ISBN 978-93-86878-06-9
11th International Conference on Data Mining, Computers, Communication and Industrial Applications (DMCCIA-2017)
Kuala Lumpur (Malaysia) Dec. 14-15, 2017

Predicting Student's Academic Performance in a MOOC Environment

Rahila Umer[1] , Teo Susnjak[1] , Anuradha Mathrani[1] and Suriadi Suriadi[2]

[1]Institute of Natural and Mathematical Science, Massey University, Auckland, New Zealand.

[2]Queensland University of Technology, Brisbane.

(r.umer@massey.ac.nz, T.Susnjak@massey.ac.nz, A.S.Mathrani@massey.ac.nz, s.suriadi@qut.edu.au)

Abstract**:** *Massive open online courses (MOOCs) provide an opportunity for students to register for courses offered by best universities around the world. There are massive enrollments in the courses offered in MOOCs; however they suffer from low completion rate and low student retention. To address these problems, it is necessary to make early prediction of students' academic performance to enable targeted and timely interventions for those students at risk of non-completion. This study proposes the use of machine learning algorithms to predict students' academic performance in a MOOCs environment and compared the predictive power of four machine learning algorithms: Logistic Regression, Naïve Bayes, Random Forest and K Nearest Neighbor. Our results show that performance of predictive models are promising and can be used for early prediction of students that are likely to fail in a course. Furthermore, Random forest classifier outperforms other classifiers in a statistical significant way.*

Keywords: *Learning analytics, MOOCs, Machine learning, Data mining, Prediction*

1. Introduction

Massive open on-line courses (MOOCs) provide an opportunity for students to register for courses offered by best universities around the world. Students enrolled in the courses have diverse goals and motivations. Anyone with access to the Internet can register for any high quality and advanced courses offered. MOOCs comprise of video lectures, quizzes, reading resources and provide a forum platform where students can do productive discussion and engage with peers anytime and from anywhere.

All activities of student's are recorded in the MOOCs environment, opening up an opportunity to analyse the large amount of data to gain evidence-based understanding of the behaviour of students in such an environment. Despite the large number of enrolments, student retention in MOOCs is low, often less than 20% [1] and is therefore heavily criticized. In order to increase the retention rate, one could exploit the massive amount of data to predict the likelihood of dropout or failure of the course, thus enabling effective early intervention strategy for those students who are struggling during the course by offering relevant and targeted help.

Learning Analytics (LA) and Educational Data Mining (EDM) are two approaches that focus on data-driven techniques to inform research on practices to mitigate issues like drop-outs, failure rates and low retention rates. Students' digital traces which they leave behind as a result of their online interactions, such as clicks, page visited, and video watched, are recorded in log history during the course [2]. Campbell and Oblinger [3] presented a 5-step method (including captures, report, predict, act and refine) as the main theme in LA. Once, data related to students' interactions with a course is captured and reported, they can be analysed to make some predictions about students performances, informing the design of the subsequent pedagogical interventions (act).

Several works [4-10] have suggested machine learning techniques as the way forward to extract features for addressing student drop-out problems and enabling the predictions of academic failure among students. Appropriate pedagogical actions can then be applied by educators to support students.

In this study, we focus on predicting students' academic performances through the online traces they leave while pursuing a course in a MOOC environment. Machine learning algorithms have been applied to the trace data for each student as they progressed through a course to help predict which students are at risk of failure. The identification of such students would then enable educators to carry out various forms of early intervention, or provide additional support to the students who are struggling.

The study is guided by the research question: "Which machine learning techniques are effective in predicting a student's likelihood to fail a course with high accuracy". In this study, some of the most widely used classifiers [11], namely Random Forest [12], Logistic Regression, Naive Bayes and K-Nearest Neighbour [13] are applied and compared to answer the posed question.

The remainder of the paper is organized as follows. In section 2, we present related work in similar fields. Section 3 discusses data sources. Section 4 describes the methods applied in our experiments. In section 5 we present answers to the research question and discuss experimental results. Finally in section 6, we make conclusion and propose future direction.

2. Related Work

Educational institutions are keen to undertake preventive actions to tackle above mentioned issues. Several researchers have used data mining methods to help in detection and prediction of students likely to fail in some given course using data-driven approaches. Most of these studies have used student academic data in combination with data obtained from other sources (e.g., field data from surveys), all of which requires extra effort to collect.

For instance, Marquez et al. [14] applied white box classifiers, such as induction rules and decision tree on a primary data set collected through surveys and interviews from middle or secondary school students in order to predict academic failures. The data collected include information about the students' background and academic information. The study analysed effects of data pre-processing approaches giving promising results for making predictions on academic performance. However, the data gathering procedure is laborious and time consuming. Moreover survey data is not temporal and does not give a chronological order of the student's learning activities.

Costaet al.[15] compared different EDM techniques to predict the students that are likely to fail in a programming course. The significance of this study is that the techniques used could predict the student's performance at early stages, which allowed educators to plan some intervention strategies which could be made to help students. This study also investigated the impact of data pre-processing methods and algorithms fine-tuning tasks on prediction results. Their results indicated Support vector machine [16] outperformed other classifiers and through data pre-processing, accuracy was further improved.

Khobragadeet al. [17] used Naïve Bayes and white box classifiers like induction rules and decision tree for prediction of student's likelihood of failure. Classification was on data about social, academic and background information of the student, all of which has been collected through surveys. Feature selection algorithms were also used. Naive Bayes provided best accuracy of above 87%. However, since the data was gathered through surveys, the process was time consuming and involved methods for making overall predictions rather than early predictions. The overall predictions are not specific and so do not allow for interventions to be made for supporting students early on into the course. The work proposed in [18] by Ahmadet al, presents an approach where EDM techniques are used to predict academic performance of first year students in computer science course. EDM techniques used are Decision tree, Naive Bayes and rule based classification. The data used during the course of study includes demographic data, previous academic records and other family related information. Rule based classifier outperformed other methods and provide prediction accuracy of 71%.

Ye and Biswas [19] have used extended standard features for MOOCs analysis with higher granularity to make more accurate predictions for dropout and performance. Analysis was made using data collected from video lectures, weekly quizzes, and peer assessments from the ten-week course. Standard features were extended using some detailed temporal features, like, when some assessment was started during the week, or when the first lecture was viewed. Findings compared with existing studies showed that these features improved the prediction accuracy. The time when a student starts the peer assessment assignment was found to be a good predictor. Once the peer assessment score was available, the prediction performance improved. Analysis shows that the students who watched video and did not take quizzes were the ones who mostly dropped out. Overall results show that more precise temporal features and more quantitative information improved early prediction accuracies and false alarm rates as compared to using only assessment score features.

3. Data Sources

In this study data from the Coursera MOOC platform for the course "Principles of Economics" offered in summer 2015 has been analysed. The dataset comprises anonymized student information, assessment details, student grades achieved, time stamps of all student interactions performed during the course, video lecture related student activities and few aspects of participant's demographic information. The course duration was of eight weeks. Dataset consists of more than 3000 students data out of which majority of the students had failed the course. Students with final score greater than 0.5 were considered passed. Following features were used in final dataset (Table 1).

TABLE I: Features Obtained from MOOCs Dataset

S. No	Feature	Explanation
1	Average Score in Weekly quiz	Average score in quizzes of particular week
2	Number of Quizzes attempted	Average attempt for quiz in particular week
3	Quiz lag	Duration between first and last activity of quiz
4	Lecture Lag	Duration between first and last activity of Lecture
5	Total lecture attended	Total lectures attended in particular week
6	Video Activity Count	Activity counts during video lecture (pause, play, stop etc.)
7	Efforts in seconds	Total time spent in a particular week

4. Experimental Design

The objective of the study is to use the machine learning algorithms for early identification of students, who are most likely to fail the course or are at risk of not completing the course. For classification we used methods that are widely used in the field of education and are well suited for imbalanced dataset. Machine learning algorithms used in experiments are: Naïve Bayes (NB), Random Forest (RF), Logistic Regression (LR) and K Nearest Neighbour (KNN). Table 2 shows the parameters used for classifiers.

TABLE II: Machine Learning Parameters

S.No	Classifier	Training Setting	Implementation Source
1	KNN	K=3	scikit-learn [20]
2	Random Forest	estimator=10	scikit-learn[20]
3	Naïve Bayes	Gaussian default setting	scikit-learn[20]
4	Logistic Regression	default settings	scikit-learn[20]

4.1 Evaluation Measures

Machine learning algorithm's performance is compared using F1-score, which is widely used in binary classification problems. Overall accuracy might be misleading, due to the imbalanced nature of dataset. F1 score is the harmonic mean between Precision and Recall.

$$\text{Precision} = TP/ (TP+FP) \quad (1)$$

$$\text{Recall}=TP/ (TP+FN) \quad (2)$$

True Positive is the number of positive instances correctly classified as positive.

False Positive is the number of negative instances incorrectly classified as positive.

False Negative is the number of positive instances incorrectly classified as negative.

$$F1 - \text{Score} = 2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall}) \quad (3)$$

Training Procedure

To estimate the generalization capability of the model for future dataset, 10-fold cross validation technique was used. Performance of the classification methods are then evaluated using F1-score. These classification methods are used for prediction of student's final score and classify it into two classes: Pass or Fail. Prediction was based on grades of students in assignments, quizzes, and time spent on various activities. Prediction was performed weekly based on the available data.

5. Experimental Results

This section first presents the experimental results. We answer the research question in the light of results of analysis. Prediction was performed using four machine learning methods. The result of prediction of student's academic performance is depicted in table 3. As the course progressed from week 1 to week 8, more data on student related activities could be availed upon to extract insights, which helped improve the prediction accuracy considerably. Results show that as time passes and more data become available, prediction accuracy improve.

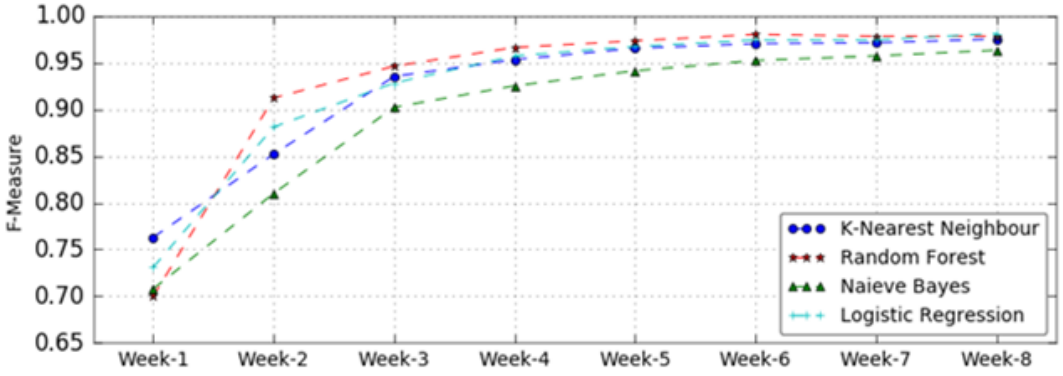

Fig. 1: Comparative Results of the Effectiveness of Machine Learning Algorithms

TABLE III: Comparative Results of the Effectiveness of Machine Learning Algorithms

Week-No	***LR***	***KNN***	***NB***	***RF***
Week-1	0.731	0.763	0.708	0.700
Week-2	0.881	0.852	0.810	0.912
Week-3	0.928	0.935	0.902	0.946
Week-4	0.957	0.953	0.925	0.966
Week-5	0.967	0.965	0.941	0.973
Week-6	0.974	0.970	0.952	0.980
Week-7	0.974	0.971	0.957	0.978
Week-8	0.981	0.975	0.963	0.978
Ranks	2.0	2.6	3.8	**1.5**

Results show that maximum F1-score obtained after week-1 is 0.76 by KNN. After week-1 accuracy increased for all classifiers and maximum F1-score reached to 0.90 for Random Forest classifier. After week-2 there is continuous growth in F1-score for all classifiers. However, growth slows after week-4. The reason for growth of F1-score prior to week-4 is due to availability of more data and for slow growth after week-4 is that majority of the students were not active after week-4. In a MOOC environment, it is often observed that students are active in the beginning of the course and later become less active or withdraw from the course. Maximum F1-score obtained is 0.98 using Random forest classifier after week-6, which then drops to 0.97 after week-7 and week-8.

The success of an early intervention pedagogical strategy depends on the accurate prediction of students who are identified to be at risk as early as possible ideally before the mid of the course. That way timely intervention can be made to help these students. Given this context, the performance of the models looks promising, whereby all classifiers managed to get >0.90 F1-score at Week-4 (mid-way into the course), while the maximum F1-score reached is 0.96 by Random forest classifier. Random forest classifier outperforms all classifiers.

In order to see the significance of above findings, we used Friedman test [21] methodology for comparison of multiple classifiers over multiple datasets. Friedman test is a non-parametric test which is used to compare observations tested on same subjects. Chi-square with k-1 degree of freedom is the test statistic for the Friedman's test; where k is the number of repeated measures. When the p-value is small (<0.05), null hypothesis is rejected. Following is our null hypothesis. H0: "there is no difference among the performance of multiple classifiers." After applying Friedman test p-value obtained for the above result is (p=0.0003) which is less than 0.05, hence null hypothesis is rejected. Therefore, we can conclude that there is significance difference between performances of the classifiers. The calculation of mean ranks of classifier (from highest to lowest), shows that the Random Forest scored highest rank and outperformed other classifiers.

6. Conclusions

The objective of this study was to predict student's academic performance through the digital traces they leave during the course. Machine learning algorithms have been used to predict student performance weekly using the data available. We conducted a comparative analysis of four machine learning algorithms: Logistic Regression, Naïve Bayes, Random Forest and K Nearest Neighbour. The results show that performance of predictive models are promising and can be used for early prediction of students that are likely to fail in a course. Random forest classifier outperforms other classifiers in a statistically significant way. The study informs on machine learning techniques to realize rich analytics and has implications for researchers and educators.

7. References

[1] R. F. Kizilcec, C. Piech, and E. Schneider, "Deconstructing Disengagement : Analyzing Learner Subpopulations in Massive Open Online Courses," Lak '13, p. 10, 2013.

[2] D. Clow, "An overview of learning analytics," Teach. High. Educ., vol. 18, no. 6, pp. 683–695, 2013.

[3] J. P. Campbell, P. B. DeBlois, and D. G. Oblinger, "Academic Analytics," Educ. Rev., vol. 42, no. October, pp. 40–57, 2007.

[4] C. Marquez-Vera, C. R. Morales, and S. V. Soto, "Predicting School Failure and Dropout by Using Data Mining Techniques," IEEE Rev. Iberoam. Tecnol. del Aprendiz., vol. 8, no. 1, pp. 7–14, 2013.

[5] C. Ye and G. Biswas, "Early Prediction of Student Dropout and Performance in MOOCs using Higher Granularity Temporal Information," J. Learn. Anal., vol. 1, no. 3, pp. 169–172, 2014.

[6] J. Bayer, H. Bydzovská, J. Géryk, T. Obsivac, and L. Popelínský, "Predicting drop-out from social behaviour of students," Proc. 5th Int. Conf. Educ. Data Min., no. Dm, pp. 103–109, 2012.

[7] L. M. B. Manhães, S. M. S. da Cruz, and G. Zimbrão, "WAVE: An Architecture for Predicting Dropout in Undergraduate Courses Using EDM," Proc. 29th Annu. ACM Symp. Appl. Comput., pp. 243–247, 2014.

[8] V. R. C. Martinho, C. Nunes, and C. R. Minussi, "Prediction of school dropout risk group using neural network," 2013 Fed. Conf. Comput. Sci. Inf. Syst. FedCSIS 2013, pp. 111–114, 2013.

[9] Simon et al., “Predictors of success in a first programming course,” Proc. 8th Austalian Conf. Comput. Educ. - Vol. 52, pp. 189–196, 2006.

[10] C. Watson, F. W. B. Li, and J. L. Godwin, “Predicting performance in an introductory programming course by logging and analysing student programming behaviour,” Proc. - 2013 IEEE 13th Int. Conf. Adv. Learn. Technol. ICALT 2013, pp. 319–323, 2013.

[11] X. Wu et al., Top 10 algorithms in data mining, vol. 14, no. 1. 2008.

[12] G. G. Moisen, “Classification and Regression Trees,” Encycl. Ecol., no. 2000, pp. 582–588, 2008.

[13] K. Hechenbichler and K. Schliep, “Weighted k-Nearest-Neighbor Techniques and Ordinal Classification,” Mol. Ecol., vol. 399, p. 17, 2004.

[14] C. Marquez-Vera, C. R. Morales, and S. V. Soto, “Predicting School Failure and Dropout by Using Data Mining Techniques,” IEEE Rev. Iberoam. Tecnol. del Aprendiz., vol. 8, no. 1, pp. 7–14, 2013.

[15] E. B. Costa, B. Fonseca, M. A. Santana, F. F. de Araújo, and J. Rego, “Evaluating the effectiveness of educational data mining techniques for early prediction of students’ academic failure in introductory programming courses,” Comput. Human Behav., vol. 73, pp. 247–256, 2017.

[16] C. Cortes and V. Vapnik, “Support-Vector Networks,” Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995.

[17] L. P. Khobragade, “Predicting Students’ Academic Failure Using Data Mining Techniques,” Int. J. Adv. Res. Comput. Sci. Manag. Stud., vol. 3, no. 5, pp. 2321–7782, 2015.

[18] F. Ahmad, N. H. Ismail, and A. A. Aziz, “The prediction of students’ academic performance using classification data mining techniques,” Appl. Math. Sci., vol. 9, no. 129, pp. 6415–6426, 2015.

[19] C. Ye and G. Biswas, “Early Prediction of Student Dropout and Performance in MOOCs using Higher Granularity Temporal Information,” J. Learn. Anal., vol. 1, no. 3, pp. 169–172, 2014.

[20] “scikit-learn user guide,” 2017.

[21] J. Demšar, “Statistical Comparisons of Classifiers over Multiple Data Sets,” J. Mach. Learn. Res., vol. 7, pp. 1–30, 2006.