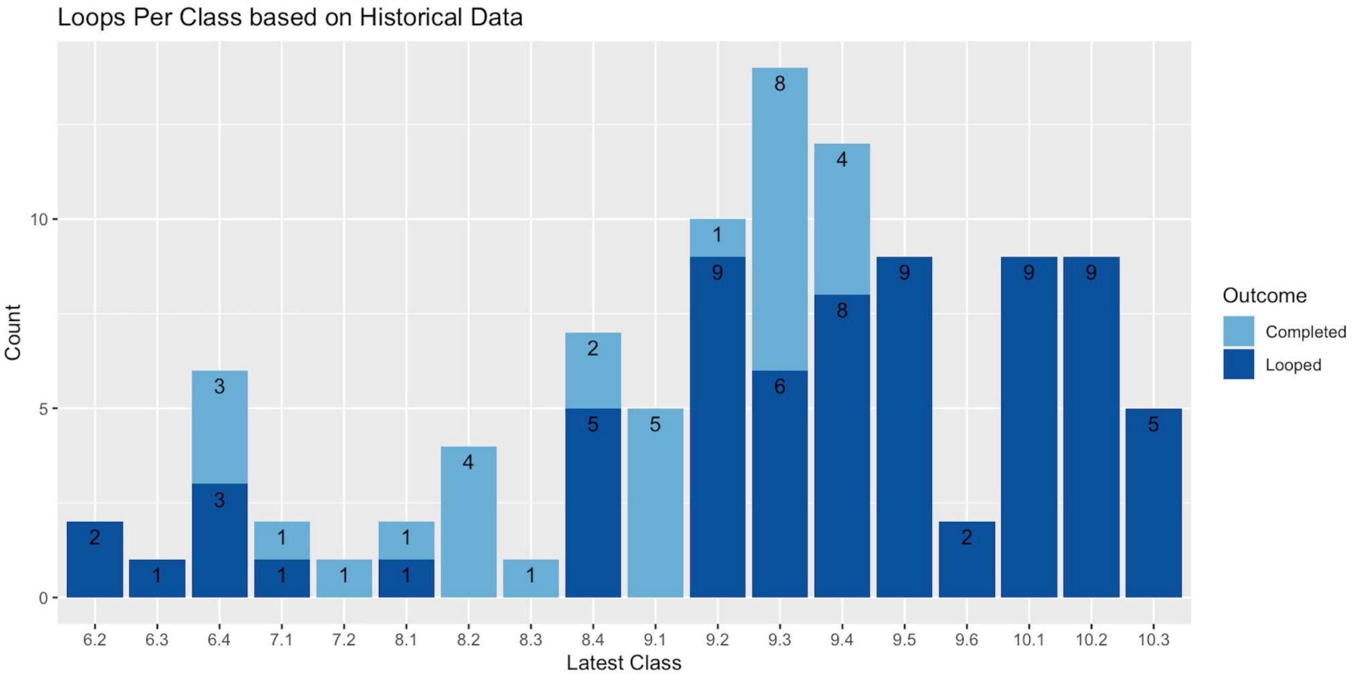
Whether or not someone was recommended to reapply was later found to be **inconsistent** for older fellow records, and hence dropped for consideration

Age and Performance (Grades) although not directly correlated, were significant enough for helping predict who loops



Latest Class is Inconsistent

Reinforced by Historical Data



Latest Class is not a solid enough attribute from which to start making inferences about our loopers. **Insufficient data is available for fellows pre-8.0**, and for some of the more recent loops, such as 10.2 & 10.3, those fellows **would have no chance to complete as their original cohorts have yet to conclude as of 2023**.

Causation

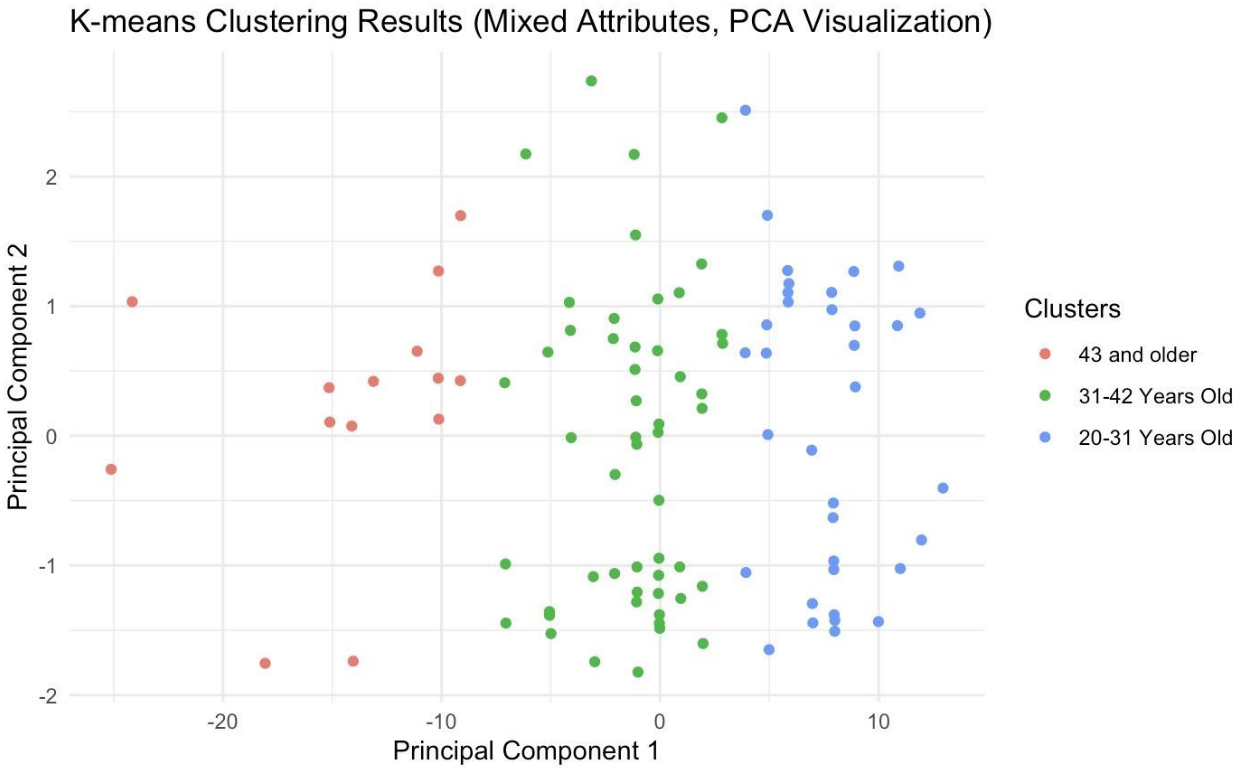
Getting Rid of the Noise

K-Means & Principal Component Analysis

Through the combination of two machine learning models: a k-means clustering algorithm and principal component analysis, we were able to understand where most of the noise comes from in our data.

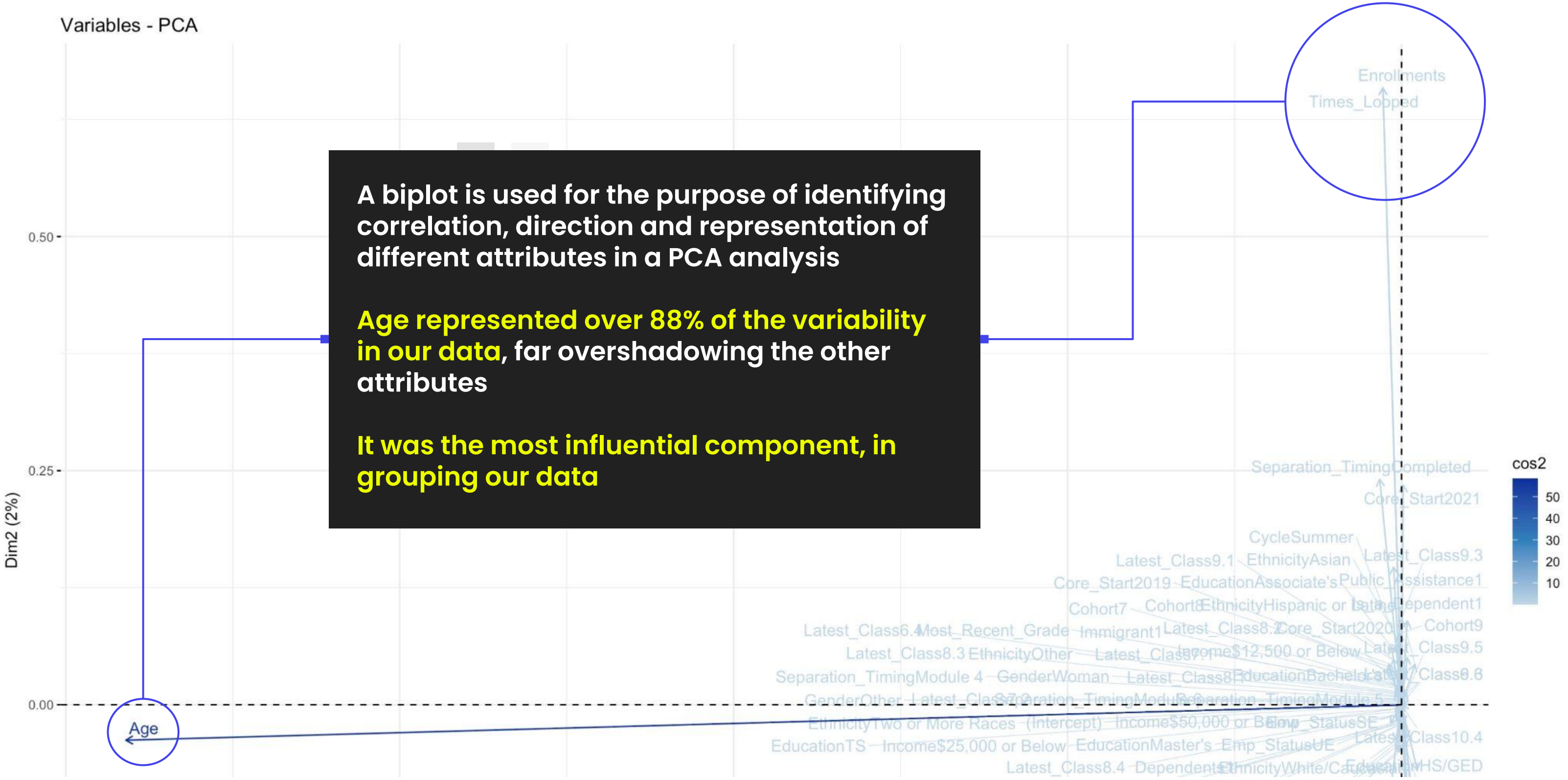
Supporting the findings of the models that had been conducted thus far, **age overwhelmingly** was identified as both one of the most well represented in the data set, as well as one of the most impactful on being able to group fellows together.

Finding this was important, but proving its causation is where we pull real value.



Age: The Most Influential Component

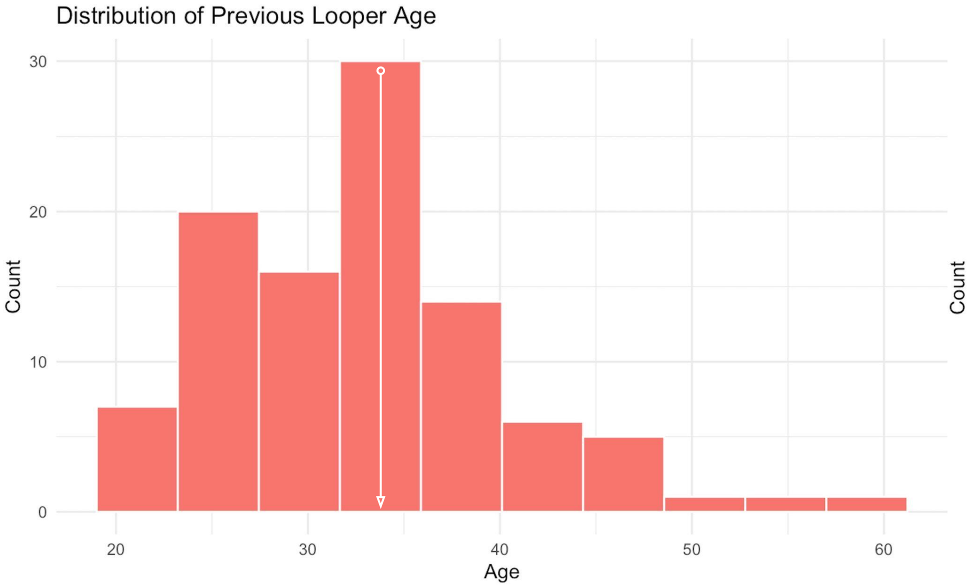
Plotting Principal Components

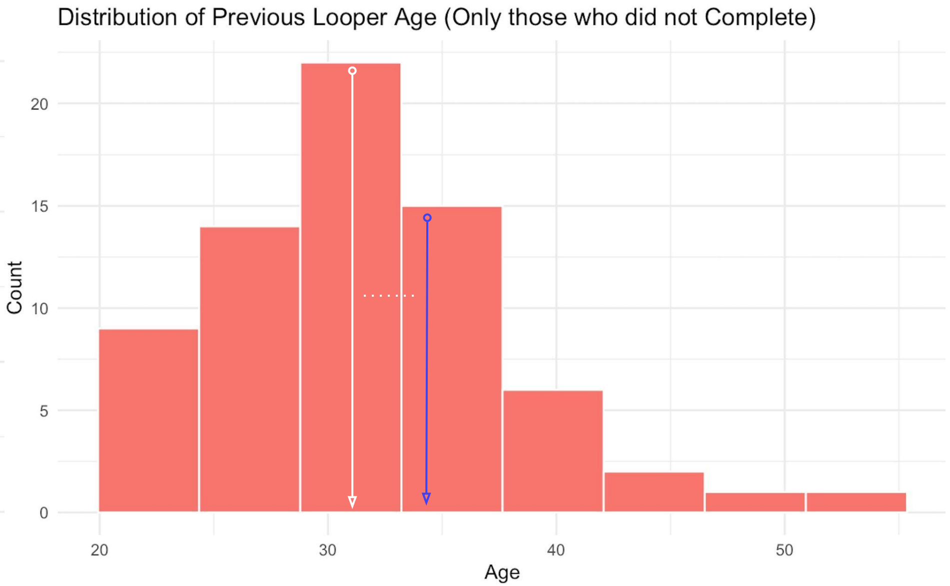


A biplot is used for the purpose of identifying correlation, direction and representation of different attributes in a PCA analysis

Age represented over 88% of the variability in our data, far overshadowing the other attributes

It was the most influential component, in grouping our data



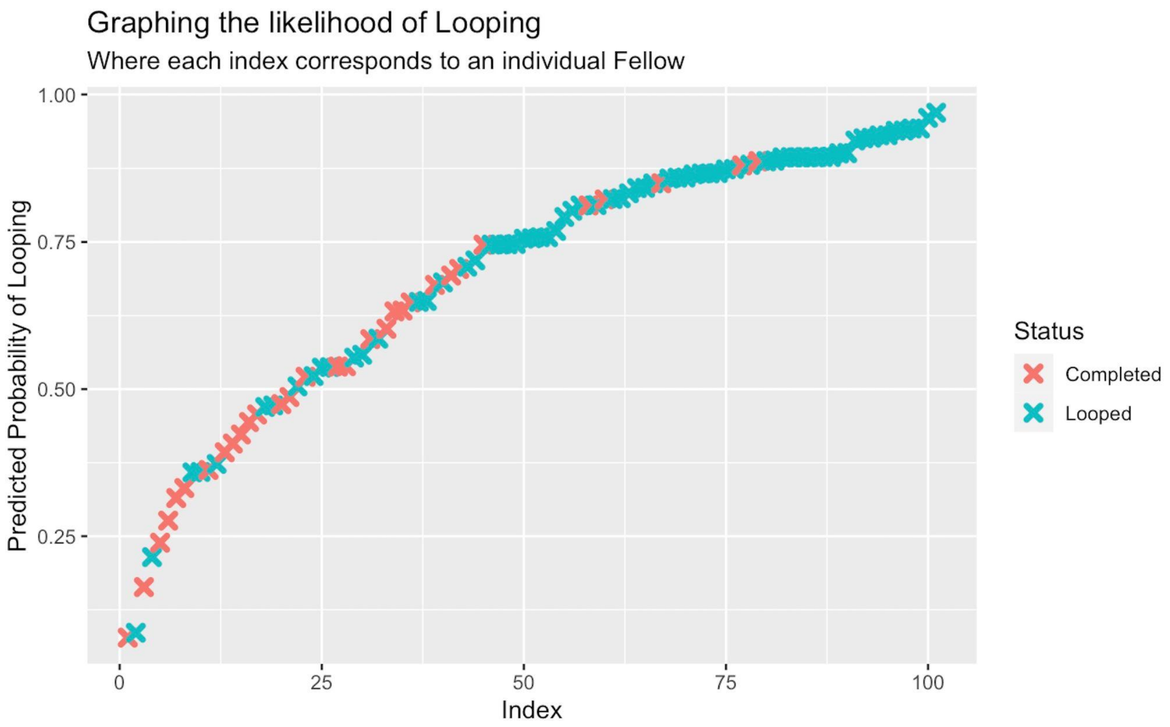


Younger fellows are the ones looping back out

Observed in the above, the median age of a looper based off of our full historical set hovers at almost 35, whereas the median for all of our loopers currently looped out (right) seems to shift left. With a slightly lower median at around 31, it is reasonable to say Age may be impacting our data inversely from the way we expect.

Consistently Significant

Logistic Regression



To further test our observance of Age and other attributes, a causal model to allow us to see if these factors were not just well represented but impacted our fellow success outcomes was key. Supplementing the findings of previous models, **An increase in Age, Number of Enrollments (Number of classes fellow has been apart of), and Most Recent Grade all contribute to lower probability of looping.**

Average probability of looping for Fellows who went on to complete core **54%**

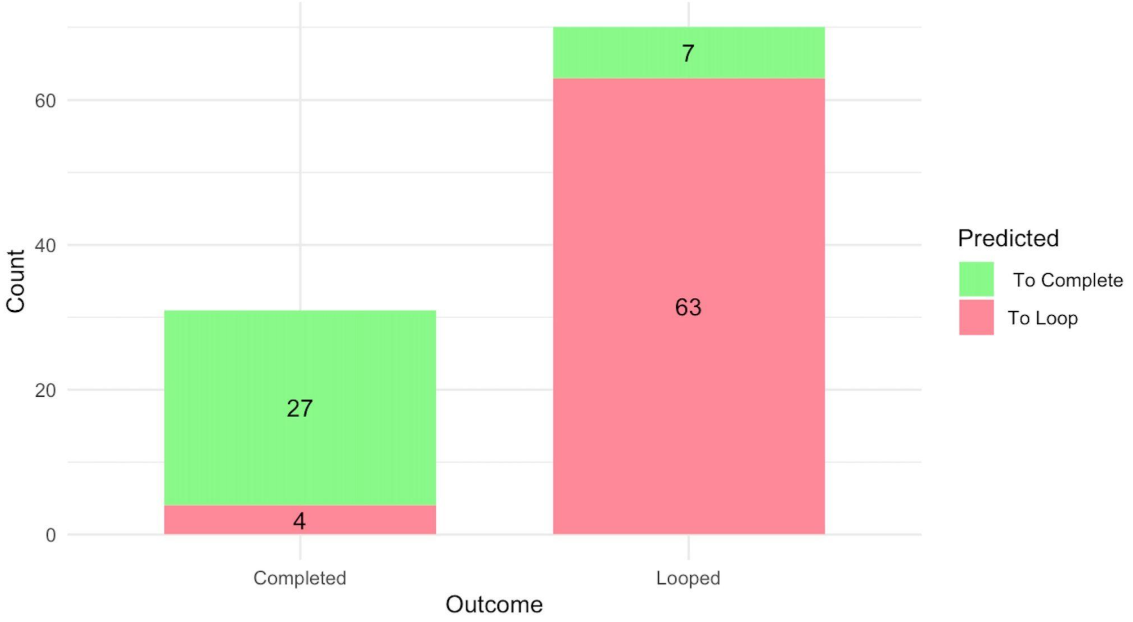
Average probability of looping for Fellows who went on to loop back out **76%**

Quantifying Grades & Performance

Naive Bayes Modeling

Confusion Matrix

Misclassification Rate for Full Model: 10.89%



Because of the noise previously observed in our data, a naive bayes model was conducted, as it is a statistical technique which assumes **all attributes have an equal effect on the one we are trying to predict.**

This model drove to findings supporting our previous identification of significant factors and **predicted a large difference in how likely both female and N+W looper fellows were to complete core. It also helped quantify our observations around grades.**

Grade threshold for marking someone at risk for looping out **< 82%**

Average score of those predicted to loop vs those who go on to complete **6 Points Lower**

The Findings & Aftermath

Summary of Outcomes

Answering our Initial 3 Questions

What are the chances of looping?

Who is looping?

Why are they looping?

The chance of a fellow completing the program after having previously looped is **30%**, once re-enrolled, fellow has about a **47% chance of completion**, of those looped **nearly 25% can potentially go on to complete core** at some future point

Based largely on Bayesian statistics, a prototypical looper is *a* ***woman in a nights and weekend class, under the age of 35***

Age, Most Recent Grade, and Number of Enrollments were the largest causal factors in determining the probability of an individual looping

With each 1 year increase in *Age*, the chances of looping *drop* by 6%

Women are 19% more likely to loop back out at 57% compared to the Men's probability of 38%

N+W fellows are 8% more likely to loop back out at 54% compared to the Daytime probability of 46%

The Predictions & Outcomes Personified

Understanding the prototypical looper, it is clear even if within this small dataset, **women and N+W fellows suffer from a lower chance of success.** Even small increments in the performance of two fellows can make a major difference in their probability of success upon re-entering Pursuit.

Repeat loopers were typically

4+ years younger

than those predicted to complete

X1, Enrolled in ____

99% Chance to Complete

1% Chance of Looping

Age → 37

Enrollments → 2

Grade → 98%

Schedule → Day

X2, Enrolled in ____

0.03% Chance to Complete

99.7% Chance of Looping

Age → 30

Enrollments → 2

Grade → 80%

Schedule → N+W

X3, Enrolled in ____

77% Chance to Complete

23% Chance of Looping

Age → 38

Enrollments → 2

Grade → 82%

Schedule → N+W

X4, Enrolled in ____

2% Chance to Complete

98% Chance of Looping

Age → 33

Enrollments → 2

Grade → 81%

Schedule → Day

The Results in Reality

Time will tell with regard to the rest of our group, but **at the time of writing two fellows who were enrolled at the start of this analysis were accurately predicted to loop out**, fitting the schema we have established thus far with regards to Age, and other pivotal factors.

X5,
Enrolled in ____

Looped

>0.01%
Chance to Complete

99.99%
Chance of Looping

Age → 37

Enrollments → 2

Grade → 100%

Schedule → N+W

X6,
Enrolled in ____

Looped

12%
Chance to Complete

88%
Chance of Looping

Age → 41

Enrollments → 2

Grade → 74%

Schedule → N+W

Why is the difference between X6 & X5's odds so great?

Although not in our predefined age-bracket, their results are in line with the parameters we have established thus far.

XXXX XX X XXX XXX X XX XX XX X XXX XXX XX XXXXXX XX XX X X XXX X XX X XXXX XX XX XX XXX XX X X XXX X XXXXXX X X X X XX XX XX

Performing quite well, X5's exit was unexpected. This is a perfect example of the need for an extension to this analysis, into separation reason in future.

Its easier for N+W fellows to loop back out, but also easier for them to get jobs?

Based on 7.1 & 8.3 Daytime Fellow Data & 7.2/8.4 Nights and Weekends Fellow Data we observe a shift in the employment rate between fellows across schedules

81
N+W Fellows

62
Daytime Fellows

Both Schedules had similar completion rates at almost 65%, and both saw 1 in 5 fellows loop, however

Nearly **50%** N+W Fellows Landed Jobs

vs.

Only **35%** Daytime Fellows Landed Jobs

A topic for further exploration, it should be noted that employability although not explicitly explored in this analysis, **historically has seen that N+W fellows are more successful in securing employment post core.** A natural progression from this analysis would be to determine the effect schedule, grade, gender, age, amongst all the other factors we have identified, on gaining employment.

Theoretical Impact

Some ‘what ifs’ to consider

These estimations assume:

- Average hired fellow salary of XXXXXX
- **XXXXXXXXXXXXXXXXXXXXXXXXXX (Based on 2023 hires)**
- XXXXXXXXXXXXXXXXXX

The Predicted

25% of Currently Looped Out Return

- XXXXXXXXXXXXX XXXXXXXXXXXXX
- XXXXXXXXXXXXXXX XXXXXXXXXX XXXX XXXXX
- XXXXXX XXXXXXXXXXXXX XXXXX XXXXXXXXXX XXX XXXXXX
- XXXXXXXXXXXX XXXXXXXXXXXXX XXXXXX

$XXXXXXX

Potential additional revenues over 4 years

The Ideal

100% Return and complete

- XXXXXXXXXXXXXXXXXXXXXXXXXX
- XXXXXXXXXXXXXXXXXXXX

$XXXXXXX

Potential additional revenues over 4 years

In an Ideal Market

XXX XX XXX XXXXX XXX XX XX X XXX XXXXXX XXX XXXXXX XXXX XXX XXX X XXX XXXXXX

These estimations assume:

- Average hired fellow salary of XXXXXX
- **XXXXXXXXXXXXXXXXXXXXXXXXXX (Based on 2021-2023 hires)**
- XXXXXXXXXXXXXXXXXX

The Predicted

25% of Currently Looped Out Return

- XXXXXXXXXXXXX XXXXXXXXXXXXX
- XXXXXXXXXXXXXX XXXXXXXXXX XXXX XXXXX
- XXXXXX XXXXXXXXXXXXX XXXXX XXXXXXXXXX XXX XXXXXX
- XXXXXXXXXXXX XXXXXXXXXXXXX XXXXXX

+XXXXX

$XXXXXXX

Potential additional revenues over 4 years

The Ideal

100% Return and complete

- XXXXXXXXXXXXXXXXXXXXXXXXXX
- XXXXXXXXXXXXXXXXXXXX

+XXXXXXX

$XXXXXXX

Potential additional revenues over 4 years

Recommendations

Operations & Process

Data Restructure & Documentation

- Admissions demographic form standardization
- Familiarizing instructors with an agreed standard of data keeping across all attributes

Core

Module 1 Pain Point Survey & Re-entry requirements review

- Module Reviews at end of each module for core improvements
- Instructor Module debriefs, reviewing potential struggling fellows, and marking loopers on regular basis
- Behavioral screening and establish re-admission criteria for returning loopers

More Further Research

Something more than Age & Performance

- Does looping reduce the likelihood of success on the job?
- For loopers who are completing in extended amounts of time, does the time to ROI justify more than a single loop?

THANK YOU

ACKNOWLEDGEMENTS

HEAD ADVISOR & MENTOR

Tim Asprec

GENERAL COUNSEL

Victor Eduardo
Rachel Liang
Davonte Williams

PROJECT OWNER & LEAD

Youssef Agour