

Identifying At-Risk K-12 Students in Multimodal Online Environments: A Machine Learning Approach

Hang Li, Wenbiao Ding, Zitao Liu^{*}
TAL Education Group
16/F, Danling SOHO 6th Danling St
Haidian District Beijing, China, 100060
{lihang4, dingwenbiao, liuzitao}@100tal.com

ABSTRACT

With the rapid emergence of K-12 online learning platforms, a new era of education has been opened up. It is crucial to have a dropout warning framework to preemptively identify K-12 students who are at risk of dropping out of the online courses. Prior researchers have focused on predicting dropout in Massive Open Online Courses (MOOCs), which often deliver higher education, i.e., graduate level courses at top institutions. However, few studies have focused on developing a machine learning approach for students in K-12 online courses. In this paper, we develop a machine learning framework to conduct accurate at-risk student identification specialized in K-12 multimodal online environments. Our approach considers both online and offline factors around K-12 students and aims at solving the challenges of (1) multiple modalities, i.e., K-12 online environments involve interactions from different modalities such as video, voice, etc; (2) length variability, i.e., students with different lengths of learning history; (3) time sensitivity, i.e., the dropout likelihood is changing with time; and (4) data imbalance, i.e., only less than 20% of K-12 students will choose to drop out the class. We conduct a wide range of offline and online experiments to demonstrate the effectiveness of our approach. In our offline experiments, we show that our method improves the dropout prediction performance when compared to state-of-the-art baselines on a real-world educational dataset. In our online experiments, we test our approach on a third-party K-12 online tutoring platform for two months and the results show that more than 70% of dropout students are detected by the system.

Keywords

Dropout, retention, multimodal learning, online tutoring, K-12 education

1. INTRODUCTION

^{*}The corresponding author.

Hang Li, Wenbiao Ding, Songfan Yang and Zitao Liu "Identifying At-Risk K-12 Students in Multimodal Online Environments: A Machine Learning Approach" In: *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, Anna N. Rafferty, Jacob Whitehill, Violetta Cavalli-Sforza, and Cristobal Romero (eds.) 2020, pp. 137 - 147

With the recent development of technologies such as digital video processing and live streaming, there has been a steady increase in the amount of K-12 students studying online courses worldwide. Online classes have become necessary complements to public school education in both developing and developed countries [31, 27, 37, 42, 35, 34, 36]. Different from public schools that focusing on teaching in traditional brick-and-mortar classrooms with 20 to 50 students, online classes open up a new era of education by incorporating more personalized and interactive experience [20, 33, 45, 9, 57].

In spite of the advantages of this new learning opportunity, a large group of online K-12 students fail to finish course programs with little supervision either from their parents or teachers. Students drop out of the class may be due to many reasons such as lack of interests or confidence, mismatches between course contents and students' leaning paths or even no immediate grade improvements from their parents' perspectives [37, 39, 25]. Therefore, it is crucial to build an early dropout warning system to identify such at-risk online K-12 students and provide timely interventions.

A large spectrum of approaches have been developed and successfully applied in predicting dropout in Massive Open Online Courses (MOOCs) [29, 47, 43, 58, 3, 5]. However, identifying dropout of K-12 students on online courses are significantly different from MOOCs based attrition prediction. The main differences are summarized as follows:

- **watching v.s. interaction:** Even though both learning are conducted in the online environment, learners' engagements on MOOCs and K-12 online platforms vary a lot [20]. In MOOCs, learners mainly watch the pre-recorded video clips and discuss questions and assignments with teaching assistants on the MOOC forums [18]. While in K-12 online courses, students frequently interact with the online tutors in a multimodal and immersive learning environment. The tutors may answer students' questions, summarize the knowledge points, take notes for students, etc.
- **spontaneous action v.s. paid service:** Learners on existing popular MOOC platforms such as Coursera¹, edX², etc. are adults, who aim at continuing

¹<https://www.coursera.org/>

²<https://www.edx.org/>

their lifelong learning in higher education and obtaining professional certificates such as Coursera’s Specializations and edX’s MicroMasters. MOOC learners are typically self-motivated and self-driven. On the contrary, most available K-12 online education choices are commercialized in service industry. Students pay to enroll online tutoring programs to strengthen their in-class knowledge levels and improve their grades in final exams. As a result, there are numerous out-of-class activities involved in K-12 online learning such as follow-ups from personal instructors, satisfaction survey and communications with students’ parents, etc. These out-of-class activities rarely appear in MOOC based learning.

- **high v.s. low dropout rate:** The dropout rate for MOOC based program is often as high as 70% - 93% [31, 52] while the dropout rate in K-12 online courses is below 20%.

Therefore, it is important to study approaches to identify at-risk K-12 online students and build an effective yet practical warning system. However, this task is rather challenging due to the following real-world characteristics:

- **multiple modalities:** K-12 online learning is conducted in an immersive and multimodal environment. Students and instructors interact with each other visually and vocally. There are a lot of multimodal factors that may influence the final decisions of dropout, ranging from interaction qualities between students and teachers, teaching speeds, volumes, emotions of the online tutors, etc.
- **length variability:** Students join and leave the online platforms independently, which results in a collection of observation sequences with different lengths. A dropout prediction system should be able to (1) make predictions for students with various lengths of learning histories; and (2) handle newly enrolled students.
- **data imbalance:** The overall dropout rate for K-12 online classes is usually below 20%, which makes the training samples particularly imbalanced.

The objective of this work is to study and develop models that can be used for accurately identifying at-risk K-12 students in multimodal online environments. More specifically, we are interested in developing models and methods that can predict risk scores (dropout probabilities) given the history of past observations of students. We develop a data augmentation technique to alleviate class imbalance issues when considering the multi-step ahead prediction tasks. We conduct extensive sets of experiments to examine every component of our approach to fully evaluate the dropout prediction performance.

Overall this paper makes the following contributions:

- We design various types of features to fully capture both in-class multimodal interactions and out-of-class

activities. We create a data augmentation strategy to simulate the time-sensitive changes of dropout likelihood in real scenarios and alleviate the data imbalance problem.

- We design a set of comprehensive experiments to understand prediction accuracy and performance impact of different components and settings from both qualitative and quantitative perspectives by using a real-world educational dataset.
- We push our approach into a real production environment to demonstrate the effectiveness of our proposed dropout early warning system.

The remainder of the paper is organized as follows: Section 2 discusses the related research work of dropout prediction in both public school settings and MOOCs scenarios. Comparisons with relevant researches are discussed. In Section 3, we introduce assumptions when building a practical at-risk student identification system and formulate the prediction task. Section 4, we describe the details about our prediction framework, which include (1) extracting various types of features from both online classroom recordings and offline activity logs (See Section 4.1); and (2) data augmentation technique that helps us create sufficient training pairs and overcomes the class imbalance problem (See Section 4.2). In Section 5, we (1) quantitatively show that our model supports better dropout predictions than alternative approaches on an educational data derived from a third party K-12 online learning platform and (2) demonstrate the effectiveness of our proposed approach in the a real production environment. We summarize our work and outline potential future extensions in Section 6.

2. RELATED WORK

Dropout prediction and at-risk student identification have been gaining popularity in both the educational research and the AI communities. Understanding the reasons behind dropouts and building early warning systems have attracted a growing interest of academics in the learning analytics area. Broadly speaking, existing research regarding dropout prediction can be categorized by learning scenarios and divided into two categories: (1) public school dropout (See Section 2.1); and (2) MOOCs dropout (See Section 2.2).

2.1 Public School Dropout

Education institutions are faced with the challenges of low student retention rates and high number of dropouts [45, 32]. For examples, in the United States, almost one-third of public high school students fail to graduate from high school each year [40, 7] and over 41% of undergraduate students at four-year institutions failed to graduate within six years in Fall 2009 [38]. Hence, research work has focused on predicting the dropout problem and developing dropout prevention strategies [40, 41, 8, 55, 13, 30, 10, 49]. Zhang and Rangwala develop an at-risk student identification approach based on iterative logistic regression that utilizes all the information from historical data from previous cohorts [59]. The state of Wisconsin creates a predictive dropout early warning system for students in grades six through nine and provides predictions on the likelihood of graduation for over

225,000 students [30]. The system utilizes ensemble learning and is built on the steps of searching through candidate models, selecting some subsets of best models, and averaging those models into a single predictive model. Lee and Chung address the class imbalance issue using the synthetic minority over-sampling techniques on 165,715 high school students from the National Education Information System in South Korea [33]. Ameri et al. consider different groups of variables such as family background, financial, college enrollment and semester-wise credits and develop a survival analysis framework for early prediction of student dropout using Cox proportional hazards model [1].

2.2 MOOCs Dropout

With the recent boom in educational technologies and resources both in industry and academia, MOOCs have rapidly moved into a place of prominence in the mind of the public and have attracted a lot of research attentions from many communities in different domains. Among all the MOOC related research questions, dropout prediction problem emerges due to the surprisingly high attrition rate [54, 19, 26, 23, 44, 56, 6, 11, 12, 20]. Ramesh et al. treat students' engagement types as latent variables and use probabilistic soft logic to model the complex interactions of students' behavioral, linguistic and social cues [43]. Sharkey et al. conduct a series of experiments to analyze the effects of different types of features and choices of prediction models [47]. Kim et al. study the in-video dropouts and interaction peaks, which can be explained by five identified student activity patterns [25]. He et al. propose two transfer learning based logistic regression algorithms to balance the prediction accuracy and inter-week smoothness [21]. Tang et al. formulate the dropout prediction as a time series forecasting problem and use a recurrent neural network with long short-term memory cells to model the sequential information among features [50]. Both Yang et al. and Mendez et al. conduct survival analysis to investigate the social and behavioral factors that affect dropout along the way during participating in MOOCs [58, 39]. Detailed literature surveys on MOOC based dropout prediction are reviewed comprehensively in [51, 5].

In this work, we focus on identifying at-risk students in K-12 online classes, which is significantly distinguished from dropout predictions in either public school or MOOCs based scenarios. In the K-12 multimodal learning environment, the learning paradigm focuses on interactions instead of watching. The interactions come from different modalities, which rarely happen in traditional public schools and MOOC based programs of higher education. Furthermore, as a paid service, K-12 online learning involves both in-class and out-of-class activities and both of them contain multiple factors that could lead to class dropouts. These differences make existing research works inapplicable in K-12 online learning scenarios. To the best of our knowledge, this is the first research that comprehensively studies the dropout prediction problem in K-12 online learning environments from real-world perspectives.

3. PROBLEM FORMULATION

3.1 Assumptions

In order to characterize the K-12 online learning scenarios, we need to carefully consider every cases in the real-world

environment and make reasonable assumptions. Without loss of generality, we have the following assumptions in the rest of the paper.

ASSUMPTION 1 (RECENCY EFFECT). *Time spans between the date of dropout and the date of last online courses vary a lot. Students may choose to drop the class right after one course or quit after two weeks of no course. Therefore, the per-day likelihood of dropout should be time-aware and the closer to the dropout date, the more accurate the dropout prediction should be.*

ASSUMPTION 2 (MULTI-STEP AHEAD FORECAST). *The real-world dropout prediction framework should be able to flexibly support multi-step ahead predictions, i.e., the next-day and next-week probabilities of dropout.*

3.2 The Prediction Problem

In this work, our objective is to predict the value of future status for the *target student* given his or her past learning history, i.e., observations collected from K-12 online platforms. More specifically, let \mathbf{S} be the collection of all students and for each student $s, s \in \mathbf{S}$, we assume that we have observed a sequence of n_s past observation-time pairs $\{\langle \mathbf{x}_j^s, t_j^s \rangle\}_{j=1}^{n_s}$, $\mathbf{x}_j^s \in \mathbf{X}^s$, and $t_j^s \in \mathbf{T}^s$, such that $0 < t_j^s < t_{j+1}^s$, and \mathbf{x}_j^s is the observation vector made at time (t_j^s) for student s . \mathbf{X}^s and \mathbf{T}^s represent the collections of observations and timestamps for student s . Correspondingly, let \mathbf{Y}^s be the collection of indicators of status (*dropout, ongoing or completion*) of student s at each timestamp, i.e., $\mathbf{Y}^s = \{y_j^s\}_{j=1}^{n_s}$. Let Δ be the future time span in multi-step ahead prediction. Time $t_{n_s+\Delta}^s$ ($\Delta > 0$) is the time at which we would like to predict the student's future status $\hat{y}_{t_{n_s+\Delta}^s}^s$.

Please note that we omit the explicit student index s in the following sections for notational brevity and our approach can be generalized into a large samples of student data without modifications.

4. THE PREDICTION FRAMEWORK

The dropout prediction for K-12 online courses is a time-variant task. A student who just had the class should have a smaller dropout probability compared with a student haven't take any class for two weeks. Therefore, when designing a real applicable approach of dropout prediction, such recency effect, i.e., Assumption 1, has to be considered. In this work, we extract both static and time-variant features from different categories to capture the factors leading to dropout events comprehensively (See Section 4.1). Furthermore, we create a label augmentation technique that not only alleviates the class imbalance problem when building predictive framework for K-12 online classes, but incorporates the recency effect into label constructions (See Section 4.2). The learning of our dropout model is discussed in Section 4.3 and the overall learning procedure is summarized in Section 4.4.

4.1 Features

In this section, we develop a distinguished set of features for at-risk student identification from the real-world K-12 online learning scenarios, which can be divided into three categories: (1) in-class features that focus on K-12 students'

online behaviors during the class (See Section 4.1.1); (2) out-of-class features that consider as much as possible real-world factors happened after the class, which may influence the dropout decisions (See Section 4.1.2); and (3) time-variant features that include both historical performance of teachers and aggregated features of student activities within fixed-size windows (See Section 4.1.3).

4.1.1 In-class Features

Different from adults who continue their learning in higher education on MOOC based platforms, K-12 students come for grade improvements. This intrinsic difference in their learning goals leads to contrasting learning behaviors. Adult learners of MOOCs study independently by various activities, such as viewing lecture videos, posting questions on MOOC forums, etc. This results in various types of in-class click-stream data, which are shown to be effective in dropout prediction in many existing research works [15, 51, 54, 6, 12, 11, 20]. However, such click based activities barely happen in K-12 online scenarios. Instead, there are frequent voice based interactions between K-12 students and their teachers. The teachers not only make every effort to clarify unsolved questions that students remain from their public schools, but are responsible for arousing students’ learning interests and building their studying habits. Therefore, we focus on extracting in-class multimodal features specializing in K-12 tutoring scenarios from the online classroom videos. We categorize our features as follows. Table 1 illustrates some examples of in-class features from different categories.

- **Prosodic features:** speech-related features such as signal energy, loudness, Mel-frequency cepstral coefficients (MFCC), etc.
- **Linguistic features:** language-related features such as statistics of part-of-speech tags, the number of interregnum words, distribution of length of sentences, voice speed of each sentence, etc.
- **Interaction features:** features such as the number of teacher-student Q&A rounds, the numbers of times teachers remind students to take notes etc.

To extract all the features listed in Table 1, we first extract audio tracks from classroom recordings on both teacher’s and student’s sides. Then we extract acoustic features directly from classroom audio tracks by utilizing the widely used open-sourced tool, i.e., *OpenSmile*³. We obtain classroom transcriptions by passing audio files to a self-trained automatic speech recognition (ASR) module. After that, we extract both linguistic and interaction features from the conversational transcripts. Finally, we concatenate all features from above categories and apply a linear PCA to get the final dense in-class features. The entire in-class feature extraction workflow of our approach is illustrated in Figure 1.

Please note that due to the benefits of online steaming, both students’ and teachers’ videos are recorded separately and hence, there is no voice overlap in the video recordings. This

³<https://www.audeering.com/opensmile/>

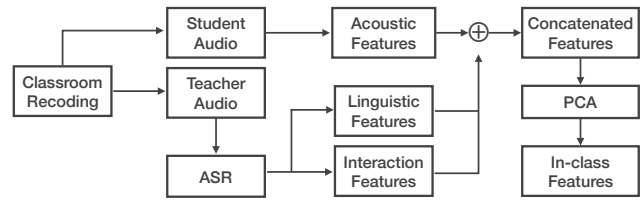


Figure 1: The workflow of our in-class features extraction. ASR is short for automatic speech recognition.

avoids the unsolved challenge of speaker diarization [2]. Similar to Blanchard et al. [4], we find that publicly available ASR service may yield inferior performance in the noisy and dynamic classroom environments. Therefore, we train our own ASR models on the classroom specific dataset based on a deep feed-forward sequential memory network, proposed by Zhang et al. [60]. Our ASR has a word error rate of 28.08% in our classroom settings.

4.1.2 Out-of-class Features

As we discussed in Section 1, personalized K-12 online tutoring is a paid service in most countries. Besides the course quality itself, there are multiple other factors in such service industry that may change customers’ minds to drop the class. Therefore, out-of-class features play an extremely important role in identifying at-risk students in real-world K-12 online scenarios, which are typically ignored in previous literatures. In this work, we collect and summarize all the available out-of-class features and divide them into the following two categories. The illustrative examples are listed in Table 1.

- **Pre-class features:** Pre-class features capture the students’ (or even their parents’) behaviors before taking the class, such as purchasing behaviors, promotion negotiations, etc. Examples: the number of rounds of conversation and negotiation before the class, how much the discount student received, etc.
- **Post-class features:** Post-class features model the offline activities in such paid K-12 online services. For examples, students and their parents receive follow-ups based on their previous class performance and give their satisfaction feedbacks. Another example is that students may request changes to their course schedules.

4.1.3 Time-variant Features

Besides in-class and out-of-class features, we manually design time-variant features to model the changes of likelihood of students’ dropout intentions. Cases like a student just had a class compared to a student had a class two weeks ago should be explicitly distinguished when constructing features. Therefore, we create time-variant features by utilizing a lookback window approach on students’ observation sequences. More specifically, for a given timestamp, we only focus on previously observed activities of each student within a period of time. The length of lookback windows varies from 1 to 30 days. Sufficient statistics are extracted as time-variant features from each lookback window. Meanwhile, we

Table 1: List of examples in in-class, out-of-class, and time-variant features.

Category	Type	Examples
In-class	Prosodic	the average signal energies of student and teacher the average loudness of student and teacher the Mel-frequency cepstral coefficients of audio tracks from student and teacher the zero-crossing rates of student and teacher ...
	Linguistic	# of sentences per class of student and teacher # of pause words per class of student and teacher average lengths of sentences per class of student and teacher voice speeds (char per second) of student and teacher ...
	Interaction	# of teacher-student Q&A rounds # of times the teacher reminds the student to take notes and summarization # of times the teacher asks the student to repeat # of times the teacher clarifies the student's questions ...
Out-of-class	Pre-class	# of days since the student places the online course order # of courses in the student's order # of conversations between the sales staff and the students (or their parents) the discount ratio of the student's order ...
	Post-class	# of follow-ups after the student took the first class # of words in the latest follow-up report # of times the student reschedules the class the follow-up ratio, i.e., # of follow-ups divided by # of taken courses ...
Time-variant	Historical performance	# of courses taught by each individual teacher in total # of courses the student had in total historical dropout rates historical average time span between classes ...
	Lookback window	# of courses taken in past one/two/three weeks # of courses the student scheduled in past one/two/three weeks # of positive/negative follow-up reports in past one/two/three weeks the average time span of classes taken in past month ...

compute historical performance features to reflect the teaching experience and performance for each individual teacher. Table 1 shows some examples of time-variant feature we use in our dropout prediction framework.

- **Lookback window features:** The lookback window features aggregate important statistics from students' observations within a fixed-length lookback window, such as the numbers of courses taken in past one, two, three weeks.
- **Historical performance features:** The historical features aggregate each teacher's past teaching performance, which represent the overall teaching quality profiles. They include total numbers of courses and students taught, historical dropout rates, etc.

4.2 Data Augmentation

According to Assumptions 1 and 2 and the problem formulation in Section 3.2, a real-world early warning system is

supposed to flexibly support multi-step ahead predictions for each student, i.e., given any future time span Δ , the system computes the probability of student's status $\hat{y}_{n_s+\Delta}^s$. The predicted probability should be able to dynamically adapt when the values of Δ get changed. The multi-step ahead assumption essentially requires the approach to make predictions at a more fine-grained granularity of $\langle \text{student, timestamp} \rangle$ pair, i.e., $\langle s, t_{n_s+\Delta}^s \rangle$, instead of student level, i.e., s . This poses a challenging question: due to the fact that only about 20% of K-12 students drop their online classes, *how do we tackle the class imbalance problem when extracting $\langle \text{student, timestamp} \rangle$ training pairs from a collection of multimodal observation sequences (either completion or dropout) in K-12 online scenarios?*

Let \mathbf{S}_1 and \mathbf{S}_2 be the set of student indices of dropout and non-dropout students, i.e., $\mathbf{S}_1 = \{i | y_{n_i}^i = \text{dropout}\}$, and $\mathbf{S}_2 = \{j | y_{n_j}^j = \text{completion}\}$. Let \mathbf{P} and \mathbf{N} be the sets of positive (dropout) and negative (non-dropout) $\langle \text{student, timestamp} \rangle$ pairs. By definition, \mathbf{P} and \mathbf{N} are constructed as follows:

$$\begin{aligned}
\mathbf{P} &= \{ \langle \mathbf{x}_{n_i}^i, t_{n_i}^i \rangle \mid i \in \mathbf{S}_1 \} \\
\mathbf{N} &= \{ \langle \mathbf{x}_k^i, t_k^i \rangle \mid i \in \mathbf{S}_1, k \in \mathbf{T}_i \setminus t_{n_i}^i \} \\
&\cup \{ \langle \mathbf{x}_k^j, t_k^j \rangle \mid j \in \mathbf{S}_2, k \in \mathbf{T}_j \}
\end{aligned} \tag{1}$$

Similar to many researches such as fraud detection [53], the sizes of \mathbf{P} and \mathbf{N} are typically very imbalanced. While in some cases the class imbalance problem may be alleviated by applying an over-sampling algorithm on the minority class sample set, the diversity of the available instances is often limited. Therefore, in this work, we propose a time-aware data augmentation technique that artificially generates pseudo positive (dropout) $\langle \text{student}, \text{timestamp} \rangle$ pairs.

More specifically, for each dropout student i in \mathbf{S}_1 , we set a lookback window with length Λ where $\Lambda \leq t_{n_i}^i - t_{n_i-1}^i$. For each timestamp \tilde{t}_l^i in the lookback window such that

$$\max(t_{n_i-1}^i, t_{n_i}^i - \Lambda) < \tilde{t}_l^i < t_{n_i}^i. \tag{2}$$

We generate its corresponding pseudo positive training pair $\langle \tilde{\mathbf{x}}_l^i, \tilde{t}_l^i \rangle$ as follows: $\tilde{\mathbf{x}}_l^i = \mathcal{F}(\mathbf{X}^s, \mathbf{T}^s)$ where $\mathcal{F}(\cdot, \cdot)$ is the generation function. The choices of $\mathcal{F}(\cdot, \cdot)$ are flexible and vary among different types of features (See Section 4.1). In this work, for in-class and out-of-class features, we aggregate all the available features till \tilde{t}_l^i and re-compute the time-variant features according to timestamp \tilde{t}_l^i . We use $\tilde{\mathbf{P}}$ to represent the collection of all positive training pairs generated from dropout students in \mathbf{S}_1 . Figure 2 illustrates how the pseudo positive training pairs are generated.

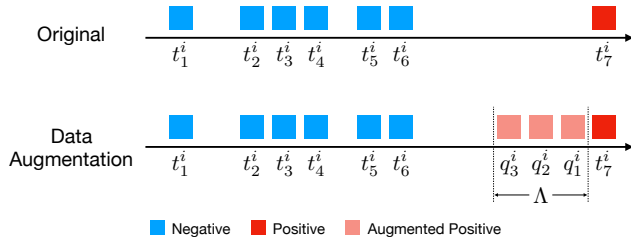


Figure 2: Graphical illustration of the data augmentation technique.

Besides, we assign a time-aware weight to each pseudo positive training pair to reflect the recency effect in Assumption 1. For each pseudo pair $\langle \tilde{\mathbf{x}}_l^i, \tilde{t}_l^i \rangle$, the corresponding weight w_l^i is computed by

$$w_l^i = \mathcal{G}\left(\frac{t_{n_i}^i - \tilde{t}_l^i}{\Lambda}\right) \tag{3}$$

where the weighting function $\mathcal{G}(\cdot)$ takes the normalized time span between each timestamp of pseudo pair and the exact dropout date as input and outputs a normalized weighting score to reflect our confidence on the “positiveness” of the simulated training pairs. The closer to the dropout date, the

larger the confidence weights should be. The choices of $\mathcal{G}(\cdot)$ are open to any function that gives response values ranging from 0 to 1, such as linear, convex or concave functions illustrated in Figure 3.

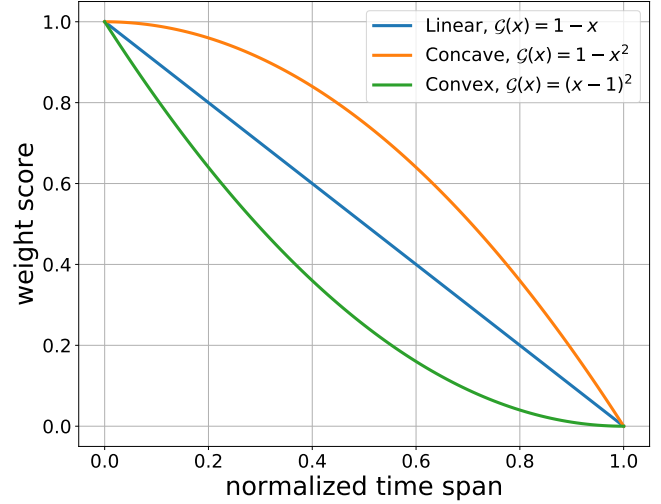


Figure 3: Graphical illustration of different weighting function of $\mathcal{G}(\cdot)$.

The effect of different choices of weighting function is discussed in Section 5.4. The augmented training set $\tilde{\mathbf{P}}$ and the corresponding time-aware weights are used in the model training in Section 4.3.

4.3 Model Learning

In the learning stage, we combine the original training set (\mathbf{P} and \mathbf{N}) with the augmented set $\tilde{\mathbf{P}}$ for model training. Even though the data augmentation alleviates the class imbalance problem, i.e., improving the positive example ratio from 0.1% to 10%, the imbalance problem still exists. Therefore, we employ the classical weighted over-sampling algorithm on positive pairs to further reduce the imbalance effect. Here, the weights of the original positive examples in \mathbf{P} are set to 1 and pseudo positive examples’ weights are computed by $\mathcal{G}(\cdot)$ in Section 4.2. Here, since the dropout datasets are usually small compared to other Internet scaled datasets, we choose to use Gradient Boosting Decision Tree⁴ (GBDT) [16] as our prediction model. The GBDT exhibits its robust predictive performance in many well studied problems [24, 48].

4.4 Summary

The overall model learning procedure of our K-12 online dropout prediction can be summarized in Algorithm 1.

5. EXPERIMENTAL EVALUATION

In this section, we will (1) introduce our dataset that is collected from a real-world K-12 online learning platform and the details of our experimental settings (Section 5.1); (2) show that our approach is able to improve the predictive performance when compared to a wide range of classic

⁴<https://scikit-learn.org/stable/modules/ensemble.html#gradient-tree-boosting>

Algorithm 1 Model learning procedure of the K-12 online dropout prediction.

INPUT:

- A set of K-12 students \mathbf{S} and their corresponding multi-modal classroom recordings and activities logs.
- The length of lookback window Λ .
- The choice of weighting function $\mathcal{G}(\cdot)$.

PROCEDURE:

- 1: // Feature extraction
- 2: Extract in-class features from multimodal recordings, see Section 4.1.1.
- 3: Extract out-of-class features from student activities logs, see Section 4.1.2.
- 4: Extract time-variant features, see Section 4.1.3.
- 5: Concatenate three types of features above.
- 6: // Label generation and augmentation
- 7: Create original positive and negative training pair sets, i.e., \mathbf{P} and \mathbf{N} , see eq.(1).
- 8: Generate the augmented pseudo positive training sets, i.e., $\tilde{\mathbf{P}}$ and the corresponding weights, see eq.(3).
- 9: // Model learning
- 10: Conduct weighted over-sampling on the union of \mathbf{P} and $\tilde{\mathbf{P}}$.
- 11: Train the GBDT model on the over-sampled positive examples and original negative examples.

OUTPUT:

- The GBDT dropout prediction model Ω .
-

baselines (Section 5.2); (3) evaluate the impacts of different sizes of lookback windows, different weighting functions in data augmentation and feature combinations (Section 5.3, Section 5.4 and Section 5.5); and (4) deploy our model into the real production system to demonstrate its effectiveness (Section 5.6).

We would also like to note that hyper parameters used in our methods are selected (in all experiments) by the internal cross validation approach while optimizing models' predictive performances. In the following experiment, we set the size of lookback window to 7 and the impact of window size is discussed in section 5.4. We choose to use the convex weighting function when conducting pseudo positive data augmentation.

5.1 Experimental Setting

5.1.1 Data

To evaluate the effectiveness of our proposed framework, we conduct several experiments on a real-world K-12 online course dataset from a third-party online education platform. We select 3922 registered middle school and high school students from August 2018 and February 2019 as our samples. All the features listed in Section 4.1 are computed and extracted from daily activity logs on the platform. In our dataset, 634 students choose to drop the class and the dropout rate is 16.16%. The average time span of the students on the platform is about 86 days, which provide us 338428 observational $\langle \text{student}, \text{time stamp} \rangle$ sample pairs in total. We randomly select 80% of students and use their corresponding $\langle \text{student}, \text{time stamp} \rangle$ sample pairs as training set and the remaining 20% of students' sample pairs for testing propose. The data augmentation technique discussed in Section 4.2 is only applied in training set.

5.1.2 Multi-step Ahead Prediction Setting

To fully examine the dropout prediction performance, we evaluate the model's predictions in terms of different multi-step ahead time spans, i.e, given a current timestamp, we predict the outcome (dropout or non-dropout) in the next X days, where X ranges from 1, 2, \dots , 14.

5.1.3 Evaluation Metric

Similar to [18, 50, 17, 15, 51, 21], we evaluate and compare the performance of the different methods by using the Area Under Curve (AUC) score, which is the area under the Receive Operating Characteristic curve (ROC) [14]. An ROC curve is a graphic plot created by plotting the true positive rate (TPR) against the false positive rate (FPR). In our dropout prediction scenario, the TPR is the fraction of the "at-risk" predicted students who truly drop out. The FPR is the ratio of the falsely predicted "dropout" students to the true ones. The AUC score is invariant to data imbalance issue and it does not require additional parameters, for models comparisons. AUC score reaches its best value at 1 and the worst at value 0.

5.1.4 Baselines

We compare our proposed approach with the following representative baseline methods: (1) Logistic Regression (LR) [28], (2) Decision Tree (DT) [46] and (3) Random Forest (RF) [22]. LR, DT and RF are all trained on the same set of features defined in Section 4.1 with our proposed method. The training set is created by using eq.(1).

5.2 The Overall Prediction Performance

The results of these models are shown in Figure 4. As we can see from the Figure 4, we have the following observations:

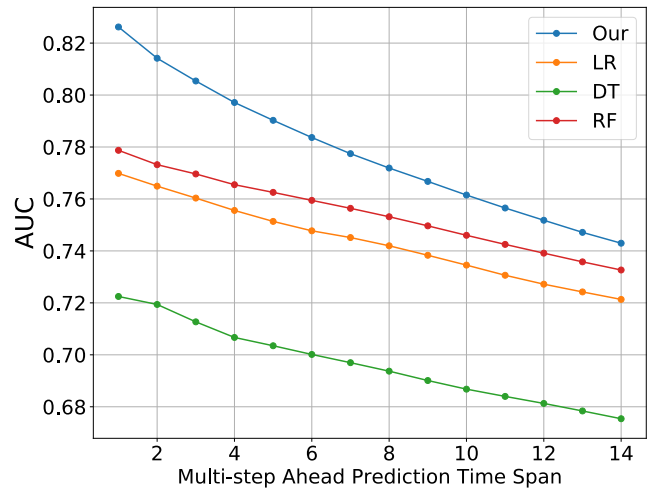


Figure 4: The overall prediction performance with different multi-step ahead time spans in terms of AUC scores.

- First, our model outperforms all other methods in terms of AUC scores on different future time spans, which demonstrates the effectiveness of our approaches with positive data augmentation. By adding more diverse pseudo positive training pairs with the corresponding decaying confidence weights, the GBDT model is able to learn the dropout patterns from multiple factors.

- Second, as we increase the lengths of time spans of multi-step ahead prediction, all the models' performances decrease accordingly. Our approach achieves AUC score of 0.8262 in the task of next day prediction while the performance downgrades to 0.7430 in the next two-week prediction task. We believe this is because of the truth negative mistakes the models make, i.e., the model thinks the students will continue but they drop classes in next two weeks. This indicates that without knowing more information from the students, the ML models have very limited ability in predicting the long-term outcomes of student status, which also reflects the fact that there are many factors that could lead to the dropouts.
- Third, comparing LR, DT, and RF, we can see, the DT achieves the worst performance. This is because of its instability. With small number of training data, the DT approach suffers from fractional data turbulence. The RF approach remedies such shortcomings by replacing a single decision tree with a random forest of decision trees and the performance is boosted. Meanwhile, as a linear model, the LR is not powerful enough to accurately capture the dropout cases.

5.3 Impact of Sizes of Lookback Windows

As we can see, the number of augmented positive training pairs is directly determined by the size of lookback window Λ . Therefore, to comprehensively understand the performance of our proposed approach, we conduct experimental comparisons on different sizes of lookback windows. We vary the window size from 3, 7, and 14. Meanwhile, we add a baseline with no data augmentation. The results are shown in Figure 5.

From Figure 5, we can see that the size of lookback windows has a positive relationship on AUC scores with the length of time span in multi-step ahead prediction. When conducting short-term dropout predictions, models trained on data augmentation with smaller size of lookback window outperform others. As we gradually increase the time span of future predictions, the more the model looks back, the higher the prediction AUC score it achieves. Overall, the model trained with 7-day lookback window has the best performance across different multi-step ahead time spans in terms of AUC scores.

5.4 Impact of Different Weighting Functions

In this section, we examine the performance changes by varying the forms of weighting functions. More specifically, we compare the prediction results of using the convex function to results of the other choices. The results are shown in Figure 6. As we can see from Figure 6, the convex option outperforms other choices by a large margin across all different multi-step ahead time spans. When computing the over-sampling weights of pseudo training examples, the convex function gives more weights to the most recent examples, i.e., examples close to the timestamp of true dropout observations. This also confirms the necessity of considering the recency effect assumption (Assumption 1) when building the dropout prediction framework.

5.5 Impact of Different Features

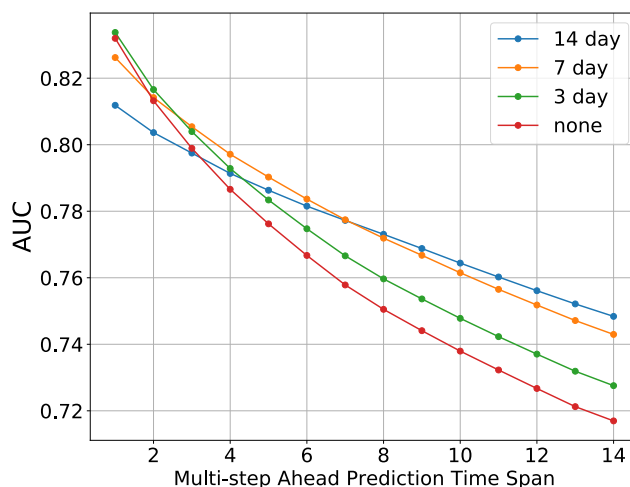


Figure 5: Models trained on data augmented by different size of lookback windows with different multi-step ahead time spans in terms of AUC scores. *none* represents the model training without any lookback data augmentation.

In this subsection, we systematically examine the effect of different types of features by constructing following model variants:

- In: only the in-class features are used.
- Out: only the out-of-class features are used.
- Time: only the time-variant features are used.
- In+Time: it eliminates the contribution of *Out* features and only uses features from *In* and *Time*.
- Out+Time: it eliminates the contribution of *In* features and only uses features from *Out* and *Time*.
- In+Out: it eliminates the contribution of *Time* features and only uses features from *In* and *Out*.
- In+Out+Time: it uses the combination of all the features from *In*, *Out* and *Time*.

Meanwhile, we also consider different multi-step ahead prediction settings, i.e., next 7-day prediction and next 14-day prediction and the prediction results are shown in Table 2. From Table 2, we observe that (1) by considering all three types of features individually, the model trained from *Out* features yields the best performance. Moreover, when comparing *In*, *Out* to *In+Time*, *Out+Time*, we obtain the significant performance improvement by adding *Out* features. These indicate the fact that dropout prediction for K-12 online scenarios are very different from MOOC based dropout prediction. The out-of-class activities and the quality of the service play an extremely important role in the prediction task; and (2) by utilizing all the sets of features, we could be able to achieve the best results in both prediction tasks.

Table 2: Experimental results of different types of features and different lengths of multi-step ahead time span in terms of AUC scores.

	In	Out	Time	In+Time	Out+Time	In+Out	In+Out+Time
Multi-step ahead time span - 7 day	0.6249	0.7764	0.6992	0.7145	0.7759	0.7768	0.7774
Multi-step ahead time span - 14 day	0.6251	0.7386	0.6766	0.6932	0.7393	0.7420	0.7430

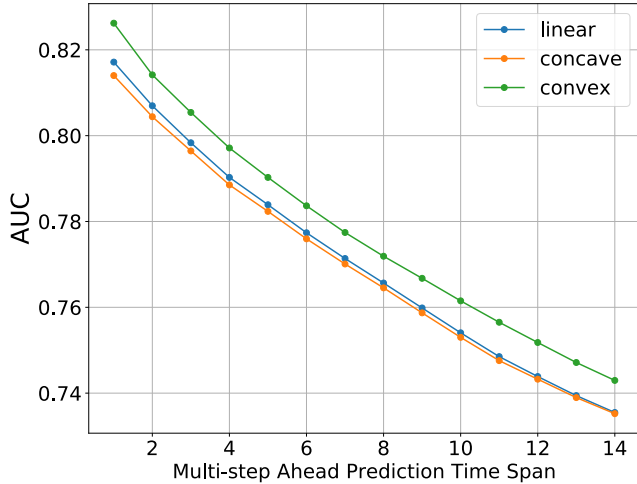


Figure 6: Models trained on data augmented by different choices of weighting functions with different multi-step ahead time spans in terms of AUC scores.

5.6 Online Performance

We deployed our at-risk student warning system in the real production environment on a third-party platform between February 2nd, 2019 to April 1st, 2019. To watch the system performance in practice, we conduct the next-day prediction task where the system predicts the dropout probability for each on-going student at 6 am in the morning. All the students are ranked by their dropout probabilities and the top 30% of students with highest probabilities are marked as at-risk students. At the end of each day, we obtain the real outcome of all the students who drop the class. We conduct the overlap comparison between the predicted top at-risk students (30% of total students) and the daily dropouts and we are able to achieve that more than 70% of dropout students are detected by the system.

6. CONCLUSION

In this paper, we present an effective at-risk student identification framework for K-12 online classes. Compared to the existing dropout prediction researches, our approach considers and focuses on the challenging factors such as multiple modalities, length variability, time sensitivity, class imbalance problems when learning from real-world K-12 educational data. Our offline experimental results show that our approach outperforms other state-of-the-art prediction approaches in terms of AUC scores. Furthermore, we deploy our model into a production environment and we are able to achieve that more than 70% of dropout students are detected by the system. In the future, we plan to explore the opportunity of using deep neural networks to capture heterogeneous information in the K-12 online scenarios to enhance

the existing prediction pipeline.

7. ACKNOWLEDGMENTS

The authors would like to thank Nan Jiang from Peking University and Lyu Qing from Tsinghua University, for their contributions to this research during their internship in TAL Education Group.

8. REFERENCES

- [1] S. Ameri, M. J. Fard, R. B. Chinnam, and C. K. Reddy. Survival analysis based framework for early prediction of student dropouts. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 903–912, 2016.
- [2] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals. Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):356–370, 2012.
- [3] G. Balakrishnan and D. Coetzee. Predicting student retention in massive open online courses using hidden markov models. *Electrical Engineering and Computer Sciences University of California at Berkeley*, 53:57–58, 2013.
- [4] N. Blanchard, M. Brady, A. M. Olney, M. Glaus, X. Sun, M. Nystrand, B. Samei, S. Kelly, and S. D’Mello. A study of automatic speech recognition in noisy classroom environments for automated dialog analysis. In *International Conference on Artificial Intelligence in Education*, pages 23–33. Springer, 2015.
- [5] I. Borrella, S. Caballero-Caballero, and E. Ponce-Cueto. Predict and intervene: Addressing the dropout problem in a mooc-based program. In *Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale*, pages 1–9, 2019.
- [6] S. Boyer and K. Veeramachaneni. Transfer learning for predictive models in massive open online courses. In *International Conference on Artificial Intelligence in Education*, pages 54–63. Springer, 2015.
- [7] J. M. Bridgeland, J. J. DiIulio Jr, and K. B. Morison. The silent epidemic: Perspectives of high school dropouts. *Civic Enterprises*, 2006.
- [8] J. S. Catterall. On the social costs of dropping out of school. *The High School Journal*, 71(1):19–30, 1987.
- [9] J. Chen, H. Li, W. Wang, W. Ding, G. Y. Huang, and Z. Liu. A multimodal alerting system for online class quality assurance. In *International Conference on Artificial Intelligence in Education*, pages 381–385. Springer, 2019.
- [10] C. Coleman, R. S. Baker, and S. Stephenson. A better cold-start for early prediction of student at-risk status in new school districts. 2020.

- [11] C. A. Coleman, D. T. Seaton, and I. Chuang. Probabilistic use cases: Discovering behavioral patterns for predicting certification. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*, pages 141–148. ACM, 2015.
- [12] S. Crossley, L. Paquette, M. Dascalu, D. S. McNamara, and R. S. Baker. Combining click-stream data with nlp tools to better understand mooc completion. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, pages 6–14. ACM, 2016.
- [13] V. Dupéré, E. Dion, T. Leventhal, I. Archambault, R. Crosnoe, and M. Janosz. High school dropout in proximal context: The triggering role of stressful life events. *Child development*, 89(2):e107–e122, 2018.
- [14] T. Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.
- [15] M. Fei and D.-Y. Yeung. Temporal models for predicting student dropout in massive open online courses. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 256–263. IEEE, 2015.
- [16] J. H. Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, 2002.
- [17] J. Gardner, Y. Yang, R. S. Baker, and C. Brooks. Modeling and experimental design for mooc dropout prediction: A replication perspective.
- [18] N. Gitinabard, F. Khoshnevisan, C. F. Lynch, and E. Y. Wang. Your actions or your associates? predicting certification and dropout in moocs with behavioral and social features. *arXiv preprint arXiv:1809.00052*, 2018.
- [19] J. A. Greene, C. A. Oswald, and J. Pomerantz. Predictors of retention and achievement in a massive open online course. *American Educational Research Journal*, 52(5):925–955, 2015.
- [20] S. Halawa, D. Greene, and J. Mitchell. Dropout prediction in moocs using learner activity features. *Proceedings of the Second European MOOC Stakeholder Summit*, 37(1):58–65, 2014.
- [21] J. He, J. Bailey, B. I. Rubinstein, and R. Zhang. Identifying at-risk students in massive open online courses. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.
- [22] T. K. Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282. IEEE, 1995.
- [23] S. Jiang, A. Williams, K. Schenke, M. Warschauer, and D. O’ Dowd. Predicting mooc performance with week 1 behavior. In *Educational Data Mining 2014*, 2014.
- [24] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, pages 3146–3154, 2017.
- [25] J. Kim, P. J. Guo, D. T. Seaton, P. Mitros, K. Z. Gajos, and R. C. Miller. Understanding in-video dropouts and interaction peaks in online lecture videos. In *Proceedings of the First ACM Conference on Learning@ Scale Conference*, pages 31–40. ACM, 2014.
- [26] R. F. Kizilcec and S. Halawa. Attrition and achievement gaps in online learning. In *Proceedings of the second (2015) ACM Conference on Learning@ Scale*, pages 57–66. ACM, 2015.
- [27] R. F. Kizilcec, C. Piech, and E. Schneider. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, pages 170–179. ACM, 2013.
- [28] D. G. Kleinbaum, K. Dietz, M. Gail, M. Klein, and M. Klein. *Logistic regression*. Springer, 2002.
- [29] M. Kloft, F. Stiehler, Z. Zheng, and N. Pinkwart. Predicting mooc dropout over weeks using machine learning methods. In *Proceedings of the EMNLP 2014 workshop on analysis of large scale social interaction in MOOCs*, pages 60–65, 2014.
- [30] J. E. Knowles. Of needles and haystacks: Building an accurate statewide dropout early warning system in wisconsin. *JEDM| Journal of Educational Data Mining*, 7(3):18–67, 2015.
- [31] D. Koller. Moocs on the move: How coursera is disrupting the traditional classroom. *Knowledge@ Wharton Podcast*, 2012.
- [32] J. E. Lansford, K. A. Dodge, G. S. Pettit, and J. E. Bates. A public health perspective on school dropout and adult outcomes: A prospective study of risk and protective factors from age 5 to 27 years. *Journal of Adolescent Health*, 58(6):652–658, 2016.
- [33] S. Lee and J. Y. Chung. The machine learning-based dropout early warning system for improving the performance of dropout prediction. *Applied Sciences*, 9(15):3093, 2019.
- [34] H. Li, Y. Kang, W. Ding, S. Yang, S. Yang, G. Y. Huang, and Z. Liu. Multimodal learning for classroom activity detection. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9234–9238. IEEE, 2020.
- [35] H. Li, Z. Wang, J. Tang, W. Ding, and Z. Liu. Siamese neural networks for class activity detection. In *International Conference on Artificial Intelligence in Education*. Springer, 2020.
- [36] Z. Liu, G. Xu, T. Liu, W. Fu, Y. Qi, W. Ding, Y. Song, C. Guo, C. Kong, S. Yang, and G. Y. Huang. Dolphin: A spoken language proficiency assessment system for elementary education. In *Proceedings of the Web Conference 2020*, page 2641–2647. ACM, 2020.
- [37] I. Lykourantzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and V. Loumos. Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers & Education*, 53(3):950–965, 2009.
- [38] J. McFarland, B. Hussar, C. De Brey, T. Snyder, X. Wang, S. Wilkinson-Flicker, S. Gebrekristos, J. Zhang, A. Rathbun, A. Barmer, et al. The condition of education 2017. nces 2017-144. *National Center for Education Statistics*, 2017.
- [39] G. Mendez, T. D. Buskirk, S. Lohr, and S. Haag. Factors associated with persistence in science and engineering majors: An exploratory study using classification trees and random forests. *Journal of*

- Engineering Education*, 97(1):57–70, 2008.
- [40] M. Monrad. High school dropout: A quick stats fact sheet. *National High School Center*, 2007.
- [41] A. M. Pallas. *School dropouts in the United States*. Center for Education Statistics, Office of Educational Research and . . . , 1986.
- [42] F. D. Pereira, E. Oliveira, A. Cristea, D. Fernandes, L. Silva, G. Aguiar, A. Alamri, and M. Alshehri. Early dropout prediction for programming courses supported by online judges. In *International Conference on Artificial Intelligence in Education*, pages 67–72. Springer, 2019.
- [43] A. Ramesh, D. Goldwasser, B. Huang, H. Daume III, and L. Getoor. Learning latent engagement patterns of students in online courses. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.
- [44] C. P. Rosé, R. Carlson, D. Yang, M. Wen, L. Resnick, P. Goldman, and J. Sherer. Social factors that contribute to attrition in moocs. In *Proceedings of the First ACM Conference on Learning@ Scale Conference*, pages 197–198. ACM, 2014.
- [45] R. W. Rumberger. High school dropouts: A review of issues and evidence. *Review of educational research*, 57(2):101–121, 1987.
- [46] S. R. Safavian and D. Landgrebe. A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics*, 21(3):660–674, 1991.
- [47] M. Sharkey and R. Sanders. A process for predicting mooc attrition. In *Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs*, pages 50–54, 2014.
- [48] J. Son, I. Jung, K. Park, and B. Han. Tracking-by-segmentation with online gradient boosting decision tree. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3056–3064, 2015.
- [49] R. Sullivan et al. Early warning signs. a solution-finding report. *Center on Innovations in Learning, Temple University*, 2017.
- [50] C. Tang, Y. Ouyang, W. Rong, J. Zhang, and Z. Xiong. Time series model for predicting dropout in massive open online courses. In *International Conference on Artificial Intelligence in Education*, pages 353–357. Springer, 2018.
- [51] C. Taylor, K. Veeramachaneni, and U.-M. O’Reilly. Likely to stop? predicting stopout in massive open online courses. *arXiv preprint arXiv:1408.3382*, 2014.
- [52] Y. Wang. Exploring possible reasons behind low student retention rates of massive online open courses: A comparative case study from a social cognitive perspective. In *AIED 2013 Workshops Proceedings Volume*, page 58. Citeseer, 2013.
- [53] W. Wei, J. Li, L. Cao, Y. Ou, and J. Chen. Effective detection of sophisticated online banking fraud on extremely imbalanced data. *World Wide Web*, 16(4):449–475, 2013.
- [54] J. Whitehill, K. Mohan, D. Seaton, Y. Rosen, and D. Tingley. Mooc dropout prediction: How to measure accuracy? In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*, pages 161–164. ACM, 2017.
- [55] L. Wood, S. Kiperman, R. C. Esch, A. J. Leroux, and S. D. Truscott. Predicting dropout using student-and school-level factors: An ecological perspective. *School Psychology Quarterly*, 32(1):35, 2017.
- [56] W. Xing, X. Chen, J. Stein, and M. Marcinkowski. Temporal predication of dropouts in moocs: Reaching the low hanging fruit through stacking generalization. *Computers in Human Behavior*, 58:119–129, 2016.
- [57] S. Xu, W. Ding, and Z. Liu. Automatic dialogic instruction detection for k-12 online one-on-one classes. In *International Conference on Artificial Intelligence in Education*. Springer, 2020.
- [58] D. Yang, T. Sinha, D. Adamson, and C. P. Rosé. Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proceedings of the 2013 NIPS Data-driven education workshop*, volume 11, page 14, 2013.
- [59] L. Zhang and H. Rangwala. Early identification of at-risk students using iterative logistic regression. In *International Conference on Artificial Intelligence in Education*, pages 613–626. Springer, 2018.
- [60] S. Zhang, M. Lei, Z. Yan, and L. Dai. Deep-fsmn for large vocabulary continuous speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5869–5873. IEEE, 2018.