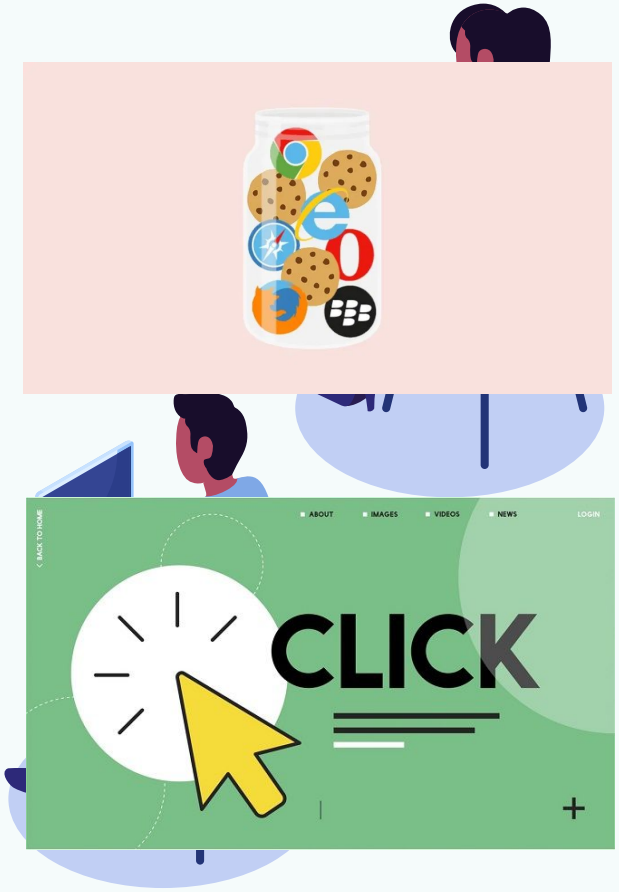
Predicting Student Outcomes in Virtual Learning Environments

Jaime DyBuncio
5/15/2020
Github

Motivation



Dataset

- Anonymised Dataset which covers **32k student outcomes from 7 Total Modules**
  - Modules presented during 4 terms: Feb 2013, Oct 2013, Feb 2014, Oct 2014
  - Avg presentation length was 8.5mos

- 7 Data Tables containing information on the:
  - Course, Assessments, Students, Registration, Scores, **VLE material**, and **VLE Logs**

Question

Can one predict who will Fail the course on the first day of class?

Potential Application

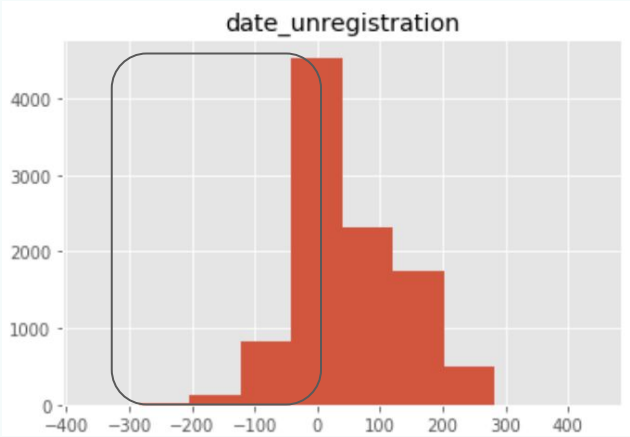
- **Educator:**
  - Early Intervention
- **Administrator**:
  - Maximize Retention -> Maximize Tuition

Optimization

- **AUC**
  - Maximize *TPR*: Predict maximum % students who Fail
  - Minimize *FPR*: Minimize % of students predicted to fail, who Pass (intervention cost)

EDA

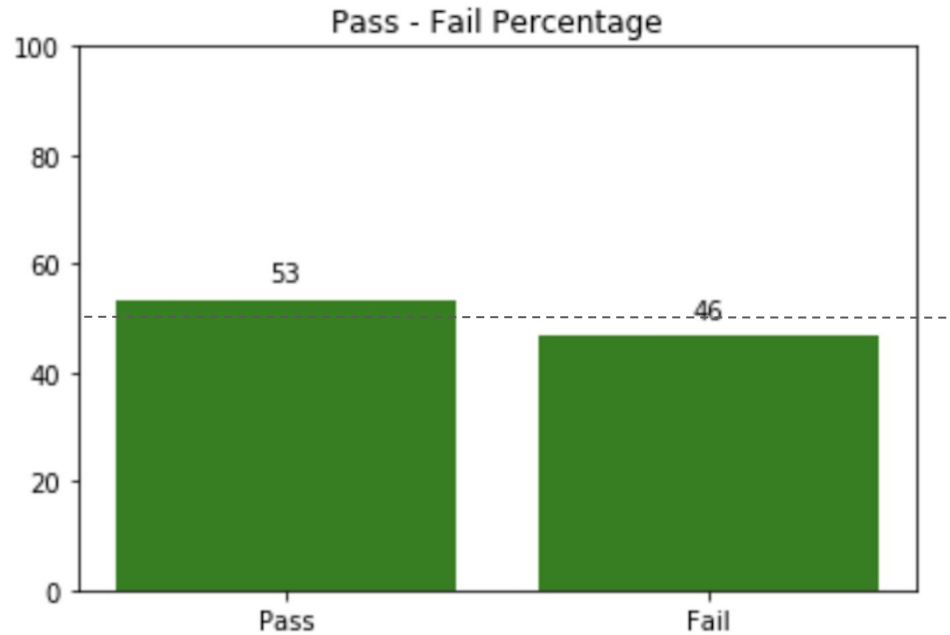
Potential Leakage - Class Balance - Relationships



- 9% of Students Unregister (and Fail) before Day 0 of the Course (**3k**)
  - Removed these Students (left with 29k students)

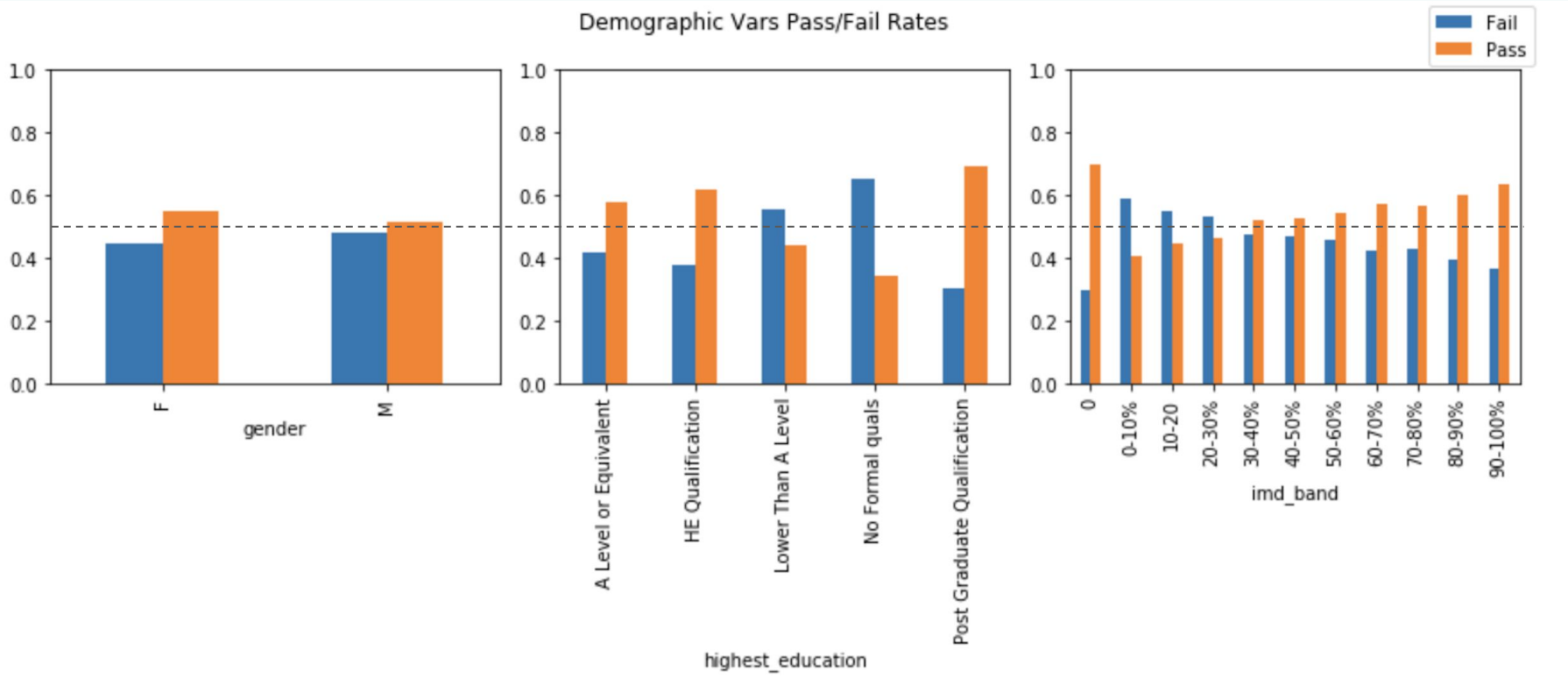
EDA

Potential Leakage – **Class Balance** – Relationships



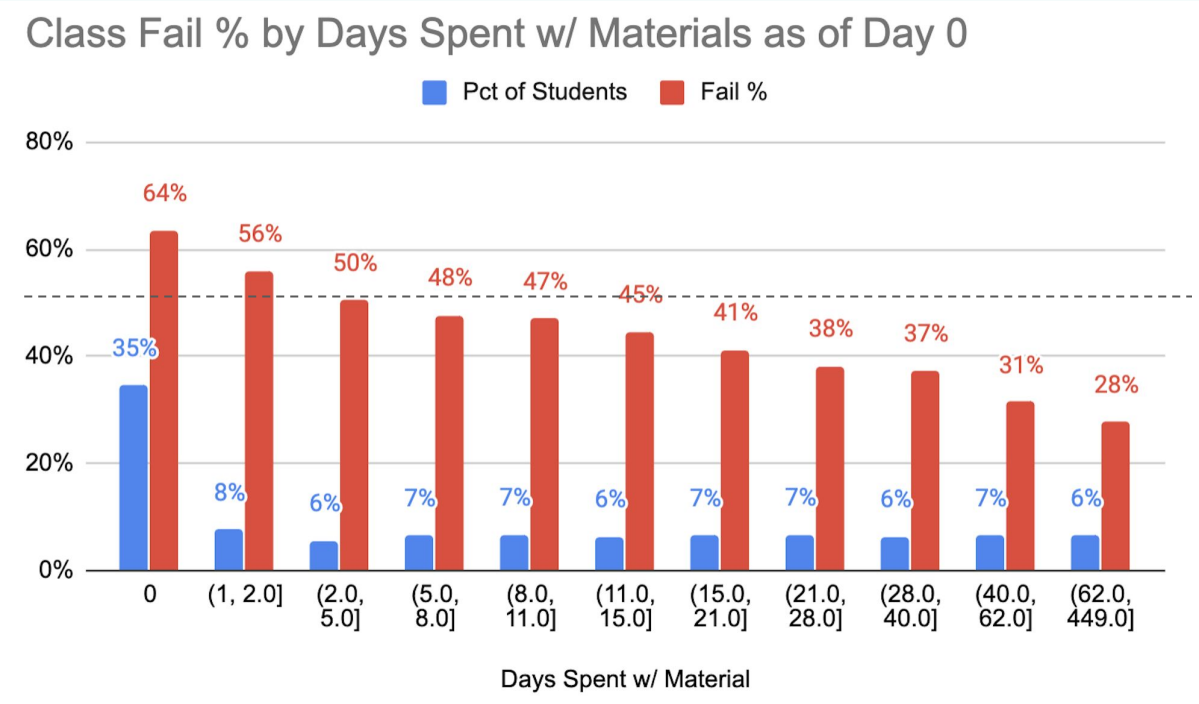
EDA

Potential Leakage – Class Balance – Relationships

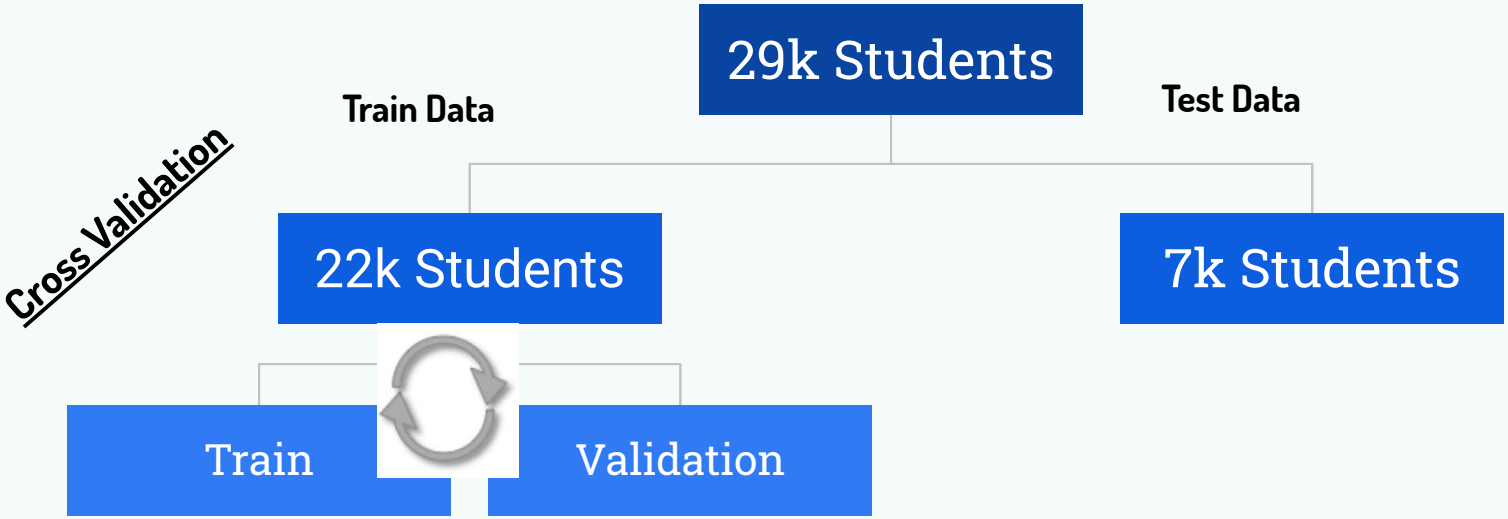


EDA

Potential Leakage – Class Balance – Relationships



Modeling Approach



107 Features (99 of which were created)

- **Demographic**
- **Click Data w/ materials before first day**

Model Progression in Cross Validation

Model	AUC (w/o Click Data)
Highest Education	0.578

Model Progression in Cross Validation

Model	AUC (w/o Click Data)
Highest Education	0.578
Logistic Regression	0.65
Random Forest	0.65
AdaBoost	0.65
Gradient Boost	0.66

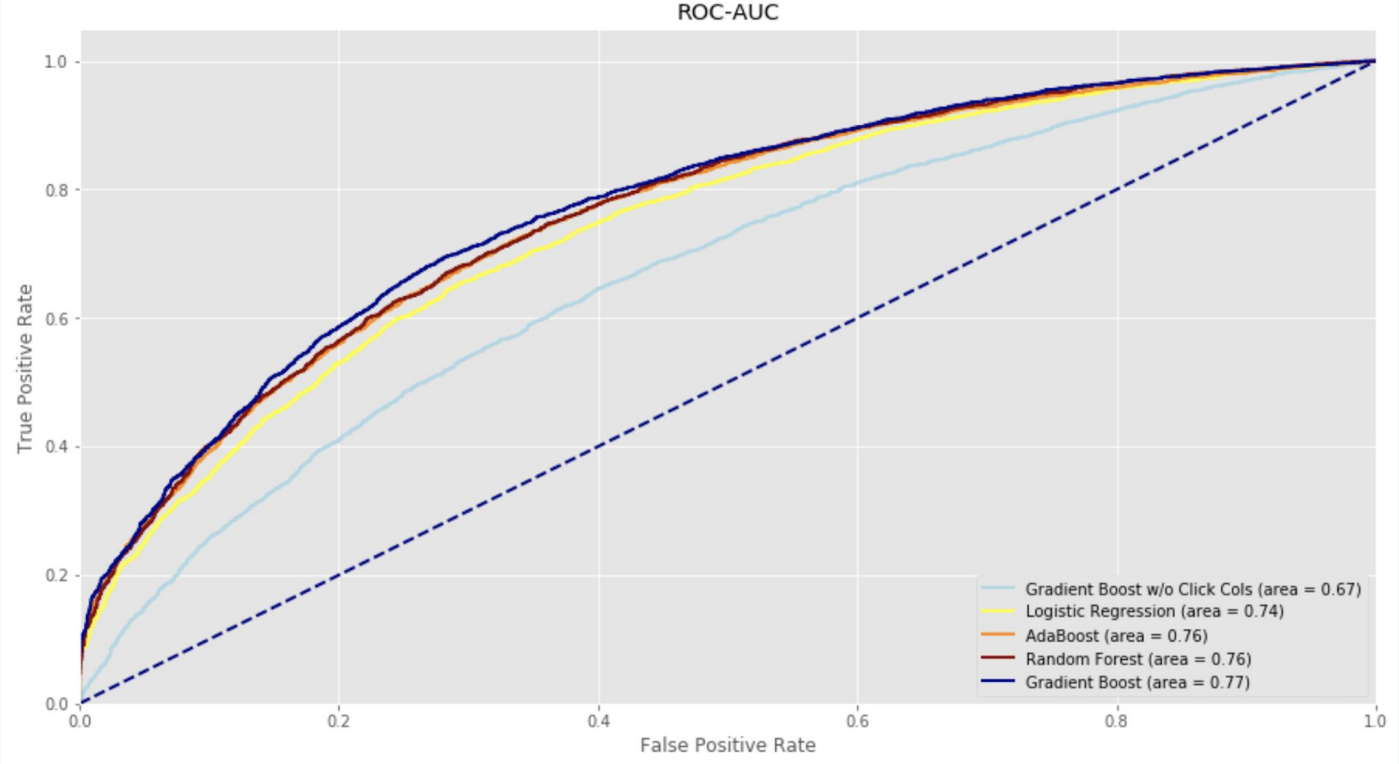
Model Progression in Cross Validation

Model	AUC (w/o Click Data)	AUC (w/ Click Data)
Highest Education	0.578	
Logistic Regression	0.65	0.764
Random Forest	0.65	0.765
AdaBoost	0.65	0.75
Gradient Boost	0.66	0.77

Model Progression in Cross Validation

Model	AUC (w/o Click Data)	AUC (w/ Click Data)	AUC (w/ Tuning)	AUC (Final after GS)
Highest Education	0.578			
Logistic Regression	0.65	0.764		
Random Forest	0.65	0.765	0.766	0.768
AdaBoost	0.65	0.75	0.766	0.766
Gradient Boost	0.66	0.77	0.776	**0.777**

Tuned Models vs. Test Data

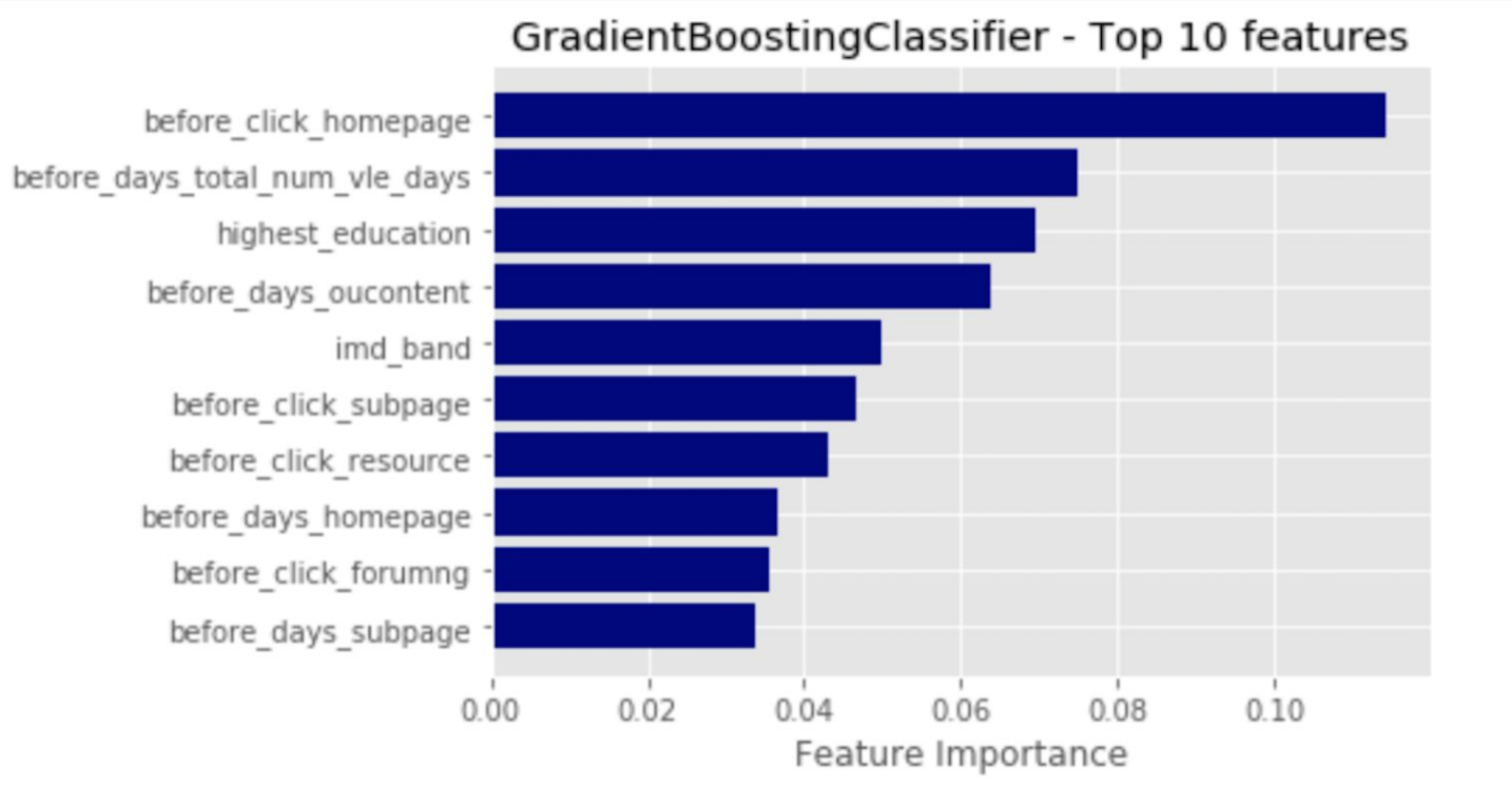


At a threshold of 0.39:

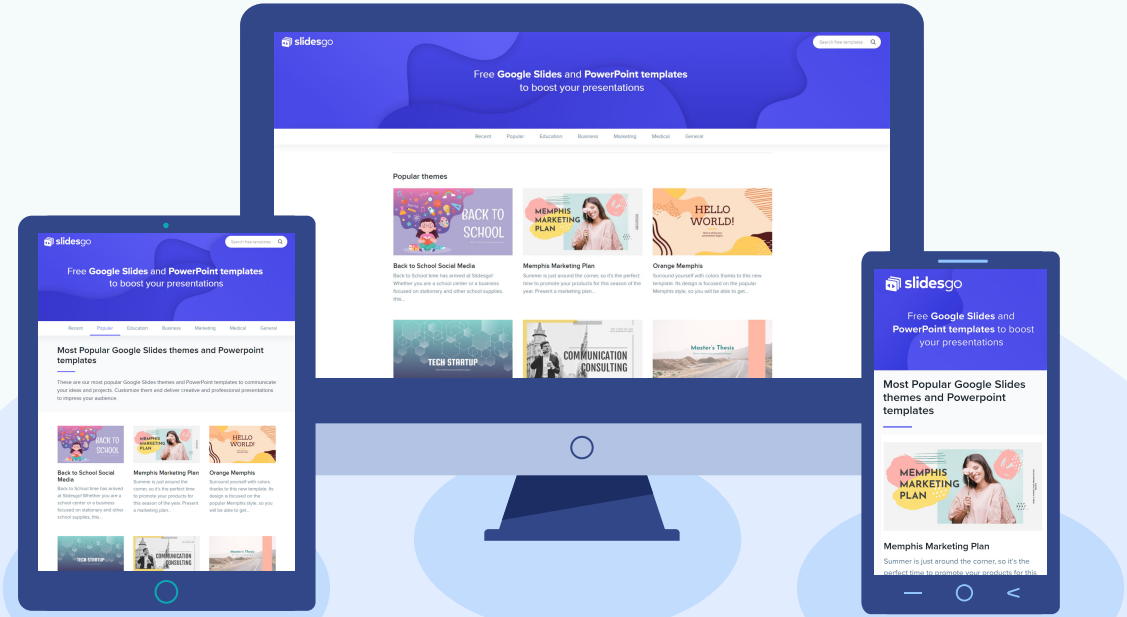
- TPR = 0.80
- FPR = 0.42

Can predict **80%** of the students who will Fail on first day of class.
But will also would predict **42%** of those who will pass, would Fail.

Feature Importance



Conclusion & Next Steps



THANK YOU


slidesgo

Motivation

The New York Times

Fearing a Second Wave, Cal State Will Keep Classes Online in the Fall

The move by the nation's largest four-year public university system comes as many other schools insist they will find a way to bring students back to campus despite the coronavirus.