# How MOOC Videos Affect Dropout? A Lightweight Pipeline Making Student Dropout Interpretable From Several Levels

Deming Sheng, Jingling Yuan*, Xin Zhang
School of Computer Science and Technology,
Wuhan University of Technology,
Wuhan 430070, China
Email: shengdeming@whut.edu.cn, yjl@whut.edu.cn, xinz@whut.edu.cn

***Abstract*—Massive Online Open Courses (MOOC) have popularized educational opportunities for students all over the world, while immensely high dropout is becoming a central challenge nowadays. Most researches predict course dropout labels through analyzing the student engagement data. However, these models have high structural complexity with high time cost and cannot provide in-depth insights into why a student is likely to drop out. We devise a lightweight pipeline to simplify the MOOC dropout problem, grasp the core features to make student behaviours interpretable at the model and instance level, visualize the changing trend of predicted label probability estimation with feature values for longitudinally interpreting the sample student behaviours. Based on qualitative insights and quantitative analysis, our main findings are that shorter videos and instructors speak fast are more engaging. Most students complete MOOC learning with a rapid speed, while a few students who watch the video slowly have a higher completion rate. When the frequency of fast-forwarding increases while the percentage of videos watching decreases, the likelihood to drop this course raises. In the end, our pipeline achieves 69.52% AUC, 0.744 $R$-squared and 0.553 $\bar{R}$-squared with 0.982s inference time on the 20238 sample student data.**

***Index Terms*—MOOC, Dropout, Student Behaviours, Lightweight Pipeline, Interpretable Model**

## I. Introduction

Massive Online Open Courses (MOOCs), such as Coursera, edX, Udacity and XuetangX, have developed rapidly in recent years and attracted wide attention of both educators and the public all over the world. As of the end of 2019, more than 900 universities around the world have provided about 11,400 MOOC courses, and in 2018 alone, there were 2,000 new courses. However, the rapid growth of the number of courses has resulted in fewer and fewer students for each course. What's more, research shows that the average course completion rate on edX is only 5% [1], [2], and 4.5% on XuetangX similarly [3]. How to effectively improve the MOOC completion rate of students has become a prominent problem.

DOI reference number: 10.18293/SEKE2021-023

Most of the students engaging MOOCs do not complete the courses and drop them out halfway, which hinders the further development of MOOCs. With the student past engagement time, many works attempt to predict whether a student will drop out or not. However, these models have high structure complexity, strict training conditions and high time cost, which inappropriate for the dynamic of MOOC courses and the diversity of attached student watching video behaviours. More importantly, they do not provide in-depth insights into why a student is likely to drop out, which can not offer analysis for instructors to refine the future versions of MOOC.

Though MOOC dropout models are progressively proposed in recent years, still there are many issues to be addressed when it comes to the real scenarios. For instance, the engagement time can not reflect whether a student is actively paying attention to this video or just playing it in the background while multitasking. A comprehensive collection and analysis of student information are costly, which also increases the burden on instructors. Besides, a wide range of features does improve the accuracy of model predictions while impairing interpretability.

We attempt to filter the original video data and extract core features, utilize a simple pipeline to explain the behaviours of student groups withdrawing from courses, and dynamically analyze the actions of individual students. Through this pipeline, instructors can intuitively predict whether a student will drop out this course from a small amount of data, and take countermeasures to help students complete the course. At the same time, instructors can also make targeted changes to the curriculum for the next semester based on historical data. In summary, our main contributions in this paper are three folds:

- We simplify the MOOC dropout problem and describe some preliminary to make the pipeline interpretable. After that, we explore data analysis, feature engineering, model selection, model metrics

and visualization as design parameters in the context of our lightweight pipeline. In the end, our pipeline achieves 69.52% AUC, 0.744 $R$-squared and 0.553 $\bar{R}$-squared with 0.982s inference time on the 20238 sample student data.
- We adopt two interpretable models to provide in-depth insights into students dropout at several levels. Besides, we visualize the probability estimation and change trend of predicted label linked to feature values to longitudinally interpret the sample student behaviours.
- We devise a deterministic finite automaton to construct the complete student behaviours, which overcomes the inherent limitation of the student engagement time. Moreover, we find that shorter videos and instructors speak fast is more engaging, which enables instructors to adjust MOOC videos for future education.

## II. Related Work

Many recent works mainly employ the clickstream as the student engagement and attempt to predict dropout thanks to the detailed records of the students' interactions with course content, including video lectures, discussion forums, assignments, and additional course content, within the MOOC platform. Kloft et al. [4] propose a Support Vector Machine (SVM) framework focusing on clickstream data, which takes the weekly history of student data into account. Liu et al. [5] employ K-means to make a quantitative analysis of the low completion in course study on MOOC platform. Al-Shabandar et al. [6] apply Decision Tree (DT) to compare in terms of their suitability for predicting the course outcome of learners participating in MOOCs.

Three are also works that considered Neural Networks (NNs) to predict the MOOC dropouts. They usually convert clickstream data into a fixed-length representation for the downstream classification models. Fei and Yeung [7] approach a recurrent neural network (RNN) model to solve this time series prediction problem. Josh Gardner and Christopher Brooks [8] present a procedure to statistically test hypotheses about their deep neural network (DNN) model performance. Wang et al. [9] propose a hybrid deep neural network dropout prediction model by combining the CNN and RNN.

However, the majority of student interaction data in MOOCs is in the form of hardly interpretable clickstreams, and Neural Networks are too sophisticated to uncover the underlying factors of student withdrawal. In this paper, We only utilize a small number of students watching MOOC video data, and propose a lightweight pipeline to quickly judge the dropout rate of students, then visualize the core features that affect the results to assist instructors in making adjustments.

TABLE I
Notations.

| Symbol | Interpretation |
|---|---|
| $\mathcal{S}, \mathcal{C}, \mathcal{V}, \mathcal{B}, \mathcal{L}$ | Set of students, courses, videos, behaviours, labels |
| $\mathcal{F}^s_{wc}, \mathcal{F}^s_{wp}$ | Feature set of the number and percentage of students watch videos |
| $\mathcal{F}^s_{wt}, \mathcal{F}^s_{pt}, \mathcal{F}^s_{st}$ | Feature set of the time to students watch the video, progress the video, stay on the webpage |
| $\mathcal{F}^s_f, \mathcal{F}^s_b, \mathcal{F}^s_p, \mathcal{F}^s_l$ | Feature set of the frequency of students fast forward, slow backwards, pause, leave the videos |
| $\mathcal{F}^v_c, \mathcal{F}^v_d, \mathcal{F}^v_s$ | Feature set of the subtitle length, duration time, speed of videos |
| $\varepsilon_f, \varepsilon_b, \varepsilon_p, \varepsilon_l$ | The upper limit parameter of the student's fast forward, slow backwards, pause, and leave behavior |
| $\eta_f, \eta_b, \eta_p, \eta_l$ | The lower limit parameter of the student's fast forward, slow backwards, pause, and leave behavior |

## III. Preliminary

**Preliminary** Before proposing our methodology, we describe some preliminary of our pipeline and corresponding notations (Table I). Given a specified student-course pair $< s \in \mathcal{S}, c \in \mathcal{C} >$, we attempt to make the dropout result $l \in \mathcal{L}$ interpretable after a series of the behaviours $\{b_1, b_2, ..., b_n\} \subseteq \mathcal{B}$ about watching these MOOC videos $\{v_1, v_2, ..., v_m\} \subseteq \mathcal{V}$.

*Definition 1.* **Dropout:** We define dropout in this paper as meaning that the student will not continue to study the remaining videos of the course after the behaviour of a certain video (not the final video) of a certain course ends.

*Definition 2.* **Student Behaviours:** We define student behaviours $\mathcal{B}$ as the behaviours of each student $s$ watching a series of videos $\mathcal{V}$ in the course $c$. We adopt the tuple ($s$, $v$, $\mathcal{F}^s_{wt}$, $\mathcal{F}^s_{pt}$) to infer whether the student $s$ has the action of fast forward or slow backwards and the tuple ($s$, $v$, $\mathcal{F}^s_{wt}$, $\mathcal{F}^s_{st}$) to infer whether the student $s$ has the action of pause or leave in this video $v$.

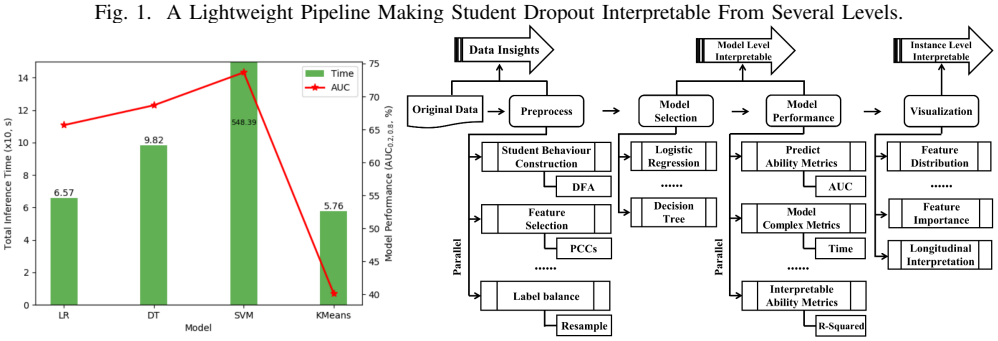
*Definition 3.* **Interpretable:** We make efforts to explain the final predictions of student dropout models for advancing the development of MOOC. In this paper, interpretability is an explanation, usually in a way that humans can understand, associating the feature values of an instance with its model predictions.

## IV. The Proposed Lightweight Pipeline

### A. Overview of the pipeline

As shown in Fig.1, our lightweight pipeline mainly consists of five components. We adopt this pipeline to obtain quantitative results and qualitative insights about the impact of MOOC videos on student dropout and utilize visualization for instructors to intuitively judge the impact of each student behaviour on the final label. We can infer from Fig.1 that Logistic Regression and Decision Tree can provide a fast and accurate prediction on the student dropout rate. And both of them have an excellent model level interpretation to uncover the underlying factors of student withdrawal.

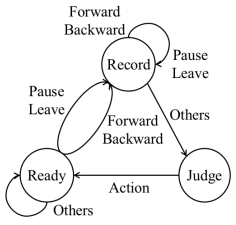
Fig. 1. A Lightweight Pipeline Making Student Dropout Interpretable From Several Levels.




### B. Data Processing and Feature Engineering

**Data Insights** Our dataset comes from the competition"MOOCCube student behaviour analysis", and data was crawled from Chinese one of the largest MOOC platform named XuetangX. XuetangX has provided over 1,000 courses and attracted more than 10,000,000 registered students, which offers abundant data to analyse students behaviours.

**Basic Features** This competition provides three types of JSON files about courses, videos and student behaviours. A series of basic information can be obtained from these original files, but some are not intuitive enough to make students behaviours interpretable, and some are inappropriate. We force on analyzing the behaviours of each student watching these videos because it is a necessary (but not sufficient) premise for learning and can be quantified by calculating the different times of each student watching these videos. Moreover, the characteristics of the course (video) itself are quite important for students' MOOC completion rate.




$$\begin{cases} f_f^{s_i} = 1, & \eta_f < f_{wt}^{s_i} - f_{pt}^{s_i} < \varepsilon_f \\ f_b^{s_i} = 1, & \eta_b < f_{wt}^{s_i} - f_{pt}^{s_i} < \varepsilon_b \\ f_p^{s_i} = 1, & \eta_p < f_{st}^{s_i} - f_{wt}^{s_i} < \varepsilon_p \\ f_l^{s_i} = 1, & \eta_l < f_{st}^{s_i} - f_{wt}^{s_i} < \varepsilon_l \end{cases}$$

Fig. 2. **Student Behaviours Construction -** The left figure illustrates the construction of complete student behaviours based on deterministic finite automaton (DFA) and the right equation shows the calculation process of these corresponding features.

**Student Behaviours Features** Video interaction data in XuetangX is recorded for each student generated event separately. These data are recorded in the form of JSON type events, which we then structure into students behaviours features form as described below. Every row in the raw JSON file records a series of times like local_watching_time, video_progress_time, video_start(end)_time and local_start(end)_time. We attempt to aggregate these into simple, interpretable features for a specified video id-student id pair: watching_count, local_watch_time, local_watch_percentage, local_progress_time, local_web_time, forward, backward, pause and leave. Each of these captures various potential factors in the consumption of video content by students, which is shown in Table II.

We employ a deterministic finite automaton (DFA) to construct the complete student behaviours. Fig.2 shows three state transitions in this DFA: Ready, Record, Judge. When the state is Ready, it stays until receives a Forword or Backword event (Pause or Leave, $\mathcal{F}_f^s = 1$ *or* $\mathcal{F}_b^s = 1$ && $\mathcal{F}_p^s = 1$ *or* $\mathcal{F}_l^s$), then the state transforms to Record. At the Record state, there is a stack. When getting a new Pause or Leave (Forward or Backward) events, it pushes all the events into the stack. If there come some other events, it goes to the Judge state. At the Judge state, we check whether the Action.json is normal data to decide whether to save it and return to the Ready state. It is worth noting that in each cycle, $\mathcal{F}_f^s$ and $\mathcal{F}_b^s$ ($\mathcal{F}_p^s$ and $\mathcal{F}_l^s$) are mutually exclusive, but $\mathcal{F}_f^s$ and $\mathcal{F}_p^s$ ($\mathcal{F}_l^s$) can exist at the same time.

**Course (Video) Features** Each course has a list of corresponding videos, and we utilize the characteristics of these videos to represent different courses. For the original video JSON file, it is difficult to directly use the text information. We make efforts to classify these videos by the video name, but because of the limitation of information, judging based on only a few characters is time-consuming and meaningless. Here we adopt a more direct and effective strategy. We care less about the content of video text itself but pay more attention to the coarse-grained features such as the duration of videos and the speed of speech. The text of all frame is counted and combined with the duration time to calculate the speed of every video. Intuitively, videos where instructors speak fairly fast and with high enthusiasm are more engaging, which is shown in Table II.

**Linear Correlation** We utilize the Pearson Correlation Coefficients (PCCs) to explore the linear correlation between different features and student dropout (Eq.1).

TABLE II
PCCs OF THE 12 FEATURES AND STUDENT DROPOUT.

| $\mathcal{F}^s_{wc}$ | $\mathcal{F}^s_{wp}$ | $\mathcal{F}^s_{wt}$ | $\mathcal{F}^s_{pt}$ | $\mathcal{F}^s_{st}$ | $\mathcal{F}^s_{f}$ |
|---|---|---|---|---|---|
| -0.2990 | -0.3147 | -0.2015 | -0.3219 | -0.2991 | -0.2332 |
| $\mathcal{F}^s_{b}$ | $\mathcal{F}^s_{p}$ | $\mathcal{F}^s_{l}$ | $\mathcal{F}^v_{c}$ | $\mathcal{F}^v_{d}$ | $\mathcal{F}^v_{s}$ |
| -0.2144 | 0.0072 | -0.2990 | -0.1697 | 0.0273 | 0.1842 |

$$\rho_{f_1,f_2} = \frac{cov(f_1,f_2)}{\sigma_{f_1}\sigma_{f_2}} = \frac{E((f_1-\mu_{f_1})(f_2-\mu_{f_2}))}{\sigma_{f_1}\sigma_{f_2}} \quad (1)$$

Where $cov$ and $\sigma$ are covariances and standard deviations of two different continuous feature variables, respectively. Each student will produce a series of actions, so the feature parameters are a list. In order to simplify the parameters, we explore the linear correlation between the maximum, minimum, median and average value of each feature and the dropout. Take feature $\mathcal{F}^s_{wp}$ as an example, $\rho_{max(\mathcal{F}^s_{wp}),\mathcal{L}} = -0.2825$ $\rho_{mid(\mathcal{F}^s_{wp}),\mathcal{L}} = -0.3147$ $\rho_{min(\mathcal{F}^s_{wp}),\mathcal{L}} = -0.0557$ $\rho_{avg(\mathcal{F}^s_{wp}),\mathcal{L}} = -0.2935$, here we will select the feature value with the largest linear correlation as our final input, that is, when the absolute value of $\rho$ is the largest.

Table II verifies our above conjecture, the completion rate of MOOC course with more video content and instructors speaking fast is higher [$\rho_{max(\mathcal{F}^v_d),\mathcal{L}} = -0.0262$ $\rho_{min(\mathcal{F}^v_d),\mathcal{L}} = 0.0273$ $\rho_{max(\mathcal{F}^v_s),\mathcal{L}} = -0.0618$ $\rho_{min(\mathcal{F}^v_s),\mathcal{L}} = 0.1842$]. Simultaneously, other features that mostly take the median and average values have a linear and negative correlation with the MOOC dropout, and we remove the low linear correlation feature $\mathcal{F}^s_p$.

In addition, we calculate the PCCs for the remaining 11 features pairwise so as to avoid the problem of multicollinearity between variables (Fig.3). We identify $\rho_{f_1,f_2} \geq 0.8$ as a highly correlated variable pair $< f_1, f_2 >$, and eliminate features with smaller $\rho_{f,\mathcal{L}}$ values.

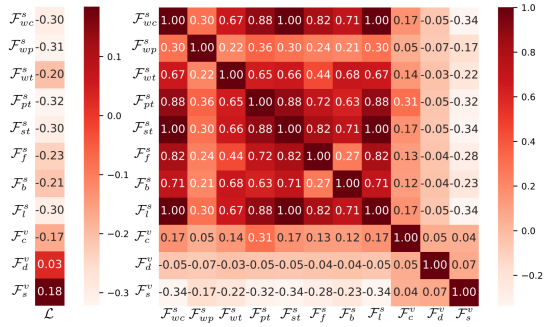



Fig. 3. **Data Insight -** PCCs heat map of features and labels.

### C. Logistic Regression Model

**Logistic Regression** A logistic regression model (LR model) predicts the label $\mathcal{L}$ as a probabilities between 0 and 1 (Eq.2). However, the interpretation of the weights in logistic regression do not influence the probability linearly, which impairs the intuitive interpretability of each input feature $f \in \mathcal{F}$. Here we reformulate the original equation for the interpretation so that only the linear term is on the right side of the formula and we employe the "odds" represents the probability of event divided by the probability of no event (Eq.3).

$$P(l^i = 1) = \frac{1}{1 + e^{-(\beta_0+\beta_1 f^i_1+\beta_2 f^i_2+\ldots+\beta_n f^i_n)}} \quad (2)$$

$$odds = \frac{P(l^i = 1)}{1 - P(l^i = 1)} = e^{\beta_0+\beta_1 f^i_1+\beta_2 f^i_2+\ldots+\beta_n f^i_n} \quad (3)$$

$$\frac{odds(f_n + \Delta f)}{odds} = e^{\beta_n(f_n+\Delta f)-\beta_n f_n} = e^{\beta_n \Delta f} \quad (4)$$

At last, An increase in a feature $f_n$ by $\Delta f$ changes the odds ratio (multiplicative) by a factor of $e^{\beta_n \Delta f}$ (Eq.4). Although interpreting the advantage ratio requires some mental arithmetic, it is much simpler than considering the log function. Particularly, for the numerical feature, if the value of feature $f_n$ increases by one unit, the estimated odds change by a factor of $e^{\beta_n}$. And for the binary categorical feature, altering the feature $f_n$ from the reference category to the other category changes the estimated odds by a factor of $e^{\beta_n}$ as well.

**Model Level Interpretation** In the above, we introduce odds to more intuitively show the impact of each feature change on the prediction results, for the interpretation of weight in the logistic regression model depends on the type of the corresponding feature. Another important measurement for interpreting models is the $R$-squared measurement, which implies the total variance of the target result is explained by the model (Eq.5).

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{m}(l^i - \hat{l}^i)}{\sum_{i=1}^{m}(l^i - \bar{l}^i)} \quad (5)$$

Where SSE is the squared sum of the error terms, which is measured by the squared differences between the predicted and actual target values. And SST is the squared sum of the data variance, which is measured by the square difference between the average and actual target value. $R$-squared increases with the number of features in the model, but it is not related to the number of instances and labels. Here, we also list the adjusted $R$-squared ($\bar{R}$-squared) values to account for the number of features used in the model. The calculation method of $\bar{R}$-squared is shown in Eq.6, where $i$ is the number of features and $n$ the number of instances.

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-i-1} \quad (6)$$

As we can see from Table III, the values of $R$ and $\bar{R}$ are both increasing with the number of input features, which indicates that the interpretability of the logistic regression model will strengthen with the richness of input features.

TABLE III

**MODEL PERFORMANCE -** THE INFLUENCE OF DIFFERENT FEATURES ON THE INTERPRETABILITY AND ACCURACY. WE ADD THEM ONE BY ONE ACCORDING TO THE IMPORTANCE OF THE FEATURE (REFER TO THE ESTIMATE AND SIGNIF.CODES COLUMNS OF TABLE IV FOR THE RELEVANT IMPORTANCE).

| Model ($\mathcal{L} \sim \mathcal{F}$) | AIC | BIC | $R$-squared | $\bar{R}$-squared |
|---|---|---|---|---|
| $\mathcal{F}^s_{wp}$ | 23780 | 23795 | 0.6985642 | 0.4879666 |
| $\mathcal{F}^s_{wp}\mathcal{F}^s_b$ | 22572 | 22569 | 0.7056793 | 0.4979337 |
| $\mathcal{F}^s_{wp}\mathcal{F}^s_b\mathcal{F}^s_f$ | 20904 | 20963 | 0.7071946 | 0.5000501 |
| $\mathcal{F}^s_{wp}\mathcal{F}^s_b\mathcal{F}^s_f\mathcal{F}^s_{pt}$ | 20451 | 20491 | 0.7214005 | 0.5203239 |
| $\mathcal{F}^s_{wp}\mathcal{F}^s_b\mathcal{F}^s_f\mathcal{F}^s_{pt}\mathcal{F}^v_s$ | 20447 | 20494 | 0.7226397 | 0.5220901 |
| $\mathcal{F}^s_{wp}\mathcal{F}^s_b\mathcal{F}^s_f\mathcal{F}^s_{pt}\mathcal{F}^v_s\mathcal{F}^s_{wt}$ | 20445 | 20500 | 0.7271775 | 0.5286474 |

However, this does not mean using all variables, because too many variables will lead to over-fitting, which will reduce the generalization of the model. In addition, complex features are not suitable for model interpretability. In this paper, we hope to find a small number of decision-making features to help MOOC instructors infer whether students will complete this course for future course adjustments.

We simplify the input to improve the interpretability of the model by eliminating linear irrelevant and collinearity features. Here we employ two indicators, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), to verify that the 8 modelling variable models we selected alleviate the problem between accuracy and overfitting.

$$AIC = -2In(L)+2i,\ BIC = -2In(L)+i*In(n) \quad (7)$$

Where $L$ is the maximum likelihood under the logistic regression model, $n$ is the number of data, and $i$ is the number of input features in this model. Table IV shows that the model we built balances the complexity of the model and the ability to describe the data set (likelihood function).

It can be seen from Table III and Table IV that the course completion rate is the most relevant to the percentage of course watching ($\mathcal{F}^s_{wp}$), followed by the behaviours of student watching videos, forward ($\mathcal{F}^s_f$) and backward ($\mathcal{F}^s_b$). What's more, the dropout rate of courses with slower video speaking rates will increase. In general, the logistic regression model gives more weight to the preprocessed features. Taking video watching time ($\mathcal{F}^s_{wt}$) and video watching rate ($\mathcal{F}^s_{wp}$) as examples, both of them can measure the degree to which a student learns a certain video. But from the actual data, we can see that $\mathcal{F}^s_{wp}$ is more sensitive, which means it is essential to extract student behaviours from basic information.

**Instance Level Interpretation** We visualize the coefficients of the logistic regression model to explain the result of the sample student's final label after a series of actions (Fig.4(b)). For individuals, basic features and student behaviours features have a linear and negative correlation with the MOOC dropout. That is to say, the higher the time or the percentage of an example student watching the video, the more frequently playing forward or backwards, the higher likelihood he is to complete this course. Video features are positively linearly correlated with the MOOC dropout, the speed of the video also affects other features. As shown in Fig.4(c), when the influence of $\mathcal{F}^v_s$ increases, the influence of other features will decrease. At this time, the student is more likely to give up the course.

TABLE IV

**MODEL PARAMETERS -** FEATURE WEIGHT AND IMPORTANCE.

| Feature | Estimate | Odds ratio | Std. Error | z value | Pr ($>\|z\|$) | Signif.codes |
|---|---|---|---|---|---|---|
| $Intercept$ | 0.7912388 | 2.2061276 | 0.0739276 | 10.703 | $<$ 2e-16 | $\star\star\star$ |
| $\mathcal{F}^s_{wp}$ | -0.6834781 | 0.5048579 | 0.0794952 | -8.598 | $<$ 2e-16 | $\star\star\star$ |
| $\mathcal{F}^s_{wt}$ | -0.0010715 | 0.9989290 | 0.0006284 | -1.705 | 0.08819 | . |
| $\mathcal{F}^s_{pt}$ | -0.0150089 | 0.9851032 | 0.0010296 | -14.578 | $<$ 2e-16 | $\star\star\star$ |
| $\mathcal{F}^s_f$ | -0.1154977 | 0.8909226 | 0.0081302 | -14.206 | $<$ 2e-16 | $\star\star\star$ |
| $\mathcal{F}^s_b$ | -0.1640974 | 0.8486593 | 0.0127081 | -12.913 | $<$ 2e-16 | $\star\star\star$ |
| $\mathcal{F}^v_c$ | 0.0000014 | 1.0000014 | 0.0000012 | 1.132 | 0.25773 | |
| $\mathcal{F}^v_d$ | -0.0004825 | 0.9995176 | 0.0004183 | -1.153 | 0.24872 | |
| $\mathcal{F}^v_s$ | 0.0005496 | 1.0005497 | 0.0001792 | 3.066 | 0.00217 | $\star\star$ |

TABLE V

**MODEL PARAMETERS -** THE IMPORTANCE OF FEATURES UNDER DIFFERENT DEPTH DECISION TREE MODELS.

| Model ($\mathcal{L} \sim c\mathcal{I}$) | Feature Importances | | | | | Model Performance | |
|---|---|---|---|---|---|---|---|
| | $\mathcal{F}^s_{wp}$ | $\mathcal{F}^s_{pt}$ | $\mathcal{F}^s_f$ | $\mathcal{F}^s_b$ | $\mathcal{F}^v_s$ | $R$-squared | $\bar{R}$-squared |
| max_depth = 3 | 0.0 | 0.942 | 0.033 | 0.0 | 0.025 | 0.725 | 0.526 |
| max_depth = 4 | 0.011 | 0.919 | 0.030 | 0.024 | 0.015 | 0.731 | 0.534 |
| max_depth = 5 | 0.026 | 0.870 | 0.032 | 0.029 | 0.043 | 0.733 | 0.537 |
| max_depth = 6 | 0.049 | 0.839 | 0.025 | 0.026 | 0.061 | 0.744 | 0.553 |

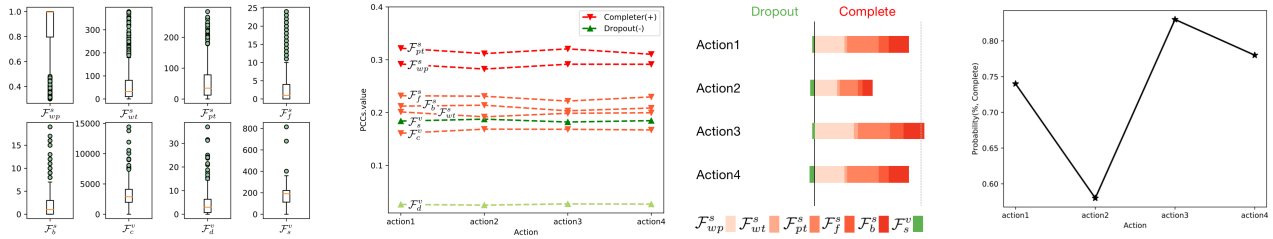
### *D. Decision Tree Model*

**Decision Tree** Decision tree models can handle situations where the relationship between features and outcome is nonlinear or where features interact with each other. Decision tree models split the data multiple times according to certain cutoff values in the features. Through splitting, different subsets of the dataset are created, with each instance belonging to one subset. The final subsets are called terminal (leaf) nodes and the intermediate subsets are called internal (split) nodes. The following equation describes the relationship between the outcome label $\mathcal{L}$ and the input features $\mathcal{F}$.

$$l_i = \sum_{m=1}^{M} c_m I\{f \in T_m\} \quad (8)$$

Each instance falls into one terminal node subset $T_m$. $I\{f \in T_m\}$ denotes the identity function which return 1 when the input features $\mathcal{F}$ is in the final subsets $T_m$, 0 otherwise. For a terminal node $t_i$, the instance outcome label is $l_i = \bar{c}$, where $\bar{c}$ denotes the average value of the whole instances in this terminal node subset $T_m$.

**Model Level Interpretation** We finally select the five most relevant features to construct the decision tree model through the previous analysis. As the max_depth of the decision tree model increases, the accuracy of the model will be improved, and the corresponding feature weights will be more balanced. But it is worth noting that the




(a) A distribution boxplot of the most linearly correlated feature values.

(b) A line graph of Pearson correlation coefficients of features and labels.

(c) The probability estimation and change trend of predicted label linked to feature values.

Fig. 4. **Instance Interpretation -** The features of the dropout prediction model have different priorities. As students' actions of watching videos increase, these features will change differently (the feature change interval is shown in subfigure (a) ), and the estimated weight of the feature will also change (the trend of feature change is shown in subfigure (b) ).

TABLE VI
**MODEL INSTANCE -** DROP AND COMPLETE INSTANCES UNDER DIFFERENT DEPTH DECISION TREE MODELS.

| | Model ($\mathcal{L} \sim c\mathcal{I}$) | Instance | Gini | Samples |
|---|---|---|---|---|
| Drop | max_depth = 3 | $\mathcal{F}^s_{pt} \geq 95.675, \mathcal{F}^v_s \leq 249.084$ | 0.039 | 2577 |
| | max_depth = 4 | $\mathcal{F}^s_{pt} \geq 121.675, \mathcal{F}^v_s \leq 249.084$ | 0.023 | 2106 |
| | max_depth = 5 | $\mathcal{F}^s_{pt} \geq 90.873, \mathcal{F}^v_s \leq 324.997$ | 0.08 | 72 |
| | | $\mathcal{F}^s_{pt} \geq 141.56, \mathcal{F}^v_s \geq 247.44, \mathcal{F}^s_f \geq 9.5$ | 0.005 | 1127 |
| | | $\mathcal{F}^s_{pt} \geq 131.481, \mathcal{F}^v_s \leq 249.725$ | 0.0 | 74 |
| | max_depth = 6 | $\mathcal{F}^s_{pt} \leq 102.824, \mathcal{F}^s_{wp} \leq 0.992, \mathcal{F}^s_f \geq 9.5$ | 0.0 | 117 |
| | | $\mathcal{F}^s_{pt} \geq 122.564, \mathcal{F}^s_{wp} \leq 0.980, \mathcal{F}^v_s \leq 260.331$ | 0.0 | 188 |
| Complete | max_depth = 3 | $\mathcal{F}^s_{pt} \leq 51.099, \mathcal{F}^v_s \geq 248.616$ | 0.452 | 626 |
| | max_depth = 4 | $\mathcal{F}^s_{pt} \leq 59.075, \mathcal{F}^v_s \geq 304.247$ | 0.229 | 114 |
| | max_depth = 5 | $\mathcal{F}^s_{pt} \leq 56.561, \mathcal{F}^v_s \geq 304.025$ | 0.013 | 100 |
| | | $\mathcal{F}^s_{pt} \geq 90.873, \mathcal{F}^v_s \leq 160.196, \mathcal{F}^s_f \geq 4.5$ | 0.0 | 18 |
| | | $\mathcal{F}^s_{pt} \leq 43.297, \mathcal{F}^v_s \geq 249.725$ | 0.0 | 48 |
| | max_depth = 6 | $\mathcal{F}^s_{pt} \geq 102.824, \mathcal{F}^v_s \leq 250.019, \mathcal{F}^s_f \geq 5.5$ | 0.0 | 25 |
| | | $\mathcal{F}^s_{pt} \geq 53.186, \mathcal{F}^s_{wp} \geq 0.967, \mathcal{F}^v_s \geq 348.699$ | 0.0 | 87 |

increase in model complexity will affect our intuitive judgment, which is contrary to the original intention of this paper. Therefore, in the following instance analysis, we will adopt a three-layer decision tree. As can be seen from Table V, $\mathcal{F}^s_{pt}$ and $\mathcal{F}^v_s$ are the most significant features in the decision tree model. Table VI shows the student instances under different depth decision tree models. We can infer intuitively whether a student will dropout from a course is closely related to the video watching time, the playing speed, and the instructor's speaking speed.

**Instance Level Interpretation** Three distinct profiles of dropout can be inferred from paths to leaves indicating label = Drop. These profiles are 1) Video progress exceeds two minutes while the instructor's speaking speed is under four words per second, 2) Students who fast forward many times, the video progress time is less than one and a half minutes, the percentage of watches is relatively low, 3) Students who fast forward many times, video progress exceeds two and a half minutes while the instructor's speaking speed is above four words. It is not difficult to infer that the completers are those who watch a series of short videos with a fast instructor's speaking speed at a fast rate, which is consistent with the experimental instances at Table VI.

## V. CONCLUSION

Our lightweight pipeline begins with interpretable features, we employ a deterministic finite automaton to construct the complete student behaviours and we adopt a direct and effective strategy to extract the course features. We do our utmost to simplify the MOOC dropout problem, grasp the core features through two interpretable models. In the end, the student's dropout behaviour is explained at several levels, which enables instructors and video producers to make the most of online videos for future education.

## REFERENCES


[1] J. He, J. Bailey, B. I. P. Rubinstein, and R. Zhang, "Identifying at-risk students in massive open online courses," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, B. Bonet and S. Koenig, Eds. AAAI Press, 2015, pp. 1749–1755.

[2] D. T. Seaton, Y. Bergner, I. L. Chuang, P. Mitros, and D. E. Pritchard, "Who does what in a massive open online course?" *Commun. ACM*, vol. 57, no. 4, pp. 58–65, 2014. [Online]. Available: https://doi.org/10.1145/2500876

[3] W. Feng, J. Tang, and T. X. Liu, "Understanding dropouts in moocs," in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 2019, pp. 517–524.

[4] 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*, 2014, pp. 60–65.

[5] T.-y. LIU and L. Xiu, "Finding out reasons for low completion in mooc environment: an explicable approach using hybrid data mining methods," *DEStech Transactions on Social Science, Education and Human Science*, no. meit, 2017.

[6] R. Al-Shabandar, A. Hussain, A. Laws, R. Keight, J. Lunn, and N. Radi, "Machine learning approaches to predict learning outcomes in massive open online courses," in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 713–720.

[7] 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)*. IEEE, 2015, pp. 256–263.

[8] J. Gardner and C. Brooks, "Dropout model evaluation in moocs," *arXiv preprint arXiv:1802.06009*, 2018.

[9] W. Wang, H. Yu, and C. Miao, "Deep model for dropout prediction in moocs," in *Proceedings of the 2nd International Conference on Crowd Science and Engineering*, 2017, pp. 26–32.