Social Network Analysis and Mining (2022) 12:86
https://doi.org/10.1007/s13278-022-00914-8

ORIGINAL ARTICLE

Survival analysis for user disengagement prediction: question-and-answering communities' case

Hassan Abedi Firouzjaei[1]

Received: 18 February 2022 / Revised: 22 June 2022 / Accepted: 23 June 2022 / Published online: 22 July 2022

Abstract
We used survival analysis to model user disengagement in three distinct questions-and-answering communities in this work. We used the complete historical data from domains including Politics, Data Science, and Computer Science from Stack Exchange communities from their inception until May 2021, including information about all users who were members of one of these three communities. Furthermore, in formulating the user disengagement prediction as a survival analysis task, we employed two survival analysis techniques (Kaplan–Meier and random survival forests) to model and predicted the probabilities of members of each community becoming disengaged. Our main finding is that the likelihood of users with even a few contributions staying active is noticeably higher than those who were making no contributions; this distinction may widen as time passes. Moreover, the results of our experiments indicate that users with more favourable views toward the content shared on the platform may stay engaged longer. Finally, regardless of their themes, the observed pattern holds for all three communities.

Keywords Question-and-answering platforms · User disengagement · Survival analysis · Stack exchange

1 Introduction

Online question-and-answering (QA) social networks[1] like Stack Overflow[2] and Quora[3] are dependent on their users' contributions for proper functioning. Arguably, the main functionality of a QA platform is to connect two types of users (Kuzmeski 2009); on one side, people who seek answers to their questions and on the other side, people who are willing to share their knowledge and expertise with others (Guan et al. 2018). Nevertheless, a user who joined and made many contributions to the community may become uninterested and then gets disengaged after a while. By disengaged, we mean the situation where users—as individuals who previously made contributions (e.g., answered questions and participated in debates)—suddenly stopped their activities (i.e., there is no sign of them even visiting the platform's web pages). Moreover, it is not known whether these users left the community or not, but they did not perform any activity on the platform for a relatively long period of time (e.g., more than a year). In this context, disengagement might have happened for various reasons; e.g., it might have occurred because disengaged users believed that the platform had an elitist or even toxic culture. Another reason could have been that user interests changed drastically over time, and the platform hosting the QA community could not adapt to the change in an agile way.

At the very least, a high disengagement rate has adverse effects on the overall quality of the service of a QA social network and platform. For example, suppose all the experts (i.e., users who post answers perceived as high quality by the community) become disengaged within a few months of joining and being active. In that case, the quality of answers might plummet, which may increase the rate of users' disengagement from the community (Pudipeddi et al. 2014; Dror et al. 2012). In the worst-case scenario, one could expect the situation where the QA platform loses the bulk of its contributors, which in turn would lead to its demise.

Survival analysis (Cox and Oakes 2018; Wang et al. 2019) is a family of statistical methods and techniques that

✉ Hassan Abedi Firouzjaei
hassan.abedi@ntnu.no

[1] Department of Computer Science, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

[1] In this work, we use the terms question-and-answering platform, social network, and community interchangeably.

[2] https://stackoverflow.com.

[3] https://www.quora.com.

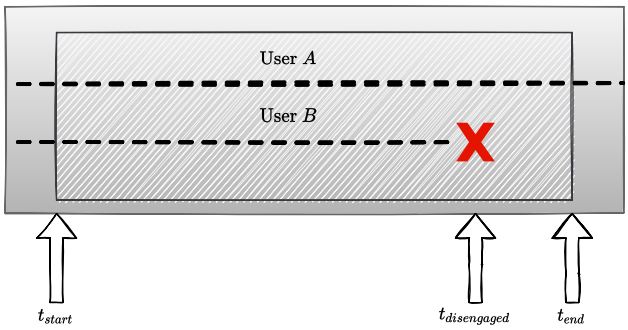


Fig. 1 User A and user B joined the platform in the past; during a period of observation which started at t_{start} and ended at t_{end}, B became disengaged at $t_{disengaged}$. A did not become disengaged during the observation, but it is not known that he will become disengaged in future or not; information about A's disengagement is censored

can help model and predict the time of the occurrence of an event of interest. Initially, it emerged out of medical research to find the probability of a patient surviving a disease such as cancer—hence the term *survival analysis*. More recently, survival analysis methods have found widespread use in new areas such as customer churn analysis (Dias et al. 2020; Rothmeier et al. 2021) and credit risk scoring (Stepanova and Thomas 2002), mainly due to their flexibility and power in accurately and reliably modelling the problems posed in these areas.

In this work, we used survival analysis to study user disengagement in three distinct QA social networks, namely, Politics, Data Science, and Computer Science Stack Exchange. Our choice allowed us to pose questions and seek answers based on the data from QA communities with the themes mentioned above. To our knowledge, this is the first work that applied survival analysis to quantify and study user disengagement using the entire historical data of online QA social networks. Figure 1 illustrates how disengagement prediction can be seen and formulated as a survival analysis task.

Following are the main contributions of our work:

- We study the factors likely to be associated with the probability that users of QA communities will stay active for an extended period. For the first time, we analyze the relationships between attributes related to users' contributions and their engagement time.
- We propose to exploit behavioural (see Table 4), and content-based user attributes (see Table 5) to estimate the engagement time on three comprehensive datasets from distinct QA communities.

The rest of this article is organised as follows. Section 2 discusses the related work. Section 3 presents preliminary concepts related to survival analysis and introduces techniques used to model and evaluate the problem of user disengagement prediction. Section 4 gives an overview of the dataset and the methodology used to represent users and the engagement time. Section 5 presents the results of the experiments and Sect. 6 discusses the results. Section 7 discusses the limitations of our work and gives an outline for the direction of future work. Finally, Sect. 8 concludes the paper.

2 Related work

QA platforms like Stack Exchange and Quora provide an accessible knowledge-sharing environment. Due to the importance of this role which earlier was played by mailing lists, newsgroups and IRC channels, the interest in studying phenomena on these platforms has exploded lately. For example, in Joyce and Kraut (2006), authors studied the continued user participation in newsgroups. They used the posts from six public newsgroups to test whether answers that users receive to their first few questions are crucial for prolongation of user participation. Their findings suggest that longer questions are more likely to receive a response. Furthermore, the quality and emotional tone, and whether the answer was in response to a question from a new user, seems not to influence of likelihood of further participation.

Authors, in Guan et al. (2018), used the data from the most popular Chinese social QA platform, Zhihu,[4] to investigate the factors related to users' motivations to participate in community activities, especially knowledge contribution. Their findings suggest that social exchange is an important factor influencing users' continuous knowledge contribution in social QA communities. Moreover, the findings show that knowledge exchange based on norms of reciprocity is an important factor affecting users' continuous contribution. For example, a user who frequently seeks knowledge is more likely to contribute knowledge to others, indicating users contribute because they expect that they can get a response to their questions in future. Similarly, Jin et al. (2015) studied the elements, based on the data from Zhihu, that were influencing user knowledge contributions in QA platforms, incorporating three theories of social capital theory, social exchange theory, and social cognitive theory in their work.

Furthermore, the use of survival analysis methods is also gaining popularity, where an analogy could be made between the problem and the task of survival analysis. In Wang et al. (2019), the authors provided a comprehensive review of two major categories of methods and techniques for survival analysis; namely, conventional and various machine learning methods for survival analysis. Their work described and discussed different related topics, including data transformation

[4] https://www.zhihu.com.

and early prediction of complex events—along with appropriate evaluation metrics.

Yang et al. (2010) used survival analysis methods to analyze and study user retention in three major QA communities: Baido Knows, Yahoo! Answers, and Naver Knowledge-iN. Their findings suggest that users who preferred answering tend to have a more extended and more active engagement period within the platform. Moreover, garnering enough questions in order to retain the experts seems essential. Additionally, users who put more effort into the average length of questions they post both tend to receive more answers and stay engaged longer. Finally, for answerers, acknowledging one's contribution by having one's answers selected as best or being commented on, was tied to a more extended stay on the platform. Although their work is similar to our work, we used the data for the whole lifespan of the communities, where their work mainly focused on a limited period.

Arguably, three of the most popular metrics to measure user engagement on a web-based platform are click-through rates, page views, and time spent by the user on the website (Dupret and Lalmas 2013). The authors, in Dupret and Lalmas (2013), used survival analysis to analyze the user engagement in a dataset of questions and answers from Yahoo! Answers in Japan— utilising user absence time (or absence time for short) which is the duration between two consecutive visits by the user—to measure engagement. The intuition is that if a user finds a website more exciting and engaging, they will return to it sooner rather than later. The study's main goal was to identify observable correlations between absence time and user engagement.

Most works in this area related to data from QA communities are mainly focused on the data from a few larger communities, such as Stack Overflow (Ortega et al. 2014). For example, in Pudipeddi et al. (2014), authors investigated the factors that correlate with user churn on Stack Overflow, including the time gap between posts, answering speed, number of answers received by the user, and reputation of the users who answered to the user's questions. Their findings suggest that the time gap between subsequent posts is the most significant indicator of an increase in their interest in staying engaged. Additionally, in Adaji and Vassileva (2015), the authors studied the problem of expert churn on Stack Overflow, formulated as a classification task. To label the users who left the community, the authors used the definition from Karnstedt et al. (2010), where a churner is defined as a user whose average activity over a specific subsequent period has dropped to less than a fraction of their average activity in a previously observed period. In other words, if a user has a noticeable drop in his activity following a period of considerable activity (e.g., answering multiple questions for a period then stopping) he is considered a churner. To that aim, they used four machine learning methods to predict the churn of expert respondents on Stack Overflow. Their result indicated that the random forest had the highest classification accuracy of the four machine mining algorithms and the highest values for the other evaluation metrics.

With the recent success of deep learning methods in tackling problems in domains such as computer vision and natural language processing, interest in the use of artificial neural networks for handling the censored data used in survival analysis has drastically increased. For example, in Yao et al. (2017), the authors proposed a deep correlational survival model (or DeepCorrSurv for short), which, in contrast with traditional survival analysis methods, can handle multimodal data. In essence, DeepCorrSurv is able to learn the complex interdependencies on multimodal patient data (e.g., the mixture of images and features). Furthermore, recurrent neural network-based approaches also have been successfully combined with survival analysis techniques to predict the events regularly occurring, such as the time to check-in by the user to a venue (Yang et al. 2018), and for content recommendation and personalization (Jing and Smola 2017).

Finally, Table 1 shows the information about the differences between the work described in this paper and in the literature. Based on the information presented in Table 1, the topic of user disengagement analysis for QA communities has been investigated using two main approaches: as a classification or as a survival analysis task. Both approaches have three major components: data, model, and disengagement criterion. Furthermore, the approaches mentioned above (i.e., classification and survival analysis) are sufficiently different in their goals and assumptions. Arguably, the most pronounced difference between the two approaches is that when the disengagement is formulated as a classification task, the time is not considered. In other words, it is assumed that the probability that a user gets disengaged is constant and independent of the time. In contrast, the central notion behind the survival analysis is that the probability of a user becoming disengaged is a function of time. Furthermore, the main goal of survival analysis is to find a good estimate of the survival function, which outputs the probability of the event of interest not happening (in our case, the likelihood that the user stays engaged) at a specific time. In this regard, the main benefit of the survival analysis approach is the possibility of taking into account the censoring of the data. These properties make the work presented in this study different from existing works that use a classification-based approach. In addition, the remaining existing works using survival analysis utilised the Cox model. Some recent studies (e.g., in Miao et al. 2015) suggest that the Cox model, compared to a random survival forests model, may have a weaker discriminative power. The main reason for this can be because the Cox model can only infer the linear effects between the target and independent variables, while RFS can

Table 1 Difference with the work in the literature

References	Data and models	Disengament/churn criteriun
Yang et al. (2010)	Authors used data from three QA communities for a period of 2 years. The main model used was Cox model (Cox 1972)	User inactivity over 100 days
Dror et al. (2012)	Authors used data from Yahoo! Answers for a period of about nine months. The churn prediction was formulated as a binary classification. Altogether, seven learners were used: the majority, naive Bayes, logistic regression, SVM, decision tree, random forests, and KNN	User inactivity after his first week of joining
Dupret and Lalmas (2013)	Authors used data from Yahoo! Answers Japan for a period of two weeks. Cox model (Cox 1972) was used	User absence time in days
Pudipeddi et al. (2014)	Stack Overflow data for a period of 4 years (from 2008 to 2012) were used. User churn prediction was formulated as a binary classification task, and three types of classifiers were used. Namely, SVM, decision tree, and logistic regression	No new post by user for six months or more
Adaji and Vassileva (2015)	Authors used the data from Stack Overflow from a period of 6 years (from 2008 to 2014). The problem of predicting the expert users' churn was formulated as a binary classification task, and four learners were used. Namely, logistic regression, multi-layer perceptron, random forests, and SVM	Decrease in user activity during a follow-up period relative to his activity during a previous observation period
The approach in this work	Data from three QA communities were used; namely, Politics, Data Science, Computer Science Stack Exchange. The data include the entire lifespan of communities. The prediction of user disengagement was formulated as a survival analysis task, and two methods were used. Namely, Kaplan–Meier and random survival forests	User absence for an extended period of time

handle nonlinearity (Wang et al. 2019). Furthermore, RFS is a nonparametric method; compared to the Cox model, which is a semi-parametric method, it is more versatile because it does not make any assumption about the underlying distribution of the data.

3 Preamble

3.1 Survival analysis

Survival analysis or time-to-event analysis (Cox and Oakes 2018) is a set of statistical models and methods for estimating the time it takes for a particular event of interest to happen. In a typical survival analysis task, a group of individuals (e.g., patients) are observed for a period. For each individual, the time when the event of interest happened is recorded. Usually, the event will not occur for all the individuals in the period of observation. The situation when the event of interest did not happen for an individual during the observation is called *censoring*. The goal of survival analysis is to find the probability of happening of the event of interest. In this regard, survival analysis is similar to regression analysis but with a major difference, where survival analysis models take into account the information related to individuals for whom the event did not take place, i.e., the censored individuals. This difference allows for obtaining more accurate estimations. Although survival analysis originated from the field of medical research, mainly for estimating the time a patient would live after being diagnosed having a deadly disease such as breast cancer, it has gained much attention in other areas such as customer churn analysis and prediction (Dias et al. 2020) and time to occurrence of a fault in a system (Widodo and Yang 2011).

Formally, $T \geq 0$ is a random variable that models the time for an event of interest to happen; $f(t)$ and $F(t)$ are its probability distribution and cumulative probability distribution, respectively.

$$F(t) = \int_{-\infty}^{t} f(x)\,\mathrm{d}x \tag{1}$$

Furthermore, $S(t)$, called the survival function, is defined as the probability that the event did not happen before time t. (Typically, when $S(t)$ is plotted, it is called the survival curve.)

$$S(t) = P[T > t] = 1 - F(t) \tag{2}$$

The hazard function $h(t)$, is the instantaneous occurrence rate of the event of interest, and is defined as:

$$h(t) = \lim_{\mathrm{d}t \to 0} \frac{P[t \leq T < T + \mathrm{d}t | T \geq t]}{\mathrm{d}t} = \frac{f(t)}{S(t)} \tag{3}$$

Survival and hazard functions can be connected via the following formula:

$$S(t) = e^{-\int_0^t h(x)\,\mathrm{d}x} \tag{4}$$

Given n individual samples, each sample $i \in [1...n]$ is represented as triplet (A^i, E^i, T^i) where:

- $A^i \in R^d$ is a d-dimensional real-valued vector of individual features (i.e., user attributes in our context);
- $E^i \in \{0, 1\}$ is the variable indicating the event of interest happened when $E^i = 1$ or not (censored) when $E^i = 0$, for individual i during the observation;
- $T^i = min(t_i, t_{end})$ is the time when the event happened for individual i during the observation period; t_{end} is the time when the observation was ended. $T^i = t_{end}$ (i.e., event did not happen) indicates sample i is censored.

The main task of the survival analysis methods is to estimate $h(t)$ and $S(t)$.

3.2 Kaplan–Meier estimator

Kaplan–Meier estimator (Kaplan and Meier 1958) is a nonparametric model that calculates the survival function $\hat{S}_{KM}(t)$ of a homogeneous cohort, i.e., the individuals in the same cohort (or group) share the same survival function. Given N individual samples in a cohort, it assumes that there are J distinct actual event times such that $t_1 < t_2 < \cdots < t_J$ when $J \leq N$, then:

$$\hat{S}_{KM}(t) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{n_j}\right), \tag{5}$$

where d_j is the individuals who experienced an event and n_j is the number of individuals that did not experience the event in time interval $[t_{j-1}, t_j]$.

Kaplan–Meier method only uses the information from E^i and T^i to estimate the survival function.

3.3 Random survival forests

Ishwaran et al. (2008) proposed the random survival forests (RSF) model, which is an extension to the random forests ensemble model (Breiman 2001) for working with censored data. The general idea for creating an RSF model for a particular dataset is as follows (Utkin et al. 2019; Ishwaran et al. 2008):

1. Bootstrap q samples from the data, where q is the number of trees. On average, each sample excludes 37% of the original data as out-of-bag (OOB) data.
2. Grow a survival tree for each bootstrap sample. At each node of the tree, select $\sqrt{m}$ (i.e., a subset of variables used during the node split) candidate variables. Then split the node using the variable that maximises the survival difference between its children nodes.
3. Furthermore, grow the tree to be full under the constraint where no leaf node should have less than $d > 0$ deaths. The value of d is a hyperparameter, similar to q, which is chosen to produce the best results.
4. Compute the cumulative hazard function (or the survival function) for each tree.
5. Use the OOB data to calculate the prediction error for the ensemble cumulative hazard function (or the survival function).

Different implementations of RSF mainly differ in their splitting rule. Ideally, the splitting rule should maximise the survival difference across two dataset partitions. In this paper, we used the implementation from PySurvival library (Fotso et al. 2019).

3.4 Concordance index

The concordance index (or C-index for short) is a generalisation of the area under the ROC curve (AUC), which supports censored data (Harrell et al. 1982). The C-index widely is used as an evaluation metric of the performance of survival models. It summarises the model's discriminatory power, which is how well a model can rank the survival times of samples. Similar to AUC, the value of the C-index ranges from 0.5 to 1, where 1 indicates the best performance.

More formally, given $S(t)$ be the survival function estimated by some survival model, let $t_1^*, \ldots, t_s^*$ be a set of fixed time points, e.g., $t_1, \ldots, t_N$ where N is a distinct time index. Then C-index is defined as:

$$C = \frac{1}{M} \sum_{i:E^i=1} \sum_{j:t_i<t_j} \mathbf{1}[S(t_i^*) > S(t_j^*)], \tag{6}$$

where M is the total number of comparable pairs and $\mathbf{1}[.]$ is a function that will return 1 if its input argument is true or 0 otherwise. Note that there are slightly different definitions for C-index in other works. In this work, we used the definition proposed by Utkin et al. (2019).

3.5 Log-rank test

The log-rank test (Mantel 1966) is a nonparametric statistical test for comparing the hazard functions, i.e., $h(t)$, of two cohorts/groups of individuals. The null hypothesis is

Table 2 Information about the datasets

Characteristic	Dataset		
	Pol	DS	CS
Number of questions	12,416	28,950	40,792
Number of users	31,242	100,582	113,434
Number of answers	25,909	32,334	46,785
Number of comments	135,648	64,244	167,038
Year founded	2012	2014	2008

that the hazard functions of two groups, e.g., group 1 and 2, are equal, i.e., $h_1(t) = h_2(t)$. The Log-rank test assumes that survival probabilities (i.e., the probabilities of not becoming disengaged in our context) stay the same over time. It is widely used to check whether the underlying survival distributions of two groups are the same or are different, essentially.

4 Data

4.1 Data description

As mentioned earlier, we used data from three online QA platforms that are Politics (Pol), Data Science (DS), and Computer Science (CS) Stack Exchange (SE). Pol SE is an ad-hoc QA community focused on politically-themed content, such as questions related to the nature of democracy and the state of human rights. DS SE covers topics concerning the widespread field of data science. And CS SE covers topics related to computer science. We chose these three communities for two reasons. Firstly, although the sizes of these communities are smaller than the sizes of some other QA communities hosted on SE like Stack Overflow, the chosen communities are thriving in their niche. Secondly, each of these communities is more or less focused on separate fields that, although they might share some topics, are different enough to be viewed as distinct. It allows us to search for possible patterns related to disengagement, regardless of the specific topics of a field.

The datasets of the three communities were downloaded from the Stack Exchange data dump available on Archive.org.[5] The data included the complete historical information about the questions and answers posted on the three QA communities from their inception until May 2021. Table 2 shows the general information about the datasets.

4.2 Community characteristics

Table 3 includes the summary statistics for users belonging to three communities whose data are used in this study. The information shown in the table was extracted from the corresponding *Users* table for each community from the Stack Exchange data dump. Based on the information presented in Table 3, for all three communities the distribution of first four attributes (i.e., *user reputation*, *profile views*, *upvotes*, and *downvotes*) seems to follow a

Table 3 Summary statistics of users in each community

Community	Attribute	Statistics				
		Mean	STD	Median	First quartile	Fourth quartile
Pol	User reputation	160.50	1377.06	101	1	101
	Profile views	10.01	107.82	0	0	1
	Upvotes	12.38	126.12	0	0	2
	Downvotes	2.35	56.70	0	0	0
	Year joined	2017	1.94	2018	2016	2019
DS	User reputation	50.78	195.70	1	1	101
	Profile views	1.58	24.44	0	0	0
	Upvotes	1.36	26.76	0	0	0
	Downvotes	0.13	10.47	0	0	0
	Year joined	2018	1.81	2018	2017	2020
CS	User reputation	67.50	962.95	3	1	101
	Profile views	3.60	119.03	0	0	1
	Upvotes	2.39	94.92	0	0	0
	Downvotes	0.31	32.94	0	0	0
	Year joined	2017	2.37	2017	2015	2019

[5] https://archive.org/download/stackexchange; the data are available under Creative Commons licences.

Table 4 Information about behavioural user attributes

User attribute	Description
A_1^i	Number of downvotes cast by user i
A_2^i	Number of upvotes cast by user i
A_3^i	Number of questions posted by user i
A_4^i	Number of answers posted by user i
A_5^i	Number of comments written by user i

heavy tail distribution due to large size of dispersion (i.e., STD) around the mean. Moreover, relative to users from the other two communities, users in Pol SE show more intensity of activity on average. This is apparent based on the observation that although the number of registered users on Pol SE is smaller than the number of registered users on the other two communities, nonetheless, the average values of the first four user attributes listed in Table 3 is noticeably larger. Additionally, regarding the trend of about the increase in the number of registered users in each community, DS SE had the fastest growth relative to two other communities. It took only 4 years for DS SE to reach half of its registered users, while in comparison, the number of years it took for Pol SE and CS SE to reach half of their registered users were 6 and 9 years, respectively. Furthermore, all three communities show a large increase in the number of users lately (i.e., around 2018 onwards) which we suspect to be due to the *tipping point* phenomenon (Singh et al. 2020).

4.3 User attributes

The bulk of users in QA platforms do not make any contributions. These users, who are referred to as lurkers in some previous work (e.g., Tagarelli and Interdonato 2018), can be differentiated from the normal users, who include the experts, by their level of contribution to the platform. In this work, two categories of user attributes were used in order to investigate the relationship between the user contributions and the probability of disengagement. Namely, *behavioural attributes* and *content-based attributes*.

4.3.1 Behavioural attributes

We identified five user attributes in the datasets that directly correspond to the level of user contribution. These attributes primarily are based on the information related to user behaviour that seems crucial to the proper functioning of the platform. Table 4 includes the name and description of these attributes.

4.3.2 Content-based attributes

In addition to behavioural attributes, we picked up a set of content-based user attributes. These attributes hint at how the contributions made by each user might have been perceived favourably by the community, i.e., other users. The primary motivation is that users can indirectly contribute to the platform, e.g., by asking a question that starts a stream of debates over a controversial topic such as *refugee crisis* in the context of the Pol SE community. And the information about this type of indirect user contribution, which is not only limited to the behaviour of a particular user, can be extracted and utilised from user content (e.g., mainly from metadata of users' posts). Table 5 includes the name and description of the content-based attributes employed in this work.

4.4 User representation

Based on the two types of user attributes mentioned earlier, we constructed a numerical vector for each user, called *user representation vector* (or URV for short). The URV of user i is defined as:

$$URV_i = (A_1^i, A_2^i, \dots, A_j^i) \tag{7}$$

Table 5 Information related to the content-based attributes

User attribute	Description
A_6^i	Average number of times questions posted by user i were viewed
A_7^i	Average number of comments written for questions posted by user i
A_8^i	Average score (i.e., sum of upvotes and downvotes given to question) of the questions posted by user i
A_9^i	Average score (i.e., sum of upvotes and downvotes given to answer) of the answers posted by user i
A_{10}^i	Average number of comments written for answers posted by user i
A_{11}^i	Average number of upvotes given to the comments posted by user i

Table 6 Sets of users defined by dichotomizing users based on their behavioural attributes corresponding to their contribution

User set	Definition
Q	Users who posted at least one question
$\overline{Q}$	Users who posted no question
A	Users who answered at least one question
$\overline{A}$	Users who did not answer any question
C	Users who commented on at least one question/answer
$\overline{C}$	Users who did not answer any question
U	Users who upvoted at least one question/answer/comment
$\overline{U}$	Users who did not upvoted
D	Users who downvoted at least one question/answer
$\overline{D}$	Users who did not downvote

where A_j^i is the corresponding user attribute from Tables 4 and 5.

4.5 Dichotomizing users

Moreover, we dichotomised users into pairs of disjoint sets (or groups) using each one of behavioural user attributes. The main idea is that users can be partitioned into two groups naturally, where the criterion for the split is whether a user has made a particular type of contribution or not. Categorising users in two disjoint sets based on the value of a behavioural attribute allowed us to investigate the importance of one specific user attribute, e.g., by comparing the survival curves of two disjoint sets of users who posted at least one question and who did not. Table 6 includes the information about each group of users and its counterpart.

4.6 Disengagement criterion

What amounts to the event of a user becoming disengaged is domain-dependent and thus can vary in different settings. For example, normally, in medical research, the event usually is the patient's death (Cox and Oakes 2018). In this work, suitable to our need, we opted to use the information about the last time a user visited the QA platform to detect disengagement. The information about the last time a user visited is available in the *LastAccessDate* column from the *Users* table in each dataset. The activity time of a user was calculated based on the difference in the number of months since the user joined the platform until the last recorded activity time of the user (i.e., *LastAccessDate* associated with the user). For user i, if the number of months since his last visit to the platform exceeded a certain threshold value, he would be tagged as a disengaged user (i.e., $E^i = 1$); otherwise, the user's state was considered still active (i.e., $E^i = 0$) which means the information about user's disengagement was censored. More precisely, let t_l^i be the last time user i been seen visiting the platform, and θ be the threshold value, then the value of E^i would be set based on the following relation:

$$E^i = f_\theta(t_l^i) = \begin{cases} 0, & \text{if } \hat{d}(t_l^i, t_d) \leq \theta \\ 1, & \text{otherwise} \end{cases} \tag{8}$$

where t_d is the last recorded time in the dataset, and $\hat{d}(t_l^i, t_d)$ is the time difference between t_l^i and t_d in months. In this work, we used two threshold values (i.e., θ) of 24 and 36 months. Subsequently, users who had not visited the platform for more than 2 and 3 years were considered disengaged.

Table 7 Sizes of each user set per dataset

User set	Dataset		
	Pol	DS	CS
$\lvert Q\rvert$	3775 (12%)	16,041 (16%)	20,841 (18%)
$\lvert\overline{Q}\rvert$	27,466 (88%)	84,540 (84%)	92,592 (82%)
$\lvert A\rvert$	3743 (12%)	7226 (7%)	7020 (6%)
$\lvert\overline{A}\rvert$	27,498 (88%)	93,355 (93%)	106,413 (94%)
$\lvert C\rvert$	6358 (20%)	12,056 (12%)	16,788 (15%)
$\lvert\overline{C}\rvert$	24,883 (80%)	88,525 (88%)	96,645 (85%)
$\lvert U\rvert$	11,025 (35%)	16,328 (16%)	20,767 (18%)
$\lvert\overline{U}\rvert$	20,216 (65%)	84,253 (84%)	92,666 (82%)
$\lvert D\rvert$	1596 (5%)	732 (1%)	1143 (1%)
$\lvert\overline{D}\rvert$	29,645 (95%)	99,849 (99%)	112,290 (99%)

5 Results

5.1 Results from Kaplan–Meier

Kaplan–Meier method was used to estimate (within the confidence interval of 95%) the survival functions (i.e., $S(t)$) of sets of users dichotomised based on the definitions shown in Table 6. The implementation from Lifeline library (Davidson-Pilon et al. 2021) was used to produce the survival curves. Figures 2, 3, and 4 show the survival curve of each pair sets of users for {Pol, DS, CS} SE datasets, respectively. Table 7 includes the information about the size and proportion of each user set dichotomised based on a single behavioural attribute. As mentioned earlier, for each user, the label indicating whether the user is disengaged or not (i.e., E^i) was censored if the difference between the user's last activity time and t_d was less than or equal to 24 and 36 months, respectively. Log-rank test (with $p < 0.005$) was performed on each pair of curves.

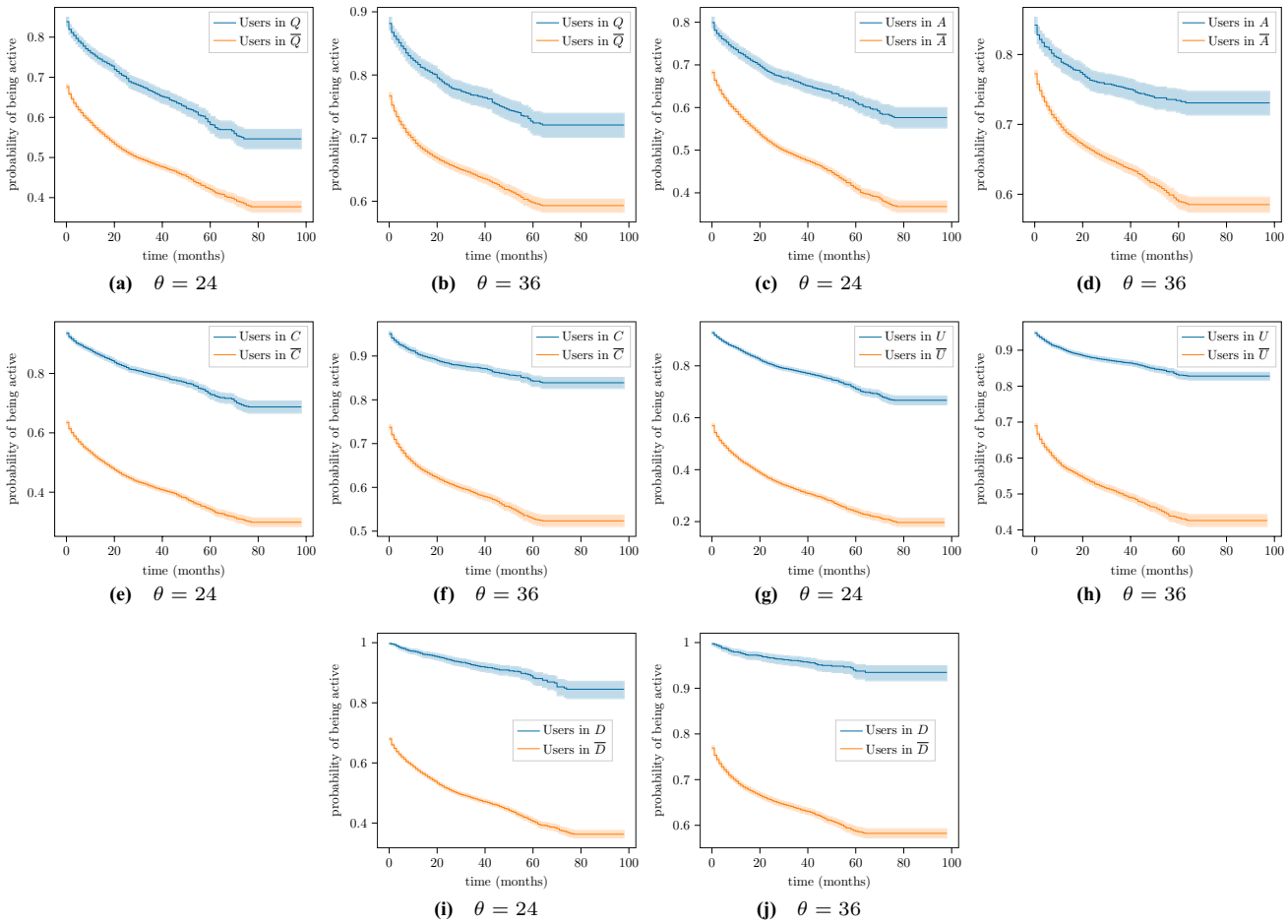


Fig. 2 Survival curves for Pol SE dataset estimated using Kaplan–Meier method

5.2 Results from RSF

We used k-fold cross-validation (with k=5) in 30 runs to make the predictions (using RSF models). The values of model hyperparameters such as the number of trees (i.e., q) have been tested in order to choose the ones that lead to the best results. We used C-index to evaluate the performance of the models. Only data for users with contributions were used to train and evaluate the RSF-based models. In other words, only the information of users belonging to $Q \cup A \cup C \cup U \cup D$ (from Table 6) was used. For each user, three URVs were constructed, using the behavioural user attributes only, content-based attributes only, and finally, a combination of both. Table 8 includes the average C-indexes computed for the RSF models over the runs.

5.3 Attribute importance

We used the permutation importance measure (Breiman 2001; Molnar 2020) present in RSF models to rank each user attribute. The attributes which permuting their values caused more significant prediction errors ranked more important. We chose permutation importance mainly because of its intuitive definition and subsequent interpretation, which is based on the idea that the importance of a variable is the increase in model error when the variable's information is destroyed via value permutation (Molnar 2020). Moreover, it provides a compact global insight into the model's behaviour. Figure 5 shows the per-dataset average permutation importance of each attribute on the prediction results of the RSF models used in this work.

6 Discussion

Based on results from the Kaplan–Meier method, we observed: (i) the underlying hazard function of each set of users seems to be different; (ii) the probability that users with even a few contributions (e.g., the user asked one question) are noticeably higher than other users who did not contribute to the platform. We observed a distinctive difference between the survival functions of the users who contributed

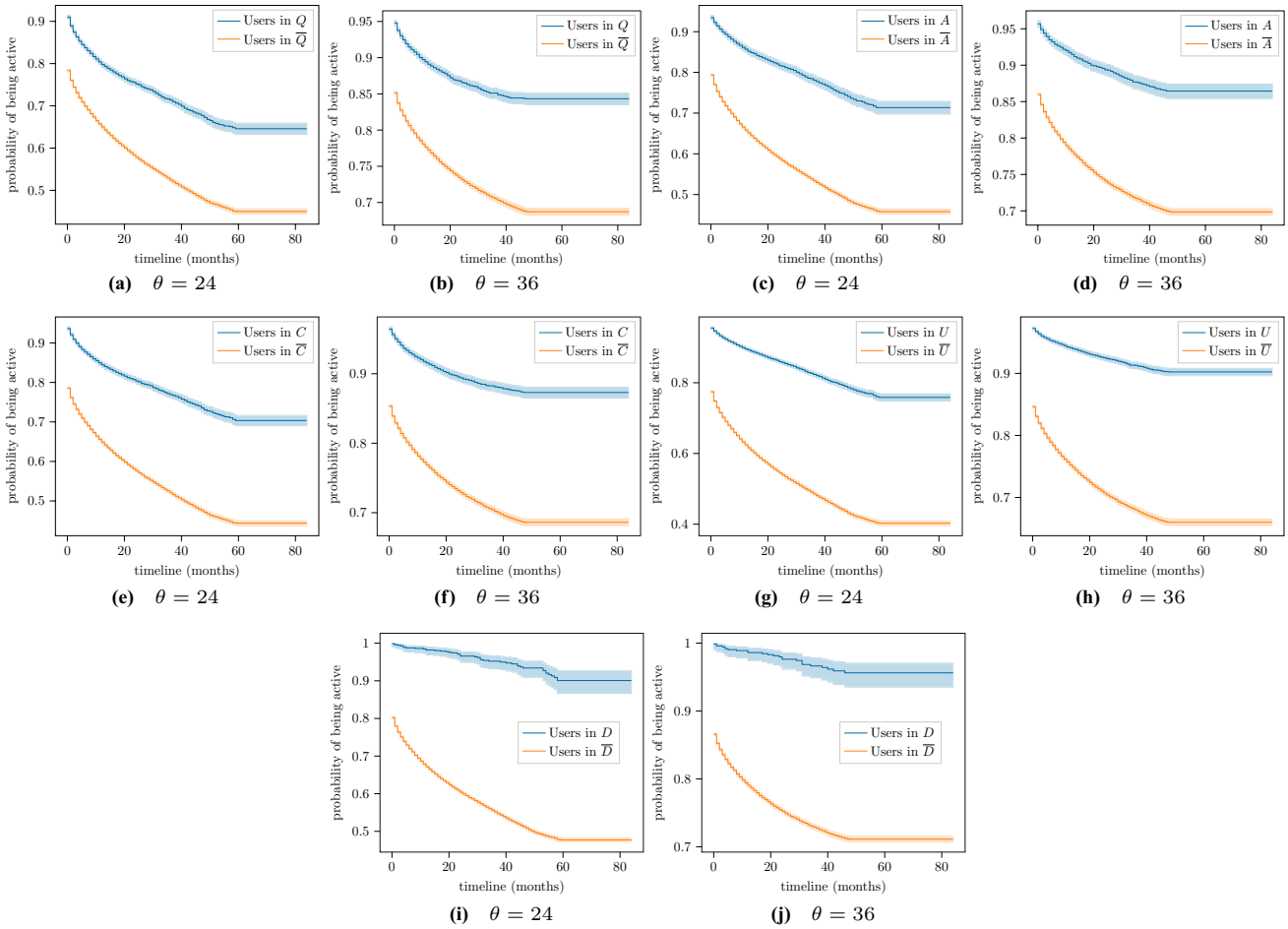


Fig. 3 Survival curves for DS SE dataset estimated using Kaplan–Meier method

to the platform and those who did not contribute. This pattern, which is present in all three datasets regardless of the community niche, confirms the finding from previous related studies such as the ones reported in Joyce and Kraut (2006) and Yang et al. (2010) that suggested that users with even a few initial contributions are more likely to stay loyal than users without any contributions. The latter make up the bulk of the users. Furthermore, the gap between the probability of disengagement of two groups seems to widen over time.

Predictions using RSF models show relatively similar patterns on all three datasets. Results (shown in Table 8) indicate that the inclusion of the information of behavioural attributes leads to better predictions compared to the use of content-based user attributes only. Furthermore, using the mixture of the information of behavioural and content-based attributes yielded a slight improvement. The value of θ does not seem to affect the overall results.

Based on the permutation importance of attributes (see Fig. 5), behavioural features play a more salient role in the output of the RSF models. On average, 4 out of 5 top attributes with the most permutation importance are from behavioural user attributes. The number of upvotes (i.e., A_2) received the highest importance in all three datasets. Subsequently, with a noticeable difference, the average number of the times user questions were viewed (i.e., A_6) received relatively high importance. We suspect that the user's higher upvotes might show that they hold a favourable view of the community (or platform) in general. On the other hand, the information about the number of downvotes did not contain much predictive information. We suspect it could be due to the small number of users with downvoting activity in the datasets.

7 Limitations and future work

There are a few limitations regarding the work done in this paper. The datasets were used only from three QA communities hosted by (the larger) Stack Exchange platform. Consequently, this work did not investigate and compare disengagement on other major platforms such as Quora. It seems interesting to compare our results with the results

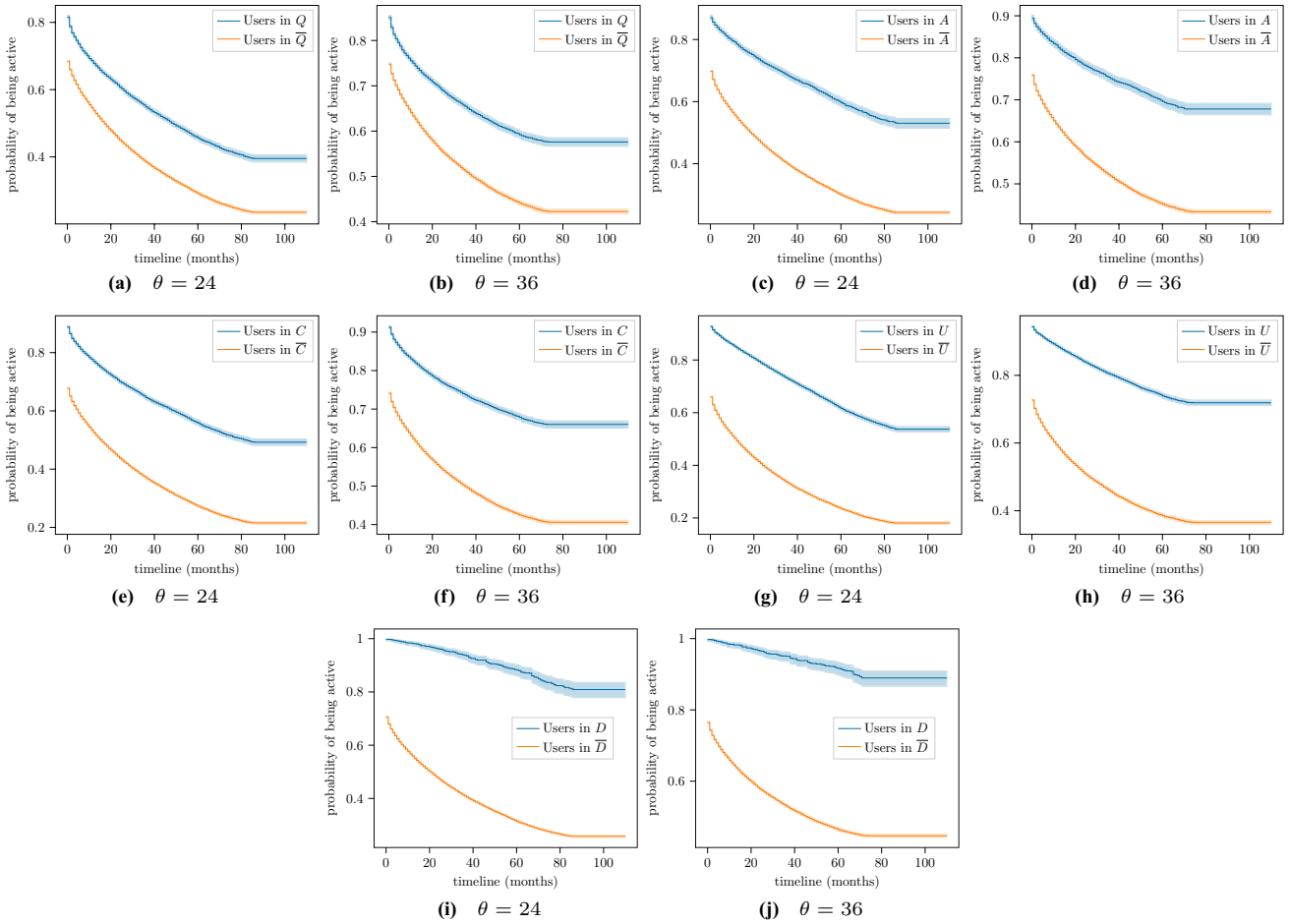


Fig. 4 Survival curves for CS SE dataset estimated using Kaplan–Meier method

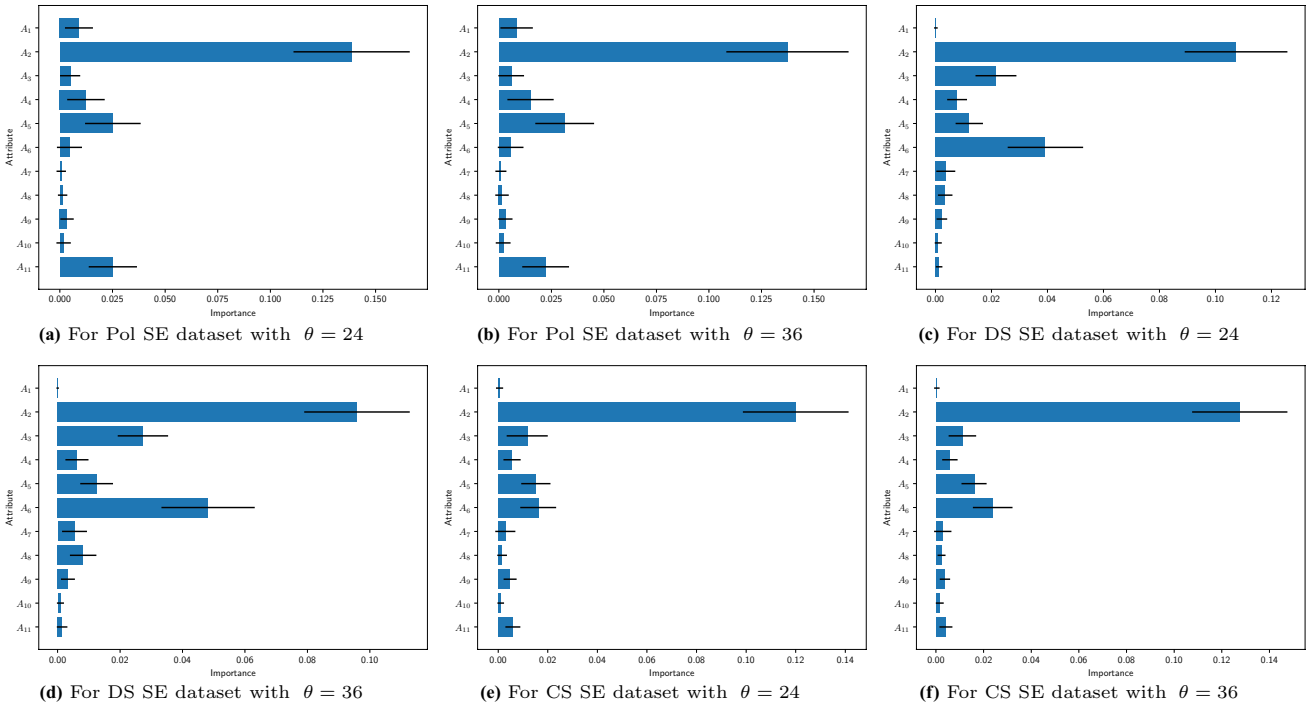
obtained with data from other major QA platforms in future. Conventional assumptions related to the application of survival analysis techniques hold over our results, e.g., the assumption that the probabilities of disengagement of censored and none censored individuals are essentially the same. We used user inactivity for an extended period (e.g., 2 years passed since the user visited the community web pages) to distinguish between disengaged and censored users. This required the use of a time threshold in which its value is set experimentally, not based on a well-defined rule. Finally, for most users, their behavioural information does not exist, making it hard to investigate further the survival probabilities of users dichotomised based on the definitions given in Table 6.

Including the data from a more numerous and diverse set of QA platforms could be interesting for future work. Furthermore, the information related to the body of the posts (e.g., the text of questions and answers) of each user could be utilised to find the probabilities of disengagement. Additionally, methods and models that do not assume the probabilities for the disengaged and censored users are the same can be used, which theoretically should lead to better predictions.

Temporal context can tentatively play an essential role in the intensity of user activities in a community and subsequently be an informative factor in the level of user

Table 8 Average C-index for RSF models using different attribute sets; higher C-index indicates better prediction

Dataset	Attributes	$\theta = 24$		$\theta = 36$	
		Mean	STD	Mean	STD
Pol	Behavioural only	0.75	0.01	0.76	0.01
	Content-based only	0.68	0.01	0.68	0.01
	Behavioural plus content-based	0.75	0.01	0.76	0.01
DS	Behavioural only	0.66	0.01	0.66	0.01
	Content-based only	0.61	0.01	0.63	0.01
	Behavioural plus content-based	0.68	0.01	0.70	0.01
CS	Behavioural only	0.68	0.00	0.68	0.01
	Content-based only	0.62	0.01	0.63	0.01
	Behavioural plus content-based	0.69	0.01	0.68	0.01



(a) For Pol SE dataset with $\theta = 24$ **(b)** For Pol SE dataset with $\theta = 36$ **(c)** For DS SE dataset with $\theta = 24$

(d) For DS SE dataset with $\theta = 36$ **(e)** For CS SE dataset with $\theta = 24$ **(f)** For CS SE dataset with $\theta = 36$

Fig. 5 Average permutation importance of each attribute; models are trained and evaluated using behavioural and content-based attributes simultaneously over the datasets shown in Table 2

engagement. By temporal context, we mean the effects of real-world events occurring within a specific time period on the behaviour of users of a QA community. Examples of such events include Brexit and the political campaigns during an election in an influential country such as the USA.

8 Conclusion

We used survival analysis to study user disengagement using the historical data from three distinct QA communities from their inception to May 2021. We employed two categories of user attributes and investigated the importance of these attributes. Our results confirm the previous findings that users with some initial contributions (e.g., questions and answers) are likelier to stay active longer than users who contributed nothing. Furthermore, based on our results, behavioural user attributes can be used to estimate the disengagement probability of each user with reasonable accuracy.

Moreover, based on the importance of attributes used to train and evaluate the models, how favourable users see the content posted on the platform seems to affect the disengagement time.

Acknowledgements I thank the reviewers and my colleague Yanzhe Bekkemoen for helping me improve the manuscript.

Funding Open access funding provided by NTNU Norwegian University of Science and Technology (incl St. Olavs Hospital - Trondheim University Hospital). This work was carried out as part of the Trondheim Analytica projecthttps://www.ntnu.edu/trondheimanalytica, supported by NTNU's Digital Transformation programme.

Data availability The data used in this study are publicly available from Archive.org under Creative Commons licences. Furthermore, for reproducibility, the code and other related artefacts, such as the preprocessed version of the data used in the experiments, are also available on GitHub.comhttps://github.com/habedi/SurvivalAnalysisQACommunities.

Declarations

Conflict of interest I do not have any conflicts or competing interests to declare.

References

Adaji I, Vassileva J (2015) Predicting churn of expert respondents in social networks using data mining techniques: a case study of stack overflow. In: ICMLA. IEEE, pp 182–189

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Cox DR (1972) Regression models and life-tables. J R Stat Soc Ser B (Methodol) 34(2):187–202

Cox DR, Oakes D (2018) Analysis of survival data. Chapman and Hall, London

Davidson-Pilon C, Kalderstam J, Jacobson N et al (2021) CamDavidsonPilon/lifelines: 0.26.0. 10.5281/zenodo.4816284

Dias J, Godinho P, Torres P (2020) Machine learning for customer churn prediction in retail banking. In: Computational science and its applications, pp 576–589

Dror G, Pelleg D, Rokhlenko O et al (2012) Churn prediction in new users of Yahoo! Answers. In: WWW, pp 829–834

Dupret G, Lalmas M (2013) Absence time and user engagement: evaluating ranking functions. In: WSDM, pp 173–182

Fotso S et al (2019) PySurvival: open source package for survival analysis modeling. https://www.pysurvival.io/

Guan T, Wang L, Jin J et al (2018) Knowledge contribution behavior in online Q &A communities: an empirical investigation. Comput Hum Behav 81:137–147

Harrell FE, Califf RM, Pryor DB et al (1982) Evaluating the yield of medical tests. JAMA 247(18):2543–2546

Ishwaran H, Kogalur UB, Blackstone EH et al (2008) Random survival forests. Ann Appl Stat 2(3):841–860

Jin J, Li Y, Zhong X et al (2015) Why users contribute knowledge to online communities: an empirical study of an online social Q &A community. Inf Manag 52(7):840–849

Jing H, Smola AJ (2017) Neural survival recommender. In: WSDM, pp 515–524

Joyce E, Kraut RE (2006) Predicting continued participation in newsgroups. J Comput Mediat Commun 11(3):723–747

Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J Am Stat Assoc 53(282):457–481

Karnstedt M, Hennessy T, Chan J et al (2010) Churn in social networks: a discussion boards case study. In: ICSC, pp 233–240

Kuzmeski M (2009) The connectors: how the world's most successful businesspeople build relationships and win clients for life. Wiley, Hoboken

Mantel N (1966) Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep 50(3):163–170

Miao F, Cai YP, Zhang YT et al (2015) Is random survival forest an alternative to Cox proportional model on predicting cardiovascular disease? In: MBEC, pp 740–743

Molnar C (2020) Interpretable machine learning. Lulu.com

Ortega F, Convertino G, Zancanaro M et al (2014) Assessing the performance of question-and-answer communities using survival analysis. arXiv preprint arXiv:1407.5903

Pudipeddi JS, Akoglu L, Tong H (2014) User churn in focused question answering sites: characterizations and prediction. In: WWW, pp 469–474

Rothmeier K, Pflanzl N, Hüllmann JA et al (2021) Prediction of player churn and disengagement based on user activity data of a freemium online strategy game. IEEE Trans Games 13(1):78–88

Singh A, Dharamshi N, Thimma Govarthanarajan P et al (2020) The tipping point in social networks: investigating the mechanism behind viral information spreading. In: BigDataService, pp 54–61

Stepanova M, Thomas L (2002) Survival analysis methods for personal loan data. Oper Res 50(2):277–289

Tagarelli A, Interdonato R (2018) Mining lurkers in online social networks: principles, models, and computational methods. Springer, Berlin

Utkin LV, Konstantinov AV, Chukanov VS et al (2019) A weighted random survival forest. Knowl Based Syst 177:136–144

Wang P, Li Y, Reddy CK (2019) Machine learning for survival analysis: a survey. ACM Comput Surv 51(6):1–36

Widodo A, Yang BS (2011) Machine health prognostics using survival probability and support vector machine. Expert Syst Appl 38(7):8430–8437

Yang J, Wei X, Ackerman M et al (2010) Activity lifespan: an analysis of user survival patterns in online knowledge sharing communities. In: ICWSM

Yang G, Cai Y, Reddy CK (2018) Spatio-temporal check-in time prediction with recurrent neural network based survival analysis. In: IJCAI, pp 2976–2983

Yao J, Zhu X, Zhu F et al (2017) Deep correlational learning for survival prediction from multi-modality data. In: MICCAI