IOP Conf. Series: Journal of Physics: Conf. Series **1487** (2020) 012032 doi:10.1088/1742-6596/1487/1/012032

Massive Learning Behaviours Influence Educational Sustainability: A Machine Learning Approach

Syed Muhammad Raza Abidi[1,a,b], Mushtaq Hussain[1], Sen Ge[1], Hu Ding[2], Wenhao Zhu[1] and Wu Zhang[1,2,c]

[1]School of Computer Engineering and Science, Shanghai University, Shanghai, 200444, China
[2]Shanghai Institute of Applied Mathematics and Mechanics, Shanghai University, Shanghai, 200444, China

[a]abidi@i.shu.edu.cn, [b]razaabdi@live.com, [c]wzhang@shu.edu.cn

Abstract. Discovering learning behavior patterns while using Massive Open Online Courses (MOOCs) or e-learning platform education is being one of the substantial aspects in the plethora of past studies. In this study, we predict the taxonomy of learner behaviors by using the data of Junyi Academy, an e-learning platform similar to Khan Academy. We evaluate the dataset of first 2,000 unique learners and applying machine learning algorithms, i.e., Logistic Regression, and Neural Network (NNET) to predict the status of learners' behavior. Our result shows that NNET autotuned is high accuracy champion predictive model as compared to Logistic Regression and NNET manually tuned hyperparameters models. The accuracy we measured based on Kolmogorov-Smirnov (KS) statistic for VALIDATION partition dataset. Furthermore, our model allows teachers to track or predict behavioral changes of the individual student and reveal insights into the content of the course material. It will also assist teachers in getting an early prediction of behaviors before finishing the course or skill, which is constructive and valuable for sustainable education or educational sustainability.

1. Introduction

The way students behave in an education system is an essential matter in educational data mining (EDM). Understanding this behavior in the education system supports to figure out how students seek, help, gain, and grasp knowledge for excellent learning. These attitudes can be assumed through both, a precise study [1], or through the system generated logged data [2].

MOOCs have considered as an unsettling modernization in an educational system and likely to assist in increasing the opportunities related to a career in rising fields, such as data sciences [3].

Numerous methods have established to examine the performance and behaviors of massive learners and content data to recommend what behaviors learner should be taken into account, for instance; watching a lecture video, solving practice questions, or seeking help, etc. [4]. By procuring the demands of each particular learner, such practice and approach can augment the learning competencies with serious attitudes [5].

Likewise, [6] showed that learners or students probably seem to be less confuse or having sincere behavior who have a great practice of perquisites mandatory knowledge and proficiencies, compared to those with little or no mastery. Furthermore, in the study of [7], explained that in predicting the behavior

or responses of a learner on a particular skill is the learner's ability and familiarity on prerequisite skills, which is an important aspect.

1.1. Existing methods for performance or behavior of learners

Attempting assignments, quizzes questions, seeking help, and practicing exercises, etc. are usually considered the learner's behavior and performance. Several models or algorithms have used to make predictions from performance data which include causal graphs, Bayesian estimation [8], probabilistic association rules, and correlation/regression analysis [9].

The final performance is measured at the end of the MOOC or mastery skill to determine how many point and proficiency earned by the student. Also, the Learner's demographic background, motivation, preceding knowledge, and session with the tutor [10] all linked to behavior and performance of the learners.

[3] analyzed clickstream data to measure how many times a learner interacts with course components, i.e., videos, discussion forums or boards, assignments, and acquires help, etc. Yang, Sinha, David, and Rose [11] identified the predictors of completion of a course from the discussion forums and further found how the student triggers conversation and what is the frequency of posting a question on forums.

1.2. Our method and contributions

In this study, we develop a methodology to extract behavioral learner's data from the large-scale Junyi Academy Problem Log which contains over 10 million records [12] using Logistic Regression and NNET machine learning models. We measure the behavioral data of learners shown in Table 1. Explicitly, our methodology consists of the following main steps:

1.2.1. Feature engineering. Firstly, we evaluate the features related to the behavior of the learner by using e-learning platform. We then analyze and transform features to form new features to help to categorize the student's behavior. These features summarize the learner's behavior in section *2.3.*

1.2.2. Modeling and inference. Secondly, we infer the parameters of our learning model through training and validation on the dataset. Our model parameters are trained by using the selected best features and predict the status of learner's behavior shown in Table 2.

In the end, we also describe how our model parameters identify the behaviors. Generally speaking, we trust that this work will encourage a new research thrust in classifying human behavior to aid sustainable development for education.

2. Relationship discovery and procedures

2.1. Data and predictors collection

Data were collected from Junyi Academy (http://www.junyiacademy.org/), similar to Khan Academy, an e-learning platform. We have initially retrieved the first two thousand records related to unique user IDs in our experimentation. Table 1 shows the description of the predictor variables used in this study.

Table 1. List of features used.

Class	Features
Numerical	user_ids, exercise, sess_w_tutr, practicing_ex, ex_suggestion, review_mode, tme_done, tme_spnd_whl_ex, hint_rqstd, count_attempts_ans, hint_used, count_hints, tme_spnd_hnt, earned_proficiency, points_earned, tutr_incorrect, tutr_correct, tutr_hint

2.2. Preprocessing and selection of predictor variables

The first vital step before applying machine learning technique is the preprocessing of a data, which assures that the desired data is harmonized, standardized, structured, cleaned, and noise-free, and it should be transformed in a way to be useful for machine learning analysis [13]. We used SAS Viya for Learners (https://www.sas.com/en_us/software/viya-for-learners.html) analytics software and used the state of the art Model Studio and used SAS Visual Data Mining and Machine Learning (DMML) environment to prepare and transform the data as per the experimentation.

After removing noise (i.e., duplication, skewness, missingness, and replacing incorrect values, etc.) from the data, we are left with a total of 1881 unique records of students. Then, we turn to choose which predictors are useful for our model and drop out those which are not suitable. We have used SAS variable selection node to filter out the most appropriate predictors as shown in Table 2.

2.3. Feature extraction and label collection

For the selection of Label (response variable), we extracted features from the list of predictors in Table 1 and combined them to form a new label variable "behavior_status" to identify the behavior of the students attempting exercises using e-learning platform. We have used below logic to create a response variable or Label.

If (practicing_exercise > Total average (practicing_exercise) Or (Total_time_sepnd_hint) > Total average (Total_time_sepnd_hint) then (Insincere_behavior) else (sincere_behavior)

The label has values "sincere_bhavior" and "insincere_behavior," which is 0, and 1 as an integer respectively.

2.4. Final feature selection for learning models

We have finally selected variables through SAS variable selection node. Varying combinations of criteria can be used to select inputs *(i.e., Unsupervised Selection, Fast Supervised Selection, Linear Regression Selection, Decision Tree Selection, Forest Selection, and Gradient Boosting Selection)*. We keep the Combination Criterion at Selected by at least 1. It means that any input selected by at least one of the selection criteria chosen is passed on to subsequent nodes as inputs. Table 2 shows the predictors verified by two criterions *(Fast Supervised Learning, and Linear Regression Selection)*.

Table 2. Features selected by fast supervised selection and linear regression selection.

Name of Input Variable	Fast Supervised Selection	Linear Regression Selection	Input	Rejected	Output Role
count_attempts_ans	Input	Rejected	1	1	Input
count_hints	Input	Rejected	1	1	Input
earned_proficiency	Rejected	Rejected	0	2	Rejected
Exercise	Input	Input	2	0	Input
ex_suggestion	Input	Rejected	1	1	Input
hint_rqstd	Input	Input	2	0	Input
hint_used	Rejected	Input	1	1	Input
points_earned	Input	Input	2	0	Input
review_mode	Rejected	Rejected	0	2	Rejected
sess_w_tutr	Rejected	Rejected	0	2	Rejected
tme_done	Rejected	Rejected	0	2	Rejected
tme_spnd_whl_ex	Rejected	Rejected	0	2	Rejected
tutr_correct	Rejected	Rejected	0	2	Rejected
tutr_hint	Input	Rejected	1	1	Input
tutr_incorrect	Rejected	Rejected	0	2	Rejected

2.5. Relationship prediction

Effectively specifying the features and relationship labels, we articulate the relationship prediction task as a logistic regression analysis as a base model, then we use the collected labels to experiment on the effects of using NNET algorithm. As our class target is categorical, so, using 10-dimension features, we choose the champion predictive model.

3. Experiment and Results

3.1. Construction of machine learning models

We have used two machine learning methods, i.e., Logistic Regression and NNET to construct the candidate models for predicting the status of the behavior of students using Junyi Academy e-learning platform. We have used SAS Visual DMML to execute above-mentioned machine learning methods and tested. SAS Viya is the high-tech software for a robust collection of statistical, data mining, and machine learning built-in environments, modules, and functionalities.

3.2. Training & evaluation of learning models

After creating the project in the model studio, we partitioned the data into a 70% – 30% ratio with the stratified method and then trained the learning algorithms. We then evaluated our models on the validation data to confirm the models are not overfitting the data which is an essential part in machine learning techniques and also verifies that the model is good enough to learn and classify correctly. We also chose and selected the champion model based on validation data using the (KS) statistic as selection criteria class and Averaged squared error as Selection Criteria Interval properties.

3.3. Results and analysis

Table 3 explains the Fit statistics of the logistic model and based on the (KS) value; it is considered to be a good model.

Table 3. Fit statistics of logistic regression model.

Data Role	Misclassification Rate	Average Squared Error	KS (Youden)	KS Cut-off	Area Under ROC
Train	0.1929	0.1338	0.5933	0.35	0.8414
Validate	0.1755	0.1328	0.5135	0.35	0.8237

ROC: Receiver operating characteristic

Then, we have applied the NNET model and modified the architecture, learning, and optimization parameters with the intent to improve performance. Table 4 explains the modification in the parameters, whereas, Table 5 reveals the Fit statistics of tuned hyper-parameterized NNET model.

Table 4. Modified hyper-parameters list of NNET.

Property	Default Value	Modified Value
Input Standardization	Midrange	Z score
Number of Hidden Layers	1	1
Use the Same Number of Neurons in Hidden Layers	Ticked checkbox	Cleared the checkbox
Number of Neurons per Hidden Layer	50	26
Common Optimization Options L1 Weight Decay	0	0.01
L2 Weight Decay	0.1	0.0001

Table 5. Fit statistics of NNET Model.

Data Role	Misclassification Rate	Average Squared Error	KS (Youden)	KS Cut-off	Area Under ROC
Train	0.1913	0.1402	0.4362	0.25	0.7815
Validate	0.1791	0.1366	0.4528	0.20	0.7782

Further, we have also built a NNET model using Autotuning and finally compared the performance of the other models already in the pipeline in SAS Visual DMML and chose the champion model on the basis of increasing (KS) value shown in Table 6 on the validation data.

Table 6. Fit statistic of learning models' comparison (logistic regression and NNET).

Data Role	Model Name	Misclassification Rate	Average Squared Error	KS (Youden)	KS Cut-off	Area Under ROC
Validate	Logistic Regression	0.1755	0.1328	0.5135	0.35	0.8237
	NNET (Tuned)	0.1791	0.1366	0.4528	0.20	0.7782
	NNET (Autotuned)	0.1773	0.1335	0.5305	0.40	0.8232

Thus, based on (KS) statistic, our champion model is NNET Autotuned which has the highest value of (KS) which shows the goodness of fit statistic that denotes the maximum separation between the model ROC curve and the baseline ROC curve. ROC of all the models used in this study, and the Neural Network (1) TRAIN and VALIDATE shown in 'figure 1' is NNET Autotuned model.

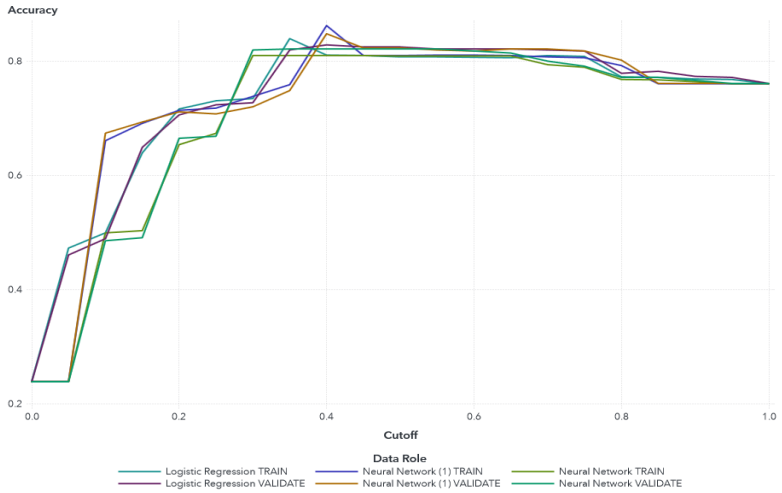

Figure 1. Accuracy performance comparison of all the learning models.

4. Conclusion and future works

In this paper, we have proposed Logistic Regression, NNET, and variants of NNET learning-based models to extract the behavior of students using e-learning platform education. We have applied our models on the Junyi Academy dataset and mine the data of first 2,000 unique learners and identify the behavioral patterns by examining various activities attempted by students. Our experiment exposed that the NNET autotuned model significantly outperforms based on Kolmogorov-Smirnov (KS) statistic on the validation data. The (KS) value is 53.05%, and a higher value is considered excellent.

There are several boulevards of future works to augment the performance of the learning models and prediction of the behavioral attitudes of the students by utilizing other techniques such as incorporating more variables into our model to allow variation in learner behaviors, also other factors, e.g., total time spent on the whole exercise, problem type, answers attempt count, and hint requested, etc. can be measured to further dig down the insights through responses of students.

Our method in this study allows for tracking the status of the behavior of an individual student, and helpful for teachers to recognize which student needs attention on which module or exercise and requires training to avoid dishonest behavior or attitude. So, by attaining the early prediction of behavioral attitudes aid to stabilize the educational sustainability or sustainable development for education. As the sustainability in education prioritizes the present needs than the future needs [14].

Further, educational sustainable development focuses on the economic, societal, and environmental surface of the formal and informal education. Also encourages and promotes learners to grasp and advance the inherent intelligence, capabilities, and expertise to become a liable person of society and execute symbolic role in sustainable development for education [15].

Acknowledgments

This research was funded by the “Program of Shanghai Municipal Education Commission (No. 2019-01-07-00-09-E00018).”

References

[1] Hutt S, Mills C, White S, Donnelly PJ and D ’Mello SK 2016 *The Eyes Have It: Gaze-based Detection of Mind Wandering during Learning with an Intelligent Tutoring System* (Proc. 9th In- t Conf Educ Data Mining Int Educ Data Min Soc.) p 86–93

[2] Lorenzen S and Alstrup S 2018 *Tracking Behavioral Patterns among Students in an Online Educational System* (Proc 11th Int Conf Educ Data Min.) p 2–7

[3] Wang Y, Baker RS and Paquette L 2017 *Behavioral predictors of MOOC post-course development* 1967 (CEUR Workshop Proc.) p 100–111

[4] Manickam I, Lan AS and Baraniuk RG 2017 *Contextual multi-armed bandit algorithms for personalized learning action selection* (ICASSP, IEEE Int Conf Acoust Speech Signal Process P- roc.) p 6344–6348

[5] Chen W, Lan AS, Cao D, Brinton C and Chiang M 2018 *Behavioral Analysis at Scale: Learning Course Prerequisite Structures from Learner Clickstreams* (Proceedings of the 11th Internati-onal Conference on Educational Data Mining) p 66-75

[6] Wan H and Beck JE 2015 *Considering the influence of prerequisite performance on wheel spinning* (Proc. 8th Int Conf Educ Data Min.) p 129–135

[7] Botelho A, Wan H and Heffernan N 2015 *The prediction of student first response using prerequisite skills* (L@S - 2nd ACM Conf Learn Scale.) p 39–45

[8] Han S-Y, Yoon J and Yoo YJ 2017 *Discovering Skill Prerequisite Structure through Bayesian - Estimation and Nested Model Comparison* (International Conference on Educational Data M-ining) p 398–399

[9] Koedinger KR, Carbonell J, Chaplot DS and Yang Y 2016 *Data-driven Automated Induction of Prerequisite Structure Graphs* (Proceedings of the 9th International Conference on Education-al Data Mining) p 318–323

[10] Hone KS and El Said GR 2016 *Exploring the factors affecting MOOC retention: A survey study* 98

(Computers and Eduction) p 157–168
[11] Yang D, Sinha T, Adamson D and Rose C 2013 *"Turn on, Tune in, Drop out": Anticipating student dropouts in Massive Open Online Courses* 22(2) (Proceedings of the NIPS Workshop on-Data Driven Education) p 1–8
[12] Chang H-S, Hsu H-J and Chen K-T 2015 *Modeling Exercise Relationships in E-Learning : A Unified Approach* (Proc. 8th Int Conf Educ Data Min.) p 532–535
[13] Jeff Thompson and Truxillo C 2019 *Machine learning using SAS Viya* [cited 2019 Jul 2] Available from: https://www.coursera.org/learn/machine-learning-sas?
[14] Abidi SMR, Hussain M, Xu Y and Zhang W 2018 *Prediction of confusion attempting algebra - homework in an intelligent tutoring system through machine learning techniques for educati-onal sustainable development* 11(1) (Sustainability) p 105
[15] Education for sustainable development | Higher Education Academy [Internet] 2018 [cited 2018 Nov 16] Available from: https://www.heacademy.ac.uk/knowledge-hub/education-sustainabl-e-development-0