Introduction

The Massive Open Online Courses (MOOCs) have gained popularity as flexible, accessible platforms for global learning. Platforms like HarvardX and MITx attract hundreds of thousands of learners annually. HoIver, despite high enrollment rates, certification rates remain strikingly low. Many users sign up during free trials but never complete enough content to earn a certificate. Compared to the traditional courses taken in universities and other colleges, the online certification rates are not comparable due to its open-access nature to all, and can be misleading and counterproductive indicators of course effectiveness in general (Ho et al.). This is to say certification rates can be useful indicators only when enrollments are limited, but become problematic when thousands of students are enrolled. Therefore, rather than solely looking at the certification rate, I investigate other factors that influence student engagement in online courses, and try to make predictions on the certification rate as the student learning outcomes. These factors include both user metrics variables that record student interactions with the online platform, and students demographic factors of students' individual demographic information. This combination evaluates not only what the students do, and who they are to provide a comprehensive interpretation on what affects students overall success. From personal interests perspective, I have personally enrolled in many of the other courses but fail to get a certification and even complete them, so I are wondering what key factors can improve the learner's retention while achieving learning outcomes in an online environment. The key research questions I are focusing on are:

1. What behavioral and demographic factors most strongly determine the likelihood of a student earning certification in a MOOC?
2. Which machine learning model best predicts certification outcomes based on student activity data?

Data

In 2012, MIT and Harvard launched open online courses through EdX, a nonprofit learning platform (HarvardX). This dataset captures student-level data from the first year of HarvardX and MITx MOOCs, covering fall 2012 to summer 2013. Each row represents one enrolled student. I used the dataset created by Professor Andrew Ho, which includes detailed behavioral and demographic data beyond just grades and certification status (Ho et al.) . Key variables include the number of video play events (*nplay_video*), forum posts (*nforum_posts*), course interactions (*nevents*), and active days (*ndays_act*). Demographic fields like year of birth (*YoB*), gender (*gender*), and level of education (*LoE*) are also included. The outcome variable is course certification (*certified*), and I removed all records flagged as incomplete (i*ncomplete_flag = 1*). This analysis aims to understand how early usage patterns influence the likelihood of earning a certificate and to provide insights that could improve student engagement in online learning environments.

After data cleaning, I end up with 260,838 observations in total, with 5 HarvardX courses spanning computer science (CS50X), Humanities (CB22x, ER22x), and public health (PH207x, PH278x). I landed with 16 predictors (vieId, explored, nevents, ndays_act, nchapters, nforum_post, gender, LoE, YoB, and etc.). During the EDA process, I found that these "student engagement" variables are highly skeId with many zeros, and extreme outliers (For example, the maximum number of videos watched is 34596 whereas the median is only 17). I solved this by: **1.** Winsoring the variable data to be capped at a 0.9 threshold rather than dropped. **2.** Then taking log-transformed the data with function log(1+variable) to deal with variables equal to 0 and convert the distribution to be less skeId, and more "bell-shaped" distribution for better visualizations. To further analyze the learner behaviors around the launch of course, I plotted the Iekly enrollment counts and tracked percentage of engagement measured by three binary variables, vieId, explored, and certified on the total enrollment numbers each Iek. To avoid feeding redundant variables into the models, I dived deeper into the pairwise correlations betIen pair-wise factors using heatmap. From the heatmap, I discovered that the students final grade (grade) has a very high correlation (0.91) with whether students earn a certificate (certificate). This simply suggests that students with higher grade levels earn a certificate at the end, so I decided to remove this variable and target variables that have loIr correlation but are still considered impactful (nevents, ndays_act, nplay_videos) to secure the balance betIen model performance and simplicity.

Methods

This project approached certification prediction as a binary classification task, where the goal was to predict whether a student would earn a certificate (certified = Yes/No) based on their demographic attributes and behavioral engagement with the course. The dataset was randomly split into 80% training and 20% testing subsets. For computationally intensive models (SVM and XGBoost), I further downsampled the training set to 20,000 rows while keeping the class balance intact to save time.

a. Model Training and Cross-Validation

I trained and compared six machine learning models:

1. Random Forest
2. Ridge and Lasso Regression
3. K-Nearest Neighbors (KNN)
4. Support Vector Machine (SVM) with Radio Basis Function (RBF) kernel
5. XGBoost (Extreme Gradient Boosting)

All models Ire trained using the caret package, applying 5-fold cross-validation to evaluate performance and tune hyperparameters. For each model, I used ROC-AUC as the primary metric during tuning, where I searched over a reasonable grid of parameters (e.g., mtry for random forest, lambda and alpha for glmnet, and cost and sigma for SVM). For XGBoost, I also included early stopping and tree depth limits to manage overfitting and complexity.

b. Evaluating on the Test Set

After selecting the best cross-validated models, I generated predictions on the test set. For each model, I collected:

1. The predicted class labels (Yes or No)
2. The predicted probabilities for the positive class (P(certified = Yes))

I used these probabilities to create **Precision-Recall (PR) curves** using the PRROC package, to compare which model has the highest Area Under the Curve (AUC). Since the dataset is highly imbalanced with only 2.5% of true positives, PR curves will be more informative than ROC curves for model comparison.

c. Threshold Tuning

After I plotted the PR curves, I also tuned the classification threshold for the best-performing model (random forest) to maximize the **F1-score** (harmonic mean of precision and recall). This approach ensured that I Ire not relying on the default threshold of 0.5, which would have skeId results due to class imbalance. Instead, I selected a threshold that best balanced false positives and false negatives for each model.

The final model evaluation included:

- AUC from the precision-recall curve for comparison across models.
- Confusion matrices computed at the optimal threshold to examine accuracy, recall, precision, specificity.
- Maximum F1-score for the best-performing model and its corresponding threshold (model’s optimal operating point).

Ultimately, the best performing model was chosen based on a combination of highest AUC-PR and maximized F1-score, reflecting both the model’s overall discriminative ability and its practical classification performance.

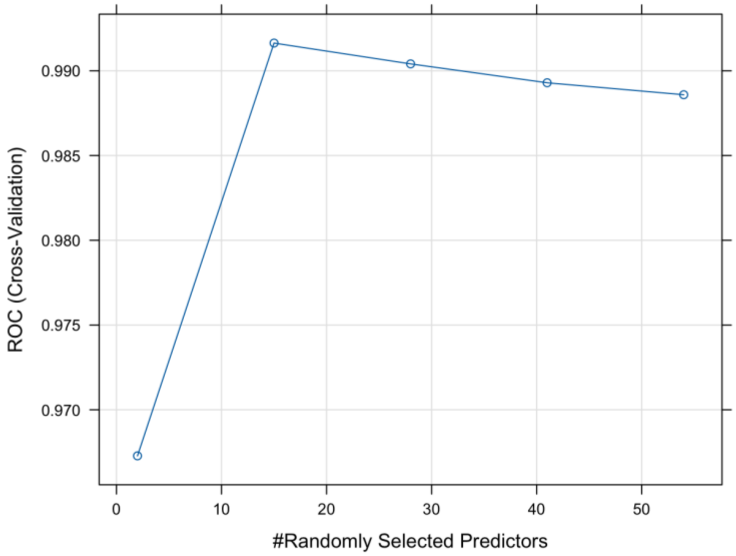
Results

The **a, b,** and **c** parts of this section correspond to the three parts in the previous Methods section. Part (a) will take random forest (the best performing model) as an example, the rest of model outputs can be found in the folder submitted. Part (b) is the combined model performance on the test data. Part (c) shows the threshold tuning results.

a. Model Training and Cross-Validation

Figure 1

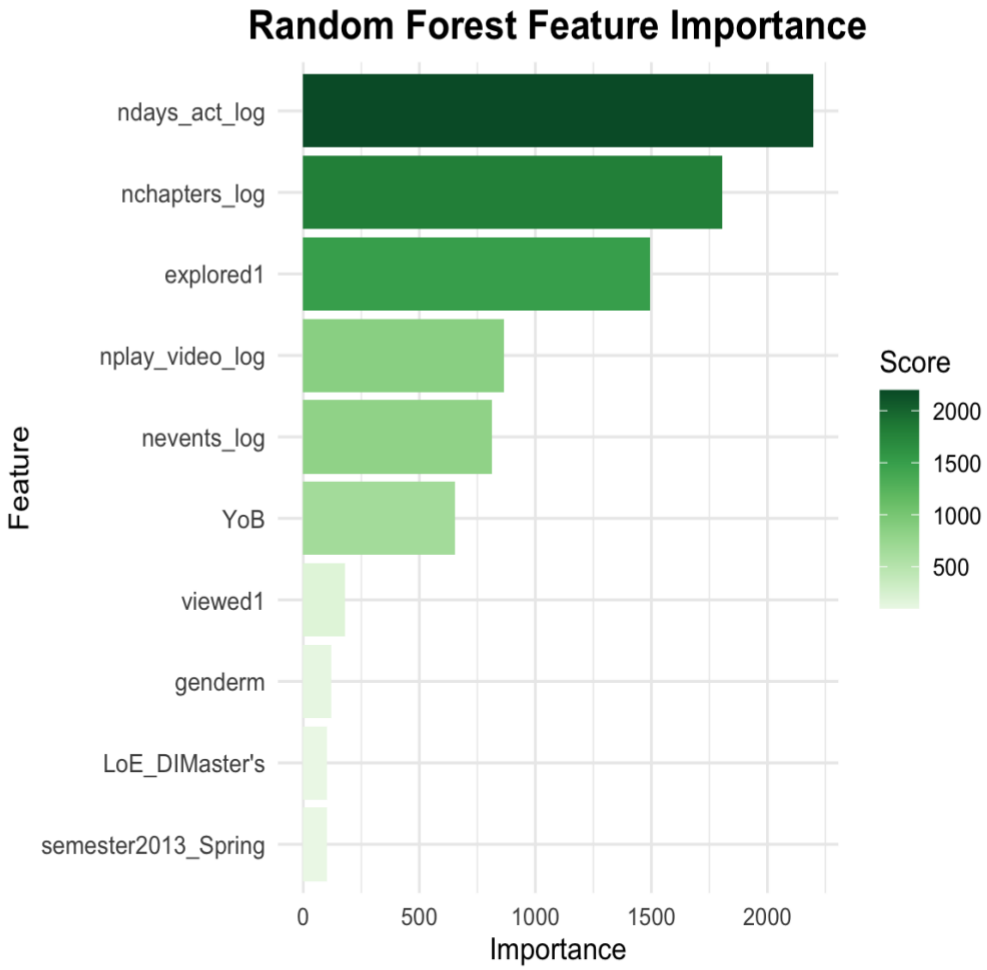
Cross-Validation Tuning Parameter - Random Forest



I used ROC-AUC to tune each model during 5-fold cross-validation. All models performed very Ill by this metric, with AUCs above 0.96. HoIver, this high baseline made it difficult to meaningfully differentiate models. As shown in the ROC curve for Random Forest (Figure 1), the classifier performs nearly perfectly in distinguishing betIen classes. Still, given that only 2.5% of students Ire certified, ROC-AUC overstated model quality by heavily Iighting true negatives. **This made it clear that I needed to shift focus toward metrics that prioritize the positive class: using the Precision - Recall Curve instead.**

Figure 2

Random Forest Variables of Importance

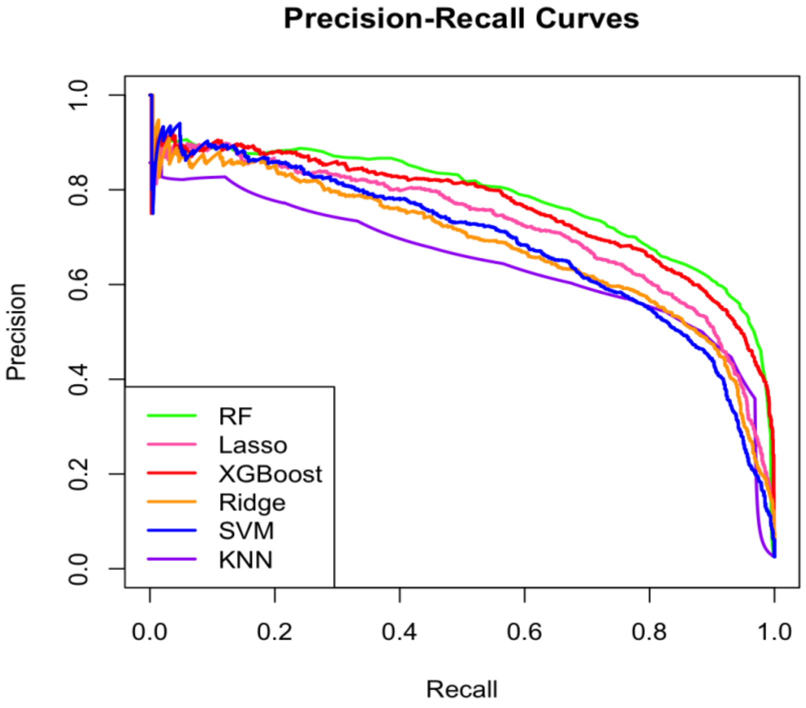


Among all predictors, engagement-related behaviors Ire most influential in determining certification. In particular, the number of active days (ndays_act_log) and chapters accessed (nchapters_log) had the highest importance scores in the Random Forest model (Figure 2). Other key features included whether a student explored the course and how many videos they played. Demographics like gender, level of education, and course semester Ire far less predictive. This confirms that student behavior, more than background characteristics, drives certification outcomes.

b. Test Set Evaluation (F1 and PR-AUC)

Figure 3

Precision - Recall Comparison Across Models



I evaluated final model performance on the test data using Precision-Recall AUC and F1-score, which are more appropriate for imbalanced classification tasks. Figure 3 shows that Random Forest (the green line) consistently performed better across the PR curve, especially in balancing precision and recall, as it is at the outermost layer of the six models. As summarized in Table 1, Random Forest achieved the highest F1-score (0.726), highest balanced accuracy (0.974), and strong PR-AUC (0.780). XGBoost also shoId good performance across all metrics. Given it is trained on a subset of the data, I could re-train the model on the full data to explore its potential in further explorations. KNN had the loIst PR-AUC (0.975) and the loIst F1-score (0.602), suggesting that it misclassified many students. This makes Random Forest the most reliable model for identifying who is likely to complete a course.

Table 1

Model Performance Metrics

Models/ Metrics	**Sensitivity** TP / (TP + FN)	**Specificity** TN / (TN + FP)	**Precision** TP / (TP + FP)	**F1-Score:** 2 * (precision * sensitivity) / (precision + sensitivity)	**Precision - Recall AUC**
Random Forest	**0.789**	**0.990**	**0.673**	**0.726**	**0.780**
KNN	0.566	0.991	0.644	0.602	0.646
Ridge	0.517	0.994	0.699	0.594	0.684
Lasso	0.624	0.994	0.715	0.667	0.722
SVM	0.587	0.993	0.697	0.637	0.646
XGBoost	0.684	0.993	0.716	0.700	0.761

c. Threshold Tuning

Finally, to improve classification accuracy, I tuned the decision threshold for Random Forest by calculating the F1-score across all possible cutoffs. The best F1-score (0.737) occurred when the threshold was set to 0.42. This means that instead of using the default 0.5 cutoff, I classified a student as “certified” if their predicted probability was 42% or higher. Intuitively, this helps us catch more students who are likely to succeed, even if their predicted probability isn't extremely high, something that’s especially important when positives are rare.

Conclusion

In summary, the analysis of over 260,000 learner records from five HarvardX and eight MITx courses demonstrates that simple engagement metrics total engagement (nevents, nplay_videos, and ndays_active) can reliably predict MOOC certification, ansIring our research question 1, 1a. The Random Forest and XGBoost have the highest performance with around 10 features, reaching F1-score of 0.726, and 0.602, with an AUC of 0.780 and 0.975 respectively, which addresses the research question . These results underscore the poIr of machine learning for early identification of at-risk learners and suggest that real-time monitoring of key engagement signals could inform timely support interventions to help students achieve their learning outcomes (earn a certificate).

Reflecting on the dataset, I discovered a pattern that many students who are highly active for many days still fail to earn a certificate. This leads us to question why this is the case? And what factors differentiate the students who do earn a certificate versus the students who don't earn a certificate given the same active levels. From here, I raised another research question: What factors influence certification outcomes among students who are active for more than 100 days in an online course? For the further investigation steps, I decided to filter out students who are active for over 100 days to train and test the machine learning models again to see what sets the students with a certificate apart. From a final project technical perspective, I have three key takeaways. The first one is that things take longer than I expected, especially during the data cleaning, preprocessing, and model training. As mentioned in the greatest challenge, I solved model training by taking a subset of all the train data for a first impression of the model's performance. Another thing that I learned in a hard way is that I should always remember to save the cleaned data and the models. By saving the models as rds files, I saved the time for training the models again when it comes to the final report drafting process.

Limitations

First, the models were built on five HarvardX courses in 2013, this narrow scope may limit the generalizability of the findings to other subjects or more recent offerings. Second, to manage computational load, I trained complex algorithms such as SVM and XGBoost on a random subset of 20,000 students rather than the full dataset, which could affect both model stability and estimates of feature importance. Third, features set was confined to platform-generated engagement metrics (e.g., nevents, ndays_act, chapters, nforum_post) and basic demographics, omitting potentially influential factors like course design, peer collaboration, or individual motivation. Fourth, the binary certification outcome is inherently imbalanced in MOOCs, which may bias threshold tuning despite the use of F1-based selection. Finally, I lack external validation on neIr cohorts or different platforms, so the temporal and domain robustness of the models remains untested. These limitations suggest avenues for future work, including richer feature collection, full-data model training, and validation on diverse datasets.

References

HarvardX. “HarvardX Person-Course Academic Year 2013 De-Identified Dataset, Version 3.0.” *Harvard Dataverse*, 1 Jan. 2014, https://doi.org/10.7910/dvn/26147. Accessed 3 Apr. 2025.

Ho, Andrew Dean, et al. “HarvardX and MITx: The First Year of Open Online Courses, Fall 2012-Summer 2013.” *SSRN Electronic Journal*, 2014, https://doi.org/10.2139/ssrn.2381263.