Survival Analytics of MOOC Learners and predicting the influencing factors of their active participation

T.Chellatamilan [a], N. Srinivasa Gupta [b], K.Santhi [c], *B.Valarmathi [d]
[a]Associate Professor, Department of Computer Applications, School of Information Technology and Engineering, Tamilnadu, INDIA
[b]Associate Professor, Department of Manufacturing, School of Mechanical Engineering, Vellore Institute of Technology, Vellore, Tamilnadu, INDIA
[c]Associate Professor, Department of Analytics, School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamilnadu, INDIA
[d]Associate Professor (Senior), Department of Software and Systems Engineering, School of Information Technology and Engineering, Tamilnadu, INDIA
E-mail: [a]chellatamilan@gmail.com, [b]n.srinivasaguptavit@gmail.com, [c]santhikrishnan@vit.ac.in , [d]valargovindan@gmail.com
*Corresponding Author

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 10 May 2021

Abstract:
MOOC is the largest online learning platform across the world. Many large universities are offering their courses with free, open and online with unlimited participation. MOOC is a learner centric environment as the learner has the full freedom of choosing the course and course content with respect to their own learning pattern or styles. MOOC intern can accommodate content through recorded lecture, videos, reading transcripts and conducting quiz or assessments. MOOC also accomplishes the discussion forum through which the learner can do their conversation to other peer learners as well as to the course owners. The discussion forum also connects the similar learners into a group from a spacious scope of experience and learning background across the earth. Furthermore, there are also few characteristics that the MOOC platform is considered. In our proposed method we have used the additional covariates in determining the survival probability of the online learners with a hybrid parametric approach which improves the prediction accuracy substantially.

Keywords: Course Owner, E-LEARNING, MOOC, Survival Analysis, Survival Probability

1. Introduction (Times New Roman 10 Bold)

Survival analytics are a time to event analysis used for forecasting the lifetime span of disciplines over a period of time with respect to various instantaneous environmental conditions. It is contributing much in the field of health care, enterprise, environment, customer relationship management, sports analytics and etc. For applying the survival analytics over the field, it is not requiring any beginning point and ending point the collection of datasets. The survival duration may be longer than the duration time of the study of the survival process.

We have decided to use this kind of survival model in the MOOC system where as many of the MOOC learners are not continuing their activities till the end. Irrespective of all these things the MOOC facilitates the following characteristics:

1. Lack of entry requirements, regardless of background, prerequisite, age or locations
2. Repetitions in running the course for more number of times.
3. High caliber and high rated education resources can be integrated
4. Feasibility Time
5. Learner centric or Self-paced learning

2. Literature Survey

Some of the major players of MOOC are Coursera, Edx, Udacity and Swayam portal. Yet there are few unavoidable challenges and limitations exist upon the usage of MOOC as listed beneath:

1. Not adaptable to the vertical diversity of learners
2. No monitoring of the active participation of learners
3. Risk of plagiarism
4. Digitization is must

5. Not based on pedagogy perspective learning

A lot of research and developments are going on towards the satisfaction of learner's expectation and engagement of online learners.

Moreover, the MOOC learners participated in an online course show interest during the beginning phase, whereas their interest has been reduced along the way towards the end (Bagarinao, 2015). There are so many factors that influence the dropout rate of the learners such as cohort, authority score and sub community. The probabilistic soft logic model has been built for capturing student domain expertise about their interaction with peer learner and outcome. From this model the student engagements towards the progress have been determined and the dropout rate has been reduced (Arti Ramesh, 2013).

The peer competitors in the online learning framework have to be identified and the role of dominant design should be neglected to increase the efficacy of advanced learning without confusion (Liang, 2016). The navigational pattern of learners was analysed through the mean number of their visits in each page and the correlation about their performance has been measured to maximize the learners learning experience through the learning transactions (Meredith Carroll, 2019). The latent engagement pattern of the MOOC learners should be determined through probabilistic model to initiate interventions and assist adaptive learners (Ramesh, 2014).Survival Analytics is helping to perform analysis over the data in each time unit until the event of interest is occurred. The output of survival analysis is always referred as survival time, event time or failure time. The incomplete observed responses are called as Censor. If there is no censor in the system, then we can use linear regression otherwise the survival model will be more useful. Sometime there are composite or multiple end points were there in the system which will decide the latent event to occur. The very first occurrence of any of such composite end points has been influenced much than others (Paola M.V. Rancoita, 2016).

The question answering system uses the survival analysis to characterize and discover the interesting communities involved in the pattern of QA platforms. The waiting time in human computer interaction is evaluated based on the temporal metaphors such as cognitive load, retention delay and the paradigm used. The objective assessment technique minimizes the influence of such factors (Ortega, Felipe & Convertino, Gregorio & Zancanaro, Massimo & Piccardi, Tiziano., 2014) (Ortega, 2017). The feedback of the learners has been acquired through the click log data and the click curve is generated which in turn predict the quality of top-rank learners (Dickman, 2008). The resources in the MOOC are designed in such a way that allows online learners to succeed their learning goals with a minimum level of prior educational experience or intention (Miaomiao Wen, 2013) (Wong JS., 2015).

Customer lifetime value is a very temporary measure to be considered for profitability and reduce risk. Customer survival curve can help the companies develop their strategies to grow the company with the right customer. In e-learning system, the customer is referred as learner whereas the company is relating the MOOC course owner running the course (Greene, 2015). The Survival analysis quantifying the participation of user patterns and then predicts the activities span over a period of time (Adamic, 2010). The defeminisation of learner's survival in early stage helps to imitate better instructor (Pega Davoudzadeh, 2015). The competing risk model is constructed from survival analysis, which will preclude the occurrences of the other events in the time to event data (Jana Fürstová, 2011).

The survival analytics model has been built by mining the collective sentiment from MOOC discussion forum to visualize learners trending opinions and estimate the dropout reasons (Wanli Xing, 2019) (Saad, 2017). The attrition happens in the online learning MOOC platform due to the emotional and social positioning factor (Rosé, 2014).

3. Survival Analytics of E-Learning Applications

The Sample set of scenario of survival analytics in the field of e-learning is given here in the form of estimation of different things.

- Estimation when a learner is passing the course / failing the course through duration modelling
- Estimate when a learner will click the link of the provided course content/materials
- Estimate the frequent access pattern of resource link for online learners

The fundamentals of survival analytics is given below. The main focus of survival analytics is to predict the time to event data from the survival data provided fully or partially (censored).

The survival analytical model is being built with the help of mathematical functions such as survival function, cumulative density function, death density function and hazard density function.

The survival function S(t) is built with probability basics that represent the probability of given instance, can survive for a longer period ‘T’ than a particular time ‘t’ as given in the Equation 1.

$$S(t) = \Pr(T \geq t) \quad \text{- Eq.1}$$

Cumulative density function F(t) is indicating the probability that the event of interest occurs before time ‘t’ as given in the Equation 2.

$$F(t) = 1 - S(t) \quad \text{- Eq.2}$$

Death density function f(t) represents the differentiation of cumulative density function with respect to the time ‘t’ as mentioned in the Equation 3.

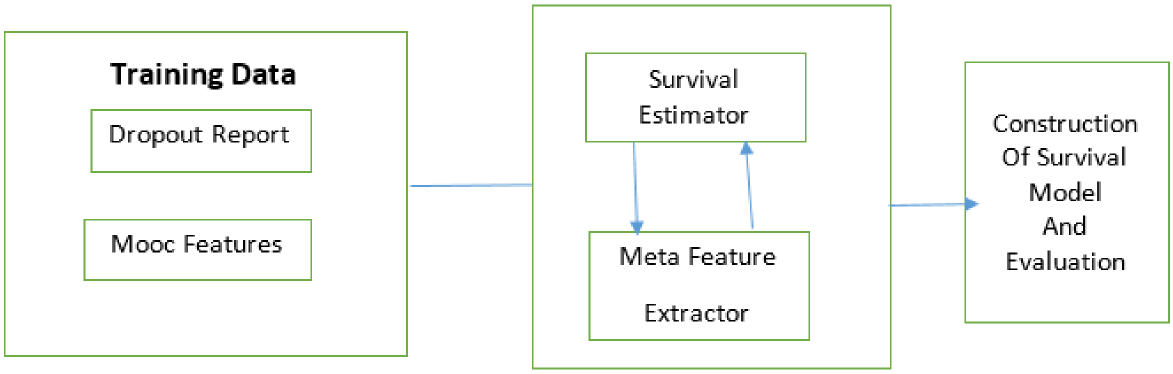
$$\mathrm{f(t)} = \frac{\mathrm{dF(t)}}{\mathrm{dt}} \quad \text{- Eq.3}$$

Hazard Function h(t) is a probability that the event of interest occurs in the next instant given survival to time ‘t’ as noted in the Equation 4.

$$\mathrm{h(t)} = \frac{\mathrm{f(t)}}{\mathrm{S(t)}} \quad \text{- Eq.4}$$

Since we are using the partial amount of data (censoring) in predicting the time to an event, it is not possible to use linear or logistic regressions with the mean squared error in survival analytics.

Figure.1. System Architecture of Survival Analytics



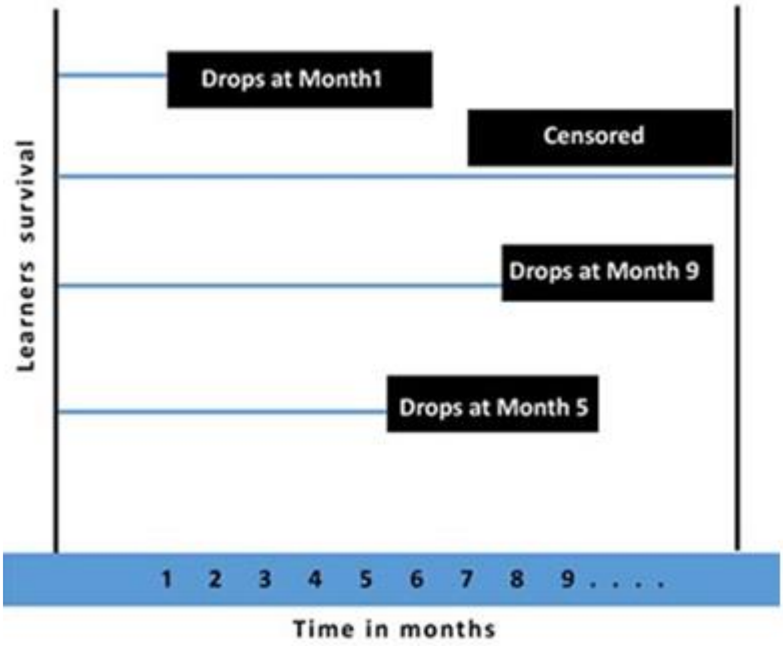
The dataset contains dropout report and MOOC features. It has 72325 rows and 27 columns like enrollment_id, access, discussion, navigate, page_close, problem, video, wiki, proccess_ period, present-day, effective time, Friday, Monday, Saturday, Sunday, Thursday, Tuesday, Wednesday, holidays, start_year, start_month, end_year, end month, course_enroll, user_enroll, course_drop_rate and dropout_prob. The feature “dropout_prob” has two values like 1 or 0. “1” means learner is dropping the course. “0” means learner is not dropping the course.

The survival estimator used here is Kaplan Meier Estimator. Kaplan Meier Estimator is utilized to assess the survival function for lifetime information. It is a non-parametric measurements strategy. It is otherwise called as far as product-limit estimator, and the idea lies in assessing the survival time for a specific time of like a significant clinical trial event, a specific season of death, failure of the machine, or any major critical occasion. Kaplan Meier Estimator is to gauge the number of leaners will leave the MOOC Course in a particular timeframe.

The meta- features, likewise called characterization measures. Meta Feature Extraction separates the characterization estimates which can describe the intricacy of datasets and assists with offering the evaluations of the algorithm performance. It likewise assisted a ton with understanding the learning inclination to be utilized in applying machine learning algorithms. These actions should have the option to predict, with a low computational expense, the exhibition of the algorithms under assessment.

The simple architecture of survival analytics is depicted in Figure 1, whereas the survival analysis features comprises of two types of variables namely output variable and input variables. The output variable is the time duration to the event/censored or its combinations. Input variable is the event features denting the duration of time till the event occurs in the span of the study and the event variant denote the value 1 to indicate the event has occurred whereas the value 0 indicates for other. The survival data is classified into four categories and they are drops at month1, censored, drops at month9, and drops at month5. The propagation of the learning processes and the occurrence of the events (Survival Data) are clearly depicted in the Figure 2. Meta Feature Extraction extracts the characterization measures which are able to characterize the complexity of datasets and helps to offer the estimates of the algorithm performance.It also helped a lot to understand the learning bias to be used in applying machine learning algorithms

Figure.2. Survival Data



4. Experiment and Results

In this study, we used Kaplan Meier estimator which allows us to estimate the survival function of the learning data. The overall distribution of time is shown in the histogram plot which uses bi-model distribution to choose the cut-off of overall learning time of the course. The next step is to apply the data to fit the Kaplan Meier estimator by passing the values of MOOC data to the survival fit function. The summary of resulting fit is shown with numerical values in the table. We can examine the values of survival fitting by plotting it on a graph. The vertical line denotes censored data and their corresponding time value at which the censoring event occurs.

A histogram is a graphical presentation of information utilizing bars of various statures. In a histogram, each bar bunches numbers into ranges. Taller bars show that more information falls around there. A histogram shows the shape and spread of persistent example information. The time is in X axis and the frequency is in Y axis. The figure 3 depicts the content of the dataset along with various learning instance events with respect to the learning progression time.

Kaplan-Meier estimate is probably the most ideal alternative to be utilized to quantify the negligible part of subjects living for a specific measure of time after treatment. The figure 4 expresses the estimates in graphical form plotted from the estimated survival probability of the MOOC learners in the Y axis and the learning progress time on the X axis. The survival probability is decreasing with respect to the time over a particular survival probability. The Figure 5 also plotted against the survival probability versus time using semi-parametric proportional hazard model. The cox model combines the additional covariates with the survival fitness function to estimate the probability of survival of online MOOC learners.

Figure.3. Histogram explores the dataset

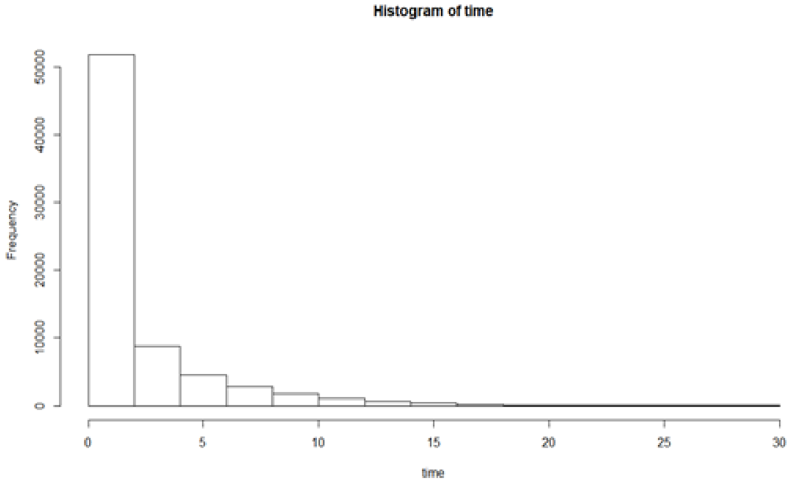
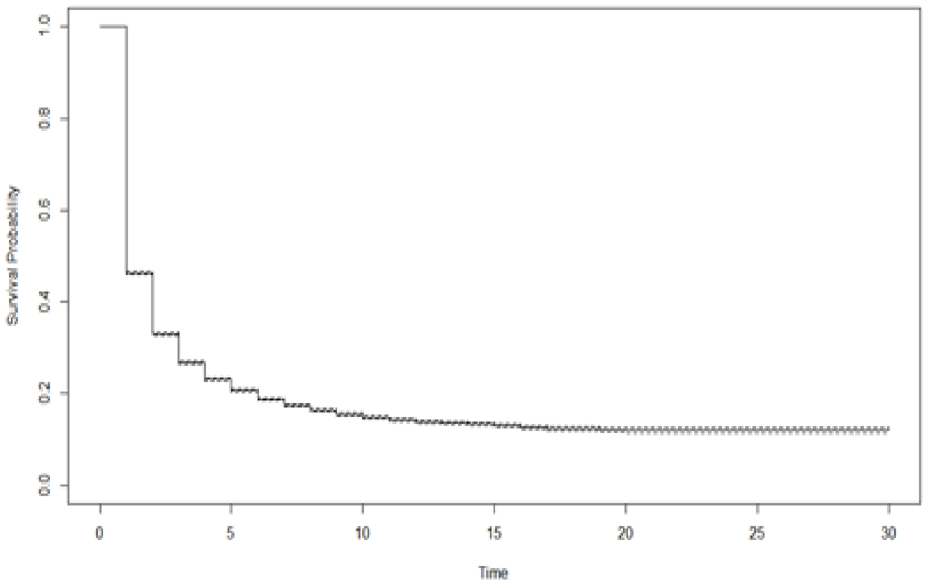


Figure.4. Kaplan Meier Estimate

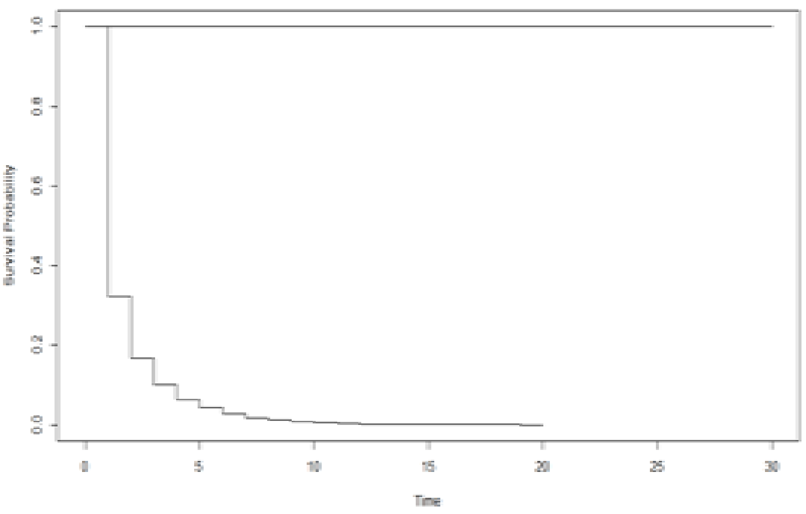


The graph plotted here shows the relationship between the predicted probability of survival and the learning time progression which also includes the vertical lines and horizontal lines. The survival duration is measured by the length of the horizontal line. The event of interest occurrences and the censored time have been visualized in the vertical lines in Figure 4 and Figure 5. The results indicated here shows that the age and location have a significant influence in the determination of survival of MOOC learners. The statistical comparison of the non parametric and semi parametric survival models have been plotted in the figures.

A parametric survival model is one in which survival time (the result) is expected to follow a known distribution. The distributions examples that are generally utilized for survival time are: the Weibull, the exponential (a unique instance of the Weibull), the log-logistic, the log-normal, and so forth.

The Cox proportional hazards model, on the other hand, is anything but a completely parametric model. Maybe it is a semi-parametric model on the grounds that regardless of whether the regression parameters (the betas) are known, the dispersion of the result stays obscure. The baseline survival (or hazard) function isn't indicated in a Cox model (we don't expect any shape or structure).

Figure.5. Semi Parametric Proportional Hazard Model



5. Conclusion

Now a day, MOOCs have become as an emerging source of global education due to its versatile features provided through the virtual learning environment. The data analysis plays very crucial role in MOOC for building predictive learner model based on the engagement of the active peer learners and producing effective outcome at the end. We suggest using survival analysis through the framework suitable for the MOOC data set. This procedure is experimentally shown on the simulated MOOC dataset. As suggested from the study of result, active participation of learners and frequent access to the relevant course material may be the factors for the successful survival of the learners towards the end.

References

Bibliography

[1] Adamic, J. Y. (2010). *Activity Lifespan: An Analysis of User Survival Patterns in Online Knowledge Sharing Communities.* ICWSM.

[2] Arti Ramesh, D. G. (2013). Modeling Learner Engagement in MOOCs using Probabilistic Soft Logic. *NIPS Workshop on Data Driven Education.* Purdue University.

[3] Bagarinao, R. T. (2015, January). Students' Navigational Pattern and Performance in an E-Learning Environment: A Case from UP Open University, Philippines. *Turkish Online Journal of Distance Education, 16*(1), 101-111.

[4] Dickman, P. W. (2008). *Biostatistics III: Survival analysis for epidemiologists Computing notes for SAS users.*

[5] Greene, J. &. (2015). Predictors of Retention and Achievement in a Massive Open Online Course. *American Educational Research Journal, 52*(2), 925–955. doi:10.3102/0002831215584621

[6] Jana Fürstová, Z. V. (2011). Statistical analysis of competing risks: overall survival in a group of chronic myeloid leukemia patients. *European Journal of Biomedical Informatics, 7*(1), 1-10.

[7] Liang, Y. S. (2016). A dynamic framework for competitor identification: A neglecting role of dominant design. *JOURNAL OF BUSINESS RESEARCH, 69*(5), 1898-1903.

[8] Meredith Carroll, S. L. (2019, July). An applied model of learner engagement and strategies for increasing learner engagement in the modern educational environment. *Interactive Learning Environments*, 1-15. doi:10.1080/10494820.2019.1636083

[9] Miaomiao Wen, D. Y. (2013). Sentiment Analysis in MOOC Discussion Forums: What does it tell us? *Educational Data Mining*, 130-137.

[10] Ortega, Felipe & Convertino, Gregorio & Zancanaro, Massimo & Piccardi, Tiziano. (2014). Assessing the Performance of Question-and-Answer Communities Using Survival Analysis. *ArXiv*, 1-10.

[11] Ortega, G. C. (2017). Toward a mixed-initiative QA system: from studying predictors in Stack Exchange to building a mixed-initiative tool. *International Journal of Human-Computer Studies*, *99*, 1-20. doi:10.1016/j.ijhcs.2016.10.008
[12] Paola M.V. Rancoita, M. Z. (2016). Bayesian network data imputation with application to survival tree analysis. *Computational Statistics & Data Analysis, 93*, 373-387. doi:10.1016/j.csda.2014.12.008
[13] Pega Davoudzadeh, M. L. (2015). Early school readiness predictors of grade retention from kindergarten through eighth grade: A multilevel discrete-time survival analysis approach. *Early Childhood Research Quarterly (ECRQ), 32*, 183-192. doi:10.1016/j.ecresq.2015.04.005
[14] Ramesh, A. a. (2014). *Uncovering Hidden Engagement Patterns for Predicting Learner Performance in MOOCs.* Atlanta, Georgia, USA: Proceedings of the First ACM Conference on Learning @ Scale Conference. doi:10.1145/2556325.2567857
[15] Rosé, C. &. (2014). Social factors that contribute to attrition in MOOCs. *Proceedings of the first ACM conference on Learning @ scale conference* (pp. 197-198). Atlanta, Georgia, USA: ACM. doi:10.1145/2556325.2567879
[16] Saad, S. B. (2017). Weekly Predicting the At-Risk MOOC Learners Using Dominance-Based Rough Set Approach. *Digital Education: Out to the World and Back to the Campus. EMOOCs. 10254*, pp. 160-169. Lecture Notes in Computer Science, Springer, Cham. . doi:10.1007/978-3-319-59044-8_18
[17] Wanli Xing, D. D. (2019). Dropout Prediction in MOOCs: Using Deep Learning for Personalized Intervention. *Journal of Educational Computing Research, 57*(2), 547-570. doi:10.1177/0735633118757015
[18] Wong JS., P. B. (2015). An Analysis of MOOC Discussion Forum Interactions from the Most Active Users. *SBP 2015: Social Computing, Behavioral-Cultural Modeling, and Prediction. 9021*, pp. 452-457. Lecture Notes in Computer Science, Springer, Cham. doi:10.1007/978-3-319-16268-3_58