# Course recommendations in MOOCs using collaborative filtering and survival analysis

**Alireza Gharahighehi**[a,b;*], **Michela Venturini**[a,b], **Frederik Cornillie**[b,c] **and Celine Vens**[a,b]

[a]KU Leuven, Campus Kulak, Department of Public Health and Primary Care
[b]Itec, imec research group at KU Leuven
[c]KU Leuven, Department of Linguistics

**Abstract.** Massive Open Online Courses (MOOCs) are becoming a complementary, or even preferred, method of learning compared to traditional education among learners. While MOOCs enable learners to access a wide range of courses from various disciplines, anytime and anywhere, a significant number of course enrollments in MOOCs end up in dropouts. To increase learners' engagement in MOOCs, they need to interact with the courses that match their preferences and needs. A course Recommender System (RS) models learners' preferences and recommends courses based on their previous interactions within the MOOC platform. Dropout events in MOOCs, like other time-to-event predictions, can be effectively modeled using survival analysis methods. The objective of this ongoing research is to evaluate the benefits of employing survival analysis in enhancing the performance of collaborative filtering-based course recommendations in MOOCs.

## 1 Introduction

Massive Open Online Courses (MOOCs) platforms offer a diverse selection of online courses to learners worldwide, promoting the concept of equitable learning by removing barriers of location and time. Despite its considerable advantages, a significant portion of MOOC enrollments end up in dropouts. It has been reported that dropout rates for courses offered by prestigious institutions like MIT and Harvard can be as high as 90% [1]. While dropouts may result from various reasons, such as accessing only the free portions of the courses, finding the course or topic irrelevant, or insufficient competencies, this information is valuable for modeling users' preferences in MOOCs.

Recommender Systems (RSs) are intelligent information retrieval algorithms that utilize users' past interactions to suggest the most relevant items to them. Generally, RSs can be categorized into two main types: Content-based filtering and collaborative filtering. Content-based filtering RSs recommend items whose features match those of items that the target user has previously liked. On the other hand, collaborative filtering RSs model users' preferences based on similarities between the past interactions of users and items.

In a MOOC platform, a collaborative filtering-based RS can be applied to recommend courses to users based on their previous enrolments in the platform. While previous enrolments are informative to model users' preferences, the dropout information is still missing. The dropout event in MOOCs is crucial as a significant portion of enrollments end in dropout. This additional information about user-course interactions can be useful to better model users' preferences or needs regarding the courses. Survival analysis (SA) comprises a set of statistical methods that model the time until an instance experiences a specific event such as death or machine failure [2]. The key characteristic of survival data is that some instances have unobserved events, referred to as censored data. The most common form of censoring in SA is right-censoring, where the target event is not observed during follow-up or the instance is lost before the end of the follow-up period. The main strength of SA is its utilization of such partial information during the learning process by considering instances with censored events, which are usually discarded in classification and regression tasks. We believe that time to dropout is highly informative in modeling users' preferences in the context of course recommendations, as it provides valuable insights regarding students' engagement in MOOCs [12].

In our previous study [5], we demonstrated that SA can improve the performance of a specific RS, namely Bayesian Personalized Ranking (BPR), when the predictions of a SA method, trained based on time to dropouts, are embedded in the BPR algorithm. In this study, we aim to generalize the usage of SA in any type of collaborative filtering-based RS. In the next section, we briefly report the existing literature around dropout in MOOCs and then in Section 3, elaborate on the research questions that we are tackling in our current study.

* Corresponding Author. Email: alireza.gharahighehi@kuleuven.be.

## 2 Related work

The task of dropout prediction in the context of MOOCs has been modeled as a classification task [3, 1]. While in this studies the task was predicting the event of dropout they ignored the time information in their predictions. SA can be used to incorporate the time information in modeling dropout in MOOCs and there are some promising examples in the literature. In [6] survival analysis was used to model dropout risk in the context of MOOCs and unveil social and behavioral feature impacts on the outcome. Xie [15] utilized survival analysis to examine the hazard function of dropout, employing the learner's course viewing duration on a course in MOOCs. Labrador et al. [7] specified the fundamental factors attached to learners' dropout in an online MOOCs platform using Cox Proportional Hazard regression. Wintermute et al. [14] applied Weibull survival function to model the certificate rates of learners in a MOOCs platform, assuming that learners "survive" in a course for a particular time before stochastically dropping out. In [10] a more sophisticated SA deep learning approach was proposed to tackle volatility and sparsity of the data, that moderately outperformed the Cox model. While SA has been already applied in literature to model dropout in MOOCs, to the best of our knowledge, such time to dropout from courses has never been incorporated in course recommendations in the context of MOOCs.

## 3 Research questions

In the ongoing study we are focusing on the merits of SA in recommendations when the time to events information is available. We are tackling the following research questions:

1. Does a SA method trained based on time to dropout have a positive impact on the performance of course recommendations in MOOCs when combined with a regular RS? In our previous paper [5] we followed the most straightforward way to combine the SA method and the RS. We enhanced the training data of the RS using the predictions of the SA method. What are the other possible ways of combining the SA method and the RS?
   (a) Is it possible to generalize the proposed idea in [5] to the other learning-to-rank recommendation approaches such as WARP [13]?
   (b) Is it possible to use the predictions of SA directly for recommendations, i.e., sort the items based on their risk scores predicted by SA? Can SA directly compete collaborative filtering approaches?
   (c) How effective it is if the predictions of SA being used to post-process the output of a RS, i.e., to adjust the predictions of RS based on the predictions of SA?
   (d) Does it make sense to model the whole prediction task in MOOCs as a multi-task learning problem where the tasks are predicting time-to-dropout or the risk scores for the SA method and ranking the courses for the recommendation task? Do these two mentioned tasks benefit from a partially shared learning process?
2. How to generalize the discussed idea to other applications?
   (a) To the best of our knowledge, there are only three publicly available MOOC datasets that can be used to validate course RSs. Is it possible to apply the same idea to other applications, for instance series recommendations, using the time to stop watching the series?
   (b) Is it possible to extend the current idea with multiple time to events, for instance course completion and course dropout in the context of MOOCs?

## 4 Experimental setup

In this ongoing study we will compare different competing approaches including the most common collaborative filtering methods such as BPR [11], WARP [13], IKNN and SLIM [9], SA-based RS, a collaborative filtering RS enhanced with SA post-processing, and a multi-task method. We will apply the competing methods on at least three publicly available datasets, namely Xuentangx [4], Canvas [8] and KDD-CUP [4] and possibly an additional dataset from another domain such as series recommendations. The competing methods will be evaluated based on main RS evaluation measures such as NDCG and recall.

## 5 Conclusion

In this extended abstract we illustrate the research objectives and questions that we are aiming in our ongoing study about the merits of survival analysis in recommendation tasks. While, in a very specific case [5], we have demonstrated that SA positively affects the performance of a learning-to-rank recommender system for a course recommendation task in three MOOCs datasets, we have not validate the idea on a more general setting where SA can assist any collaborative filtering algorithm in modeling users' preferences and ranking the items when there is time to event information.

## Acknowledgements

The authors acknowledge support from the Flemish Government (AI Research Program).

## References

[1] Jing Chen, Bei Fang, Hao Zhang, and Xia Xue, 'A systematic review for mooc dropout prediction from the perspective of machine learning', *Interactive Learning Environments*, 1–14, (2022).
[2] Taane G Clark, Michael J Bradburn, Sharon B Love, and Douglas G Altman, 'Survival analysis part i: basic concepts and first analyses', *British journal of cancer*, **89**(2), 232–238, (2003).
[3] Fisnik Dalipi, Ali Shariq Imran, and Zenun Kastrati, 'Mooc dropout prediction using machine learning techniques: Review and research challenges', in *2018 IEEE global engineering education conference (EDUCON)*, pp. 1007–1014. IEEE, (2018).
[4] Wenzheng Feng, Jie Tang, and Tracy Xiao Liu, 'Understanding dropouts in moocs', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 517–524, (2019).
[5] Alireza Gharahighehi, Michela Venturini, Achilleas Ghinis, Frederik Cornillie, and Celine Vens, 'Extending bayesian personalized ranking with survival analysis for mooc recommendation', in *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, pp. 56–59, (2023).
[6] Niki Gitinabard, Farzaneh Khoshnevisan, Collin F Lynch, and Elle Yuan Wang, 'Your actions or your associates? predicting certification and dropout in moocs with behavioral and social features', *arXiv preprint arXiv:1809.00052*, (2018).
[7] Manuel Medina Labrador, Gustavo R Garcia Vargas, Jorge Alvarado, and Mario Caicedo, 'Survival and risk analysis in moocs', *Turkish Online Journal of Distance Education*, **20**(4), 149–159, (2019).
[8] Canvas Network, 'Canvas Network Person-Course (1/2014 - 9/2015) De-Identified Open Dataset'. Harvard Dataverse, (2016).
[9] Xia Ning and George Karypis, 'Slim: Sparse linear methods for top-n recommender systems', in *2011 IEEE 11th international conference on data mining*, pp. 497–506. IEEE, (2011).
[10] Feng Pan, Bingyao Huang, Chunhong Zhang, Xinning Zhu, Zhenyu Wu, Moyu Zhang, Yang Ji, Zhanfei Ma, and Zhengchen Li, 'A survival analysis based volatility and sparsity modeling network for student dropout prediction', *PloS one*, **17**(5), e0267138, (2022).
[11] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme, 'Bpr: Bayesian personalized ranking from implicit feedback', in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 452–461, (2009).
[12] Marili Rõõm, Marina Lepp, and Piret Luik, 'Dropout time and learners' performance in computer programming moocs', *Education Sciences*, **11**(10), 643, (2021).
[13] Jason Weston, Samy Bengio, and Nicolas Usunier, 'Wsabie: Scaling up to large vocabulary image annotation', in *Twenty-Second International Joint Conference on Artificial Intelligence*, (2011).
[14] Edwin H Wintermute, Matthieu Cisel, and Ariel B Lindner, 'A survival model for course-course interactions in a massive open online course platform', *PloS one*, **16**(1), e0245718, (2021).
[15] Zheng Xie, 'Modelling the dropout patterns of mooc learners', *Tsinghua Science and Technology*, **25**(3), 313–324, (2019).