The application of survival analysis on user-generated content networks

Masters Thesis

Anna Pisna

Spring Term 2017

Advisors:
Prof. Florian Stahl
Dr. Andreas Lanz

Chair of Quantitative Marketing and Consumer Analytics
L5, 2 - 2. OG
68161 Mannheim
www.quantitativemarketing.org

Table of contents

Abstract

This study examines the potential of survival analysis as a tool to predict relationship duration in user-generated content networks on the example of SoundCloud. Survival analysis is a highly versatile method with the main advantage of overcoming censoring biases in user interaction data. This research focuses on three key objectives: exploring user relationship dynamics of SoundCloud; defining, testing, and interpreting survival models for SoundCloud user relationships; and finally, creating a general framework for the application of survival analysis on user-generated content networks in practice. The most challenging part of this research is defining the end of a relationship between two users on an online social media platform, since this event is rarely observable and abrupt. However, with a clearly defined survival event, survival analysis is a powerful tool that can assist in decision making within and beyond the domain of user-generated content networks.

Keywords: survival analysis, user-generated content networks, online social media, online relationships.

1. Introduction

In the age of online media and big data, user-generated content has become a conventional way of self-expression for almost everyone who uses online social media. With numerous platforms, social media users are able to share their opinions, make friends, do businesses, shop, work and create etc. just with a click of a button. The rise of social media did not only transform everyone's daily habits, but also changed the way people interact and build relationships.

In addition to that, companies finally have an opportunity to get to know their customers in ways they never could before. With the development of technology and the integration of online habits into people's routines, it is now possible to track and collect data on customers' social interactions, interests, purchases, hobbies, media behavior etc. However, together with this opportunity, there comes a great complexity in organizing this highly granular information, analyzing the data, and deriving actionable insights to improve the business performance.

This paper discusses online relationships on the example of SoundCloud user data, and suggests an alternative method to estimate duration of user relationships – survival analysis. Interaction between users on a platform is a key to success for an online social network: strong long-lasting user relationships are a foundation not only of the platform's sole existence, but also of the company's sustainable profits. While maintaining user engagement and interaction is the basis, knowing how users interact is what enables most efficient online marketing strategies.

The purpose of this study is to explore the potential of survival analysis as a tool in analyzing behavior in user-generated content networks. This highly versatile method can predict the probability that the relationship between two users will end within a predefined period of time, and in more complex specifications it allows to explain factors that contributed to the end of user relationship. The idea of application of survival analysis in marketing is not new (Kelsen, Schmittlein 1993 p.409; Li 1995 p.333). However, this method is new to the field of user-generated content networks.

Considering the lack of existing research on similar topics, this study has three key objectives. First objective is to explore user relationships on SoundCloud and to discover interaction patterns if there are such. Second objective is to test and compare several survival model specifications based on available data, and to prepare foundation for further research in this field. Last objective is to derive concrete managerial implications and create a framework on how to optimally use user relationship data, survival analysis, and its results to improve businesses.

The research process consisted of theoretical and empirical parts, as well as summarizing key findings in this paper. The theoretical part of this study was dedicated to the literature research of existing scholarship on related topics in fields of marketing, online social media, psychology, sociology, customer relationship management, epidemiology, and statistics. The empirical part of this study consisted of four stages: data cleaning, data exploration, data analysis, and results interpretation. Different stages of the empirical process required different software. Therefore, data cleaning was executed using Microsoft SQL Server, while the rest of the analysis was conducted in R.

The paper is structured as follows: this introductory section provides the study overview, as well as discusses research backgrounds, novelty and importance of the subject, purpose, key objectives, and methodology. Section 2 is dedicated to the foundations of survival analysis and its key definitions. It also explains origins and advantages of the method, as well as gives an overview of most common types of survival models. Next three sections are dedicated to the three respective objectives of this study. Section 3 focuses on relationships in user-generated content networks: subsection 3.1 provides theoretical discourse on this topic, and subsection 3.2 discloses first results of empirical research, and describes SoundCloud user relationship characteristics. Section 4 is an empirical-focused part, which puts survival analysis to practice on a SoundCloud data sample. It consists of three subsections. Subsection 4.1 discusses challenges and decisions of defining survival variables in the context of online relationships, in subsection 4.2 different sets of definitions are being applied in survival, and subsection 4.3 discusses the results. In Section 5, key findings of previous sections are evaluated from a managerial perspective, and a general framework of the application of survival analysis for practitioners is formulated.

2. Foundations of survival analysis

The term survival analysis refers to a broad spectrum of methods, which can be characterized as a "collection of statistical procedures for data analysis for which the outcome variable of interest is time until an event occurs" (Kleinbaum and Klein 2006, p.4). The event of interest can vary depending on the field of research. For example, in medicine and biology it is often death, disease occurrence or recovery; in sociology, it could be a divorce, a criminal's rearrest; in marketing a new product takeoff, etc. In this study, an event of interest is the end of user's relationship, which will be discussed in more detail in Section 4.1.

The time from the beginning of a survival study and until the event occurrence can be measured in years, months, days or minutes depending on the subject, in this study, survival time is going to be measures in days. Alternatively, it can be referred to as the age of an individual at the moment of occurrence of the event. Typically, only one event is defined as an event of interest, even though events can occur several times. For example, if the event is a marriage, the individual can be married more than once within a certain timeframe, but only one of these events will be considered as the event of interest.

In survival analysis literature, it is commonly accepted to refer to the time until event occurrence as the survival time, and the event of interest is typically called a failure. However, in some cases term “failure” refers to a positive outcome, such as a recovery from a disease or a project success.

One of the major advantages of survival analysis over other methods such as a linear regression is in its ability to deal with censored observations, are characterized by incomplete survival information (Kleinbaum and Klein 2006, p.7). Survival data is mostly a subject to the right-side censoring, which can happen due to the three following reasons: the event of interest did not occur within the study period (1); the individual is lost to follow-up (2); withdrawal from the study (3). Left-side censoring is also possible; however, it is not typical for survival data. For instance, if the event occurred before it was detected within the study it is the case of left-side censoring. Failure is a binary variable, meaning that it can only have two values: ‘1’ if the failure occurred within the timeframe of a study or ‘0’ if failure did not occur, which implies that the observation was censored.

Survivor functions, denoted by *S(t)* expresses the probability that an individual will survive within the timeframe *t*, or in other words, that the survival time denoted by *T* will exceed the time *t*. The equation below describes the survivor function analytically:

$$P(t) = \Pr(T > t)$$

Survival functions can be conveniently visualized with a figure of survival time plotted against probability of survival. While in theory they are supposed to be smooth curves, in practice survivor options are step functions due to limited samples. However, with large sample size like in SoundCloud user data, the survivor function plot looks more similar to the theoretical one.

Unlike survivor function, the hazard function denoted by *h(t)* focuses on the occurrence of the event. It represents the ‘instantaneous potential per unit time for the event to occur, given that the individual has survived up to time t (Kleinbaum and Klein 2006, p.11). Mathematically hazard function can be described in the following way:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t | T \ge t)}{\Delta t}$$

where the hazard function can be interpreted as a conditional failure rate, which is always non-negative and has no upper bound. Hazard functions can appear in different specifications, depending on the respective study, and are always related with the survivor function.

The goal of survival analysis usually includes one or several of the following objectives: estimation and interpretation of hazard and/or survival functions from survival data; comparison of these functions; assessment the relationship of the explanatory variables to the survival time (Kleinbaum and Klein 2006, p.15). For this paper two first objectives are more relevant, since there is limited access to explanatory variables.

The most common methods to analyze the survival data include parametric, semi-parametric, and non-parametric models. Parametric models underlie the assumption of a certain known pattern of the hazard model distribution, the best known of which are constant hazard or the exponential model, increasing and decreasing Weibull and lognormal distribution (Kleinbaum and Klein 2006, p.12-13). These models, as well as the semi-parametric regressions can be estimated using the maximum likelihood methods (Kay 1977, p.228).

The Kaplan Meier method is a commonly used non-parametric method in survival analysis, which estimates survival as a function of time. Its formula allows computing the survival probability at the failure time as a multiple of probability of surviving the previous time period and the conditional probability of surviving the previous time period (Kaplan, Meier 1958, p.362).

In this study, the empirical part will be dedicated to estimation and interpretation of Kaplan-Meier curves, as well as comparison for different model specifications. The Log-Rank test is usually used to compare the Kaplan Meier curves for different groups, which checks whether they are statistically equivalent (Kleinbaum and Klein 2006, p.57). Fundamentally, it is a large sample chi-square test that estimates whether the Kaplan-Meier curves are overall different.

Another widely used model of survival data analysis is the Cox proportional hazards regression model, which belongs to the category of the semiparametric models. Unlike parametric models, it does not require multiple assumptions, such as shape of a baseline hazard function. However, this model has an important underlying assumption of multiplicative relationship between the hazard function and its predictor variables, which is referred to as the proportional hazard assumption. Once this assumption holds, the information of baseline hazard pattern is not needed to estimate the effects of the explanatory variables on the hazard function (Kleinbaum and Klein 2006, p.94). This method will be tested for SoundCloud user relationship in Section 4, even though the focus of this study will be on non-parametric models.

In the following sections, survival analysis will be discussed in the context of its applications on user-generated content networks, and applied on the sample of SoundCloud user data. Typically for real user data samples, it is limited up to a certain date, which implies right-side censoring, especially for these user relationships that were established not long before the last date of sample. Key challenge is to define the event variable for survival analysis based on other characteristics of user interaction data, which will be discussed in more detail in Section 4 of this paper. Knowing the patterns of user interaction would help network's management to better understand drops in user activity, as well as optimize their customer relationship management strategies, and predict customer churn.

3. Interaction in user-generated content networks

Over past few decades, Internet has tremendously changed the ways of communication and information behaviors of people worldwide. User-generated content can be defined as such that fulfills three following requirements (Kaplan, Haenlein 2010, p. 61). Firstly, it has to be published either on a publicly accessible website or be accessible to a selected group of users. Secondly, user-generated content requires a certain amount of creative efforts, and thirdly, it should be “created outside of professional routines and practices” (Kaplan, Haenlein 2010, p. 61). According to this definition, user-generated content is a foundation of any social media website or network. Today it has become a new way of self-expression and communication with others online (Smith, Fischer, Yongjian 2012, p. 102), especially among younger users. User-generated content is not a new phenomenon: first blogs were created as early as in 1980s, however, since then it has transformed and became more powerful due to accessibility of Internet and the rise of online social media.

There are immense differences in ways users interact on different platforms. While the existing studies focus mostly on the social media giants such as Facebook, YouTube, Twitter, their comparison, and their implications for brands (Smith, Fischer, Yongjian 2012, p. 103), this paper explores and analyses user interaction on SoundCloud. SoundCloud is a music sharing platform, where artists can upload and promote their tracks and albums, collaborate with other players of the industry, and keep in touch with their followers. On the other side, music listeners are able to discover new artists through browsing, their friends’ activity or by directly connecting with the artists they already know. Within this network, both artists and music fans are able to follow, favorite, comment or share specific albums or tracks, as well as exchange messages.

This section discusses the nature of relationships in computer-mediated environment in subsection 3.1, and focuses on specifics of interaction on SoundCloud in subsection 3.2. In order to have a profound understanding of qualitative characteristics behind the SoundCloud data, not only relationships in user-generated content networks, but also a broader scope of relationship definitions will be explored.

3.1. Nature of user-generated content networks relationships

There are several ways of looking on the user interaction in user-generated content networks: on one hand, they are primarily relationships between two people, but on the other hand, the environment of an online platform has a strong impact on the nature of these interactions. In this subsection, interactions in user-generated content networks will be discussed from both points of view.

In terms of characteristics of the user relationships, they can best be compared to real-life casual friendships: several studies have examined the online friendships as a special case of friendship (Ellison, Steinfield, Lampe 2007 p. 1144, Hoffstetter et al. 2009, p.2) and compared it to the offline friendships(Chan, Cheng, 2003 p. 307). According to another study (Bowker 2004, p. 85-86), friends can be distinguished from non-friends based on time spent together, as well as social activities in which two people engage. For both online and offline friendships, their qualitiyy is expected to incease over time (Chan, Cheng, 2003 p. 316). Some academic frameworks differentiate between friendship types, based on either level of intimacy (Bhardwaj et al. 2016, p. 552; Bowker 2004, p. 87; Becks et al., 2009 p. 351) and by defining best, close, and casual friends, as well as considering geographic distance between friends (Becks et al. 2009, p. 350). It is common that users' primary audience in the online networks comes from their offline connections.(Ellison, Steinfield, Lampe 2007, p. 1144).

Friendship quality, sometimes referred as stability or strength can be defined by its duration, frequency of interaction, and various types of emotional or intellectual stimulation determinants (Bowker 2004, p. 89, Becks et al. 2009, p. 350; Rose 1983 p. 271-273, Ogolski 2012, p. 344). Given the objectives of research and available data, the focus of this paper will be on the relationship duration and the frequency of interaction. Nevertheless, it is important to consider that other factors could also account for changes in friendships.

Making friends in computer-mediated environment is easy, and it is usually completed by confirming a friend request or clicking the “follow” button. In the online social networks people have more control over the information flow than offline, due to advanced sharing and following settings, as well as being in control of who becomes their online friend and who does not (Bryant, Marmo 2012, p. 1021).

Friendship termination is a subject of a special interest in this paper, since one of the key challenges is to define the end of the relationship between two users for further analysis. Unlike romantic relationships, only few friendships end with a breakdown (Rose 1983, p. 367), and it is typically a continuous and less radical process. In online networks a lot depends on whether a friendship is exclusively online or also existing offline – in later case, abruptly unfriending or unfollowing is not common, just like with friendships in general.

However, user activity in user-generated content networks goes far beyond interaction or friendship. In fact, most activities in online networks are not about interacting with other users: the share of time users are interacting with others on Facebook is around 10%, while on Twitter it is 16%. (Hall, 2016, p.8).

3.2. User interaction on SoundCloud

This subsection describes user activity data sample from SoundCloud, based on data collected by the company about users who joined the platform between 01.01.2009 and 19.03.2014. It includes information about affiliations, favoritings, comments, messages and reposts of 35956 users who started using the network within the given timeframe. Each of 5 original datasets contained three variables: user ID – user who undertakes an activity, for example, comments a post; owner ID – user who owns the content towards which the activity was made; and date of this interaction.

In this study, only pairs of mutually affiliated users are analyzed, making sample somewhat more even in terms of proportion of creators and non-creators. Such sample choice makes data more similar to other online networks, with more users being content creators compared to a typical music sharing platform, and key findings could be more applicable for platforms other than SoundCloud. The chosen sampling decision also leaves only user relationships of a friendship type, that were discussed in previous subsection.

This sample excludes fan-type interactions, for example, when a fan is liking or commenting artist's posts without being followed back. Exploring this type of relationship could be of high relevance for the new artists and labels, however, it is less scalable for other user-generated content networks, and thus intentionally omitted.

The final sample of 810818 user pairs was further split into two subsamples: active and non-active relationships. Active relationships are defined as those that have had at least one interaction after the relationship was established, and includes 147637 user pairs. Non-active relationships are those whose only interaction was establishing a relationship by following each other, consists of 663181 users. Similarly, to other networks (Hall 2016, p.8), only 18.2% of all established relationships were active in over 5 years, which shows an overall interaction pattern in this type of network. While the majority of users never even interacted, the active users interactions are also in reality not that active: 66% of active users have had only one interaction over their relationship lifetime. This doesn't mean that the majority of the users are not active on a platform, and it again confirms a study (Hall 2016, p.8) which states that interaction between users is just a small fraction of possible activities on online platform.

Since the subject of this study is relationship between two users rather than the dates of relationship establishment and online activities in broader sense, the rest of this article refers only to active relationships on SoundCloud. In survival analysis, all the non-active relationships would be simply defined as these that have ended, thus the numerical results of survival analysis would be significantly different from the case where all relationships are be taken into account. In terms of survival analysis, all the non-active relationships would be defined as these with 100% probability of failing, and thus would not be relevant for resolving the central challenge of survival analysis in user-generated content networks in definition of survival variables.

It is important to note, that the dates of relationship establishment are not evenly distributed over time. Overall, the number of newly established relationships was growing over time, and can be generally explained by growing user base of a platform over time, except for a dip in late 2012. It also means that most relationships are relatively young, thus duration cannot fully access their quality or survival time. Figure 1 illustrates distribution of the dates of the new relationships establishment.

"Insert Figure 1 about here"

One of the objectives of this research was to explore whether there are any patterns in user interactions over the relationship lifetime, and if they exist, whether they differ among the different interaction types. According to several studies in field of customer relationship management (Dagger, Danaher, Gibbs, 2009 p. 373-374; Bolton, 1998, p. 63), relationship duration has a moderating effect on the relationship strength, and most relationships decrease in number of interactions as the relationship evolves. Even though this study is dedicated to customer relationship and not to friendship, this relationship pattern confirmed to be truth for the relationships in a user-generated content network.

In order to examine the relevance of such pattern in user-generated content networks, all dates of interactions are plotted against the relationship start on the graphs in Figure 2. It can be clearly seen that the vast majority of user interactions happened within first days of relationships. Figure 2 illustrates distribution of interaction over relationship time for favoritings, comments, messages, and reposts respectively. For all types of interaction, the modal value of the day when an interaction occurred was first day after establishing a relationship.

"Insert Figure 2 about here"

Even though all types of user interactions on SoundCloud follow similar pattern, several differences between interaction types were discovered. Most popular type of interaction on SoundCloud is favouriting, which accounts for 44% of all user interactions on the platform. As can be seen on Figure 2, the curve peaks in the beginning of the relationship and steadily decreases over time. Messages, which are the second most popular type of interaction accounting for 34%, have slightly different pattern with very high count on first day after relationship establishment, and dramatically decreasing after it. Comments follow a similar pattern as likes, and constitute 19% of all user activity. Reposts have the smallest sample, and account only for less than 1% of all activities, and follow a slightly different pattern than the others: it has least the steep decrease in activity.

It is important to note that the share of interactions which happened later in the relationship can be slightly underestimated in this figures, since many relationships in the sample are young. However, given that most interactions occurred in the very beginning of a relationship, and less than 10% of all observed relationships were created 6 month prior to the sample date rage end, it does not affect the validity of the interaction pattern,

Understanding how users interact on a particular platform on different stages of their relationship is important for managerial decisions. Combining this knowledge with survival model would be valuable both for SoundCloud as a company, as well as for potential advertisers who are willing to cooperate with the platform.

4. Empirical analysis and results

This section is dedicated to the definition of survival variables based on SoundCloud data paterns and relationship theories, as well as to the application of these variables in conducting survivsl analysis. Given that there was no previous research on the applications of survival analysis in user-generated content networks, there is no expected survival pattern or variable definitions to use as a foundation. Thus, in this section the author relies on the data-driven approach rather than on theoratical assumptions in construction of two essential variables: the survival event and the survival time.

4.1. Defining survival variables

According to the academia, relationship duration is often considered the key determinant of the relationship strength (Bowker 2004, p. 85, Liu, Ginther, Zelhart 2002, p. 79; Reinartz, Kumar 2003, p. 81; Dagger, Danaher, Gibbs 2009, p. 373), and the longer the relationship has already lasted, the longer it is likely to keep going.

SoundCloud user data is sufficient to calculate two key relationship strength determinants (Dagger, Danaher, Gibbs 2009, p. 373): relationship duration and relationship quality. In this study, the latter will be referred to as the frequency of interaction, and defined as the count of times two users have interacted with each other after the relationship was established. The chart below illustrates the relationship between the duration and the frequency of interaction:

"Insert Figure 3 about here"

It can be concluded from visual analysis that the frequency of interaction and the relationship duration are not linearly related, as stated in the study on the strength of customer relationships (Dagger, Danaher, Gibbs 2009). This means, that in case of SoundCloud users, the frequency of interaction is less dependent from the relationship duration. On the other hand, frequency of interaction can be possibly tested as a parameter that explains the survival outcome in a non-parametric survival model like Cox proportional hazard. This example shows that the nature of relationship in user-generated content networks is unique, and conclusions from other fields cannot be directly transmitted on them.

The greatest challenge in defining the survival variables for user relationships is definition of the end of the reationship, or the survival event. Unlike in natural sciences, where the survival analysis originates from, events are not easily observable: as was discussed in Section 3, both online and offline friendships rarely end with a breakdown. Having a clear definition of the end of users' interaction is crucial for applications of survival analysis. While for most relationships it is challenging to define this event due to censoring, some realtionships are sufficient to state as failed or successful with a high degree of certainty.

The issue with definition of the failure event is the primary reason of choosing survival analysis framework over other methods for this study: SoundCloud data is a subject to right-type censoring, which makes other more commonly used statistical methods less reliable. However, in order to have meaningful results, all observations cannot be censored, thus, the assumption of the end of a relationshipis needed.

As was stated in Section 2, survival analysis event has to be a binary value, that is being attributed a value of "1" in case the relationship is considered as such as ended, and "0" otherwise. For example, due to censoring, it is important to make sure that "0" is attributed to the newer relationships regardless of their duration and frequency of interaction. Fundamentally, the definition of which variables are censored is going to affect the outcome of survival analysis more than any other assumptions. Given that the survival event is a binary variable, the rule of thumb in assigning values is whether the event is more likely to happen or not to happen. For simplicity, if the event has over 50% chance of happening, it should be assigned “1” and “0” otherwise. This rule will be used for assigning the thresholds for different types of conditions without assuming a predefined survival distribution.

Since some of the interactions are most likely to be censored, and some are most likely to be defined as such where the survival event has occurred, the event variable can be defined as a construct of several conditions. Further in this section, a set of rules and assumptions will be developed to be able to state whether each of the relationships has ended or not. Table 1 includes the descriptions of the full list of variables which were used in the empirical analysis, and are used below to describe the survival event conditions.

“Insert Table 1 about here”

As can be seen Figure 1 that was discussed earlier in this paper, 50% of the realtionships were created after 10/12/2011, and 10% of all relationships were established after 12/06/2013, which is slightly more than 6 months before the end of the observation. This means that the actual duraction of such relationship is unknown because it is most likely to end after the end of the observation period, and these relationships are subject to the right side censoring. This observation leads to the first condition for a survival event definition.

Condition 0: if ‘created_at’ > 27/07/2013, ‘Relationship_end’ = ‘0’

The average duration of a relationship is 264 days, but given that many of them were created late in the sample, only 50% of relationships lasted more than 150 days. This means, that there is more than 50% chance that the relationship established later than that will be censored and should be assigned ‘0’. On the other hand, unless the assumption of relationship duration as the major determinant of survival holds, there is over 50% chance that the relationship will not become active and is over, if there is no interaction in over 150 days.

Condition 1: if ‘last_interaction’> 27/07/2013, ‘Relationship_end’ = ‘1’

In case it is assumed that the relationship past duration affects its future survival, it is important to make sure that the long-term active relationships don’t get filtered out as the ones that ended. In order to do so, the event value of “0” will be attributed to 10% of relationships with the longest duration, where time between their first interaction and their last interaction is more than 652 days.

Condition 2 (optional): if ‘Duration’ > 652, ‘Relationship_end’ = ‘0’

Figure 4 illustrates the density function of the relationship duration, and it can be concluded that a rather small fraction of all relationships account for the longest lasting ones, especially after the turning point exactly at 652 days.

“Insert Figure 4 about here”

Another group of relationships that can be defined as the ongoing, are the relationships with the highest number of interactions: only top 5% of all relationships had more than 15 interactions within the observed period.

Condition 3 (optional): if ‘Count_of_Interactions’ > 15, ‘Relationship_end’ = ‘0’

The threshold of the relationships with high interactions number is higher, because number of interactions density is more skewed to the left than for the duration, as illustrated in Figure 4. Overall, SoundCloud users tend to have low activity, so that the relationships that are most active should not be defined as these that are over. In the following section, the survival model will be tested both with this and without this assumption.

The rest of the relationship observations are going to be censored since there is no high certainty in either of these relationship outcomes. With shifting threshold of failed relationships, the proportion of censored observations shall be shifting as well, and for different data thresholds should differ as well.

Based on granular SoundCloud user data, there can be several possible approaches to defining the survival time variable. Firstly, it is nessesary to define the beginning of a relationship: in SoundCloud case, there is a precise date of establishment of a relationship, all of which have happened after a specific date. This also means, that data about SoundCloud is not a subject for left-type censoring. Secondly, the date of the end of a relationship is needed, which is less straight-forward than the relationship start. It can be extracted from the database as the date of last interaction between two users, however, this measure could be misleading if the relationship is newly established.

Another approach to measuring the duration of user relationships would be to assume right censoring for a user relationship unless there was no interaction for a predefined period of time. Then, it would be reasonable to define the survival time as the time between establishing a relationship and assume left censoring, and define a time period between the last interaction and the end of a study as a threshold that defines an ended realtionship. Thus, in order to define the survival time, the conditions of the survival event are of crucial importance.

4.2. Estimation and interpretation of survival curves

This subsection is dedicated to application of the survival analysis to SoundCould data using survival package in R. Since survival variables definitions were based on a list of conditions discussed in section 4.1, Kaplan-Meier survival estimates in this section will be executed for several sets of conditions. First model, S0 will be based on conditions 0 and 1, model S1 will have survival variables defined by conditions 0,1, and 2, and finally model S3 will include basic conditions 0 and 1, as well as the optional condition 3.

Due to the limited availability of data about quality of user interaction, no reliable information on the expected survival pattern, and only one potential independent control variable, which is the number of interactions, the application of parametric and semi-parametric models is hardly possible in SoundCloud sample. Thus, Kaplan-Meier non-parametric model seems to be the ideal estimation method for the purposes of this study. Figure 5 below illustrates the plots of the survival functions of user relationships on SoundCloud based on respective conditions set models.

“Insert Figure 5 about here”

Based on the relationship end model S0, the probability for the relationship not to end after 1 more day is 92.23%, with confidence interval of [0.9210;0.9237]. As illustrated on Figure 5, the survival function has two kinks at around days 55 and 900 – once these dates are reached, the probability of surviving for the next period decreases at a slower pace. On day 150, which was used as a threshold for end of a relationship definition, the probability of survival is 56.6% with confidence interval of [0.563;0.568], which is higher than the percentage of relationships that reached the 150-day relationship duration in the original sample. Table 2 summarizes survival estimates for the respective models.

“Insert Table 2 about here”

Model S1 uses same variable conditions S0 except for an additional assumption that relationship duration has a stronger impact on the probability of the relationship survival. Thus, in this specification if the relationship is long-term, is doesn’t matter when was the last interaction, and for these relationships the event value is always “0”. The results are visualized in Figure 5.

The shape of the survival curve is similar to the one of S0, however, S1 is noticeably flatter after day 652. According to this variables specification, if the relationship lasts for 652 days more, after this point its probability of survival doesn’t decrease with time. While statistically this model is correct, such development of user relationship is not quite plausible, given duration patters illustrated in Figure 5. This model showcases the importance of reliable assumptions for defining survival variables.

Third model, S2, is adding an assumption of the influence of frequency of interactions into the basic conditions model S0. Alternative estimation of a model based on variable set S0, that also accounts for frequency of interaction using Cox proportional hazard, will be discussed further in this section. The survival curve of S2 is visually more similar to S0 than S1.

According to S2, the probability of survival on day 1 is 91.19%, and the probability of relationship to survive after day 150 is 60.14% with confidence interval of [0.599;0.604]. Apparently, when taking into account only top 5% of most active users, survival probability for same day significantly increases.

According to Figure 3 previously in the text, the frequency of interaction is not linearly dependent from duration of a relationship, and according to the data in Table 2, it makes a significant difference when incorporated in the definition of survival variables. In order to explain the impact of the frequency of interaction on survival function, variable specification S0 will be used for survival time and event, and frequency of interaction will be an explanatory variable in Cox proportional hazard model. The summary of a model can be found in Figure 6.

"Insert Figure 6 about here"

These results answer the question on how big is the impact of frequency of interactions on survival curve: the R-squared value for this variable is 5.8%. This means that 5.8% of variation in survival curve can be explained by frequency of interaction.

4.3. Results and discussion

Overall, the results of survival analysis for SoundCloud user interaction data are plausible and in line with the expectations. This analysis did not only return an overall survival pattern, and the probability values of user relationship survival for a certain period of time, but also gave a valuable insight on the impact of frequency of user interactions, as well as highlighted the importance of an accurate definition of survival variables.

While survival models in the previous subsection focused on the carefully constructed data-driven variable assumptions, these assumptions have heavily influenced the results of the model. This was proven just by adding an additional assumption to model S0, to get entirely different results in S1.

On one hand, this shows the survival analysis' flexibility, versatility, and opportunity to test multiple assumptions and condition constructs, but on the other hand, this also appears to be the biggest limitation of the method used in this study. The model is sensitive to underlying assumptions, and building up these assumptions is highly a time- and effort consuming process, which requires in-depth knowledge of the survival object.

Given the available data, as well as typically fuzzy definition of the end of user relationships, adequate variable definition is the biggest challenge of this study. However, it doesn't have to be. This problem can be resolved with adding one variable into the original SoundCloud dataset: the date of "unfollowing". Technologically, tracking the unfollowing is not more complex than tracking any other event on a website, but it could make survival analysis an easy and reliable tool without need of any assumptions.

Attributing the end of a relationship to the unfollowing requires further research, as in different networks there are different "rules of friendship" (Bryant, Marmo 2012, p. 1021): while unfriending on Facebook can be perceived as socially unacceptable and rude, unfollowing users on other platforms could be completely normal. Collecting data on user unfollowing activity and testing it for a particular user-generated network is the only way to find out whether unfollowing is a plausible assumption of an online friendship relationship end, and an adequate survival event.

Another possible direction of further research is to include additional user data into research sample, such as demographics, interests, online behavior etc. In current research, there was little opportunity to apply Cox model due to limited data and different scope of the study, however, it has an immense potential in exploring factors that influence the end of the online relationship.

5. Managerial implications

Today's digital landscape is continuously changing at a fast pace, so that the strategies and methods that worked yesterday turn out to be irrelevant tomorrow. The objectives of this paper were to explore a method of analysis which is new to the user-generated content networks, to test it on SoundCloud sample, and based on results, to make suggestions about whether survival analysis is suitable for this type or data, and how to best apply it. This section discusses the relevance of the results of survival analysis for SoundCloud, possible improvements and extensions, and suggests a general framework for implementation on user-generated content networks in practice.

According to several studies (Bhardwaj et al. 2016 p. 552; Ransbotham, Kanne, Lurie 2012, p. 389), the more relationships and interactions the user has within a network, the more embedded this user is, and the higher is the benefit this user has from being active in this network. Fundamentally, each relationship that user has on a social network is a tie that makes user stay on the platform, thus, individual relationships on a user-generated network site contributes to the relationship of user and a site (Bhardwaj et al. 2016 p. 552). For the online user-generated content platforms like SoundCloud, this means that strong relationships between users will make them loyal and long-term users of the online network.

Analysis in Section 4 refers to a relatively small sample of SoundCloud users: pairs of mutual followers who engaged in at least one interaction between 2009 and 2013. The numerical results clearly cannot be used to make any kind of managerial decisions. Nevertheless, the methodology can be successfully incorporated into the company's decision-making practice. The models described earlier can serve as a prototype of a SoundCloud customer churn model: if the relationship between two users is a proxy of user's relationship with a social network, survival analysis can be an early indicator of which users go and which users stay, and which relationships are worth investments.

Application of survival analysis on user-generated content network is not limited to predicting the end of user relationship on the online platform, but goes far beyond the instance in this paper. Survival analysis is a versatile tool, and even though the initial effort of getting to understand and apply it can be high, it can provide with valuable and accurate information on social network users. General framework of application of survival analysis consists of the following steps: identifying the objective, defining survival variables, including explanatory or control variables, executing survival analysis, and finally, incorporating results into a strategy of achieving the objective.

Before considering survival analysis, is important to define the overall objective: what needs to be achieved that is essential to a company or what can be done to improve its performance. Survival analysis is best used to answer temporal questions: when to reach out to the user, when to launch a campaign, when to stop investing in customer relationship etc. In case of SoundCloud, survival analysis could be an early indicator of a certain artist's success, and generally a tool to detect most engaging content on the platform. Alternatively, it can also increase targeting efficiency about the optimal time to communicate opportunity of account upgrade to the existing customer.

Depending on the objective, defining the survival variables can either be the most challenging part of the analysis, like in predicting end of a relationship or user churn, or rather obvious. For example, if the goal is to identify new artists with highest potential that would increase user engagement on the platform, the survival event can be defined as reaching a certain amount of followers or sales. In case of predicting customer churn, the event can be user inactivity on SoundCloud for a predefined period of time. It is highly advised to internally define thresholds that would serve as a rule of thumb on whether the survival event occurred or not; defining multiple thresholds for several alternative survival events would be reasonable if the management is uncertain about which threshold to use. On the other hand, some events are easy to track: for example, unfollowing a particular user, or upgrading user's account.

Control and explanatory variables are not necessary in survival analysis, but can be extremely useful for learning more about user behavior. Including them into the model might increase complexity, time spent on model evaluation, and overall decision making, but it could give valuable insights on users from certain location, age, gender, music preferences, account type etc. Typically, there is no prior knowledge on the resulting survival distribution, and in case control or explanatory variables are included, it is advised to use Cox proportional hazard model. In case it is not possible to include them due to time scarcity, little data availability, or complexity restrictions, it is recommended to use non-parametric Kaplan-Meier survival estimation, as it was done in Section 4 of this paper.

Execution of survival analysis itself includes preparing survival variables and running analysis itself, and is rather a technical task. It starts with survival variables preparation: assigning values of "1" or "0" to an observation, and transforming time variable into number of days until event format. In case of large datasets, in is recommended to do data cleaning and preparation in specialized database management software. In this study survival analysis was executed in R software, however, survival tools are also available in most statistical software, like SPSS, SAS, Stata etc, and each of these options has advantages and disadvantages.

Last step in successful implementation of survival analysis is incorporating of results into decision making, which is extremely objective-specific. For example, having discovered that customer is most likely to stop using SoundCloud at a certain time, it is worth stop actively investing in this relationship. Finding out that a user is most likely to upgrade their account at a certain day after registration, the company would efficiently communicate upgrading when user is most likely to be interested; early discovering a new artist that is most likely to become successful can give collaboration opportunities.

While most of this section focused on implications for SoundCloud and giving some possible platform-specific examples, this framework is scalable not only for most user-generated content networks, but also for other businesses that have access to the detailed user data. Survival analysis is a powerful and versatile tool, that takes into account uncertainty and has a potential to assist in decision making in business, as well as to improve company's performance.

Table 1 – Full list of variables used in survival analysis

Variable name	Type	Description	Value range [min..max]
User_id	SoundCloud import	ID of a user who follows (and gets followed by owner)	[2..68685447]
Owner_id	SoundCloud import	ID of a user who gets followed (and follows back the user)	[2..68685447]
Count of Interactions	Calculated	Number of times an interaction occurred between user and owner: either favorite, comment, repost, or message	[1..937]
Duration	Calculated	Number of days between day of establishment of a relationship and the last observed interaction	[0..1903]
Created_at	SoundCloud import	Date of establishment of a relationship	[2009-01-01..2013-12-27]
Last_date	SoundCloud import	Date of the last observed interaction	[2009-01-01..2013-05-13]
Start_date	SoundCloud import	Date of the beginning of the experiment	2009-01-01
Experiment_end	SoundCloud import	Date of the end of experiment	2014-03-19
Total_duration	Calculated	Days between first interaction and the end of experiment	[82..1903]
Relationship_end	Calculated conditional	Binary variable indicating whether the end of the relationship occurred	[0] or [1]
Survival_time	Calculated conditional	Number of days that relationship survived before the end of observation	Depends on the Relationship_end assumptions

Table 2 – Summary tables of Kaplan-Meyer survival estimation by model

Summary S0						
time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	134584	13422	0.9089	0.000749	0.9075	0.9104
60	94634	32453	0.6807	0.001233	0.6783	0.6832
150	73872	15529	0.5656	0.001327	0.563	0.5682
365	45369	19514	0.4051	0.001363	0.4025	0.4078
652	14826	20842	0.1885	0.001229	0.1861	0.1909
900	5584	5601	0.1056	0.001093	0.1035	0.1078
1800	58	2079	0.0449	0.001228	0.0426	0.0474
Summary S1						
time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	134584	8865	0.94	0.000622	0.938	0.941
60	94634	20808	0.781	0.001131	0.779	0.783
150	73872	9882	0.694	0.0013	0.692	0.697
365	45369	11535	0.571	0.0015	0.568	0.574
652	14826	10689	0.387	0.001837	0.383	0.391
900	5584	2417	0.304	0.002119	0.3	0.308
1800	58	290	0.282	0.002329	0.278	0.287
Summary S2						
time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	136168	13004	0.9119	0.000738	0.91	0.9133
60	102348	30815	0.7014	0.001195	0.699	0.7037
150	82801	14254	0.6014	0.001285	0.599	0.604
365	49455	19237	0.4451	0.001363	0.442	0.4478
652	18279	19023	0.2428	0.001337	0.24	0.2455
900	6418	4545	0.1666	0.001343	0.164	0.1692
1800	164	2073	0.0852	0.001683	0.082	0.0886

Figure 1 – Distribution of relationship start dates over time

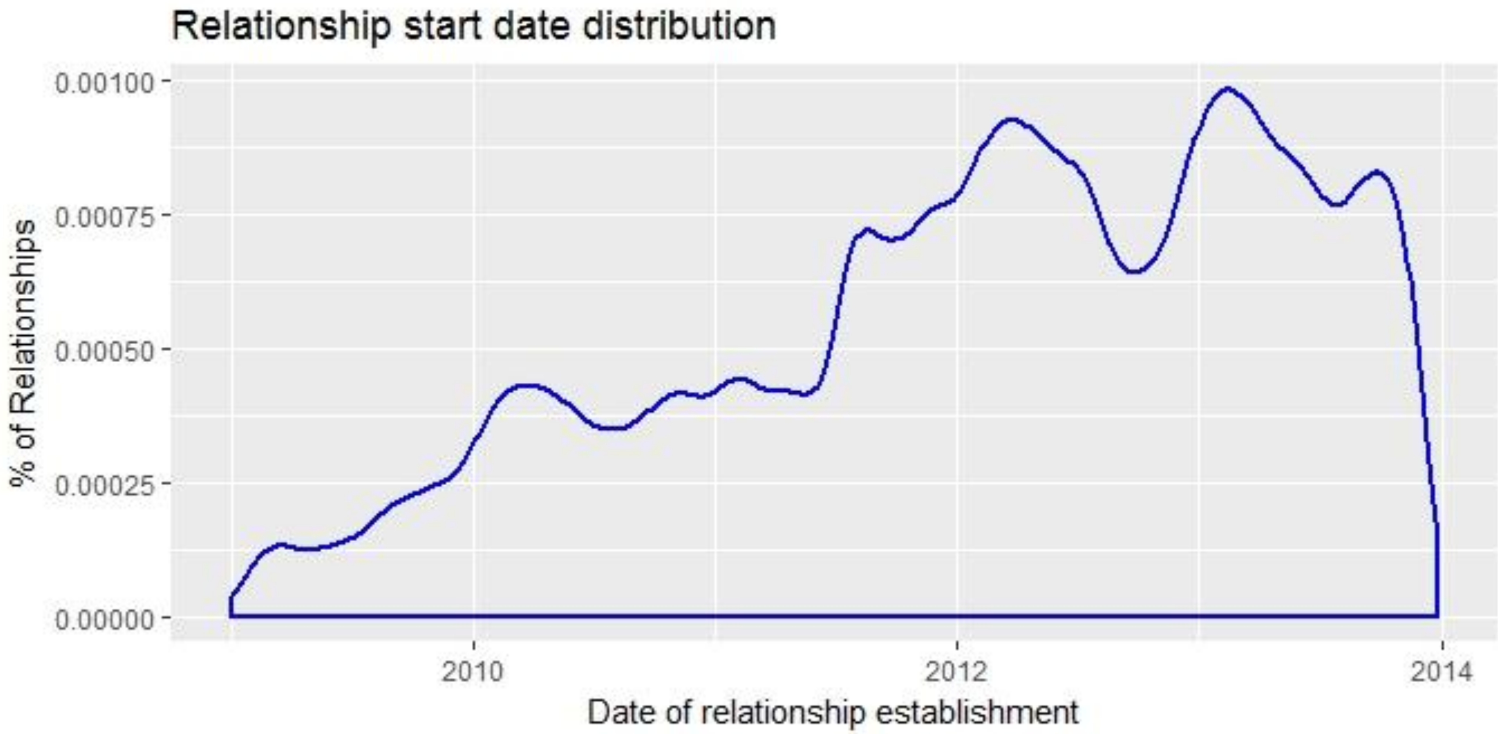

Figure 2 – User interactions over relationship lifetime by interaction type

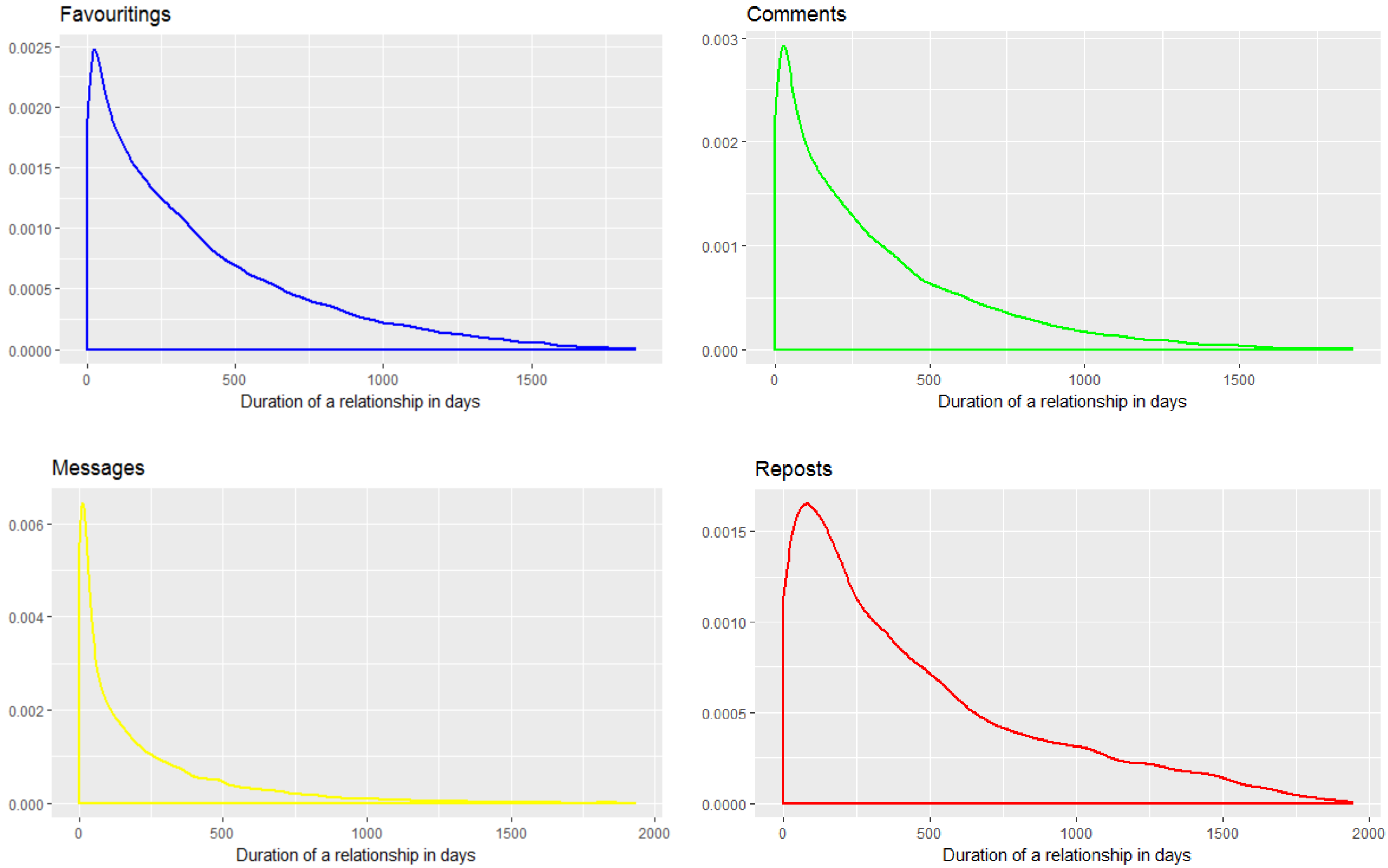

Figure 3 – Relationship between relationship duration and number of interactions

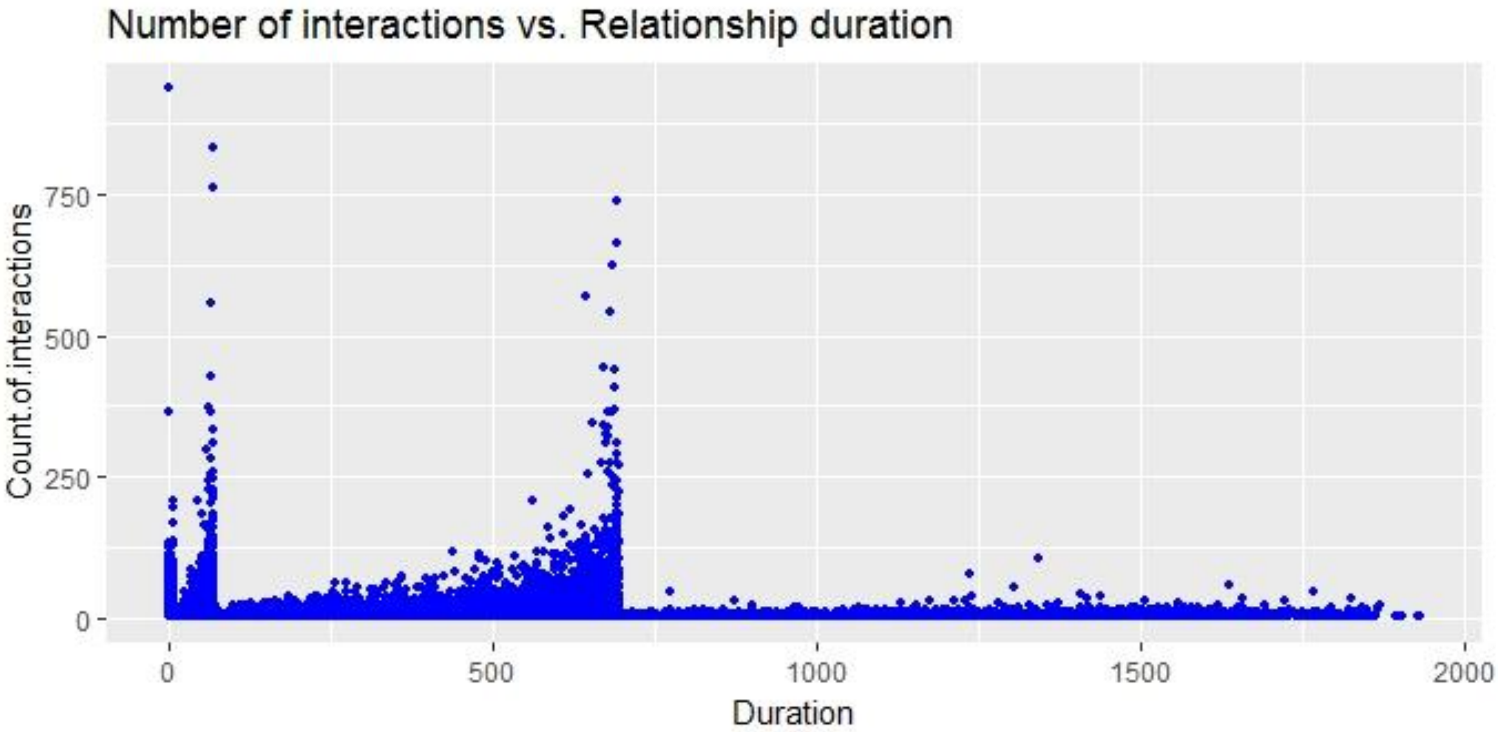

Figure 4 – Density of relationship duration

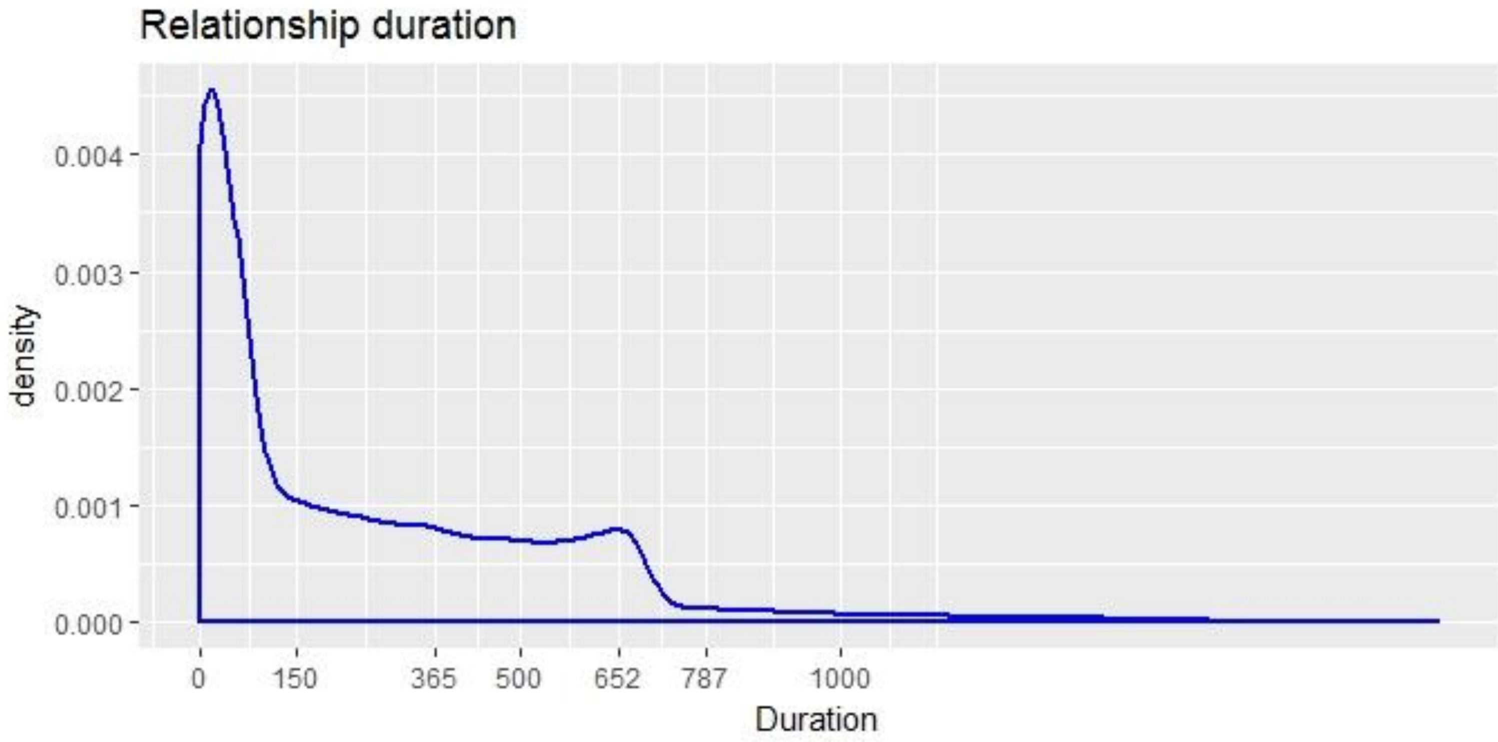

Figure 5 – Kaplan-Meyer survival curves by model

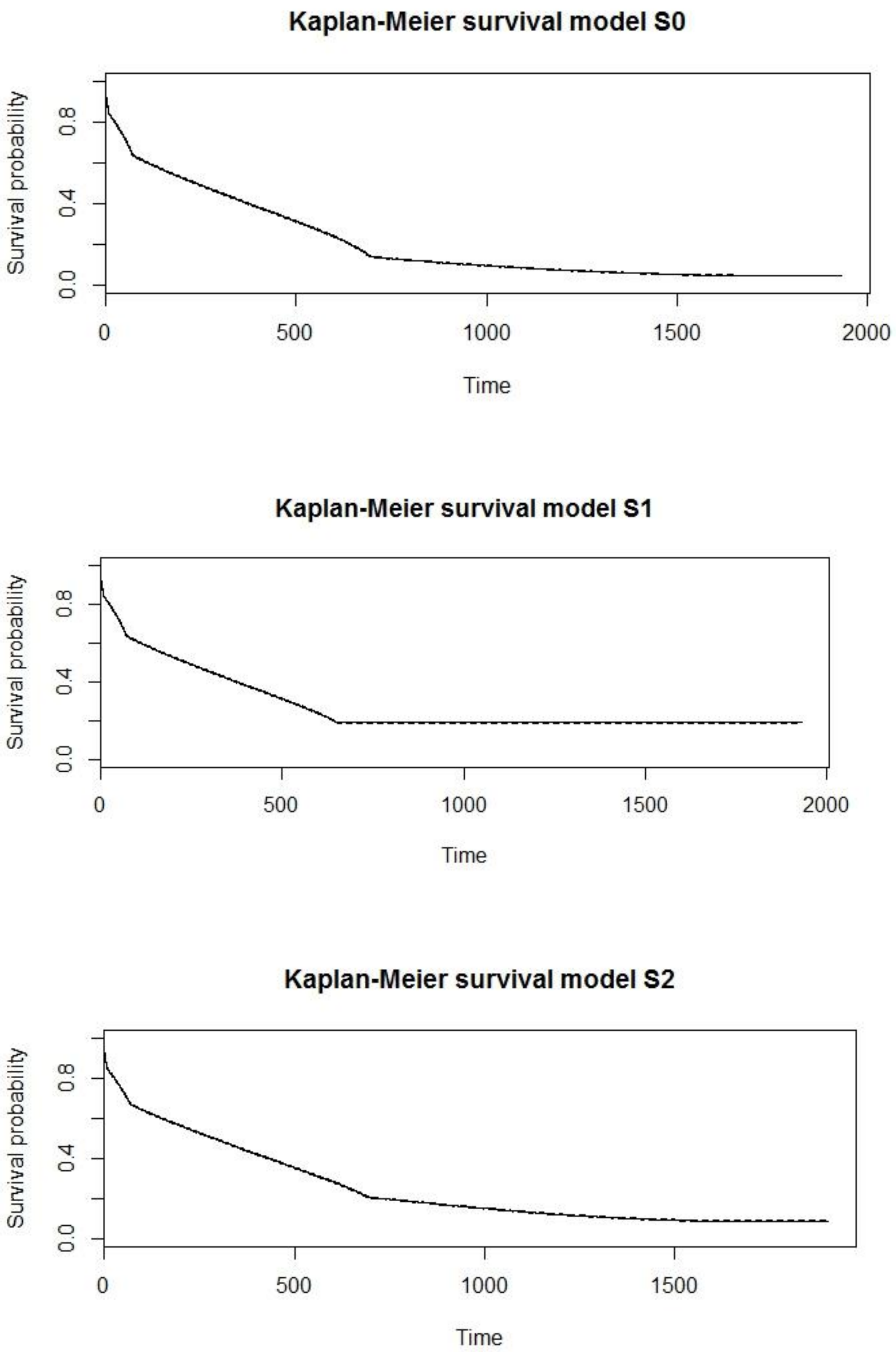

Figure 6 – Summary of Cox proportional hazard survival estimation

n= 147637, number of events= 102951				
	Coef	exp(coef)	se(coef)	z
Active_Full$Count.of.interactions	-6.34E-02	9.39E-01	8.69E-04	-72.97
Active_Full$Total_Duration	3.53E-05	1.00E+00	7.63E-06	4.63
Pr(>|z|)				
Active_Full$Count.of.interactions < 2e-16 ***				
Active_Full$Total_Duration 3.66e-06 ***				

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				
	exp(coef)	exp(-coef)	lower .95	upper .95
Active_Full$Count.of.interactions	0.9386	1.065	0.937	0.9402
Active_Full$Total_Duration	1	1	1	1.0001
Concordance= 0.549 (se = 0.001)				
Rsquare= 0.059 (max possible= 1)				
Likelihood ratio test= 8906 on 2 df, p=0				
Wald test = 5324 on 2 df, p=0				
Score (logrank) test = 2757 on 2 df, p=0				

Appendix

Appendix A - Literature Review

Relationship characteristic	Author (Year)	Research field	Key findings
Duration of customer relationship	Bolton (1998) Marketing Science	Customer Relationship	• Focuses on customer-provider relationships • Relationship duration is used as a proxy for customer retention
Customer lifetime	Li 1995	Customer Relationship	• Proportional hazard model in services - identifying factors that affect relationship duration. • The study suggests building profile for customers with long and short lifetime
Customer relationship strength	Dagger, Danaher, Gibbs (2009) Journal of Service Research	Customer Relationship	• Relationship quality is defined as contact frequency • Relationship duration moderates the effect of contact frequency • For young relationship frequency of contact is more important than for an established one.
Duration in marketing	Helsen, Schmittlein (1993) Marketing Science	Marketing	• The article summarized duration variables of interest in marketing • Hazard models are more complex in application than regression models, but they are able to deal with censor biases, time-variate objects, and generally have less restrictions

Customer relationship characteristics	Reinartz, Kumar (2003) Journal of Marketing	Marketing, Customer reationship	• Model uses time-varying variables for explaining the impact of relationship characteristics on profitable lifetime duration • One of the biggest challenges is detecting the end of a customer relationship when the customer never signs off • Both long-term and short-term customers can be profitable and it's important to focus on both • Link between lifetime duration and profitability • If the sum of the expected discounted future contribution margin is smaller than a currently planned marketing intervention, establish the death event for the customer
Online friendships	Ellison, Steinfield, Lampe (2007) Journal of Computer-Mediated Communication	Online social media	• Using Facebook is acossiated with creating human capital • Primary audience comes from users offline connections
Frequency and recency of messaging	Liu, Ginther, Zelhart (2002) Social Science Computer Review	Online social media	• Frequancy and recency of messaging are used as determinants of forming personal impressions in an empirical study • Frequency and duration of contact are more important to form social interaction
Social ties in UGC	Hoftettter et al. (2009) Net Institute working paper	Online social media	• Social ties have an effect on user's content generation • Significant positive feedback for user-generated content

Interaction is social media	Hall (2016) new media and society	Online social media	• Activities on social media are not always interactions: it is combination of interacting and observing • Social media users assume their audience to be rather limited to their connections • Users spend less than 10% of time on liking/sharing on Facebook, and around 16% retweeting
User-Generated content	Ransbotham, Kanne, Lurie (2012) Marketing Science	Online social media	• Discusses effects of collaborative creation in User-generated content networks • Embeddedness into a network is a determinant of collaboration • Relationships, interactions, and flows define network embeddedness
Online friendship rules	Bryant, Marmo (2012) Journal of personal and Social Relationships	Online social media, Social sciences	Formulates a set of socially acceptable rules for Facebook friendship and sharing behaviour based on empirical study
Network ties	Bhardwaj et al. (2016) Group and Organization Management	Organizational structure, Networks	• Distinguishes between long-lasting and short term ties • Focuses on self-monitoring personalities and how it affects forming ties
Relationship fragility	Becks et al. (2009) Journal of Social and Personal Relationships	Phycology, Social sciences	• Application of turning points analysis, where turning points are "moments in a relationship’s history when the pressures of dialogic interplay are of sufficient intensity" • Differentiates between best, close, and casual friendships

End of a friendship	Rose (1983) Journal of Social and Personal Relationships	Phycology , Social sciences	• Termination of friendship is a less radical event than end of romantic relationship • Four reasons for a friendship end model: Dislike, change of liking, displacement, and deviation from pleasure/cost ratio • Few friendships ended with a breakdown
Relationship maintenance	Ogolski (2012) Journal of Social and Personal Relationships	Phycology , Social sciences	• Assisting in preserving a relationship is relationship maintenance • Relationship duration is used as an indicator of relationship outcome
Characteristics of online and offline friendships	Chan, Cheng (2004) Journal of Social and Personal Relationships	Phycology, Social sciences	• Quality of both online and offline friendship increases over time • For long-lasting relationships differences between online and offline becomes smaller
Friendship stability	Bowker (2004) Journal of Early Adolescence	Social sciences	• Difference between friends and non-friends: spending more time together engaging in more intense social activities • Duration of a friendship has positive impact on its stability
Nature of UGC networks	Smith, Fischer, Yongjian (2012) Journal of Interactive Marketing	UGCN, online marketing	• UGC networks are a medium for both self-expression and communication • Application of content analysis • Focus on YouTube, Facebook, and Twitter - there is difference between interaction in different channels • Strongest brand-related influence of USG is on YouTube

References

Becker J.A., Johnson A.J., Craig E.A., Gilchrist E.S., Haigh M.M., Lane L.T. (2009), "Friendships are flexible, not fragile: Turning points in geographically-close and long-distance friendships," *Journal of Social and Personal Relationships*, Vol. 26(4), 347–369.

Bhardwaj A., Qureshi I., Konrad A.M., Lee S.H. (2016),"Two-Wave Study of Self-Monitoring Personality, Social Network Churn, and In-Degree Centrality in Close Friendship and General Socializing Networks," *Group & Organization Management*, Vol. 41(4) 526 –559.

Bolton R. N. (1998), "A Dynamic Model of the Duration of the Customer's Relationship with a Continuous Service Provider: The Role of Satisfaction," *Marketing Science*, Vol. 17(1), 45-65.

Bowker A. (2004), "Predicting Friendship Stability During Early Adolescence," *Journal of Early Adolescence*, Vol. 24(2), 85-112.

Boyd D.M, Ellison N.M. (2008),"Social Network Sites: Definition, History, and Scholarship," *Journal of Computer-Mediated Communication*, Vol. 13, 210–30.

Bryant E.M., Marmo J. (2012), "The rules of Facebook friendship: A two-stage examination of interaction rules in close, casual, and acquaintance friendships," *Journal of Social and Personal Relationships*, Vol. 29(8), 1013–1035.

Chan D., Cheng G. (2004), "A comparison of offline and online friendship qualities at different stages of relationship development" *Journal of Social and Personal Relationships*, Vol. 21(3), 305–320.

Dagger T.S., Danaher P.J., Gibbs B.J. (2009), "How Often Versus How Long: The Interplay of Contact Frequency and Relationship Duration in Customer-Reported Service Relationship Strength," *Journal of Service Research*, Vol.11(4), 371-388.

Ellison N.B., Steinfield C., Lampe C. (2007), “The Benefits of Facebook ‘‘Friends:’’ Social Capital and College Students’ Use of Online Social Network Sites,” *Journal of Computer-Mediated Communication*, Vol. 12, 1143–1168.

Hall J.A., (2016), “When is social media use social interaction? Defining mediated social interaction,” *new media & society*, 1 –18.

Helsen K., Schmittlein D.C. (1993), “Analyzing Duration Times In Marketing: Evidence For The Effectiveness Of Hazard Rate Models,” *Marketing Science*, Vol. 11(4), Fall 1993 p. 395-215.

Hofstetter R., Shiver S.K., Nair H.S., Miller K. (2013), “Social Ties and User Generated Content: Evidence from an Online Social Network,” *Management Science*, Vol. 59(6), 1425 – 1443.

Kaplan E. L., Meier P. (1958) “Nonparametric estimation from incomplete observations,” *Journal of the American statistical association*, Vol. 53(282), 457-481.

Kaplan, Andreas M. and Michael Haenlein (2010),“Users of the World, Unite! The Challenges and Opportunities of Social Media,” *Business Horizons*, Vol. 53, 59–68.

Kay, R. (1977) “Proportional hazard regression models and the analysis of censored survival data,” *Applied Statistics*, Vol.26(3), 227-237.

Kleinbaum D. G., Klein M. (2006), “Survival analysis: a self-learning text” *Springer Science & Business Media*.

Li, J. (1995), "Foreign entry and survival: Effects of strategic choices on performance in international markets," *Strategic management journal*, Vol. 16(5), 333-351.

Liu Y., Ginther D., Zelhart P. (2002), “An Exploratory Study of the Effects of Frequency and Duration of Messaging on Impression Development in Computer-Mediated Communication,” *Social Science Computer Review*, Vol. 20(1), 73-80.

Ogolsky B.G., Bowers J.R. (2012), “A meta-analytic review of relationship maintenance and its correlates,” *Journal of Social and Personal Relationships,* Vol. 30(3), 343–367.

Ransbotham S., Kane G.C., Lurie N.H. (2012), “Network Characteristics and the Value of Collaborative User-Generated Content,” *Marketing Science,* Vol. 31(3), 387–405.

Reinartz W.J., Kumar V. (2003), “The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration,” *Journal of Marketing,* Vol. 67, 77–99.

Rose S.M (1983), “How Friendships End: Patterns among Young Adults,” *Journal of Social and Personal Relationships,* Vol. 1(3), 267-277.

Smith A, Fischer E., Yongjan C. (2012), “How Does Brand-related User-generated Content Differ across YouTube, Facebook, and Twitter?,” *Journal of Interactive Marketing,* Vol. 26, 102–113.