Predicting Performance on MOOC Assessments using Multi-Regression Models

Zhiyun Ren
George Mason University
4400 University Dr,
Fairfax, VA 22030
zen4@masonlive.gmu.edu

Huzefa Rangwala
George Mason University
4400 University Dr,
Fairfax, VA 22030
rangwala@cs.gmu.edu

Aditya Johri
George Mason University
4400 University Dr,
Fairfax, VA 22030
johri@gmu.edu

ABSTRACT

The past few years has seen the rapid growth of data mining approaches for the analysis of data obtained from Massive Open Online Courses (MOOCs). The objectives of this study are to develop approaches to predict the scores a student may achieve on a given grade-related assessment based on information, considered as prior performance or prior activity in the course. We develop a personalized linear multiple regression (PLMR) model to predict the grade for a student, prior to attempting the assessment activity. The developed model is real-time and tracks the participation of a student within a MOOC (via click-stream server logs) and predicts the performance of a student on the next assessment within the course offering. We perform a comprehensive set of experiments on data obtained from two openEdX MOOCs via a Stanford University initiative. Our experimental results show the promise of the proposed approach in comparison to baseline approaches and also helps in identification of key features that are associated with the study habits and learning behaviors of students.

Keywords

Personalized Linear Multi-Regression Models, MOOC, Performance prediction

1. INTRODUCTION

Since their inception, Massive Open Online Courses (MOOCs) have aimed at delivering online learning on a wide variety of topics to a large number of participants across the world. Due to the low cost (most times zero) and lack of entry barriers (e.g., prerequisites or skill requirements) for the participants, large number of students enroll in MOOCs but only a small fraction of them keep themselves engaged in the learning materials and participate in the various activities associated with the course offering such as viewing the video lectures, studying the material, completing the various quizzes and homework-based assessments.

Given, this high attrition rate and potential of MOOCs to deliver low-cost but high quality education, several researchers have analyzed the server logs associated with these MOOCs to determine the factors associated with students dropping out. Several predictive methods have been developed to predict when a participant will drop out from a MOOC [4, 5, 6, 14].Using self reported surveys, studies have determined the different motivations for students enrolling and participating in a MOOC. Participants enroll in a MOOC sometimes to learn a subset of topics within the curriculum, sometimes to earn degree certificates for future career promotion or college credit, social experience or/and exploration of free online education [8]. Students with similar motivation have different learning outcomes from a MOOC based on the number of invested hours, prior education background, knowledge and skills [4].

In this paper, we present models to predict a student's future performance for a certain assessment activity witin a MOOC. Specifically, we develop an approach based on personalized linear multi-regression (PLMR) to predict the performance of a student as they attempt various graded activities (assessments) within the MOOC. This approach was previously studied within the context of predicting a student's performance based on graded activities within a traditional university course with data extracted from a learning management system (Moodle) [3]. The developed model is real-time and tracks the participation of a student within a MOOC (via click-stream server logs) and predicts the performance of a student on the next assessment within the course offering. Our approach also allows us to capture the varying studying patterns associated with different students, and responsible for their performance. We evaluate our predictive model on two MOOCs offered using the OpenEdX platform and made available for learning analytics research via the Center for Advanced Research through Online Learning at Stanford University [1].

We extract features that seek to identify the learning behavior and study habits for different students. These features capture the various interactions that show engagement, effort, learning and behavior for a given student participating in studying; by viewing the various video and text-based materials available within the MOOC offering coupled with student attempts on graded and non-graded activities like quizzes and homeworks. Our experimental evaluation shows accurate grade prediction for different types of homework as-

[1]datastage.stanford.edu

sessments in comparison to baseline models. Our approach also identifies the features found to be useful for predicting an accurate homework grade.

2. RELATED WORK

Several researchers have focused on the analysis of education data (including MOOCs), in an effort to understand the characteristics of student learning behaviors and motivation within this education model [11]. Brinton et. al. [1] developed an approach to predict if a student answers a question correct on the first attempt via click-stream information and social learning networks. Kennedy et. al. [7] analyzed the relationship between a student's prior knowledge on end-of-MOOC performance. Sunar et. al. [12] developed an approach to predict the possible interactions between peers participating in a MOOC. Elbadrawy et. al. [3] proposed the use of personalized linear multi-regression models to predict student performance in a traditional university by extracting data from course management systems (Moodle). Our study focuses on MOOCs, which presents different assumptions, challenges and features in comparison to a traditional university environment.

Most similar to our proposed work, Pardos et. al. proposed a model "Item Difficulty Effect Model" (IDEM) that incorporates the difficulty levels of different questions and modifies Bayesian Knowledge Tracing (BKT) model [2] by adding an "Item" node to every question node. By identifying the challenges associated with modeling MOOC data, the IDEM approach and extensions that involve splitting questions into several sub-parts and incorporating resource (knowledge) information [9] are considered state-of-the-art MOOC assessment prediction approaches and referred as KT-IDEM. However, this approach can only predict a binary value grade. In contrast, the model proposed in this paper is able to predict both, a continuous and a binary grade.

3. METHODS

3.1 Personal Linear Multi-Regression Models

We train a personalized linear multi-regression (PLMR) model [3] to predict student performance within a MOOC. Specifically, the grade $\hat{g}_{s,a}$ for a student s in an assessment activity a is predicted as follows:

$$\begin{aligned}\hat{g}_{s,a} &= b_s + p_s^t W f_{sa} \\ &= b_s + \sum_{d=1}^{l} (p_{s,d} \sum_{k=1}^{n_F} f_{sa,k} w_{d,k}),\end{aligned} \tag{1}$$

where b_s is bias term for student s, f_{sa} is the feature vector of an interaction between student s and activity a. The features extracted from the MOOC server logs are described in the next Section. n_F is the length of f_{sa}, indicating the dimension of our feature space. l is the number of linear regression models, W is the coefficient matrix of dimensions $l \times n_F$ that holds the coefficients of the l linear regression models, and p_s is a vector of length l that holds the memberships of student s within the l different regression models [3]. Using lasso [13], we solve the following optimization problem:

$$\underset{(W,P,B)}{minimize}\ L(W, P, B) + \gamma(\|P\|_F + \|W\|_F), \tag{2}$$

where W, P and B denote the feature weights, student memberships and bias terms, respectively. The loss function $L(\cdot)$ is the least square loss for regression problems. $\gamma(\|P\|_F + \|W\|_F)$ is a regularizer that controls the model complexity by controling the values of feature weights and student memberships. Tuning the scalar γ prevents model from over-fitting.

3.2 Feature Description

We extract features from MOOC server logs and formulate the PLMR model to predict real-time assessment grade for a given student. Figure 1 shows the various activities, generally available within a MOOC. Fig 1 (a) shows that each homework has corresponding quizzes, each of which has its corresponding video as resources for learning. Fig 1 (b) shows that while watching a video, a student can have a series of actions. Fig 1 (c) shows that while studying using a MOOC, a student can have several login sessions. In order to capture the latent information behind the click-stream for each student, we extract six types of features: (i) session features, (ii) quiz related features, (iii) video related features, (iv) homework related features, (v) time related features and (vi) interval-based features. These features constitute the feature vector f_{sa} for a student and a homework assessment. The description of these features are as follows:

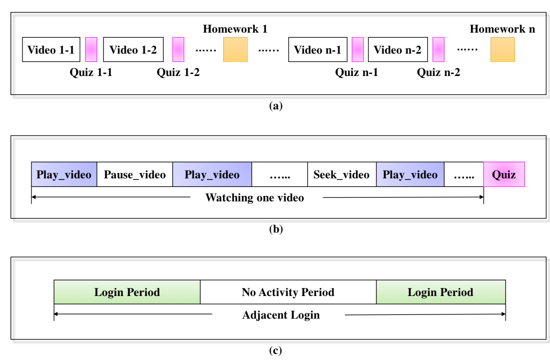

Figure 1: Different activities within a MOOC.

(i) Session features:.

A single study session is defined by a student login combined with the various available study interactions that a student may partake in. Since, students do not always log out of a session, we assume that a "no activity" period of more than one hour constitutes a student logging out of a session. We show a "no activity" period for a student between two consecutive sessions in Fig 1 (c).

- **NumSession** is the the average number of daily study sessions a student engages in, before a homework attempt.
- **AvgSessionLen** is the average length of each session in minutes. We calculate the average study time of a study session by

$$AvgSessionLen = \frac{Total\ study\ time}{NumSession}. \tag{3}$$

- **AvgNumLogin**. Students are free to choose when to login and study in a MOOC environment. We consider a day as a "work day" if a student logs into the study system; and a day as "rest day" if a student does not. The rate of "work" and "rest" can capture a student's learning habits and engagement characteristics.

$$AvgNumLogin = \frac{\#\ of \text{“}work\ day\text{”}}{\#\ of\ \text{“}work\ day\text{”}\ + \#\ of\ \text{“}rest\ day\text{”}}. \tag{4}$$

(ii) Quiz Related features:

- **NumQuiz** is the number of quizzes a student takes before a homework attempt. This feature reflects the student's dedication towards the course material and a factor towards performance in a homework.
- **AvgQuiz** is the average number of attempts for each quiz. The MOOCs studied in this paper allow unlimited attempts on a quiz.

(iii) Video Related features:

- **VideoNum** denotes the number of distinct video sessions for a student before a homework attempt.
- **VideoNumPause** is the average number of pause actions per video. There are several actions associated with viewing videos, including "pause video", "play video", "seek video" and "load video". Tracking these actions allows for capturing a student's focus level and learning habits.
- **VideoViewTime** is the total video viewing time.
- **VideoPctWatch**. In a large amount of cases, students do not finish watching a full video. As such, we calculate the average percentage of the watched part of a video.

(iv) Homework Related features:

- **HWProblemSave** is the average number of "save answer" actions for each homework assessment. Before submitting answers for a homework, students are allowed to save their answer sheet and check as many times as they need. This feature is more valuable when the MOOC provides only one chance for a homework answer submission.

(v) Time Related features:

- **TimeHwQuiz** is the time between a homework answer submission and the last quiz attempt.
- **TimeHwVideo** is the time between a homework answer submission and the last video watching activity.
- **TimePlayVideo** is the percentage of study sessions with video watching activity over all the study sessions.
- **HwSessions** is the number of sessions that have homework related activities (save and submit).

(vi) Interval-Based features:.

It is expected that there will be some changes in study activities once the students know the former homework's grade. They may study harder if they don't get a satisfactory score. The interval-based features are aiming to represent different activities between two consecutive homeworks.

- **IntervalNumQuiz**: denotes the number of quizzes the student takes between two homeworks.
- **IntervalQuizAttempt**: is the average number of quiz attempts between two homeworks.
- **IntervalVideo**: is the number of videos a student watches between two homeworks.
- **IntervalDailySession**: is the average number of sessions per day between two homeworks.
- **IntervalLogin**: is the percentage of login days between two homeworks.

We also use the cumulative grade (so-far) on quizzes and homeworks for a student as a feature and denote it by **Meanscore**. For our baseline approach we only consider the averages computed on the previous homeworks.

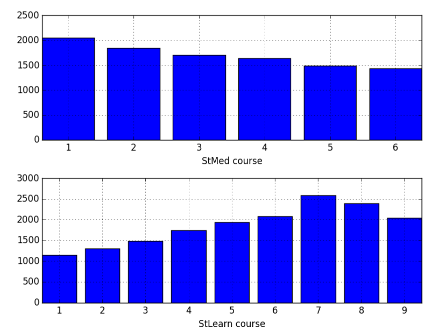

Figure 2: Distribution of students attempting each Assessment. StMed and StLearn had 6 and 9 assessments, respectively.

4. EXPERIMENTS

4.1 Datasets

We evaluated our methods on two MOOCs: "Statistics in Medicine" (represented as StMed in this paper) taught in Summer 2014 and "Statistical Learning" (represented as StLearn in this paper) taught in Winter 2015.

StMed: This dataset includes server logs tracking information about a student viewing video lectures, checking text/web articles, attempting quizzes and homeworks (which are graded). Specifically, this MOOC contains 9 learning units with 111 assessments, including 79 quizzes, 6 homeworks and 26 single questions. The course had 13,130 students enrolled, among which 4337 students submitted at least one assignment (quiz or homework) and had corresponding scores, 1262 students have completed part of the six homeworks and 1099 students have attempted all the homeworks. 193 students attempted all the 79 quizzes and six homeworks. This course had 131 videos and 6481 students had video related activity.

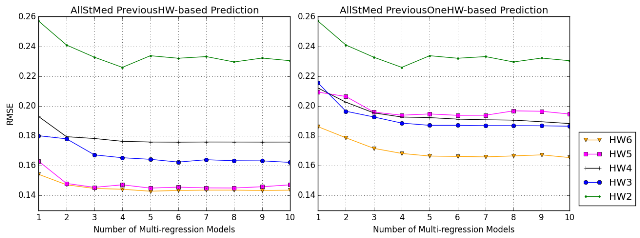

Figure 3: AllStMed Prediction Results. RMSE (↓ is better).

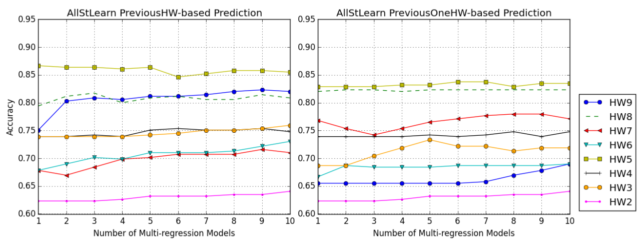

Figure 4: AllStLearn Prediction Results. Accuracy (↑ is better).

StLearn: This course had ten units. Except the first one, all units have quizzes and end of unit homeworks, which add up to 103 assessments in total. 52,821 students enrolled in this course, and 4987 students had assessment activities, 3509 students attempted a subsets of the available homeworks while 346 students attempted all the 9 homeworks, and 118 students attempted all the 103 assessments. The key difference between the homeworks in the StLearn in comparison to the StMed is that homeworks have only one question which a student can either get correct or incorrect. As such, scoring in this MOOC is binary instead of continuous. To predict whether a student answers a question correctly, we reformulate the regression problem as a classification problem using a logistic loss function. Figure 2 shows the distribution of students attempting the different assessments available across the two MOOCs studied here.

4.2 Experimental Protocol

In order to gain a deep insight of students' performance in a MOOC, we perform two types of experiments. Given n, homework assessments represented as $\{H_1, \ldots, H_n\}$ our objective is to predict the score a student achieves in each of the n homeworks. Depicting the most realistic setting, for the i-th homework, H_i we define the training set as all homework and student pairs who attempt and have a score for all homeworks up to the H_{i-1}. For predicting the score for H_i for a given student, we use all the features extracted just before attempting the target homework H_i. We refer to this as **PreviousHW-based Prediction**. Secondly, for the predicting i-th homework H_i's score, we use training data of student-homework pairs restricted from only the previous one homework i.e., H_{i-1}. This experiment is referred by **PreviousOneHW-based Prediction**. Note, in these cases we cannot make any prediction for the first homework (H_1) since, we do not have any training information for a given student.

4.3 Data Partition

We partition the students for StLearn and StMed into two groups: the group of students who attempt *all* the requested homeworks, and the group of students who finish *few* of the homeworks. This allows us to consider the different motivations and expectations of students enrolling in a MOOC. For example, the students who aim to learn in a MOOC may choose watching videos over taking all homeworks. While, the students who want to achieve a degree certificate may focus on the homework completeness. We refer to the first group by "Partial homeworks accomplished group", and the second group by "All homeworks accomplished group". We evaluate our models on the two groups for the **AllStMed** and **AllStLearn** datasets. Specifically, we name the four group of students as **AllStMed**, **AllStLearn**, **PartialStMed** and **PartialStLearn** based on their group and MOOC class.

HW#	PLMR	Meanscore
2	**0.230**	0.248
3	**0.162**	0.176
4	**0.176**	0.196
5	**0.144**	0.156
6	**0.143**	0.150
Avg	**0.171**	0.185

Table 1: PreviousHW-based RMSE Performance (RMSE) comparison for AllStMed.

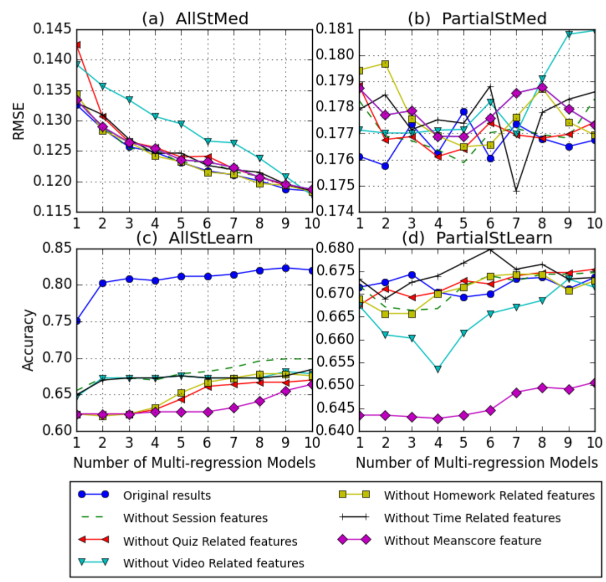

Figure 5: Predictive Performance with Removal of Feature Types.

4.4 Evaluation Metrics

StMed course has continuous scores for a homework, which are scaled between 0 and 1. However, the homework score is binary in the StLearn course, indicating whether the student answers a question correctly or incorrectly. For StLearn, we use a logistic loss and formulate a classification problem

HW#	Accuracy (↑)			F_1 (↑)		
	PLMR	Baseline		PLMR	Baseline	
		Meanscore	KT-IDEM		Meanscore	KT-IDEM
2	0.641	**0.646**	0.623	0.775	**0.777**	0.768
3	**0.760**	0.580	0.681	**0.821**	0.805	0.810
4	**0.754**	0.710	0.739	0.838	0.706	**0.850**
5	**0.867**	0.809	0.829	**0.920**	0.880	0.906
6	**0.730**	0.678	0.667	**0.808**	0.776	0.800
7	0.716	0.675	**0.730**	**0.887**	0.878	0.844
8	**0.817**	0.762	0.817	**0.903**	0.849	0.886
9	**0.823**	0.794	0.777	**0.864**	0.856	0.853
Avg	**0.764**	0.707	0.759	**0.852**	0.816	0.848

Table 2: PreviousHW-based prediction performance comparison for AllStLearn group.

instead of the regression problem as done for the StMed course. To evaluate the performance of our approach, we use the root mean squared error (RMSE) as the metric of choice for regression problem. For classification problem, we use accuracy and the F1-score (harmonic mean of precision and recall), known to be a suitable metric for imbalanced datasets.

4.5 Comparative Approaches.

In this work, we compare the performance of our proposed methods with two different competitive baseline approaches.

(i) **Average grade of the previous homeworks.** We calculate the mean score of a given student's previous homeworks to predict their future performance and is denoted as Meanscore. We use this method to compare our prediction results on StMed.

(ii) **KT-IDEM [10].** KT-IDEM is a modified version of original BKT model. By adding an "item" node to every question node, the model is able to identify different difficulty levels of each question. Since this model can only predict a binary value grade, we use this model to compare our prediction results on StLearn.

5. RESULTS AND DISCUSSION

5.1 Assessment Prediction Results

Figures 3 and 4 show the prediction results with varying number of regression models for the AllStMed and AllStLearn MOOCs, respectively. Analyzing Figure 3 we observe that as the number of regression models increases the RMSE metric goes lower and use of five models seems to be good choice for all the different homeworks. Comparing the PreviousHW- and PreviousOneHW-based results, we notice that predictions for all the homeworks (HW3, HW4, HW5, and HW6) benefits from using all the available training data prior to those homeworks i.e., to predict grade for H_i it is better to use training information extracted from $H_1 \dots H_{i-1}$ rather than just H_{i-1}. Similar observations can be made while analyzing the prediction results for the AllStLearn cohort which includes nine homework correct/incorrect binary assessments. Figure 4 shows the accuracy scores (higher is better) for the three experiments. For the PreviousOneHW- and PreviousHW-based experiments HW5 shows the best prediction results. This suggests that in the middle of a MOOC, students tend to have stable study activities and the performance is more predictable than other phases. Also, some homeworks thrive well with just using training data from the previous homework (PreviousOneHW-based, e.g. HW3).

5.1.1 Comparative Performance

Table 1 shows the comparison between baseline approach (Meanscore) and the predictive model for the PreviousHW-based experiments for the AllStMed group. We cannot report results for the KT-IDEM model since, it solves the binary classification problem only. Table 2 shows the comparison of the accuracy and F1 scores of the AllStLearn groups with baseline approaches. We notice that for predicting the second homework, which only uses the information from HW1, the predictive model is not as good as the mean baseline, which reflects that under the situation of lack of necessary amount of information, linear regression models cannot always outperform the baseline. But as the dataset gets larger, our approach outperforms the baseline due to the availability of more training data. From Table 2, we also notice for some homework, KT-IDEM has better performance than PLMR (HW7 and HW4). This could be due to unstable academic activities during these two study periods, which can effect the performance of PLMR.

5.1.2 Feature Importance

We test the effect of each feature set in predicting the assessment scores by training the models under the absence of each feature group. For the StLearn course, since there is no limit on homework attempts, we do not add Interval-Based feature groups to the predictive model. Figure 5 shows the comparison of each prediction result for AllStMed, PartialStMed, AllStLearn and PartialStLearn cohorts. Analyzing these results we observe that for the StLearn MOOC, meanscore is a significant feature and removing it leads to a substantial decrease in accuracy for both All and Partial-cohorts. For the AllStMed, the removal of video related features leads to the most decrease in performance (i.e., increased RMSE). This suggests that features related to the video watching are crucial for predicting the final homework scores. For the PartialStMed, the use of all feature types or a subset does not show a clear winner. This could be due to the varying characteristics of students within these group.

Another way to analyze feature importance is to exclude the influence of the dominant feature, which is meanscore in our study. The evaluation formula of the importance of the i_{th} feature (excluding meanscore feature) is as follows:

$$I_i = \frac{1}{N}\sum_{n=1}^{N}\frac{\sum_{d=1}^{l}|p_{n_S,d}f_{n_S,i}w_{d,i}|}{\sum_{d=1}^{l}|p_{n_S,d}\sum_{k=1}^{n_F}f_{n_S,k}w_{d,k}|}, \quad (5)$$

where N is number of test samples, n_S is the student number corresponding to the n_{th} test sample. $f_{n_S,i}$ is the feature value of an interaction between student n_S and activity i. n_F is the number of features. l is the number of linear regression models. $w_{d,i}$ is the coefficient of d_{th} linear regression model with i_{th} feature, and $p_{n_S,d}$ is the membership of student n_S with the d_{th} regression model. We calculate each feature's importance by calculating the percentage contribution of each feature to the overall grade prediction. Figure 6 shows the feature importance on the AllStMed group, excluding Meanscore feature. We can see **NumQuiz** and **VideoPctWatch** are the most important for AllStMed group besides **Meanscore** feature.

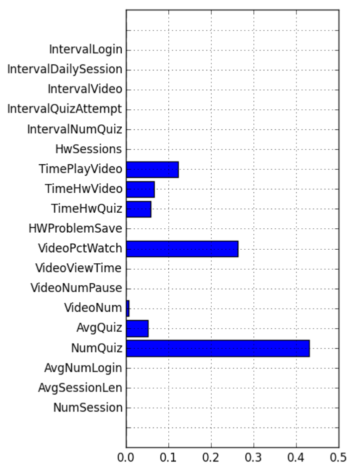

Figure 6: Feature importance for AllStMed.

6. CONCLUSION AND FUTURE WORK

In this work we formulated a personalized multiple linear regression model to predict the homework grades for a student enrolled and participating within a MOOC. Our contributions include engineering features that capture a student's studying behavior and learning habits, derived solely from the server logs of MOOCs. We evaluated our framework on two OpenEdX MOOC courses provided by an initiative at Stanford University. Our experimental evaluation shows improved performance in terms of prediction of real time homework scores compared to baseline methods. We also studied on different groups of student participants due to their motivation. Features associated with engagement (logging multiple times), studying materials (viewing videos and attempting quizzes) were found to be important along with prior homework scores for this prediction problem.

7. ACKNOWLEDGEMENTS

Funding was provided by NSF Grant, 1447489.

8. REFERENCES

[1] Christopher G Brinton and Mung Chiang. Mooc performance prediction via clickstream data and social learning networks. *To appear, 34th IEEE INFOCOM. IEEE*, 2015.

[2] Albert T Corbett and John R Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4):253–278, 1994.

[3] Asmaa Elbadrawy, Scott Studham, and George Karypis. Personalized multi-regression models for predicting students performance in course activities. *UMN CS 14-011*, 2014.

[4] Jeffrey A Greene, Christopher A Oswald, and Jeffrey Pomerantz. Predictors of retention and achievement in a massive open online course. *American Educational Research Journal*, page 0002831215584621, 2015.

[5] Glyn Hughes and Chelsea Dobbins. The utilization of data analysis techniques in predicting student performance in massive open online courses (moocs). *Research and Practice in Technology Enhanced Learning*, 10(1):1–18, 2015.

[6] Suhang Jiang, Adrienne Williams, Katerina Schenke, Mark Warschauer, and Diane O'dowd. Predicting mooc performance with week 1 behavior. In *Educational Data Mining 2014*, 2014.

[7] Gregor Kennedy, Carleton Coffrin, Paula de Barba, and Linda Corrin. Predicting success: how learners' prior knowledge, skills and activities predict mooc performance. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, pages 136–140. ACM, 2015.

[8] Daniel FO Onah and Jane Sinclair. Learners expectations and motivations using content analysis in a mooc. In *EdMedia 2015-World Conference on Educational Media and Technology*, volume 2015, pages 185–194. Association for the Advancement of Computing in Education (AACE), 2015.

[9] Zachary Pardos, Yoav Bergner, Daniel Seaton, and David Pritchard. Adapting bayesian knowledge tracing to a massive open online course in edx. In *Educational Data Mining 2013*, 2013.

[10] Zachary A Pardos and Neil T Heffernan. Kt-idem: Introducing item difficulty to the knowledge tracing model. In *User Modeling, Adaption and Personalization*, pages 243–254. Springer, 2011.

[11] Alejandro Peña-Ayala. Educational data mining: A survey and a data mining-based analysis of recent works. *Expert systems with applications*, 41(4):1432–1462, 2014.

[12] Ayse Saliha Sunar, Nor Aniza Abdullah, Susan White, and Hugh C Davis. Analysing and predicting recurrent interactions among learners during online discussions in a mooc. *Proceedings of the 11th International Conference on Knowledge Management*, 2015.

[13] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[14] Jacob Whitehill, Joseph Jay Williams, Glenn Lopez, Cody Austun Coleman, and Justin Reich. Beyond prediction: First steps toward automatic intervention in mooc student stopout. *Available at SSRN 2611750*, 2015.