Final presentation

Analyze and improve online student retention
PJ Shetty

Presentation video: https://youtu.be/5BIduHIdqJs

Context and scope

Online education has grown tremendously over the years.

Free online courses, including Massively Open Online Courses(MOOCs), have great potential to increase the inclusiveness of education.

Online classes especially MOOCs have a low retention rate when compared to traditional classes.

Understand key features that influence online student retention and build a ML model to predict student success. The model can be used by educators to identify at risk students whose retention rate can be improved by timely interventions.

Related work

There has been various studies done to understand student retention but many of these studies are related to traditional schools and on ground classes.

Student engagement would be defined differently for online students. Students online activity such as course views, videos watched, number of chapters with which the student interacted, number of posts added to the discussion forum would be considered as engagement for online students.

Data, data, data!!

The model can be useful if it can provide predictions on student retention early on in the term. To achieve this, we need to train the model on a time series dataset.

A time series dataset would include multiple records for a student during different times in a term. The model would then be able to predict student retention at any given point in the term, giving early information of at risk students to educators.

I was not able to find a time series online student retention dataset. Such datasets are not made public by schools due to FERPA(Family Educational Rights and Privacy Act).

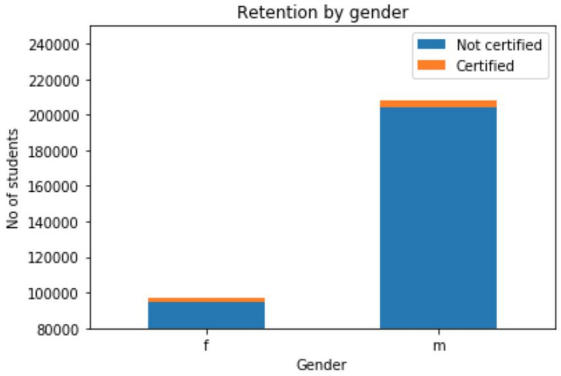
MOOC data

The second best thing I could find was the Harvardx MOOC dataset.

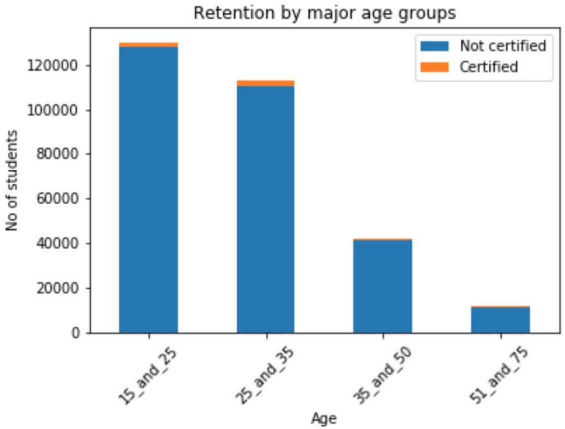
The dataset is at the level of one row per-person, per-course. It included data from some of the online courses available during Fall 2012, Spring 2013, and Summer 2013.

The dataset includes student demographic features along with student activities. It also includes a feature that indicates if a student got certified. This would translate to retention and we would use this feature as our target variable to train the model.

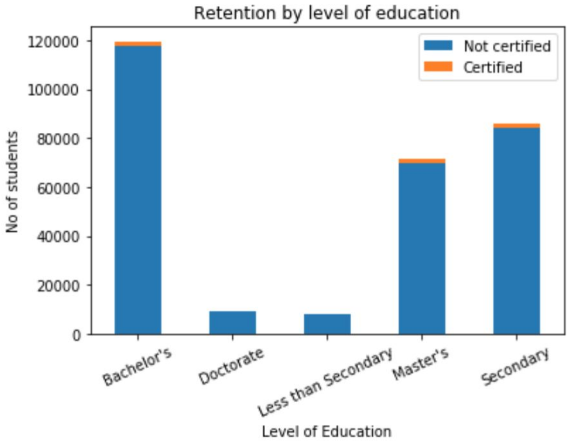
Data visualization



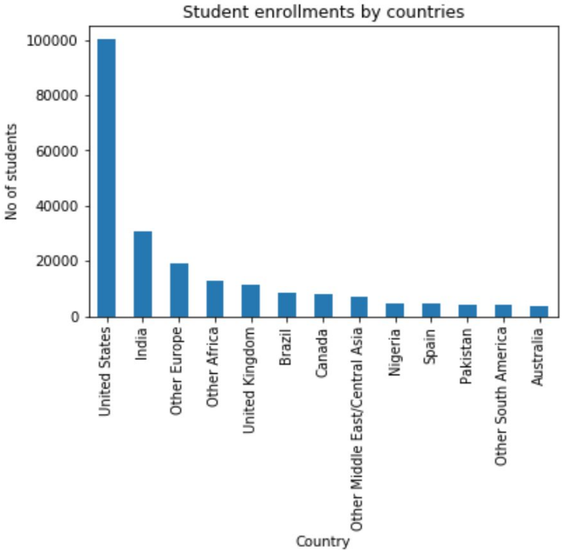
Retention by gender



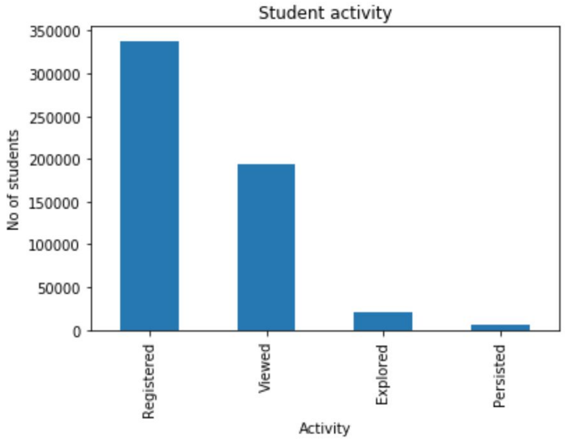
Retention by major age groups



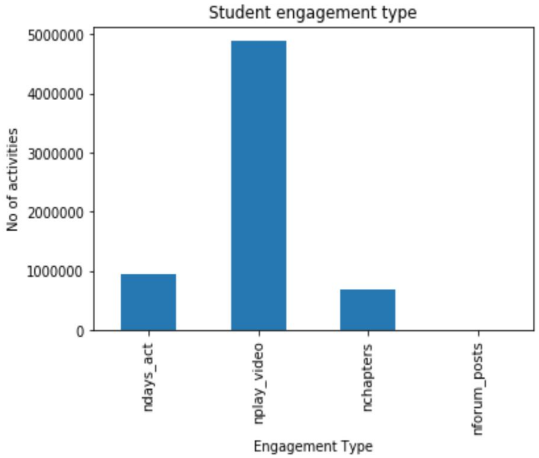
Retention by level of education



Student enrollments by countries



Student activity



Student engagement type

Hypothesis testing

Hypothesis testing show demographic features have no impact on retention while student activity features do impact retention.

Hypothesis	**P-value**	**Result**
Age has no impact on retention	0.53	We accept the null hypothesis
Gender has no impact on retention	0.37	We accept the null hypothesis
Student activities has no impact on retention	Close to 0	We reject the null hypothesis

ML models

Since we have a imbalanced dataset, ie. the number of students who got certified are far less than students who didn't certify, we need to make sure the minority class accuracy is correct. Hence we rely on test parameters like precision and recall. F1 score represents both these parameters and hence we use this to evaluate our models.

Model	Accuracy	Precision	Recall	F1 score
Logistic Regression	99.7%	98.3%	85.1%	89.9%
Random Forests	99.8%	99.1%	94.6%	96.7%
Support Vector Machines	99.7%	98.4%	84.4%	89.4%
XGBoost	99.9%	99.8%	94.3%	96.8%

Model usage

The model helps identify at risk students

Educators can use this information to send interventions to these students and help improve student retention

Interventions can include sending nudges to improve motivation, providing specific resources to at risk student population, counseling, improving time management and providing study aids.

Tools used to execute Hypothesis testing and build model

I used Anaconda Jupyter notebook to do my analysis.

You can find the notebook on GitHub.

Dataset can be found at harvard data.