Analyze and improve online student retention

PJ Shetty
pshetty35@gatech.edu

Abstract**—**This paper looks at various factors that affect online student retention. We use hypothesis testing on MOOC (Massive Open Online Course) dataset released by Harvard and MIT (Massachusetts Institute of Technology) to analyze the key features for online student retention. We then build a machine learning model to predict student retention. The goal of this project is to identify key risk students early on in the course and help educators intervene and improve student's level of engagement, thereby increasing student success.

1 INTRODUCTION

Online education has grown tremendously over the past few years. The increased accessibility of the internet and the World Wide Web has created vast opportunities for non-traditional education through this medium (Karber, 2003). The access to technology has made it possible for teachers to teach outside the traditional classroom and has also provided learners with easy access to course material. Free online courses, including Massively Open Online Courses, have great potential to increase the inclusiveness of education. However, one of the main challenges with online education is the high dropout rate. Studies show that, although online class registrations are increasing, up to 40-80% of online students drop out. My research would look at the key indicators that affect retention and build a machine learning model to predict student outcomes.

2 RELATED WORK

There have been multiple studies conducted to understand the key features to determine student success outcomes. Studies by Le et al (2016) has shown past term GPA, high school GPA, test scores and gender as some of the factors that can help predict persistence in traditional schools. Buenaño-Fernández, Gil & Luján-Mora (2019) studies have shown that intermediate grades can be a strong feature to predict the final grade of a student. Studies by Rodríguez-Muñiz, Bernardo, Esteban & Díaz (2019) shows full time vs part time, age as strong indicators that can predict dropout rate. Conijn, Van den Beemt & Cuijpers (2018) researched that LMS (Learning Management System) activities which indicates the student activity, is a good feature to use in a student success outcome. This has been also confirmed by Hlosta et al (2018) who show that student activity is a key feature in determining student outcomes. Engaged students tend to succeed more in online courses.

Tinto (1975, 1993) built the Student Integration Model (SIM) that explained the drop out of traditional students in a residential higher education setting. Tinto's model dealt on factors associated with the educational institution and student experience. Bean and Metzner (1985) built the Student Attrition Model (SAM) that looked at non-traditional students. It built upon Tinto's work and explained the attrition of students that do not live on-campus, do not belong to campus-affiliated social groups or organizations, typically have families, and attend part-time. Bean and Metzner's SAM model places additional emphasis on factors external to the learning institution. For MOOC students, the retention is significantly low. We will build upon the above work and look at factors that affect online students. We will leverage these factors to build a prediction model using machine learning to forecast student outcomes.

3 THE SOLUTION

I look at key features that affect online retention. This was accomplished by a series of hypothesis testing. I build a prediction model by using various ML models like Logistic Regression, Random Forests, Support Vector Machines(SVM) and XGBoost. I measured the results using accuracy, precision, recall and F1 score as the test parameters. The solution and test methodologies are explained in detail under the Methodology section below. The entire analysis from data cleaning to ML model testing was performed in python using an Anaconda jupyter notebook.

4 METHODOLOGY

4.1 Data Description

The dataset is at the level of one row per-person, per-course. The original dataset included students taking MITx or HarvardX courses during Fall 2012, Spring 2013, and Summer 2013. However, the revised dataset on which the analysis was done only includes HarvardX courses. The data has been anonymized to protect student identity and meet the requirements of FERPA(Family Educational Rights and Privacy Act).

Table 1—Feature descriptions

Feature Name	Feature Type	Description
course_id	Categorical	Institutional course ID
userid_DI	Categorical	Anonymized student ID
registered	Numeric	Indicates if a student is registered for a course
viewed	Numeric	Indicates if a student has viewed videos, problem sets or exams
explored	Numeric	Anyone who has accessed at least half the chapters
certified	Categorical	Anyone who earned a certificate
final_cc_cname_DI	Categorical	Student country name
LoE	Categorical	Highest level of education
YoB	Categorical	Year of Birth
gender	Categorical	Student gender
grade	Numeric	Final grade in the course, between 0 and 1
start_time_DI	Categorical	Date of course registration
last_event_DI	Categorical	Date of last interaction with course
nevents	Numeric	Number of interactions with the course
ndays_act	Numeric	Number of unique days student interacted with the course
nplay_video	Numeric	Number of videos played in course
nchapters	Numeric	Number of chapters with which the student interacted
nforum_posts	Numeric	Number of posts added to the discussion forum

4.2 Data cleaning

For multiple students, a number of numeric features were empty. I imputed a value of zero to replace the empty fields. For the machine learning model to work, I had to replace the categorical features with numeric values. I did this with label encoding - a process to replace each unique categorical value in a feature with a numeric value starting with zero. Any row in the dataset missing a value in the last_event_DI column was replaced with the value in start_time_DI. Also, since these 2 columns are possibly correlated, I added a new column which is a difference of these 2 columns and got rid of the original 2 columns.

4.3 Exploratory data analysis

The goal of my exploratory analysis is to obtain a visual sense of how features relate to retention. We first look at the student demographic features.

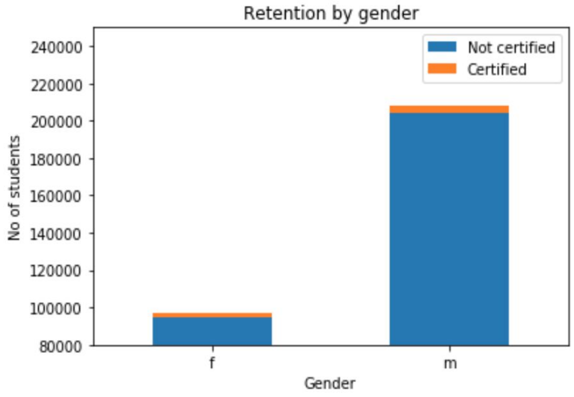

Figure 1—Retention by student gender

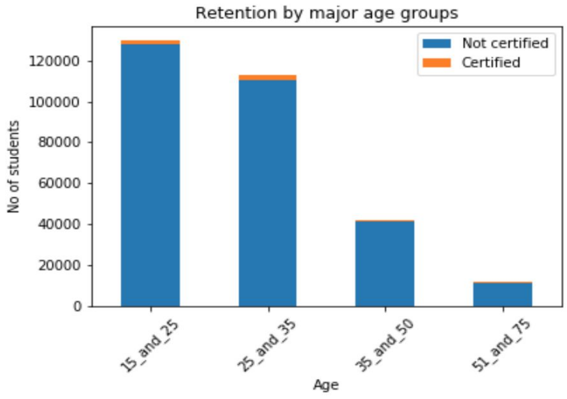

Figure 2—Retention by age groups

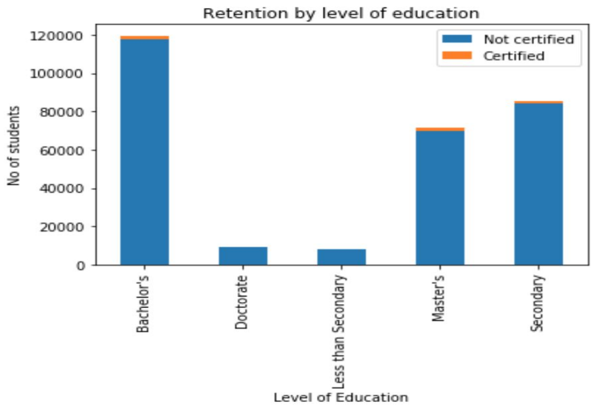

Figure 3—Retention by level of education

From the charts, we don’t see significant separation between certified and non certified students indicating demographics features may not be good indicators to predict student retention.

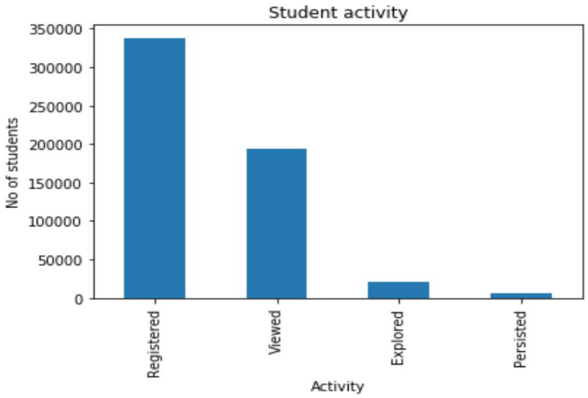

Figure 3—Count of student activities

Multiple studies in the past have shown low completion rates in online education. The above chart confirms this. We can see that a large number of students register but only a fraction of the students complete the course.

4.4 Hypothesis testing

Some of the key hypothesis tests are listed below. A random sample of 1,000 certified and 1,000 non certified students were selected for this testing. Mean value of the feature for both these groups were calculated and t-testing was performed to get the probability value or the P-value. A P-value helps us determine if the test is statistically significant. A P-value with a value less than 0.05 indicates that the hypothesis test is false and should be rejected.

4.4.1 *Age has no impact on retention*

Mean age of certified students is 29.01

Mean age of non certified students is 28.45

P-value is 0.13. Since the P-value is greater than 0.05 we accept the null hypothesis.

4.4.2 *Gender has no impact on retention*

To calculate the mean, men were given a value of 1 and women a value of 0.

Mean gender value of certified students is 0.65

Mean gender value of non certified students is 0.68

P-value is 0.2. Since the P-value is greater than 0.05 we accept the null hypothesis.

4.4.3 *Student engagement has no impact on retention*

Average number of activities of certified students is 2564.2

Average number of activities of non certified students is 122.89

P-value is close to 0. Since the P-value is less than 0.05 we reject the null hypothesis.

4.5 Machine learning models

I used supervised learning ML methods to train and test the model since we have a labelled dataset. Our target is the certified feature which determines student retention, and we will train our model to predict this in datasets with unknown retention.

The preprocessed data is normalized before it is used to train the model. There could be features whose values differ in magnitudes by a large extent. The normalized data helps put data on one scale. Normalized data will have a mean of 0 and a standard deviation of 1.

The scaled data is trained and tested using four different ML models. Logistic Regression, Random Forests, Support Vector Machines and XGBoost. For each supervised learner, 5 cross validated models were run, splitting the data in each iteration into a training and testing set. For each iteration of cross validation, the training set contains 80% of the data while the test set contains 20% of the data.

Our dataset is an imbalance dataset. The non certified students far exceeds the certified students. This is not surprising as online students in general, have a low retention.

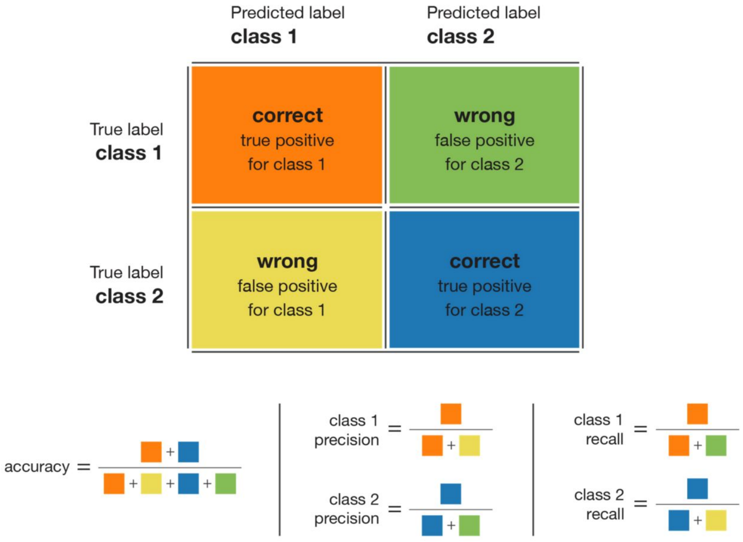

Figure 4—Confusion matrix, accuracy, precision and recall. Source: TDS

Since the data is classified into certified and non certified students, we can use confusion matrix, accuracy, precision and recall parameters to test our models. For an imbalanced dataset, accuracy won't be very useful. We can get a high accuracy, but this can be due to the model correctly predicting the majority class i.e. in this case the non certified students. But we would also like to know how the model predicted the minority class which is the certified students. We can use both precision and recall for this. Precision = True Positives / (True Positives + False Positives) and Recall = True Positives / (True Positives + False Negatives). We can also use F1 score which considers both precision and recall defined as 2*((precision * recall) / (precision + recall))

4.5.1 *Logistic regression(LR)*

LR is used when the dependent variable(certified) is dichotomous in nature. It gives the probability of the outcome. In this case it gives a probability that a student is certified. In LR, the outcome follows the Bernoulli probability distribution. The LR model gave a F1 score of 89.9% indicating its strong predictive power. The key features based on model coefficients are grade, viewed, explored, country and nevents.

4.5.2 *Random Forest(RF)*

RF classifier uses a number of decision trees to predict the outcome. For our dataset, each tree will vote to give a classification, i.e. where a student will get certified or not. The forest chooses the classification having the most votes. We used 100 trees to train the model. The RF model gave a F1 score of 96.7% which is better than LR. The key features are grade, nevents, nchapters, ndays_act and explored.

4.5.3 *Support Vector Machines(SVM)*

A SVM is a supervised machine learning model that uses classification algorithm methods on data like our MOOC dataset. For the 2 types of data points, in our case certified and non certified students, SVM builds a hyperplane that best separates the data points. The best hyperplane or the decision boundary is the one that maximizes the margin for each category i.e. the hyperplane whose distance to the nearest element for each category, is the largest. SVM works with both linear and non linear data. The SVM model gave a F1 score of 89.4%. The top features are grade, explored, course_id, nevents and country.

4.5.4 *XGBoost(XGB)*

Like RF, XGB is also an ensemble learning method in that it uses a number of decision trees to predict by combining the outputs from individual trees. However, XGB builds trees one at a time, where each new tree helps to correct errors made by previously trained trees. XGBoost builds on weak learners to generate a final strong learner. The XGBoost model gave a F1 score of 96.7%. The top features are grade, course_id, nevents, nchapters and nplay_video.

5 RESULTS

The table shows how each ML algorithm stacks up against the others. Based on F1 score, RF and XGBoost have the best performance.

Table 2—Model test results based on 5 fold cross validation

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

From each model’s top features, we found that student engagement plays a significant role in student retention. These features are key in separating the certified vs non certified students. The more activities a student has, the more likely he or she to succeed. This is also reflected in our hypothesis testing - section 4.4.3 - where we had to reject the hypothesis due to the low P-value. Demographic features like age and gender are not significant in determining a student’s retention which is also in line with the hypothesis testing.

6 LIMITATIONS

My goal was to build a model which would help educators identify at risk students early on in the course, thereby providing the required assistance a student needs to succeed. To build such a model, you would need to train on a time series dataset. The dataset would include multiple records for a student giving a full picture of how a student performs during the course of the 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. Unfortunately, I was not able to find such a dataset. Student data in general is not readily available due to FERPA limitations. This doesn't mean that the current model is not useful. We can still use this model to get predictions during the tail end of the term.

The dataset used in this project is from MOOC courses. The findings cannot be generalized for online courses offered by traditional schools. Students taking online courses in traditional schools require significantly more investment from a student's part. The online courses from traditional schools could be more expensive, require more time commitment and provide a degree which can provide different motivations factors for a student.

7 CONCLUSION

The focus of this study was to understand key factors that impact online student retention and, build a ML predictive model that can help educators identify at risk students. Hypothesis testing identified student engagement as the key factors that affect student success. In an online class, accessing course materials, watching videos, participating in class discussions by posting in forums are all considered to be student engagement. Various ML models also confirmed that student activities are the top factors that separate certified vs non certified students. ML models showed a high predictive power in predicting student retention. Due to the limitations of the dataset, the models cannot provide predictions early on in the term. Nevertheless, the current model can still be useful to educators in identifying at risk students towards the tail end of the term. Educators can nudge at risk students and motivate them, thereby improving student success.

8 FUTURE WORK

Having access to more data, in this case a time series dataset, would help build a better model. Using this dataset, we can build a model that can identify at risk students early on in the term, which would be beneficial to educators in improving student success. With FERPA, it is unlikely that this kind of data will be made publicly available. An institution like Georgia Tech can collaborate with their internal staff engineers to build this model. Including additional features in the dataset like midterm grades will also help improve the model to generate predictions early on in the term.

9 ACKNOWLEDGEMENTS

I would like to thank my mentor Christine McManus for her consistent feedback throughout the project. I would like to thank David Joyner for creating this course and thereby giving me an opportunity to do this project. Finally, I would like to thank all my peers who gave me valuable feedback throughout the course.

10 REFERENCES

1. Bean, J., & Metzner, B. (1985). A Conceptual Model of Nontraditional Undergraduate Student Attrition. Review of Educational Research, 55(4), 485-540.
2. Buenaño-Fernández, D., Gil, D., & Luján-Mora, S. (2019). Application of machine learning in predicting performance for computer engineering students: A case study. Sustainability (Switzerland), 11(10), Sustainability (Switzerland), 1 May 2019, Vol.11(10).
3. Conijn, R., Van den Beemt, A., & Cuijpers, P. (2018). Predicting student performance in a blended MOOC. Journal of Computer Assisted Learning, 34(5), 615-628.
4. Hlosta, M., Herrmannova, D., Vachova, L., Kuzilek, J., Zdrahal, Z., & Wolff, A. (2018). Modelling student online behaviour in a virtual learning environment.
5. Karber, D. J. (2003). Comparisons and contrasts in traditional versus on-line teaching in management. Higher Education in Europe, 26: 533-536.

6. Le, K., Wanger, Stephen P., Mendez, Jesse P, Moore, Tami L, & Romans, John S. C. (2016). Factors Affecting Student Persistence at Public Research Universities in Oklahoma.
7. Rodríguez-Muñiz, L. J., Bernardo, A. B., Esteban, M., & Díaz, I. (2019). Dropout and transfer paths: What are the risky profiles when analyzing university persistence with machine learning techniques? PLoS ONE, 14(6), E0218796.
8. Tinto, V. (1975). Dropout from Higher Education: A Theoretical Synthesis of Recent Research. Review of Educational Research, 45(1), 89-125.