A Novel Student Risk Identification Model using Machine Learning Approach

Nityashree Nadar[1]
Bharathiar University, Coimbatore

Dr.R.Kamatchi[2]
Amity School of Engineering and Technology
Panvel, Mumbai

***Abstract*—This research work aim at addressing issues in detecting student, who are at risk of failing to complete their course. The conceptual design presents a solution for efficient learning in non-existence of data from previous courses, which are generally used for training state-of-art machine learning (ML) based model. The expected scenario usually occurs in scenario when university introduces new courses for academics. For addressing this work, build a novel learning model that builds a ML from data constructed from present course. The proposed model uses data about already submitted task, which further induces the issues of imbalanced data for both training and testing the classification model. The contribution of the proposed model are: the design of training the learning model for detecting risk student utilizing information from present courses, tackling challenges of imbalanced data which is present in both training and testing data, defining the issues as a classification task, and lastly, developing a novel non-linear support vector machine (NL-SVM) classification model. Experiment outcome shows proposed model attain significant outcome when compared with state-of-art model.**

Keywords—Classification; imbalanced data; machine learning; virtual learning environment

I. Introduction

Student dropout is an important problem across various levels such primary school, higher secondary, graduation level and the scenario is much worse in Massive Open Online Courses (MOOCs). As per the research conducted in [1], [2], the number of student not completing graduation in USA is 20% and in Europe it is around 20% to 50% fail to finish their studies on time [3]. For online or distance education, these statistics are even worse with 78% of students not completing the graduation [4]. Further, it gets even worse for student who gets registered with MOOCs, the percentage of student who enrolled and successfully finished the course is only 5% as reported in [5] or 15% as reported in [6]. The issues of identifying student that re expected to fail the course has been extensively analyzed across various research community in recent times [7], [8], [9]. It was also a major subject of the KDD'CUP 2015 competition that mainly aimed on forecasting student withdrawing from online courses.

Establishing student, who are at chance or risk of withdrawing or failing from their respective course, is the initial step towards provisioning them with remedial (material) support. Generally, supportive measures are carried out by instructor/professor, who obtains the information/outcome of forecasting [7], [8]. In other way, the forecasting model may build email messages that communicate directly to the student [10]. The preliminary objective is to enhance the student learning, to keep student engaged in course, and aid them completing the research or study programs.

In distance or online courses, most material are delivered through Virtual Learning Environment (VLE). In VLE each and action are recorder and stored. Along with, student information such as assessment, task results, and demographic information etc, are also kept. These data are cleansed and ML is applied to build a forecasting/predictive model. These model are then used to offers online course provider to forecast student at-risk of completing it on time. A generic way of building a predictive model is to train the models using legacy data from a history or previous task submitted information of the course [8]. Further, it is applied to the present presentation. However, adopting these methods will not be efficient when applied to new type of courses that has no history. For such case, it is important to find new solution.

From extensive survey carried out by MOOCs [11] and Higher Education (HE) courses [8] shows that the highest amount of dropout occurs during first year's courses, and many student dropouts even with a month, first few weeks of the course presentation. The cause may be also due to fee payment toward courses. Therefore, the objectives are to establish or find student who are at-risk of dropping out or failing to complete on time as early as possible. It must also be noted that the same behavior or pattern may not be same across different university/education institution or course design, rapid student dropping out of course may also arise in late stage of course [9].

For overcoming research challenges this work, this work aimed at designing a forecasting model that identify student at-risk of failing or completing on time by presenting a novel non-linear support vector machine (NLSVM) classification model.

The contribution of work is as follows

- Presenting a non-linear enhanced support vector machine classification model for identifying student risk of failure. The NLSVM can be used as both binary classifier as well as multi-level classifier.
- Our model attain good accuracy performance when compared with state-of-art model.
- Experiment outcome shows good performance in terms of ROC, F-measure, and precision and recall.

The paper is organized as follows: In section II, extensive survey is presented. Experimental study are discussed in section III. Finally section IV the paper is concluded and future work of research is described.

II. Litterature Survey

Machine learning technique is composed of supervised, semi-supervised and non-supervised is widely applied and used across various state-of-art models [11], [12], [13], [14], [15], [16], [17], [18], [20], and [21] for identifying risk of student failing to complete course on time. The basic conception is to utilize legacy data to learn the forecasting models and to utilize these approaches to perform forecasting on current courses. The data can aid the course provider who is aiming to address or build policies to enhance the student performance (student retention rate) and student dropping out of courses or failing to finish on time. In [12], the approaches for finding failure or success of student were trained using data of their prior study result. it can be seen that forecasting failure for the first term of courses is very important, since the dropout rate is generally higher but with suitable policies or strategies (help) many student can be saved [19].

Behavior of students [20], [21] in the VLE can be used to construct forecasting models for online courses. These could be just simple summary statistics [15]. When neither the students virtual learning environment activities nor the student prior study results are available, demographic information can be used as the major foundation of information [16].

The proposed learning model is constructed using state-of-art models at the OU [13], [17], [18], [19]. Initially, using decision tree that is trained using data labeling student behavior in the virtual learning environment complemented by the scores of the past assessments/tasks [17]. Further, [18] used demographic features for enriching the input data for training model. The significant discovery in [19] was the prominence of the early establishment or finding of students at risk, even prior to the first task/assessment in the course. The students who do not submit or fail to complete the assessment are very likely to fail or withdraw the entire course. Further, number of approached [7], [22], and [23] for solving problem of classification with presence of imbalanced data in forecasting or identifying student at-risk of failing. However, they neglected student who haven't shown any interest in performing tasks and only focused on active students. For overcoming research challenges this work present a novel non-linear based supervised classification model.

III. Proposed Non Linear Support Vector Machine based Student Risk Identification Learning Model

This manuscript present a novel learning design that use data from running presentation for training forecasting model. The fundamental objectives is to use the information of students who have already completed and submitted the future task and analyze the behavior pattern of find the students who are at risk of failing to submit the assignment. It is assumed that the behavior pattern of student who are about to submit are expected to follow identical behavior pattern as those who already completed the completed and submitted the task similarly, the behavior pattern will different for student who don't complete or submit their task. Number of machine learning based classification model is available to utilize and attain efficient learning model. However, in this work, we present a classification model as a binary classification problem. However, it can work even for solving multi-label classification problem. That is, for a given day (present), which is k days before deadline data, the objective of this work is to build a binary classification algorithm that forecast whether the student will submit the assignment or not on/before time (i.e., within the future k days). If $k = 0$, forecasting are done on the deadline day. Only students that are enrolled in course and haven't finished the task yet are considered for the forecasting.

A. System Model

Let's consider the deadline data and the date when the forecasting is done, which is $k$days prior to the deadline day, as forecasting date. For able to construct a forecasting model for period [forecasting date; deadline date] such that d deadline date is equal to forecasting date. The k forecasting date and d deadline day can be established as a template forecasting and deadline days, respectively.

Here, the deadline is within three days from the present day and we want to forecast if set of student submit their assessment or task either today or within next 5 days. The information for the present day are inaccessible, so the training data will come from the days [presentation initilaizaed+5] = 10 with the labels of submission in [present+4; present + 1] = [9; 6].

It shows the virtual view of the days for training and testing data, day = 0 depicts the present day, negatives keys shows to known information and positive keys to new/unknown data. This aids, that we have more days vacant when applying the forecasting model, some previous/older days cannot be utilized as they were not present in training stage.

B. Long-Term vs Short Term Labelling Window Tradeoff Modelling

Based on system model described, using long-term history means the window sampling for labels is growing. The more days prior to the deadline date, the more days is required for training labels. The condition for the present day being 0 to 5 days prior to the deadline date. For k days prior to the deadline, the size of the window for both training and testing labels will be $k + 1$.

C. Feature Selection for Learning

The data available for efficiently learning is composed of information such as activities and demographics in the virtual learning environment. For extensive analysis this work carried out it can be seen the demographic data is static in nature, it is important to carryout transformation of these information, such as standardization for numerical data and vectorization of categorical data. Similarly, the virtual learning environment data are generally are composed of very rich information such as daily click events clustered by precise action, i.e., student X has viewed 15 times a particular document or presentation research material. All the events/actions are clustered into actions types such as video, resources, blogs, etc. For a given day (present) when the algorithm (model) is learned, the virtual learning environment features are aligned in reverse with respect to time on a particular days, i.e., day 0 is the present day, day 1 is depicted as yesterday and so on. The oldest day utilized for training is the day that the course is initialized. In addition to virtual learning environment daily counts, it’s likely to obtain various statistical information of student behavior pattern in the virtual leaning environment, such as the how long (days) a student is active in the virtual learning environment (i.e., when a person (student) has last accessed or has logged in).

D. Addressing Imbalanced Data Problem in Classification

The ML algorithm are generally modelled to learn objective parameter from data when the classes in the training information are balanced. However, considering real-world environment, the data are generally imbalanced (i.e., some classes data will have significantly less data than other classes). As a result, the state-of-art algorithm [24], [25], [26], [27], and [28] performs very poor in identify probability of risk of failure of student that has been modelled so far [10]. For addressing the problem of imbalanced data the following two stages must be considered such as: **Algorithm stage: -** on-class or linear classification models, cost-sensitive learning, and various kind of ensemble algorithm model are some of the design are generally used. **Data stage:-**by applying sampling window for modifying the class label distribution in such way the training data becomes more balanced. The key functionality of cost-sensitive based learning model is to penalize the cost parameter error on marginal class variable during training stage, which is done by using a cost matrix. However, for attaining fine-grained binary classification model it is better to fix the weight parameter for minority classes (i.e., by considering weight of majority class will be 1). Further, number of approached [7], [22], and [23] for solving problem of classification with presence of imbalanced data in forecasting or identifying student at-risk of failing. However, they neglected student who haven’t shown any interest in performing tasks and only focused on active students.

E. Forecasting Model using Machine Learning Model

For training the learning algorithms and for evaluation of our model, this work conducted survey of various exiting machine learning based classification models such as logistic regression, Naïve Bayes, support vector machine, Tree Boosting XGBoost [29] and so on. Further, number of approach used XGBoost for forecasting student’s dropout as described in KDD-CUP15. However, these model are not efficient when the data is linearly non-separable. As a result, incur degradation in accuracy of classification performance. Further, very few algorithm provision probabilistic forecasting. As this aids in ordering students based on their likeliness to fail, and then use the resources constraint.

For overcoming research challenges, this work present non-linear support vector machine (NLSVM) classification model. The NLSVM first extract features of student behavior considering different assessment which are labeled (Completed: 0, and Failed: 1). If the task or assessment is successfully completed by student the class is labelled as 0 or if fails to submit on time then it is labeled as 1. The NLSVM then trains itself using labelled parameter and obtains support vector among parameters that maximize the distance among varied classes. Lastly, the NLSVM constructs a decision boundary (DB) from the support vectors. If the computed outcome from the DB is different from its known label, the decision is considered as training error. For such cases, this work considers soft-margin support vector machine which can fix boundary even when the datasets are mixed and cannot be disjointed. This work introduced slack parameters to maximizing the margin and reduce training error. For computing support vector using proposed NLSVM model is obtained as follows

$$\min \mathbb{D} \sum_{z=1}^{\mathbb{O}} \mu_z + \frac{1}{2}\langle u_t, u_t \rangle, \tag{1}$$

Such that $\mathbb{U}_z(\langle u_t, z_z \rangle + \mathbb{c}_{\mathfrak{t}}) \geq 1 = \mu_z$ for $z = 1,2,3, \dots, \mathbb{O}$, and $\mu_z \geq 0$

Where $\mathbb{O}$ is the number of vectors, $\mathbb{D}$ is regulation variable, u_t is the weight vector, μ_z is the slack parameter and $<.,.>$ is the inner product function. The $\mathbb{U}_z$ is the z^{th} target parameter, $\mathbb{c}_{\mathfrak{t}}$ is the bias, and z_z is the z^{th} input parameter. The SVM decision boundary $\mathbb{G}_z$is expressed as follows

$$\mathbb{G}_z = \langle u_t^*, z \rangle + \mathbb{c}_{\mathfrak{t}}^* = 0 \tag{2}$$

Where $\mathbb{c}_{\mathfrak{t}}^*$ represent bias and u_t^* represent weight vector, and z is the input feature. By transpiring the z_z and z term to $z_z \rightarrow (z_z)$ and $z \rightarrow (z)$, the non-linear support vector machine can be transformed to linear based support vector machine as follows

$$\mathbb{U}_z(\langle u_t, (z_z) \rangle + \mathbb{c}_{\mathfrak{t}}) \geq 1 \tag{3}$$

Further, to perform classification using non-linear support vector machine, a kernel function $\mathbb{L}_{\mathfrak{t}}(.,.)$, which is a dot-product in the transpired feature vector space as follows

$$\mathbb{L}_{\mathfrak{t}}(z_z, z_{z'}) = \langle (z_z), (z_{z'}) \rangle \tag{4}$$

where$z_{z'} = 1,2, \dots, \mathbb{O}$.

Further, the proposed NLSVM model is evaluated considering various ICT data which is composed of different behavior of different student data considering various task or assessment. Our model accurately classify these imbalanced data compared to state-of-art model which is experimentally shown in next section below.

IV. Experimetal Result and Analysis

This section evaluates performance evaluation of proposed student risk identification leaning model over state-of-art models. For experiment analysis various experiment are considered and date used for experiments are publically available. The experiment is conducted using windows 10 operating system, Intel I-5 class 64 bit processor, 16 GB RAM, 4GB Nvidia CUDA enabled GPU. For experiment analysis this work used publicly available dataset obtained from OULAD [30], [31] which composed of different courses with student enrollment around 1200 to 2500. The objective of this work is to forecast the submission of first assessment of a particular course within deadline time around 20 to 30 days. The course is composed of wide variety of fields such math’s, history, engineering and so on. For completing the course, the student has to attain some minimum scores for a given task or assessment and then pass the final exam. The proposed student risk learning model for forecasting dropout has been aimed at attaining following objectives, Firstly, carryout analysis daily using ML algorithm to evaluate classification model. Secondly, analyses and identify the effects and problems of imbalanced data. Then, compare our proposed model over state-of-art model trained using legacy data. Fourthly, experiment is conducted for different k and courses and evaluate performance attained by proposed model over existing model in terms of precision, recall, F-measure, and ROC. The Table I, shows performance outcome attained by proposed model over existing model in terms of precision, and recall. Further, fig. 1 shows F-measure attained by both proposed and existing model. The overall result attained shows proposed learning model improves F-measure score by 24.45%, 26.65%, and 18.96% over existing learning model. An average improvement of 23.35% is attained by proposed learning model over existing model. Further, experiment are conducted to evaluate ROC performance attained by proposed NL-SVM over exiting SVM and XGBoost as shown in Fig. 2. The outcome shows NL-SVM attain an ROC performance improvement of 35.33%, 25.56% over SVM and XGBoost, respectively.

TABLE I. PERFORMANCE EVALUATION OF PROPOSED MODEL OVER EXISTING CLASSIFICATION MODEL

Top K	Precision SVM	Precision NL-SVM	Recall SVM	Recall NL-SVM
5	0.5751	0.88	0.8093	0.9
10	0.5044	0.89	0.8848	0.9
25	0.6698	0.88	0.8847	0.89

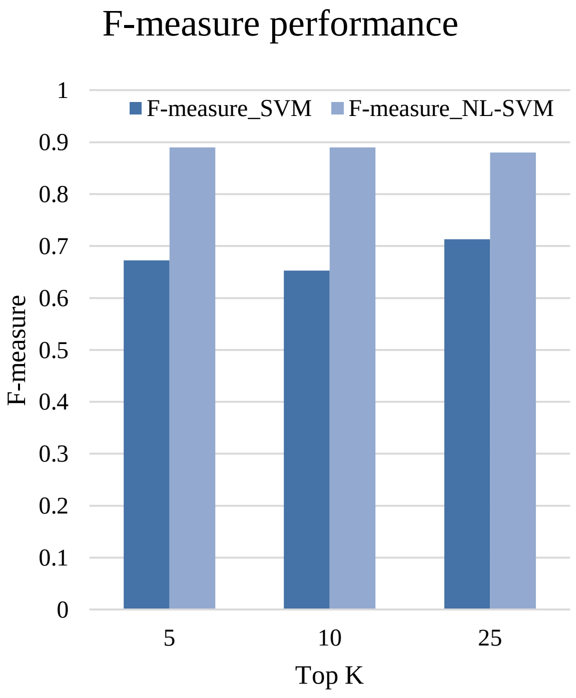

Fig. 1. F-Measure Performance for Different Top K.

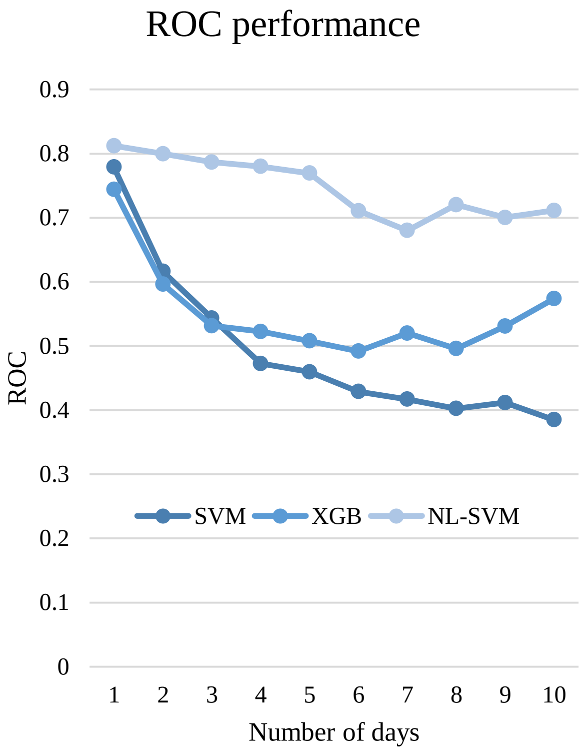

Fig. 2. ROC Performance for Number of Days.

V. CONCLUSION

This manuscript introduced a novel design for early finding of student who are at risk of failing or completing the course on time without using legacy data. The proposed model uses the significance factor of fist task being important factor in the progress of course work. The best way is to extract the student behaviour who already submitted their task and learn its pattern. This work defines the problem as a binary classification task with objective to learn and forecast daily using forecasting window. The proposed model is evaluated using publicly available OULAD dataset. The outcome shows the proposed model can predicts accurately even for early day (i.e. for 0 and 1 days), also predicts efficiently for later days of course completion, and attain better outcome that training using legacy data. From overall experiment analysis, it can be seen feature selection VLE is important for forecasting student at risk of failing. The proposed NL-SVM based classification model attains good recall, precision, and F-measure performance. An average F-measure improvement of 23.35% is attained by proposed learning model over existing model. Further, NL-SVM attains an ROC performance improvement of 35.33%, 25.56% over SVM and XGBoost, respectively. The future work we would consider experiment analysis considering different dataset and also considering building a hybrid model for enhancing forecasting model.

REFERENCES

[1] Peter J. Quinn. Drop-out and completion in higher education in europe among students from under-represented groups. Technical report, European Commission, Oct 2013.

[2] H. Vossensteyn, A. Kottmann, B. Jongbloed, and F. Kaiser. Drop-out and completion in higher education in europe executive summary. Technical report, European Commission, 2015.

[3] G. Kena, J. W. X. R. A. Musu-Gillette, Laurenand Robinson, J. Zhang, S. Wilkinson-Flicker, A. Barmer, and E. D. V. Velez. The condition of education 2015. Technical Report 2015-144, NCES, May 2015.

[4] O. Simpson. 22% - can we do better? In The CWP Retention Literature Review, 47, 2010.

[5] K. Jordan. Mooc completion rates: The data. http://www.katyjordan.com/MOOCproject.html, 2015. Accessed: 2017-10-10.

[6] D. Koller, A. Ng, C. Do, and Z. Chen. Retention and intention in massive open online courses: In depth. EDUCAUSE, http://www.educause.edu/ero/article/retention-and-intention-massive-open-online-courses-depth-0, Jun 2013.

[7] S. M. Jayaprakash, E. W. Moody, E. J. M. Lauria, J. R. Regan, and J. D. Baron. Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative. Journal of Learning Analytics, 1(1):6{47, 2014.

[8] A. Wol_, Z. Zdrahal, D. Herrmannova, J. Kuzilek, and M. Hlosta. Developing predictive models for early detection of at-risk students on distance learning modules. In Machine Learning and Learning Analytics workshop at LAK14, 24-28 March 2014, Indianapolis, Indiana, USA, 4, 2014.

[9] H. He and E. A. Garcia. Learning from imbalanced data. IEEE Trans. on Knowl. and Data Eng., 21(9):1263{1284, Sep 2009.

[10] 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, 60{65, 2014.

[11] C. Taylor, K. Veeramachaneni, and U. O'Reilly. Likely to stop? predicting stopout in massive open online courses. CoRR, abs/1408.3382, 2014.

[12] Huang, S. & Fang, N. Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. Computers & Education, Issue 61, pp. 133-145, 2013

[13] Hlosta, M. et al., 2014. Modelling student online behaviour in a virtual learning environment. Indianapolis, LAK 2014.

[14] Pandey, M. & Sharma, V. K. A Decision Tree Algorithm Pertaining to the Student Performance. Analysis and Prediction. International Journal of Computer Applications, 61(13), pp. 1-5, 2013.

[15] Romero, C., López, M., Luna, J. & Ventura, S., 2013. Predicting students' final performance from participation in on-line discussion forums. Computers & Educaton, Issue 68, pp. 458-472.

[16] Wladis, C., Hachey, A. C. & Conway, K., 2014. An investigation of course-level factors as predictors of online STEM course outcomes. Computers & Education, Issue 77, pp. 145-150.

[17] Wolff, A., Zdrahal, Z., Nikolov, A. & Pantucek, M. Improving retention: predicting at-risk students by analysing behaviour in a virtual learning environment. s.l., LAK 2013.

[18] Wolff, A., Zdrahal, Z., Herrmannova, D. & Knoth, P., 2013. Predicting student performance from combined data sources. In: A. Peña-Ayala, ed. Educational Data Mining: Applications and Trends. Verlag: Springer International Publishing, pp. 175-202, 2013.

[19] Wolff, A. et al., 2014. Developing predictive models for early detection of at-risk students on distance learning modules. Indianapolis, LAK 2014.

[20] Wang, Rui & Chen, Fanglin & Chen, Zhenyu & Li, Tianxing & Harari, Gabriella & Tignor, Stefanie & Zhou, Xia & Ben-Zeev, Dror & T. Campbell, Andrew. StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. UbiComp 2014 - Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing. 10.1145/2632048.2632054, 2014.

[21] Wang, Rui & Chen, Fanglin & Chen, Zhenyu & Li, Tianxing & Harari, Gabriella & Tignor, Stefanie & Zhou, Xia & Ben-Zeev, Dror & T. Campbell, Andrew. StudentLife: Using Smartphones to Assess Mental Health and Academic Performance of College Students. 7-33. 10.1007/978-3-319-51394-2_2, 2017.

[22] J. He, J. Bailey, B. I. Rubinstein, and R. Zhang. Identifying at-risk students in massive open online courses. In AAAI, 1749{1755, 2015.

[23] N. Thai-Nghe, A. Busche, and L. Schmidt-Thieme. Improving academic performance prediction by dealing with class imbalance. In Ninth International Conference on Intelligent Systems Design and Applications, ISDA 2009, Pisa, Italy , November 30-December 2, 2009, 878{883, 2009.

[24] C. Ye and G. Biswas. Early prediction of student dropout and performance in moocs using higher granularity temporal information. Journal of Learning Analytics, 1(3):169{172, 2014.

[25] H. Lakkaraju, E. Aguiar, C. Shan, D. Miller, N. Bhanpuri, R. Ghani, and K. L. Addison. A machine learning framework to identify students at risk of adverse academic outcomes. 1909{1918, 2015.

[26] E. Aguiar, H. Lakkaraju, N. Bhanpuri, D. Miller, B. Yuhas, and K. L. Addison. Who, when, and why: A machine learning approach to prioritizing students at risk of not graduating high school on time. In LAK '15, 93{102, New York, NY, USA, ACM,2015.

[27] J. Bainbridge, J. Melitski, A. Zahradnik, E. Lauria a, S. M. Jayaprakash, and J. Baron. Using Learning Analytics to Predict At-Risk Students in Online Graduate Public Affairs and Administration Education. The JPAE Messenger, 21(2):247{262, 2015.

[28] S. Jiang, M. Warschauer, A. E. Williams, D. ODowd, and K. Schenke. Predicting mooc performance with week 1 behavior. In EDM14, 273{275, 2014.

[29] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. CoRR, abs/1603.02754, 2016.

[30] Kuzilek, M. Hlosta, and Z. Zdrahal. Open university learning analytics dataset. In Data literacy for Learning Analytics workshop at LAK16, 26th April 2016, Edinburgh, UK, 9, 2016.

[31] Kuzilek, Jakub & Hlosta, Martin & Zdráhal, Zdenek. Open University Learning Analytics dataset. Scientific Data. 4. 170171. 10.1038/sdata.2017.171, 2017.