Predicting Fall to Fall Retention for Incoming Community College Students

Dan Larson

Director of Research and Analytics
Delaware Technical Community College

Essentially all models are wrong, but some are useful.

- George E. P. Box

Follow me on social media: @datadanlarson

Objective

Today

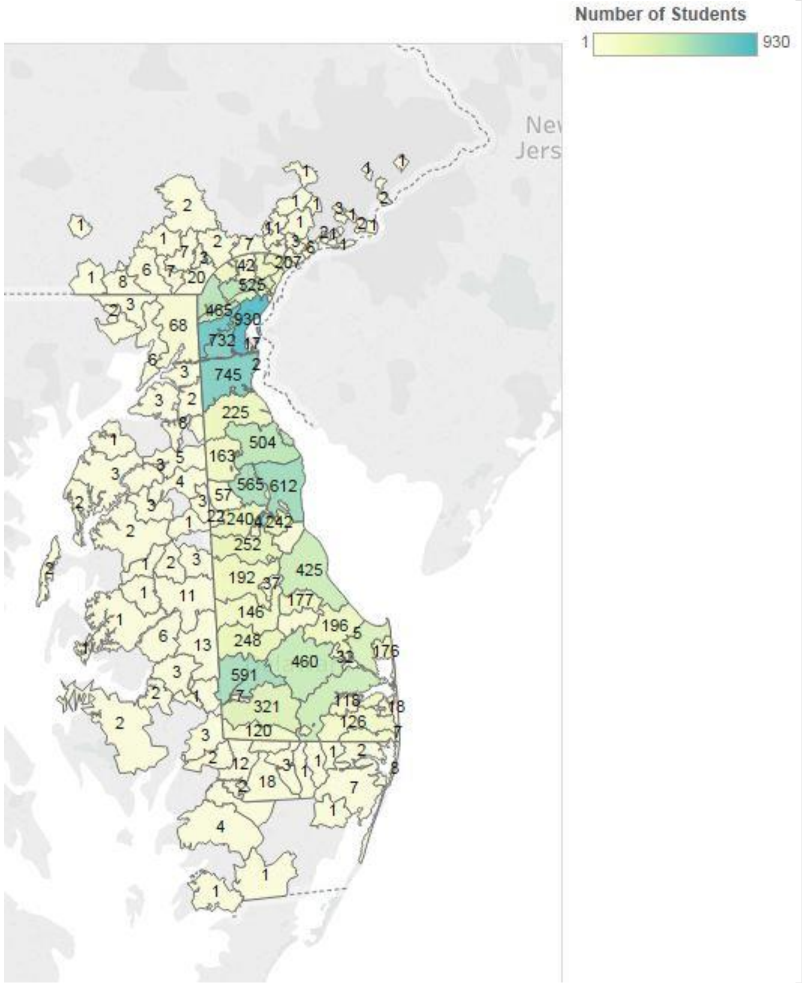
- Share an effort to utilize predictive analytics.
- Get feedback on how to make it better.

Project

- Determine which students have the highest and lowest probability to return to the College for their second year.

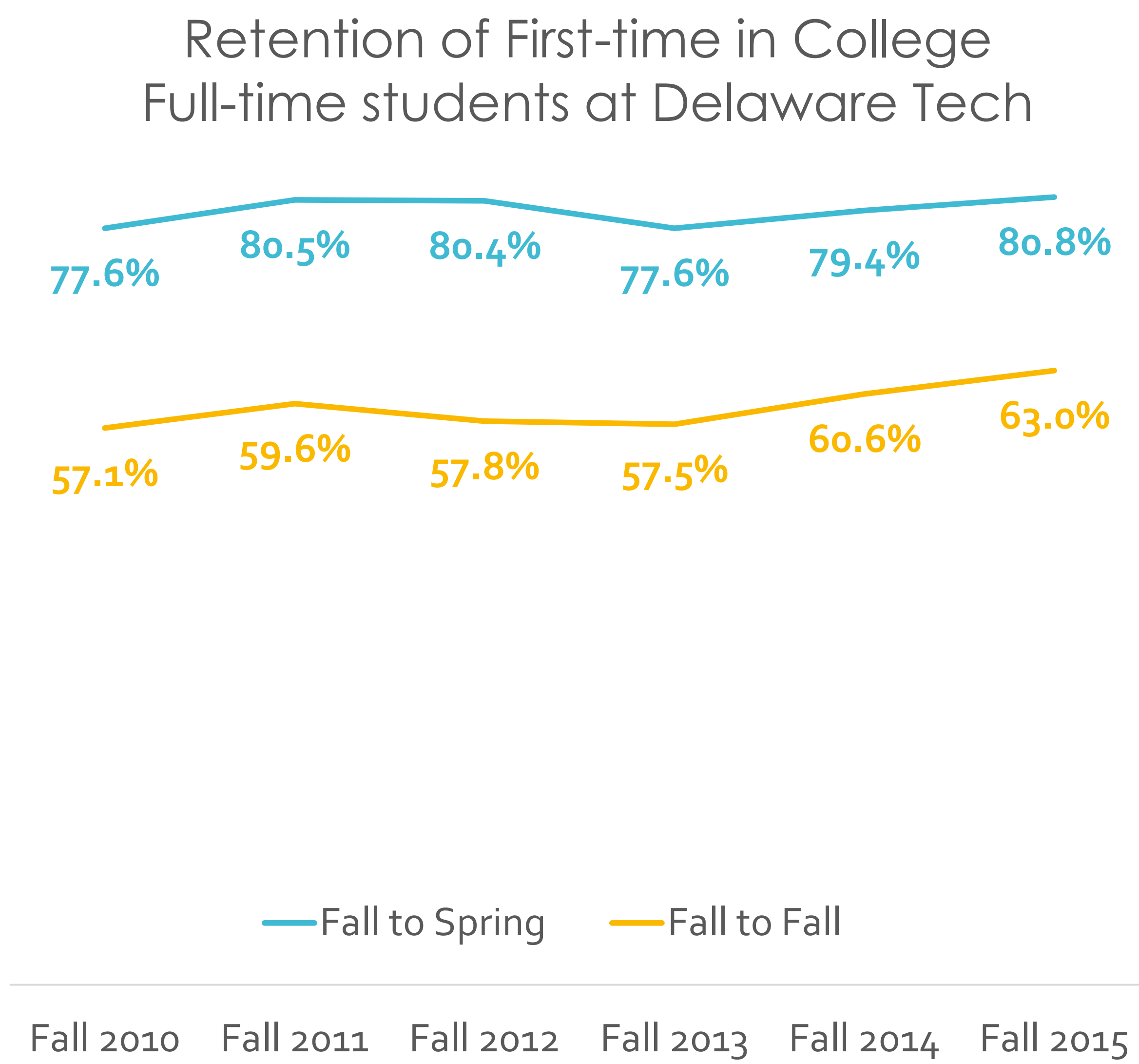
About DTCC

- 14,000 Students
- Four Campuses
- Offer 86 unique programs
- AAS Degree
- SEED Scholarship (Free College)
- 30% Minority
- 60% Female
- 70% of students require at least 1 developmental course



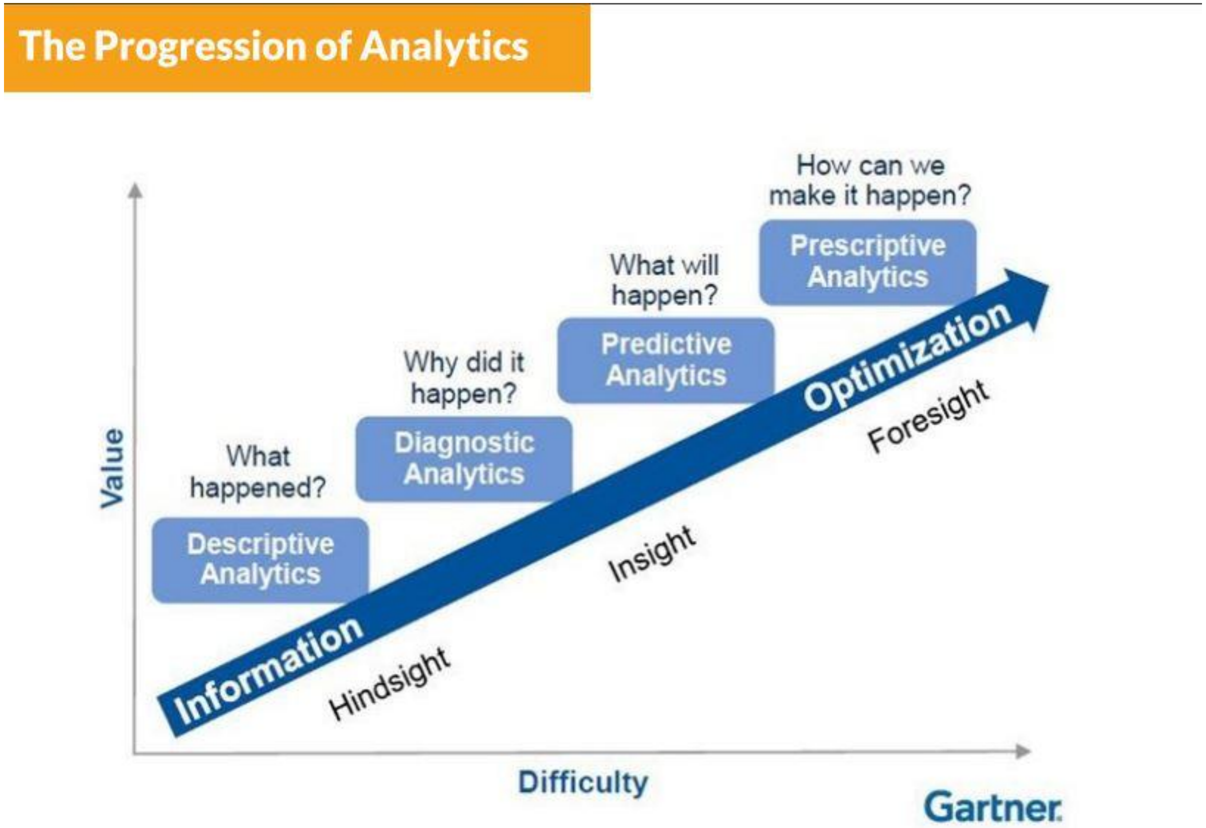
Follow me on social media: @datadanlarson

Retention Trend



Follow me on social media: @datadanlarson

Objective



Why make retention predictions?

Resource Allocation

Test the outcomes of experimental initiatives

Make smarter decisions

Strengths and Limitations

STRENGTHS

- Complete Control of the Model
- Develop custom data that may not be collected in traditional SIS or LMS
- Ability to deploy it in the most cost effective way
- Agility to change it as needed

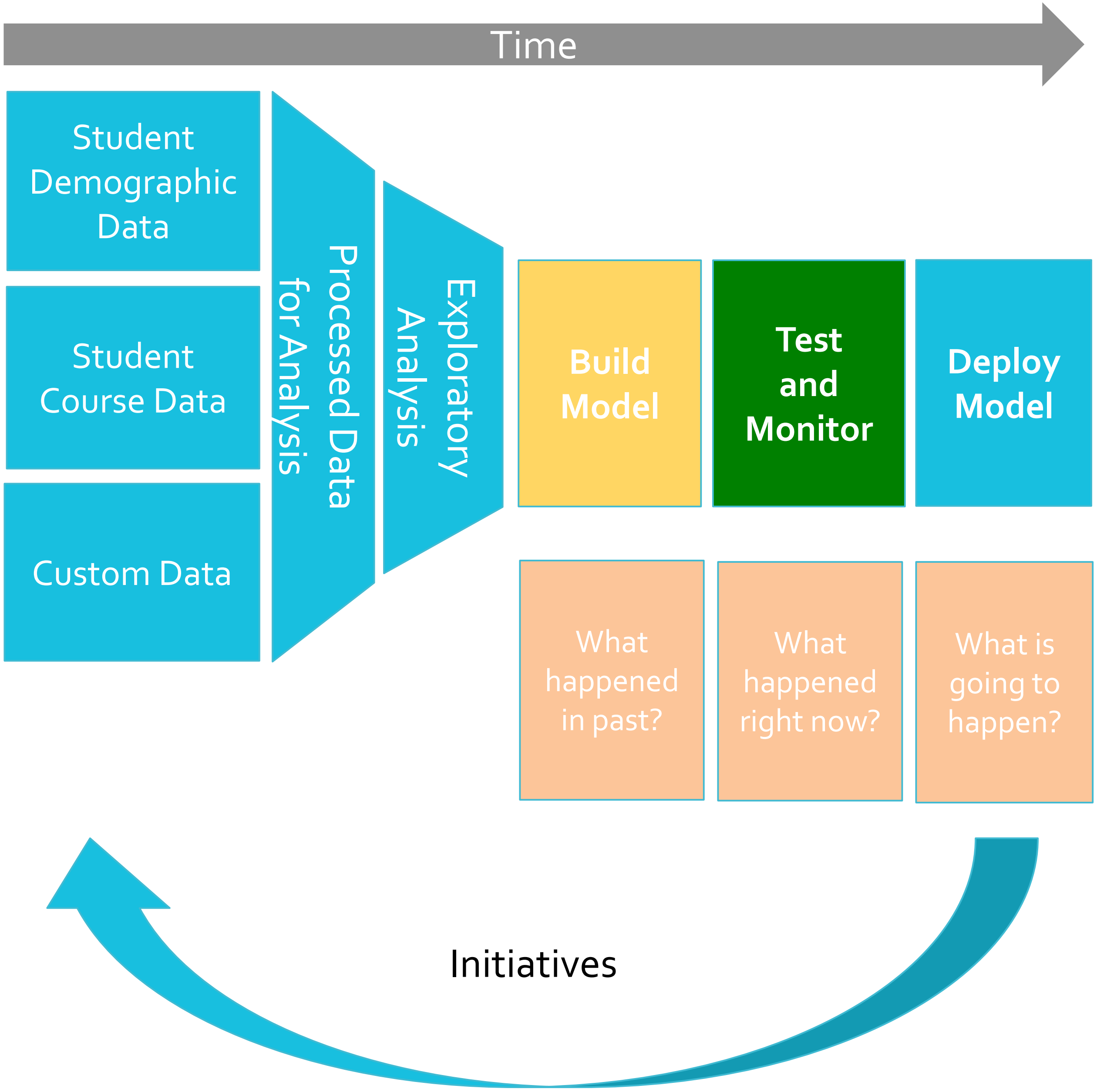
LIMITATIONS

- Limited to data we can process
- Scalability

Background Research

- Tinto (2006) – Ways students interact with social and academic environments influences whether or not they withdraw
- Bogard 2011 - Data throughout four points of the semester to predict student success. Each shift in time increased the predictive ability of the model
- Mack Sweeney et al. have attempted to utilize a recommendation system to predict student course success and retention (Sweeney 2016).
- Herzog 2006 - utilized a decision tree model and Neural Networks to predict student degree completion time. His study found that Decision Trees and Neural Networks performed at least as well as regression models
- Herzog also found that these algorithmic approaches did better at predicting more experimental variables such as time to completion than the regression models. They found that a combination of random forest and factorization machines was able to accurately predict student grades for new and returning students.

Methodology



Follow me on social media: @datadanlarson

Data Wrangling

"Data Scientists spend 60 percent of their time as *digital janitors*"

	Data Set 1	Data Set 2	Data Set 3	Data Set 4
	Student Demographic and Outcomes	Student Course Success	Student Tutoring Lab Interactions	Student Educational Plan
Number of Variables Created	27 Variables	7 Variables	3 Variables	19
Processing Method	SQL	SQL, R	SQL, R	SQL
Raw Data Dimensions	11,381 rows 30 columns	544,933 Rows 37 columns	14,800 rows 21 columns	11,361 rows 20 columns

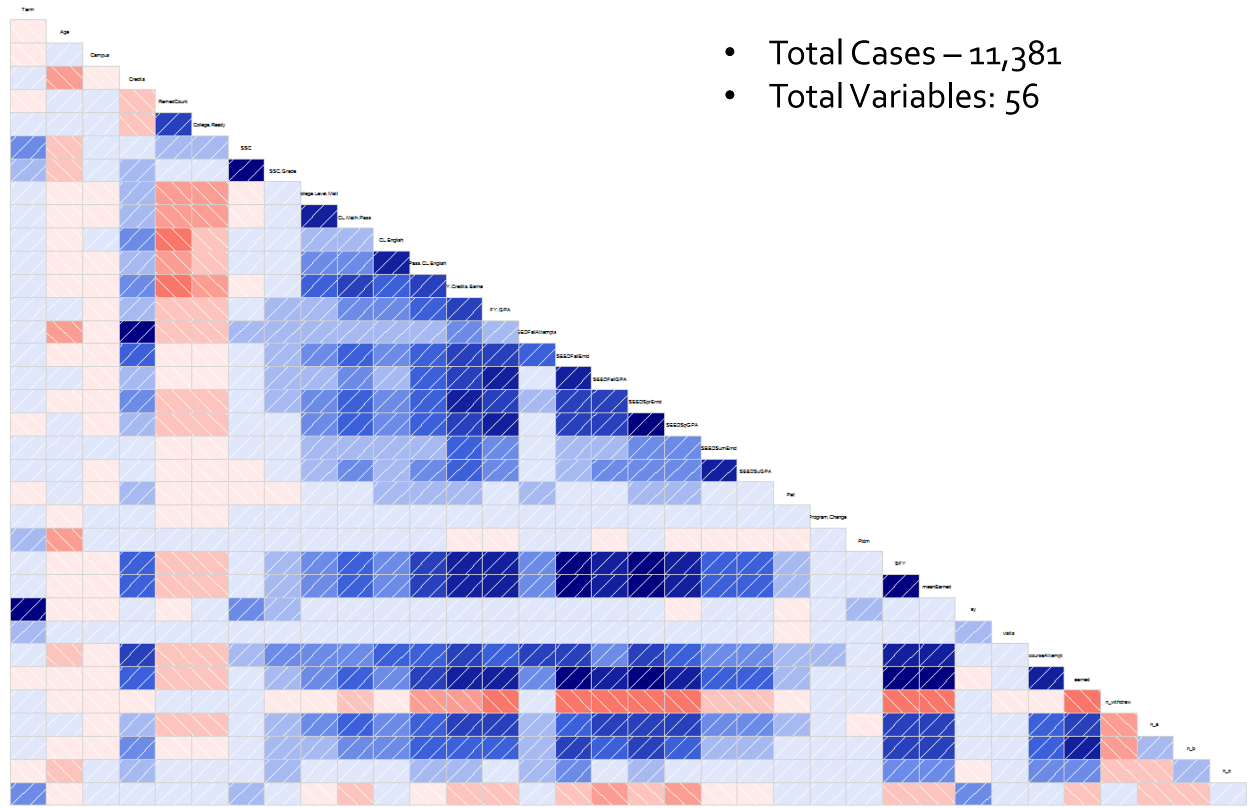
Follow me on social media: @datadanlarson

Variable Selection

Program	Fall Earned Credits	Caring for Dependents	Computer Access at home
Start Term	Fall GPA	Transportation Issues	
Gender	Spring Earned Credits	Concerned about Paying for School	Number of Withdraws
Race/Ethnicity	Spring GPA	Commitments Outside of School	Time Status
Age	Summer Earned Credits	Work Demands	Attempted Credits including Developmental
Campus	Summer GPA	Limited Timeframe to Complete	
First Term Credits Attempted	Pell Amount Received	Developmental Course Work	
Remedial Count	Number of Program Changes	Concerned about Academic Ability	
College Ready	Complete Students Educational Plan	In Reading	
Attempted Student Success Course	First Year Credits Earned (With Developmental)	In Writing	
Student Success Course Grade	Visits to Tutoring Center	In Math	
Attempted College Level Math	Time Spent with Tutor	First Generation Student	
Completed College Level Math	Average Time with Tutor	Uncertain about Major	
Attempted College level English	Number of course attempts	Uncertain about decision to attend college	
Completed College Level English	Number of As, Bs, Cs, and Fs	Internet Access at home	

Exploratory Analysis

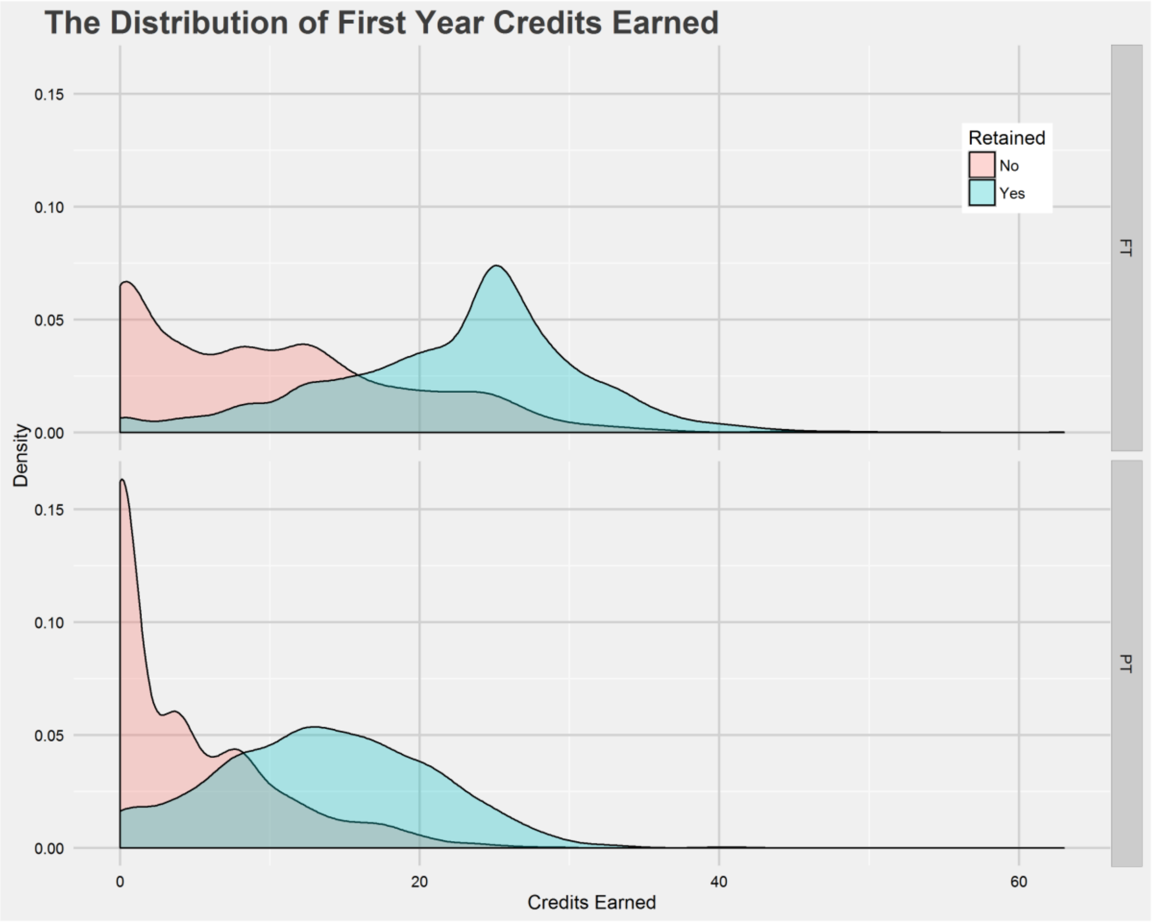
Student Retention Data



- Total Cases – 11,381
- Total Variables: 56

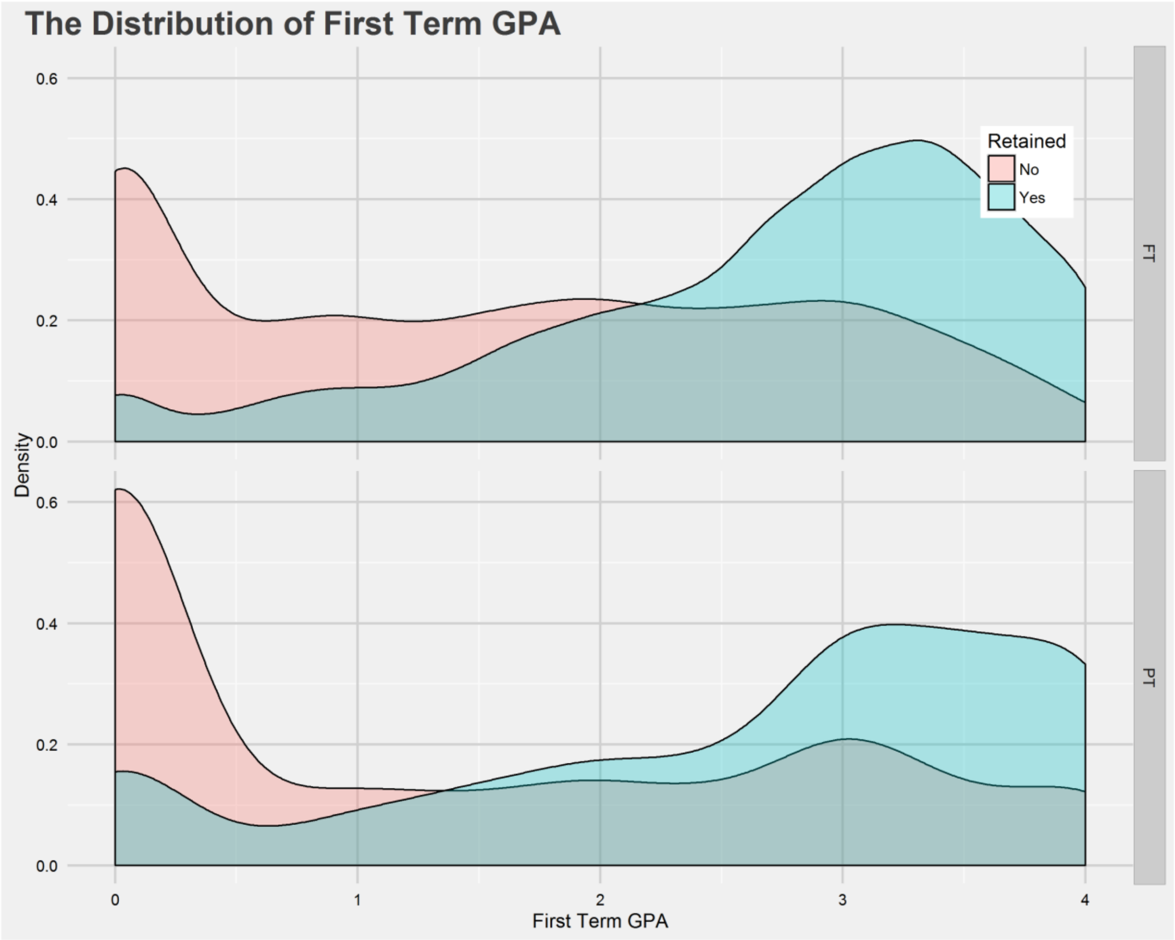
Follow me on social media: @datadanlarson

Variables that Matter



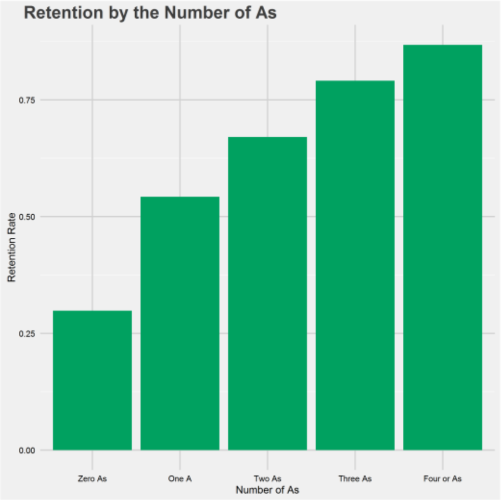
Follow me on social media: @datadanlarson

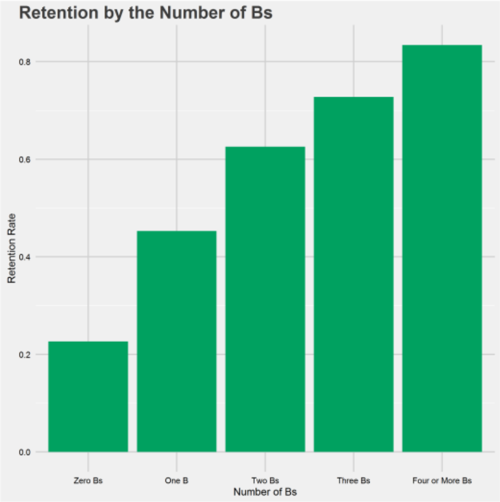
Variables that Matter

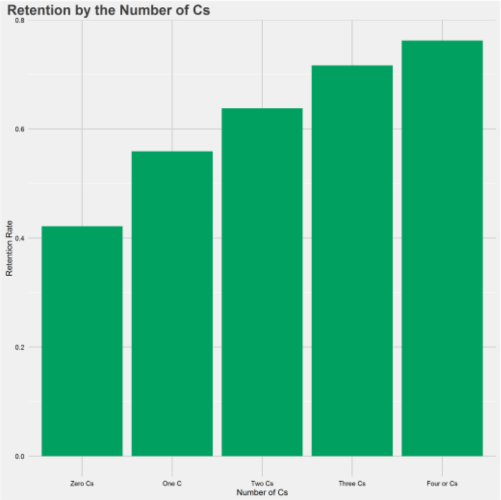


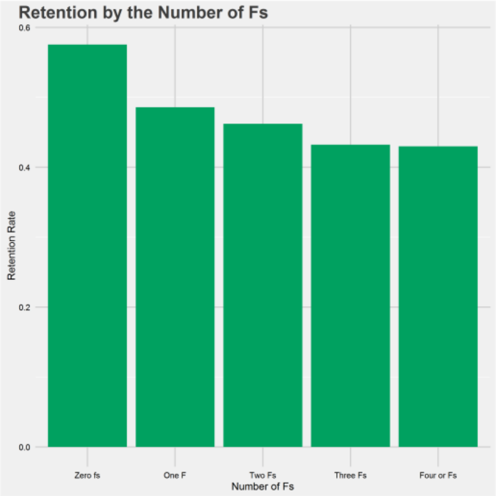
Follow me on social media: @datadanlarson

Variables that Matter



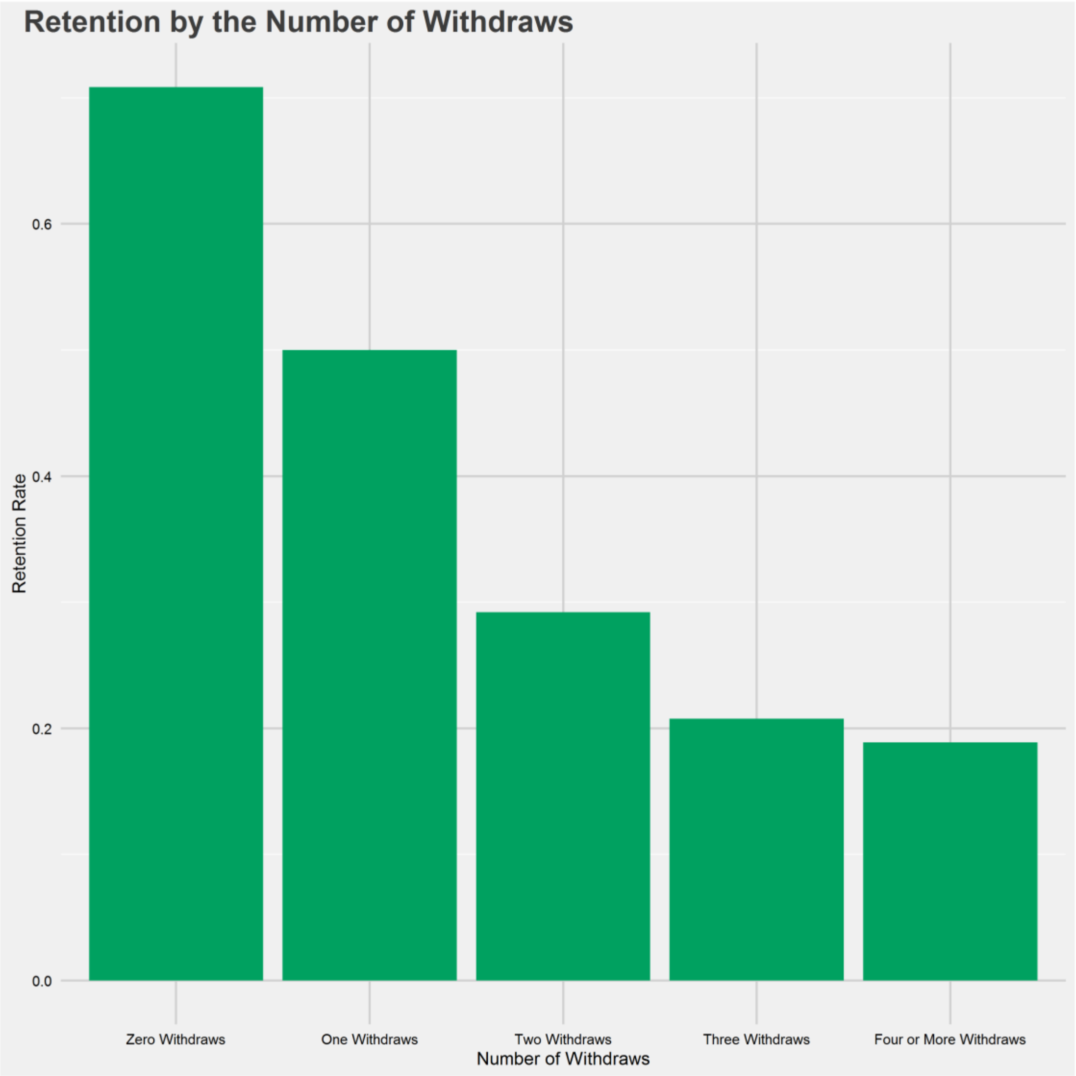






Follow me on social media: @datadanlarson

Variables that Matter



Follow me on social media: @datadanlarson

Random Forest

- To improve the decisions tree model, many machine learning experts utilize Ensemble Learning methods that generate classifiers and aggregate their results. (Liaw and Wiener 2002).
- The Random Forest method, reduces the number of variables used by creating multiple trees and aggregates them thereby providing a more accurate predictions (Leo Breiman 2001).

Model 1: Using Pre-registration info

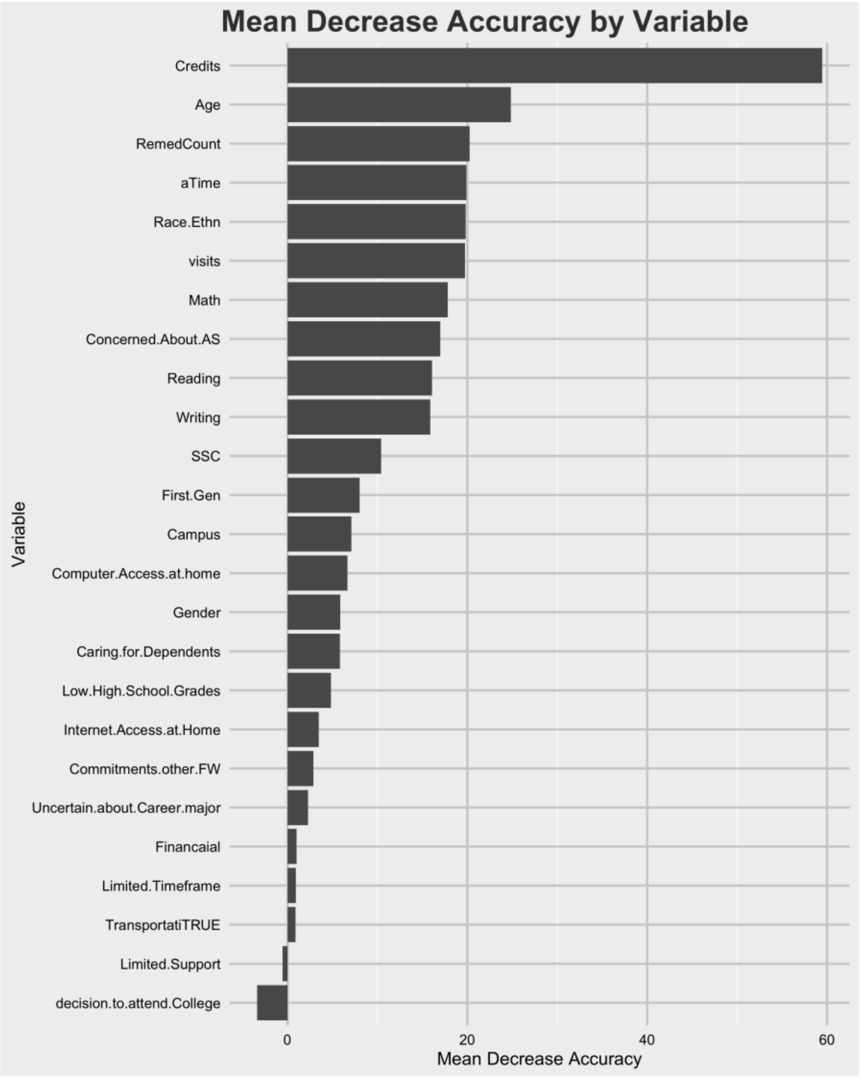
```
model <- randomForest(as.factor(retained)~Gender+Race.Ethn+Age + Campus +
                              Credits + RemedCount + SSC + Caring.for.Dependents +
                              TransportatiTRUE + Financaial + Commitments.other.FW +
                              Limited.Timeframe + Low.High.School.Grades +
                              Concerned.About.AS + Reading + Writing + Math +
                              First.Gen + Limited.Support + Uncertain.about.Career.major +
                              decision.to.attend.College + Computer.Access.at.home+
                              Internet.Access.at.Home + visits + aTime,
                    data=train2,importance=TRUE, ntree=500,mtry=3,type="prob")
```

- Type of random forest: classification
  - Number of trees: 500
  - No. of variables tried at each split: 3
  - OOB estimate of error rate: 38.07%
- Confusion matrix:

	Predict Leave	Predict Retain	Class Error
Leave College	1,987	2020	50.4%
Retained	1,211	3,270	27.0%

Follow me on social media: @datadanlarson

Variables Selected



Follow me on social media: @datadanlarson

The Second Model: End of First Term

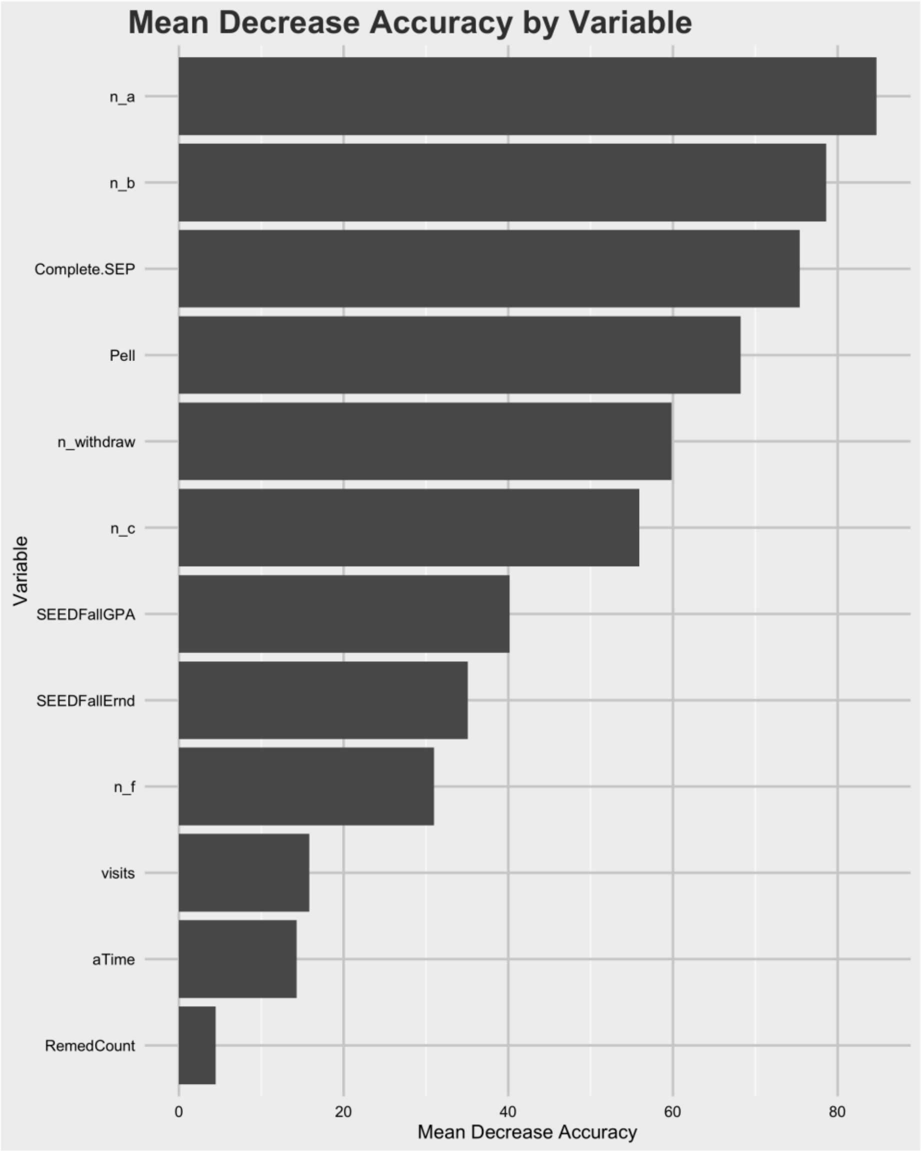
```r
fit <- randomForest(as.factor(retained) ~ Pell + SEEDFallGPA + Complete.SEP +
                         SEEDFallErnd +n_withdraw + n_a + n_b + n_c+ n_f +
                              visits + aTime ,
                         data=train, importance=TRUE, ntree=500,mtry=3,type="prob")
```

- Type of random forest: classification
  - Number of trees: 500
  - No. of variables tried at each split: 3
  - OOB estimate of error rate: 21.07%
- Confusion matrix:

	Predict Leave	Predict Retain	Class Error
Leave College	2,983	1024	.2556
Retained	764	3717	.1705

Follow me on social media: @datadanlarson

Variables Selected



Follow me on social media: @datadanlarson

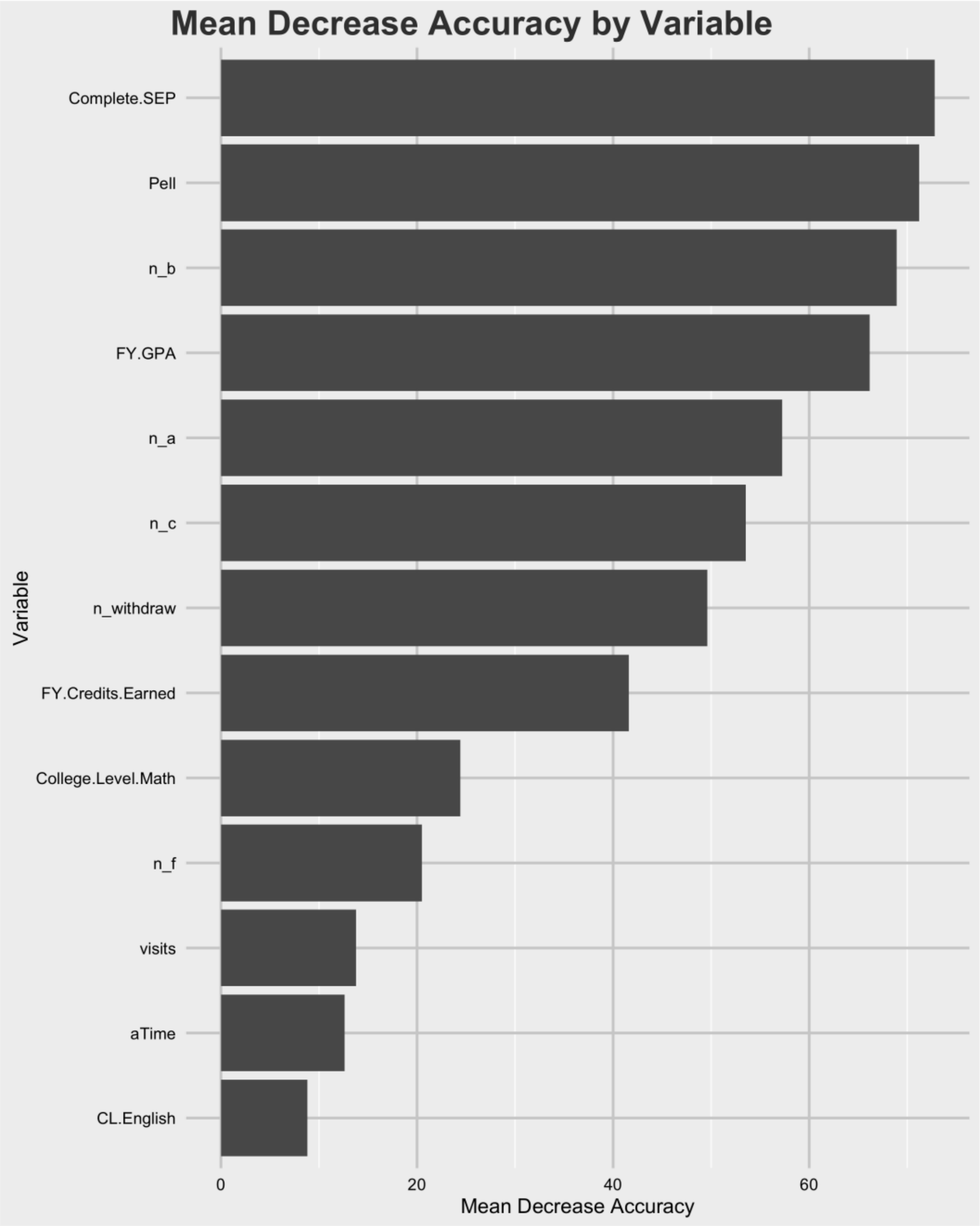
The Third Model: End of First Year

```
model2 <- randomForest(as.factor(retained) ~ Pell + FY.GPA + Complete.SEP +
                                FY.Credits.Earned + College.Level.Math + CL.English +
                                n_withdraw + n_a + n_b + n_c+ n_f +
                                visits + aTime ,
                          data=train, importance=TRUE, ntree=500,mtry=3,type="prob")
```

- Type of random forest: classification
  - Number of trees: 500
  - No. of variables tried at each split: 3
  - OOB estimate of error rate: 20.98%
- Confusion matrix:

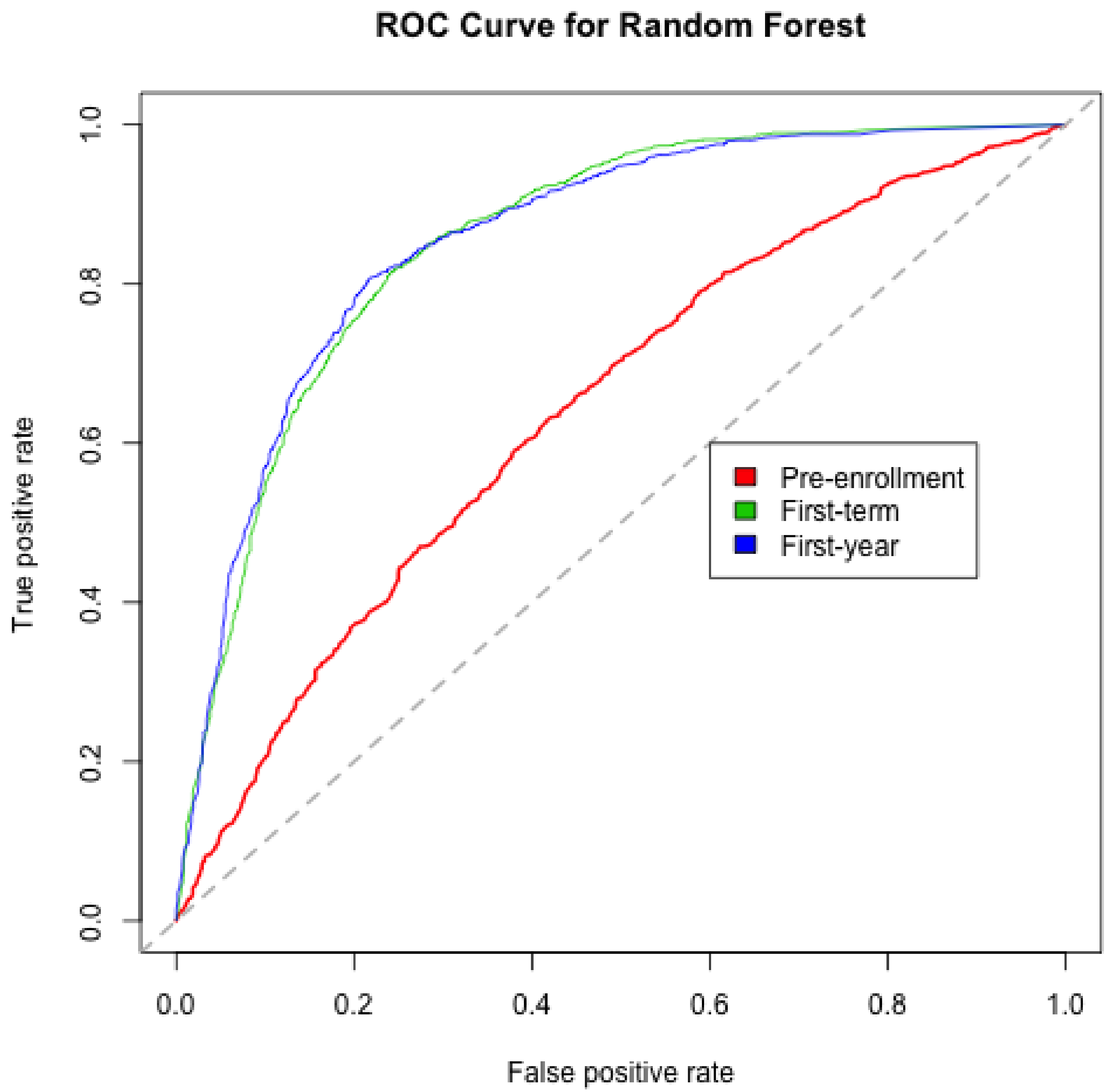
	Predict Leave	Predict Retain	Class Error
Leave College	3,049	958	.2391
Retained	823	3,658	.1837

Variables Selected



Follow me on social media: @datadanlarson

The Test Data



Follow me on social media: @datadanlarson

Future Steps

- Deploy the model for use by college advisors and faculty
  - Monitoring/adjusting model as needed
  - Collect data on communications with students
- Transition from predicting fall to fall retention to predicting term to term retention
- Integrating Learning Management System data

Learning R or Python

Follow me on social media: @datadanlarson

@datadanlarson