MOOCs Paid Certification Prediction using Students Discussion Forums

Mohammad Alshehri[1] and Alexandra Cristea[1]

[1] Department of Computer Science, Durham University, Lower Mountjoy, South Rd, Durham DH1 3LE, UK.
{mohammad.a.alshehri,alexandra.i.cristea}@durham.ac.uk

Abstract. Massive Open Online Courses (MOOCs) have been suffering a very level of low course certification (less than 1% of the total number of enrolled students on a given online course opt to purchase its certificate), although MOOC platforms have been offering low-cost knowledge for both learners and content providers. While MOOCs discussion forum rich numeric and textual data are typically utilised to address many MOOCs challenges, e.g., high dropout rate, identifying intervention-needed learners, analysing learners' forum discussion and interaction to predict certification remains limited. Thus, this paper investigates *if MOOC discussion forum-based data can predict learners' purchasing decisions (certification)*. We use a relatively large dataset of 23 runs of 5 FutureLearn MOOCs for temporal (weekly-based) prediction, achieving promising accuracies, 76% on average., in this challenging task across the five courses.

Keywords: MOOCs, Certification Prediction, Discussion Forums.

1 Introduction

Digital learning has been revolutionising and changing the means of modern education. Consequently, Several MOOC platforms have been introduced over the last decade, especially in 2012, coining 2012 as "the year of the MOOCs" [1, 2]. Since then, several platforms, such as FutureLearn, edX, Udemy and Coursera, have introduced free and paid online educational content to the public, targeting learners worldwide [3, 4]. This paper proposes a forum-based predictor of learners' financial decisions (course certificate purchase). Specifically, this paper attempts to answer the following research questions:

- *RQ1: Can MOOC discussion forum data predict course purchase decisions (certification)?*

We use multidisciplinary course data from the less analysed studied platform of FutureLearn to temporally predict financial certification. To the best of our knowledge, our method in predicting MOOC learners' financial decisions (purchasing a course certificate) using learners' discussion forums has never been applied before.

2 Related Work

Looking through the literature, few studies have explored certification in MOOCs. Their used data Sources, number of courses and students as well as the dtat types used varies as explained in **Table 1** below. Data used included Click Stream (CS), Forum Posts (FP), Assignments (ASSGN), Student Information Systems (SIS), Demographics (DEM) and Surveys (SURV).

Table 1. Certification Prediction Models versus our Model.

Ref.	Data Source	#Courses	#Students	Data Description
[5]	Coursera	1	37,933	ASSGN; FP; SIS
[6]	HarvardX	9	79,525	DEM; SURV
[7]	edX	1	43,758	CS
[8]	Coursera	1	84,786	FP
[9]	Coursera; edX	1	65,203	CS; FP
Our Model	FutureLearn	5	245,255	FP

Unlike previous studies on certification, our proposed model aims to predict the financial decisions of learners on whether to purchase the course certificate. Also, our work is applied to a less frequently studied platform, FutureLearn (Table 1). Another key novelty of our study is predicting the learner's actual financial decision on buying the course and gaining a certificate where most current course purchase prediction models identify certification as an automatic consecutive step to the completion, making them not different from completion predictors.

3 Methodology

3.1 Data Collection and Preprocessing

The current study is analysing data extracted from 23 runs spread over 5 MOOC courses, on four distinct topic areas, all delivered through FutureLearn, by the University of Warwick. These courses were delivered repeatedly in consecutive years (2013-2017); thus, we have data on several '*runs'* for each course [10-12]. The Textual Data (student comments) preprocessing involved several essential tasks such as: eliminating irrelevant data generated by organisational administrators, removing unwanted characters, such as HTML/XML, punctuations, non-alphabet characters. The last step contained removing stop-words, lowering the cases of characters, reforming contractions into the original words and grammar correction. Also, learner comments have been classified *positive*, *neutral* and *negative* using MOOCSent sentiment classifier[13].

The current study applied three shallow and one deep classification and regression algorithms to predict MOOC learners' purchasing behaviour: ExtraTree (ET), Logistic Regression (LR), XGBoost (XGB) and Multi-layer Perception (MLP). To deal with our

imbalanced dataset, we used the Balanced Accuracy (BA) to report our results, besides the commonly used metric of accuracy (Acc), which is defined as the average of recall obtained on each class.

4 Results and Discussion

As the courses analysed spanned over different weeks, we have examined the first-week only data, and compared it to the data starting from the first week, until the middle of the course. The results explore how our raw and processed (computed) features can temporally distinguish course purchasers from non-paying learners, based on their discussion forum data.

Table 2. Learner classification results distributed by course at two time points of the course; where class 0 = non-paying learners, class 1 = certificate purchasers.

Course	Classifier	1st Week only			1st - Mid Week		
		Rec_0	**Rec_1**	**BA**	**Rec_0**	**Rec_1**	**BA**
BIM	*ET*	0.82	0.63	**0.73**	0.88	0.75	**0.82**
	LR	0.87	0.57	0.72	0.89	0.63	0.76
	XGB	0.97	0.03	0.50	0.86	0.30	0.58
	MLP	0.81	0.61	0.71	0.82	0.71	0.77
BD	*ET*	0.83	0.53	0.68	0.94	0.57	0.76
	LR	0.80	0.57	0.69	0.88	0.66	**0.77**
	XGB	0.99	0.04	0.52	0.91	0.57	0.74
	MLP	0.83	0.59	**0.71**	0.92	0.61	**0.77**
SC	*ET*	0.83	0.50	0.67	0.95	0.8	0.88
	LR	0.84	0.53	0.69	0.90	0.67	**0.79**
	XGB	0.98	0.04	0.51	0.94	0.50	0.72
	MLP	0.83	0.57	**0.70**	0.93	0.60	0.77
SP	*ET*	0.81	0.60	**0.71**	0.91	0.64	0.78
	LR	0.82	0.56	0.69	0.93	0.72	**0.83**
	XGB	0.99	0.06	0.53	0.92	0.36	0.64
	MLP	0.82	0.58	0.70	0.91	0.62	0.77
TMF	*ET*	0.83	0.55	**0.69**	0.94	0.57	0.76
	LR	0.85	0.52	**0.69**	0.88	0.77	**0.83**
	XGB	0.98	0.02	0.50	0.93	0.35	0.64
	MLP	0.82	0.56	**0.69**	0.93	0.65	0.79

This MOOC prediction task is considered highly challenging, compared to other MOOC tasks, such as predicting dropout, completion and learner characteristics. The reason is the severe data imbalance of the binary class, where course certificate purchasers form less than 1% of the total number of enrolled students,

5 Conclusion

This study compared four tree-based and regression classifiers, to predict course purchasability, using discussion forum data from five MOOCs. Our proposed model achieved various balanced accuracies, 0.76 on average, using only the first half of course data. Thus, it is able to predict relatively early on if a purchase of a certificate will take place or not. Further planned improvement of our our model is to use deep models and included more student data e.g. demographics and click stream logs.

References

1. Alshehri, M., et al., *Towards Designing Profitable Courses: Predicting Student Purchasing Behaviour in MOOCs.* International Journal of Artificial Intelligence in Education, 2021. **31**(2): p. 215-233.
2. Alshehri, M., A. Alamri, and A.I. Cristea. *Predicting Certification in MOOCs Based on Students' Weekly Activities*. in *17th International Conference on Intelligent Tutoring Systems (ITS)*. 2021. Univ W Attica, ELECTR NETWORK: Springer International Publishing Ag.
3. Alamri, A., et al. *Predicting MOOCs dropout using only two easily obtainable features from the first week's activities*. in *International Conference on Intelligent Tutoring Systems*. 2019. Springer.
4. Cristea, A.I., et al. *Earliest predictor of dropout in MOOCs: a longitudinal study of FutureLearn courses.* 2018. Association for Information Systems.
5. Jiang, S., et al. *Predicting MOOC performance with week 1 behavior*. in *Educational data mining 2014*. 2014.
6. Reich, J., *MOOC completion and retention in the context of student intent.* EDUCAUSE Review Online, 2014. **8**.
7. Coleman, C.A., D.T. Seaton, and I. Chuang. *Probabilistic use cases: Discovering behavioral patterns for predicting certification*. in *Proceedings of the second (2015) acm conference on learning@ scale*. 2015.
8. Joksimović, S., et al. *Translating network position into performance: importance of centrality in different network configurations*. in *Proceedings of the sixth international conference on learning analytics & knowledge*. 2016.
9. Gitinabard, N., et al., *Your actions or your associates? Predicting certification and dropout in MOOCs with behavioral and social features.* arXiv preprint arXiv:1809.00052, 2018.
10. Alshehri, M., et al. *On the need for fine-grained analysis of Gender versus Commenting Behaviour in MOOCs*. in *Proceedings of the 2018 The 3rd International Conference on Information and Education Innovations*. 2018. ACM.
11. Cristea, A.I., et al., *How is Learning Fluctuating? FutureLearn MOOCs Fine-Grained Temporal Analysis and Feedback to Teachers.* 2018.
12. Cristea, A.I., et al. *Can Learner Characteristics Predict their Behaviour on MOOCs?* in *10th International Conference on Education Technology and Computers (ICETC 2018)*. 2018. Tokyo Inst Technol, Tokyo, JAPAN: Assoc Computing Machinery.
13. Alsheri, M.A., et al. *MOOCSent: a Sentiment Predictor for Massive Open Online Courses*. 2021. Association for Information Systems.