Adopting Automatic Machine Learning for Temporal Prediction of Paid Certification in MOOCs

Mohammad Alshehri[1], Ahmed Alamri and Alexandra Cristea[1]

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

Abstract. Massive Open Online Course (MOOC) platforms have been growing exponentially, offering worldwide low-cost educational content. Recent literature on MOOC learner analytics has been carried out around predicting either students' dropout, academic performance or students' characteristics and demographics. However, predicting MOOCs certification is significantly underrepresented in literature, despite the very low level of course purchasing (less than 1% of the total number of enrolled students on a given online course opt to purchase its certificate). Additionally, the current predictive models either choose conventional learning algorithms randomly, or fail to finetune algorithms to enhance their accuracy. Thus, this paper proposes, for the first time, *deploying automated machine learning (AutoML) for predicting the paid certification in MOOCs*. Moreover, it uses a temporal approach, with prediction based on first-week data only, and the first half of the course activities. Using 23 runs of 5 courses of FutureLearn, our results show that the AutoML technique achieves promisingly better results. We conclude that the dynamicity of AutoML in terms of automatically finetuning the hyperparameters is promising to identify the best classifiers and parameters for paid certification in MOOCs prediction.

Keywords: MOOCs, Certification Prediction, AutoML, Auto-sklearn.

1 Introduction

Online courses have been revolutionising and reforming education for decades. More recently, massive open online courses (MOOCs) were explicitly introduced, to democratise access to education and reach a massively unlimited number of potential learners from around the world. The first official emergence of MOOCs was with the launch of Stanford's Coursera in 2011 [1, 2], although the following year was coined as "the year of the MOOCs" when many of todays' successful platforms, such as *FutureLearn*, *edX*, *Udemy* and *Coursera* were introduced [3, 4], offering scalable world-wide online courses to the public [5, 6].

Although MOOCs have been successful, attracting many online learners, the staggeringly *low completion and certification rates* are still one of the more concerning

aspects to date, a funnel with students "leaking out" at various points along the learning pathway [7, 8]. While the high dropout rate has been the focus of many studies, the race towards identifying precise predictors of completion, as well as the *predictors of course purchasing*, continues. Importantly, although MOOCs have started being analysed more thoroughly in the literature, few studies have investigated the characteristics and temporal activities for modelling learners' certification decision behaviours.

Another objective this study attempts to address is examining the extent to which AutoML can help achieve competitive performance in predicting certification in MOOCs. With machine learning becoming more mainstream in the field of data science, there has been an increasing demand on automated tools that can automate the process of designing and optimising machine learning pipelines, with less human intervention [9]. In response to this demand, many AutoML frameworks have been introduced [10-12].

Considering the recent MOOCs' transition towards paid macro-programmes and online degrees, with affiliate university partners, along with the advancements in the automation and explanation of learners' activities prediction, this paper presents an automated predictor of MOOCs paid certification. Specifically, this paper attempts to answer the following research question:

- *To what extent can AutoML predict MOOC learners' purchase decisions (certification)?*

It is worth mentioning that the contribution of this study goes beyond randomly comparing different classifiers on predicting paid certifications in MOOCs to *proposing a stable, comprehensive automated model for dynamically optimising hyperparameters during the learning process*. Additionally, we are *investigating the classification performance temporally*, using different periods (early and middle) during the course. This is the *first study that employs AutoML to predict paid certification in MOOCs* to the best of our knowledge.

2 Related Works

While several studies have predicted learners' behaviours in MOOCs, the number of studies that use AutoML for this purpose remains relatively low. Concerning the previous studies that used AutoML to predict or classify learners, [13] investigates the potential of Auto-Weka (one of the standard AutoML systems) in early predicting learning outcomes (pass/fail) based on learners participation on the Moodle e-learning platform. The study limited the experiment to tree-based and rule-based models for more transparent and interpretable results, using data from 591 students over three courses. For the purpose of initial comparison, one predictor of each main category of learners (Bayes classifiers, rule-based, tree-based, function-based, lazy and meta classifiers) have been randomly chosen to compare against Auto-Weka performance. The results show that the latter significantly achieved consistently better results on the classification task.

[14] proposes a generic automated weak supervision framework (AutoWeakS), using reinforcement learning, to build a MOOC course recommender for job seekers. The framework allows training multiple supervised ranking models and automatically searching for the best combination of supervised and unsupervised models. With experiments on 1951 course descriptions of different disciplines obtained from XuetangX[1], a Chinese MOOC platform, the model significantly outperforms the classical unsupervised, supervised and weak supervision baseline.

Recently, [15] assisted the impact of adopting an AutoML strategy on feature engineering, model selection, and hyperparameters tuning in predicting student success. The researchers replicate a previous experiment to involve hyperparameter tuning via an AutoML technique for hyperparameters tuning with the data cleaning, preprocessing, feature engineering and time segmentation approach from the previous experiment as-is. The study shows significant general improvement with specific classifiers (Decision Tree, Extra Tree, Random Forest) performing the best. This is another indicator that AutoML can outperform even carefully planned educational prediction models. However, none of the previous works has addressed the issue of the low certification rate in MOOCs using AutoML. Unlike previous studies, 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 [16, 17]. Our study additionally identifies the most representative factors for certification purchase prediction. It also proposes an AutoML-based collection of tree-based and regression classifiers to predict MOOC purchasability using relatively few input features.

3 Methodology

3.1 Data Collection and Preprocessing

The current study is analysing data extracted from a total of 23 runs spread over 5 MOOC courses, on four distinct topic areas, all delivered through FutureLearn, by the University of Warwick. These topic areas are Literature, Psychology, Computer Science and Business [18].

These courses were delivered repeatedly in consecutive years (2013-2017), thus we have data on several '*runs'* for each course.

The dataset obtained went through several processing steps to be prepared and fed into the learning model. Since some students were enrolled on more than one run of the same course, the run number was attached to the student's ID, to avoid any mismatch during joining student activities over "several runs" with their current activities. Additionally, we eliminated irrelevant data generated by organisational administrators (455 admins across the 23 runs analysed) and applied other standard pre-processing.

[1] http://www.xuetangx.com

3.2 AutoML Systems

The fundamental purpose of AutoML systems is reducing human intervention via automating feature preprocessing, hyperparameters finetuning and best-performing algorithm selection, with the ultimate goal of maximising classification accuracy on a supervised classification task [9]. Auto-sklearn is a scikit-learn-based framework that uses 15 classifiers, 14 feature preprocessing methods, and 4 data preprocessing methods, giving rise to a structured hypothesis space with 110 hyperparameters. It improves on other existing AutoML methods by automatically considering the past performance on similar datasets and constructing ensembles from the models evaluated during the optimisation.

3.3 Setting the Auto-sklearn Hyper Parameters

Although AutoML systems automatically optimise pipelines with less human intervention, there are some Auto-sklearn-specific hyperparameters that master the overall learning process and already have default values for a higher level of automation. However, these parameters can be manually finetuned to improve the pipeline's performance further.

After training and testing the models, Auto-sklearn automatically nominated the best performing models and set of hyper-parameters for each one of the five courses. Our best performing classifiers include: *Bernoulli_nb, Adaboost, Extra_trees, Decision_tree, Libsvm_svc (C-Support Vector Classification), Random_forest, Linear Discriminant Analysis (LDA), Gradient_boosting, Multinominal_nb, Passive_aggressive* and *Sgd (stochastic gradient descent) learning.*

4 Results and Discussion

We demonstrate that using AutoML technique, each dataset has its own features, and thus, even the most common classifiers adopted among MOOC researchers may not be the best performing on each dataset. Our previous experiment [4], using the most commonly classifiers has reached satisfactory results. However, the results below outperform the current MOOC Paid certification state-of-art and introduce a promising approach to adopting AutoML in modelling learners' behaviours prediction in MOOCs.

Table 1 shows the result of AutoML-based predicting certification using the first week logged data only versus the first half of the course. It can be seen that although some courses results (BA), such as Supply Chain (SC), were relatively high, the difference in recall score of class 0 and class 1 is high across the five courses. This means that the model is highly biased towards class 1; hence the first-week data may not accurately predict certification.

Also it can be seen that the performance improved between 1% to 9% across the five courses when further data were edded. The SC course has shown the lowest improvement. Nevertheless, both class recalls participated almost equally in the second experiment. It is seen that the gap between the two classes recalls has shrined when further weekly activities have been included.

Table 1. Best optimised Pipelines by Auto-sklearn distributed by course using the first-week-only activities versus the first half of the course, class 0 = non-paying learners, class 1 = paid learners. Metrics rounded to 2 decimal places.

C.	Classifier	1st Week only			Classifier	1st Half of the Course Only		
		Rec_0	Rec_1	BA		Rec_0	Rec_1	BA
BIM	*Ber_NB*	0.63	0.95	0.78	*AdBoost*	0.78	0.9	0.84
	AdBoost	0.62	0.95	0.78	*RF*	0.78	0.9	0.84
	EXT	0.6	0.95	0.77	*DT*	0.8	0.89	0.84
	DT	0.6	0.95	0.77	*LIBSVM_SVC*	0.8	0.89	0.84
BD	*AdBoost*	0.76	1.00	0.88	*RF*	0.87	0.98	0.92
	LIBSVM_SVC	0.77	0.98	0.88	*DT*	0.86	0.98	0.92
	RF	0.77	0.98	0.87	*EXT*	0.86	0.98	0.92
	DT	0.75	1.00	0.87	*GrBoost*	0.86	0.98	0.92
SC	*EXT*	0.84	1.00	0.92	*LIBSVM_SVC*	0.9	0.93	0.92
	RF	0.84	1.00	0.92	*PA*	0.9	0.93	0.92
	LDA	0.83	1.00	0.91	*GrBoost*	0.9	0.93	0.92
	GrBoost	0.82	1.00	0.91	*RF*	0.9	0.93	0.91
SP	*DT*	0.59	0.99	0.79	*RF*	0.79	0.97	0.88
	Mul_NB	0.57	1.00	0.78	*Ber_NB*	0.79	0.97	0.88
	PA	0.56	1.00	0.78	*PA*	0.79	0.97	0.88
	LIBSVM_SVC	0.56	1.00	0.78	*LIBSVM_SVC*	0.79	0.97	0.88
TMF	*EXT*	0.68	0.98	0.83	*EXT*	0.83	0.95	0.89
	PA	0.68	0.98	0.83	*LIBL_SVC*	0.82	0.96	0.89
	DT	0.68	0.98	0.83	*Ber_NB*	0.82	0.96	0.89
	SGD	0.68	0.98	0.83	*LIBSVM_SVC*	0.82	0.96	0.89

5 Conclusion and future work

There are few studies on using AutoML techniques to predict MOOC learners' activities. Thus, this paper *proposes, for the first time, automated machine learning (AutoML) for predicting paid certification in MOOCs*. Our results show that the AutoML technique achieved promisingly better results, compared to the traditional approach of randomly selecting best-in-class predictive algorithms. In our subsequent work, we will further investigate the reason behind having different classifiers in each one of the temporal scenarios. It is known that each classifier initially has its own capability based on the data fed (here, the number of weekly features), but a deeper investigation of a range of parameters configuration is needed, in order to understand these varying results.

References

1. Ng, A. and J. Widom, *Origins of the Modern MOOC (xMOOC).* Hrsg. Fiona M. Hollands, Devayani Tirthali: MOOCs: Expectations and Reality: Full Report, 2014: p. 34-47.

2. 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.
3. Gardner, J. and C. Brooks, *Student success prediction in MOOCs.* User Modeling and User-Adapted Interaction, 2018. **28**(2): p. 127-203.
4. 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.
5. 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.
6. Cristea, A.I., et al. *Earliest predictor of dropout in MOOCs: a longitudinal study of FutureLearn courses*. 2018. Association for Information Systems.
7. Clow, D. *MOOCs and the funnel of participation*. in *Proceedings of the third international conference on learning analytics and knowledge*. 2013. ACM.
8. Breslow, L., et al., *Studying learning in the worldwide classroom research into edX's first MOOC.* Research & Practice in Assessment, 2013. **8**: p. 13-25.
9. Olson, R.S. and J.H. Moore. *TPOT: A tree-based pipeline optimization tool for automating machine learning*. in *Workshop on automatic machine learning*. 2016. PMLR.
10. Bergstra, J., D. Yamins, and D.D. Cox. *Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms*. in *Proceedings of the 12th Python in science conference*. 2013. Citeseer.
11. Kotthoff, L., et al., *Auto-WEKA: Automatic model selection and hyperparameter optimization in WEKA*, in *Automated Machine Learning*. 2019, Springer, Cham. p. 81-95.
12. Feurer, M., et al., *Auto-sklearn: efficient and robust automated machine learning*, in *Automated Machine Learning*. 2019, Springer, Cham. p. 113-134.
13. Tsiakmaki, M., et al., *Implementing AutoML in educational data mining for prediction tasks.* Applied Sciences, 2020. **10**(1): p. 90.
14. Hao, B., et al., *Recommending Courses in MOOCs for Jobs: An Auto Weak Supervision Approach.* arXiv preprint arXiv:2012.14234, 2020.
15. Drăgulescu, B. and M. Bucos. *Hyperparameter tuning using automated methods to improve models for predicting student success*. in *International Conference on Information and Software Technologies*. 2020. Springer.
16. Cristea, A.I., et al., *How is Learning Fluctuating? FutureLearn MOOCs Fine-Grained Temporal Analysis and Feedback to Teachers.* 2018.
17. 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.
18. 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.