Analyzing the Survival Time of Crowdfunding Campaigns:

A case study on www.gofundme.com

Muath Ibaid and John Mohoang

Prof. Vittorio Addona

Math 453 - Survival Analysis

Macalester - Fall 2016

12/16/2016

Final Paper

Introduction

In today's world, obtaining funds is critical to people who need to achieve important goals. This need for funds has led to online business models for soliciting money from a variety of people. This form of funding falls under the big umbrella term of crowdsourcing. Crowdsourcing simply means dividing a big task into smaller portions for many people to accomplish individually. In more technical terms, the term 'crowdsourcing' was coined by Jeff Howe and Mark Robinson in June 2006 and they described it as "a new web-based business model that harnesses the creative solutions of a distributed network of individuals through what amounts to an open call for proposals" (Brabhan, 2008).

Some notable online platforms that are based on crowdsourcing are Amazon's Mechanical Turk and Wikipedia. A subcategory of crowdsourcing that is gaining momentum in popularity is crowdfunding. Crowdfunding helps people obtain funds to pursue particular ideas. The most famous website in this area is Kickstarter (www.kickstarter.com), which helps collect funds for creative startup projects.

The rise in popularity of such platforms has made it an area of research. In the literature review section of this paper we summarize some of what the literature concludes. Unlike their research that mostly focuses on business oriented crowdfunding websites, we study crowdfunding for humanitarian reasons. We want to analyse the survival of a post/campaign on a humanitarian crowdfunding website. Our case study is gofundme (www.gofundme.com). According to the Wall Street Journal, this online platform was started in 2010 as a way to help individuals and small charities raise money for good causes (Macmillan and Tan, 2015). We therefore think it would be a good case study to explore what people react to when it comes to charitable actions and what factors determine the duration of a certain campaign.

We examine these factors by performing a survival analysis using parametric distribution models (Weibull, Log-normal, Exponential and Normal) as well as non-paramteric (Kaplan-Meier) and Cox Proportional Hazard's semi-parametric model. We use data that we collected via Python web-scraping techniques on around 9,000 campaigns listed on gofundme.

The data used includes information on each campaign, such as the category it is listed under, the date it was created, the goal of the campaign, and the amount reached, in addition to other useful details. Our analysis concludes that the campaign category, its goal, number of social media shares, and the ratio of positive to negative sentiments of story are all factors that play a significant role in deciding the time taken until the campaign is funded. In section 2 of our paper we provide a review of the literature related to the topic. Section 3 describes the data used. We state our hypotheses in section 4, our results and discussion in section 5, and we finally conclude our study in section 6.

Literature Review

There is a fair amount of research that has been done in this topic. Motivations and deterrents to participation in crowdfunding were studied by Elizabeth Gerber (Gerber, 2013). This was a qualitative study conducted through semi structured interviews. In this research some motivations to participate in crowdfunding were the desire to expand awareness of work, form connections, gain approval, help others and be part of a community. Some of the failures in the funding campaigns included fear of public failure and exposure, inability to attract supporters, and distrust of creator's use of funds. This means we should expect successful projects to have higher aspects of motivation and lower aspects of deterrents.

Other studies have focused on the failures of crowdsourcing projects. Greenberg points out that the consequences of failure, which amount to about 50% of all Kickstarter campaigns, have real social impacts between creators of a campaign and community from which they are trying to leverage funds from (Greenberg, 2015).

On the more quantitative side of research, Li et al. studied the success time distribution of crowdfunding data and showed that logistic distributions are a natural choice for gaining insights in this area (Li, 2016). For comparison, they used Cox proportional hazard models, Tobit regression, Buckley-James estimation and Boosting concordance index. On average, Li et. al had about 70% success in success prediction. It is noteworthy that they too were using KickStarter as their case study.

Contrary to the conclusion reached by Li et al, a study done by Greenberg et al. found that simple classification algorithms such as decision trees outperformed the more complex ones (Greenberg, 2013). The aim of the study was to build a tool to predict how well a crowdsourcing project will do. We use similar variables to those used by Greenberg et al. These variables include Goal of Campaign, Project Category and Duration of the Campaign. The key difference is that Greenberg et al. focus on the algorithms that give them the best predictions whereas our focus is on the variables that have impact on the survival of a campaign.

In terms of methodology, we employ some of the algorithms used by Stam (2016) in our dataset. His study also aimed at developing machine learning methods to predict when Kickstarter projects achieve their funding goal to classify whether or not a project will be successful. Stam used K-Nearest Neighbors and Support Vector Machines. In order to predict the number of days a project needs to reach its goal, he used survival regression, in particular, Cox Proportional hazards model because it is able to deal with missing data as well as incorporating covariates into the model. He was, however, not able reach a conclusion because his model was equivalent to random guessing.

Data

We collected 8539 data points from GoFundMe website by automating a web-scraping procedure using Python. After identifying our fields of interest, we scraped all available records on the website by those fields. Based on our data, our study focuses on the variables listed in Table 1 which are included in our models in a later section. Appendix A has a list of all the other variables that we obtained from the gofundme website.

From the data, we calculated the duration of a campaign to be the number of days between the date when the campaign was created on gofundme and the date of last activity. In this way, the duration is therefore not skewed by campaigns that are still up on the website but have not seen any activity in months, sometimes even years. Following this, the censored campaigns are then regarded as campaigns that have not reached their financial goal because they still have that potential passed the date when the data was collected. In Table 1, the censoring variable is shown as "status".

The mostlyPositive variable in Table 1 refers to whether or not the campaign's text story has a mostly positive sentiment. We obtained this variable from the result of calculating the ratio of positive to negative words in the story text. The story text is a way in which the creators of a campaign communicate their needs and persuade people to donate. We used Vader to calculate the percentage of positive and negative words in the story text. Vader is a simple rule-based model for general sentiment analysis. It uses a combination of qualitative and quantitative methods, first to construct and empirically validate a gold-standard list of lexical features (along with their associated sentiment intensity measures) and then specifically attune the gold-standard to sentiment in microblog-like contexts (Hutto and Gilbert, 2014). Vader was most suited to the study of gofundme campaigns because it generalizes more favourably across online contexts than other sentiment analysis models.

Table 1: Description of variables used

Variable	Description
durationOfCampaign	The number of days since the campaign was created till that date if its last donation
amountReached	Amount of donations reached in USD
goalOfCamp	Monetary goal amount of a campaign set by creator in USD
goalPercentage	Percentage amount reached (amountReached/goalOfCamp)
status	This is the censoring variable. 1 = reached goal, 0 = did not reach goal i.e censored
category	Different classifications under which a campaign can be created. Popular categories are Family, Events, NewlyWeds, Travel, Animals, Sports and Medical
December	Whether or not a campaign started in December. This is a proxy for holiday season which is christmas and approaching new year's eve.
socialMediaShares	Number of times the campaign was shared on social media
mostlyPositive	Whether or not a campaign has a higher percentage of positive sentiment over negative sentiment: $\frac{Percentage\ of\ positive\ sentiment}{Percentage\ of\ negative\ sentument}$

Hypotheses

Given our data constraint, we try to test the following hypotheses:

- *Hypothesis 1***:** The duration of the campaign until it is funded depends on the category it's listed under. People react differently to different categories. A medical campaign, is expected be funded faster than a sports campaign or a Newlyweds campaign
- *Hypothesis 2:* The duration of the campaign until it is funded varies depends on whether it was created during the holiday season (December). This is based on the assumption that people have a more altruistic behavior during the winter holiday season and hence we expect these campaigns to have shorter survival (USA Today, 2014). We define
- *Hypothesis 3:* The duration of the campaign until it is funded depends on the financial goal that the creator sets. We hypothesise that campaigns with large goals (for instance, $1M) take a longer time to get funded
- *Hypothesis 4***:** The duration of the campaign until it is funded depends on the number of times people interact with it. The campaigns are expected to be funded more quickly if more people interact with it. We measure this by the number of times this campaign was shared on social media websites such as Twitter and Facebook
- *Hypothesis 5:* The duration of the campaign until it is funded depends on the story text that the creator writes. We expect a campaign with a higher percentage of positive sentiments to be funded faster. People feel better about donating to a story written in a positive language as opposed a negative one.

Based on tests of these hypotheses we believe that we can make a conclusion on what variables make a campaign more successful. The success of a campaign is based on how fast it got funded. The less time the campaign 'survives' until it gets fully funded, the more successful it is. The following section explain the statistical methods we use.

Statistical Methods and Results

We test our hypotheses in a survival analysis mathematical framework using the R programming language. The response variable is duration of the campaign measured by the number of days. The incident that we model our data based on is whether the campaign meets its goal or not. This is calculated as following:

$$Percentage\ funded\ =\ \frac{amount\ reached}{goal\ of\ campaign}$$

This yields values between 0 and 1. We consider a campaign to be completed if it reaches 1, meaning that it has been fully funded. However, since it's difficult to achieve the exact value that

the creator has asked for, we broaden this definition to include values:

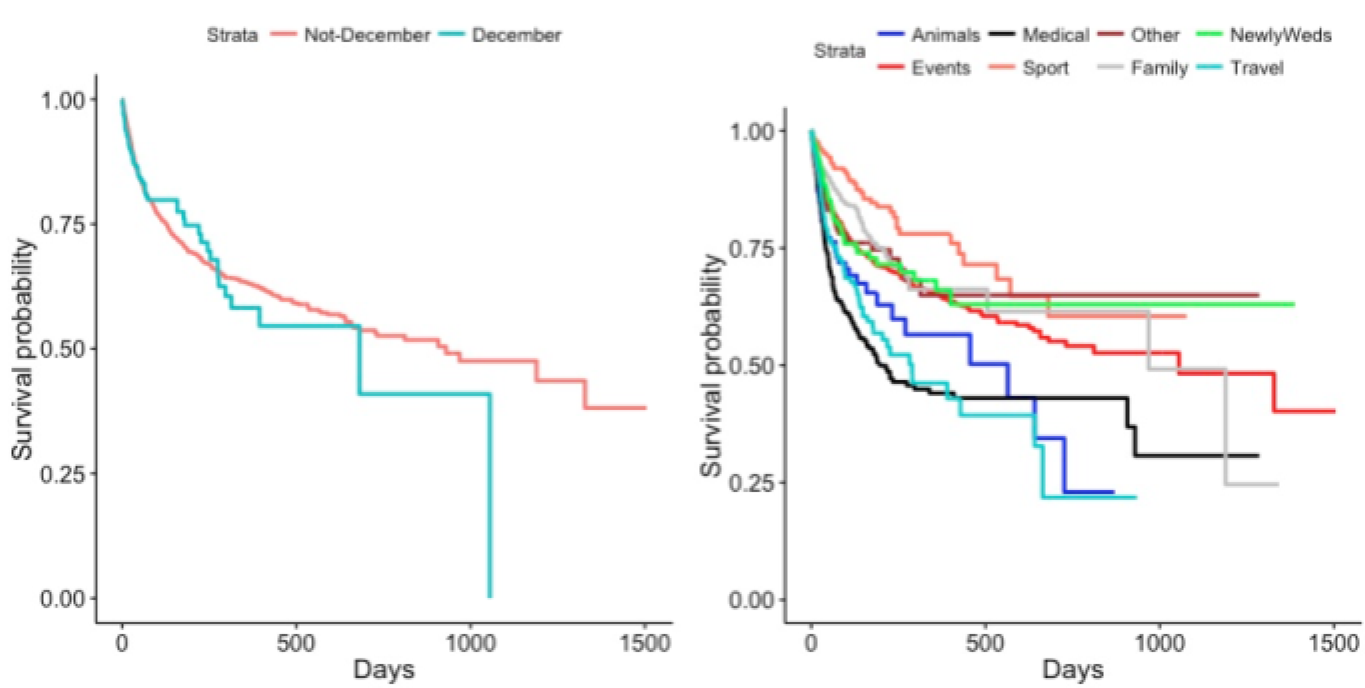
$$0.85\ < Percentage\ funded\ <\ 1.20$$

We think that this choice is justified since it won't be fair to consider campaigns that get very close to their goals as unsuccessful. Additionally, changing the lower bound to be 0.90, for instance, doesn't significantly influence our broad conclusions.

We first show the effect of our covariates on the survival time of a campaign. We use Kaplan-Meier (KM) curves to represent these differences graphically. Figure 1 shows KM curves for two categorical covariates: Category of Campaign and the Season in which the campaign was created:

Figure 1 shows that our assumption about the effect of the December variable on the survival of the campaign isn't necessarily accurate. It seems like all categories follow a similar trend up to 500 days. After that, campaigns started in December seem to drop at a faster rate until 1000 days, but the trend changes afterwards. For the category covariate, the KM curves support our hypothesis. Campaigns listed under the Medical category, on average, have a shorter survival time. Around 60% of the campaigns listed under 'Medical' get funded within the first 500 day period while only 25% of the campaigns listed under "Sport" get funded within the same period. This is justified as

Figure 1: Kaplan-Meier plot for the survival of campaigns by category and season



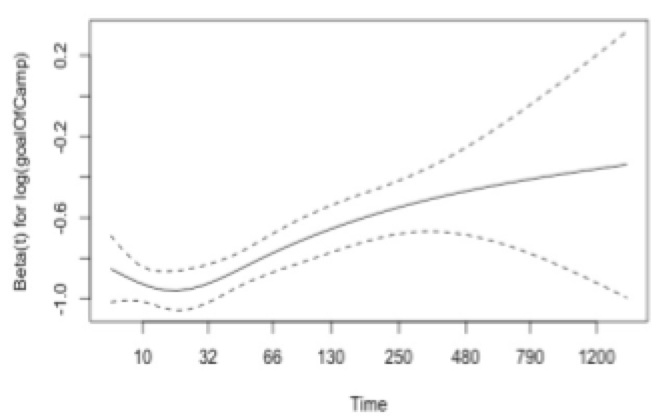
people tend to react faster to medical emergencies. However, these results are shown without controlling for other covariates.

The second approach we take is to fit an Accelerated Failure Time (AFT) model for our data to test it more formally. We considered different parametric distributions (Log-normal, Normal, Exponential and Weibull) to represent variance in our dataset. Table 2 in the appendix shows the regression coefficients for all of these models. We then measured the relative quality of these models using Akaike Information Criterion (AIC) and conclude that the Weibull distribution is the most adequate model. Based on this model, *categoryMedical* has a coefficient of -0.796 with a p-value significant at a 0.05 level. We hence can conclude that, after controlling for other variables, listing a campaign under the "Medical" category decreases its duration by a factor of 0.463 similar to what our hypothesis and KM curve predicted. According to this Weibull model, starting the campaign in the December increases its survival (coefficient = -0.222). However, this coefficient has a p-value of 0.254 which is far from being significant and does not allow us to make a conclusion.

The advantage of using AFT models is that it enables us to test the effect of categorical and continuous variables while controlling for other factors. In relation to Hypothesis 3, the model shows a similar conclusion to what we predicted: larger financial goals of campaigns are associated with a longer survival. A one unit increase in log(goal of campaign) is associated with increasing its duration by a factor of 3.26, with a very significant p-value. Number of social media shares also has a coefficient of -0.511 with a significant p-value implying that sharing the campaign more on social media reduces its survival. This result is also in accordance with Hypothesis 4. Our AFT model shows a negative and significant coefficient on the *mostlyPositive* covariate and this agrees with Hypothesis 5; Campaigns with a mostly positive sentiment in their story have significantly shorter survival time. This result was, however, not robust when changing the definition of the covariate. Using the raw percentages of negative and positive words in the story text yields an opposite conclusion[1].

Our final method of analysis is one that makes no assumption about the distribution of the data. We investigated our covariates' relationship to the duration of the campaign using a Cox Proportional Hazard (Cox PH) semi-parametric model. Given the possibility of having time-varying covariates, we tested the proportional hazard assumption using the Schoenfeld residuals. Figure 2 shows the residuals of the only covariate that violate the PH assumption.

Figure 2: Schoenfeld residuals violating Cox PH assumption



According to this graphs and the official test, we notice that the covariate *goalOfCampaign* violates the Cox PH assumption. We see a clear upward trend in the residuals implying that the model is overestimating for the first portion and underestimating for the rest. We adjust for this by stratifying the covariate in the model using the 'strata' command in R. The final outcome of our Cox PH model is shown in the second column of Table A.2 in the appendix. In most cases, we have

[1] The sentiment analysis yields three percentages that add up to 1.00: Positive, Neutral, Negative

the same conclusion as in the AFT model. It is important to keep in mind that this model is testing the hazard and not the survival, hence we expect the signs to be reversed in comparison to those of the AFT coefficients. According to the Cox PH coefficients we conclude that in agreement with Hypothesis 1, the 'Medical' category makes the campaign's duration shorter[2]. A campaign started in the December has an insignificant shorter survival time according to both the Cox PH and the Adjusted Cox PH models. Goal of campaign from Hypothesis 3 cannot be seen in Cox PH because the adjusted model is stratified based on its value. In the non-stratified Cox PH model, the goal of campaign increases the survival of the model as we predicted. However, we need to acknowledge that this was a time-varying variable as the Schoenfeld residuals in Figure 2 show. Moreover, in accordance with Hypothesis 4, increasing social media shares shortens the survival of the campaign significantly. A one unit increase in the log(socialMediaShares) decreases the hazard by a factor of 1.45. With Hypothesis 5, both the Cox PH and adjusted Cox PH models predict that a story with a mostly positive sentiment makes the campaign get funded faster with a significant p-value.

Discussion

From the previous section, we notice that our hypotheses were supported, in most cases by the statistical analysis. The AFT Weibull model showed the results we expected for all hypotheses except for Hypothesis 2. Starting a campaign in December doesn't affect the survival significantly. This can be explained by the nature of urgency in the campaigns posted on gofundme. A person creates a campaign if the need for it arises, regardless of the timing. In regard to Hypothesis 5, we concluded that writing a description with a mostly positive sentiment reduces the survival of the campaign. To illustrate this with an example, we tested the sentence; "My son is sick, and dying. Please help!" on the Linguistic Inquiry and Word Count (LIWC) webpage (http://liwc.wpengine.com/) and it came out with a score of 12.5 for negative sentiment and a score of 0 for positive sentiment. We also tested the sentence; "I would really love and appreciate your help to my sick and dying son." and it came out with a 7.1 score for negative sentiment and 14.3 for positive sentiment. The results of these two sentences show how one message can be presented in two ways: one that has a mostly positive sentiment and the other that has a mostly negative sentiment. Based on our results, people would donate more to the latter sentence than the former.

The results of our methods are supported by some of the literature. As mentioned above, although the focus of the literature has been on business focused projects, some conclusions are comparable. For instance, Mollick (2013) concludes that personal networks are associated with the success of a crowdfunding project. Our variable measuring personal networks here is the number of social media share which showed an inverse relationship with the survival of the campaign in most of our models. The humanitarian aspect of this analysis has also been emphasized by the literature. Marquis, Christopher et. al. (2015) conclude that the introduction of the online donations platform has encouraged more chinese people to engage with humanitarian and medical emergencies. This agrees with our first hypothesis in that people are more likely to react faster and engage more with a medical emergency.

[2] This result is reversed and insignificant after adjusting the Cox PH model signaling a possible correlation between goal of campaign and medical category

Conclusion

In this paper, we examined multiple survival analysis models to investigate the determinants of the duration of a crowdfunding campaign. We used gofundme.com as our case study and tested five hypotheses related to the data available. These hypotheses examine the assumption that the length of the campaign is dependent on the category it's listed under, whether or not it was created in December's holiday period, the financial goal of the campaign, the number of times the campaign was shared on social media and the sentiment of the description accompanying the campaign.

We used Weibull distribution and Cox PH models to derive coefficients on each of the mentioned variables to inform us on their significance and effect on the duration measured in days. We noticed that a campaign is likely to take shorter time if it was listed under medical, has a reasonably small goal, is shared many times on social media and has a story text whose sentiment is mostly positive. These were the results that most of our methods concluded to be significant. Whether or not a campaign was created in December wasn't significant in most cases. We also observed a time varying nature of the covariate measuring the financial goal of the campaign. We then adjusted for this by stratifying the model in an adjusted Cox PH model.

The conclusions we reached in this paper can be used to help guide how people can enhance their campaigns on sites like gofundme. For instance, if a person's campaign can be listed under medical or family, we'd advise the person to list it under medical as people would react faster to it. Similarly, people should put significant emphasis on sharing their campaign on social media. The popularity that results from social media shares seems to make a campaign significantly shorter. In relation to writing the description or the story of the campaign, it seems that people are more willing to donate to a story that is more positive. This is helpful to individuals and organizations that are humanitarian guided because they can exploit the benefits of online campaigning as a way to promote their causes. This research could also serve to give insight on people's behavior when it comes to charities in their community. The idea of a community insurance related to the mentioned themes is becoming more popular. Our findings can be used to get an insight on what makes people react faster and what people feel more obligated to react to.

Our study differs from literature by the nature of the crowdfunding website we considered. We wanted to investigate humanitarian causes and the way people react to them. The variables we considered for the success of crowdsourcing projects agree with what others in the literature found, though their work was focused on business oriented causes. These are usually related to the importance of the social network in the success of the campaign. A possible expansion of this research would be to acquire more accurate data officialy from the website for a longer time frame. It would also be beneficial to do a comparison across multiple websites using our same methods. Another interesting area to examine is the representation of the goal of the campaign as a number. Recent research in behavioral economics has shown that people react differently to numbers (Packard, 2014). A number ending in 9, for instance, is associated to different emotion in people's brains than a number ending in 0. It would be interesting to investigate whether exact values (for example, $19,569) have a shorter survival time than rounded numbers ($20,000) as they give the notion that the person who started the campaign is more considerate and aware of their own needs.

References

Brabham, Daren "Crowdsourcing as a model for problem solving an introduction and cases." Convergence: the international journal of research into new media technologies 14.1 (2008): 75-90.

Etter, Vincent, Matthias Grossglauser, and Patrick Thiran. "Launch hard or go home!: predicting the success of kickstarter campaigns." Proceedings of the first ACM conference on Online social networks. ACM, 2013.

Gerber, Elizabeth M., and Julie Hui. "Crowdfunding." ACM Transactions on Computer-Human Interaction20.6 (2013): 1-32. Web.

Greenberg, Michael D., et al. "Crowdfunding support tools: predicting success & failure." *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2013.

Greenberg, Michael D. "Public Online Failure With Crowdfunding." Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition - C&C '15 (2015): n. pag. Web.

Hutto, Clayton J., and Eric Gilbert. "Vader: A parsimonious rule-based model for sentiment analysis of social media text." *Eighth International AAAI Conference on Weblogs and Social Media*. 2014.

Li, Yan, Vineeth Rakesh, and Chandan K. Reddy. "Project Success Prediction in Crowdfunding Environments." Proceedings of the Ninth ACM International Conference on Web Search and Data Mining - WSDM '16 (2016): n. pag. Web.

Marquis, Christopher, Yanhua Zhou, and Zoe Yang. "The Emergence of Subversive Charities in China." (2015).

MacMillan, Douglas, and Gillian Tan. "GoFundMe Founders to Reap a Fortune in Buyout." *The Wall Street Journal*. Wsj.com, 24 June 2015. Web. 21 Nov. 2016.

Mollick, Ethan R., The Dynamics of Crowdfunding: An Exploratory Study (June 26, 2013). Journal of Business Venturing, Volume 29, Issue 1, January 2014, Pages 1–16. Available at SSRN: https://ssrn.com/abstract=2088298 or http://dx.doi.org/10.2139/ssrn.2088298

Network, Eleanor Mueller. "Charitable Giving Increases as Holiday Season Ramps up." USA Today. Gannett, 11 Nov. 2014. Web. retrieved: 13 Oct.2016. http://www.usatoday.com/story/news/nation/2014/11/11/charitable-giving-holiday-season/18822153/

Packard, Daniel. “Emotional Reactions to Number.” FRAUD-Magazine. June 2014. Web retrieved: 15 December. http://www.fraud-magazine.com/article.aspx?id=4294982438

Stam, Maurice. "Crowdfunding Success Prediction: From Classification to Survival Regression and back." (2016).

Tan, Douglas MacMillan and Gillian. "GoFundMe Founders to Reap a Fortune in Buyout." WSJ. Wsj.com, 24 June 2015. Web. 14 Oct. 2016.

Appendix

Table A.1: Summary Statistics of the data

Variable	Sub-category	N	Mean	Median	Sd	Max	Min
durationOfCamapaign		8539	85.55	32.00	147.32	1505.00	0.00
amountReached		8539	4471.59	1175.00	85131.16	7853140.00	5.00
goalOfCamp		8539	503166.10	5000.00	20173180.00	1000000000.00	35.00
percentageFunded		8539	0.42	0.33	0.34	1.20	0.00
peopleDonated		8953	59.88	18.00	1316.10	119497.00	0.00
socialMediaShares		8872	432.82	145.00	3402.81	244000.00	2.00
category	Family	907	-	-	-	-	-
	Events	817	-	-	-	-	-
	NewlyWeds	808	-	-	-	-	-
	Travel	504	-	-	-	-	-
	Animals	483	-	-	-	-	-
	Sports	475	-	-	-	-	-
	Medical	449	-	-	-	-	-
	Other	4096	-	-	-	-	-
Sentiment	Negative	8539	0.04	0.03	0.35	0.25	0.00
	Neutral	8539	0.80	0.81	0.06	1.00	0.52
	Positive	8539	0.16	0.15	0.06	0.46	0.00

Table A.2: Coefficients of parametric and semi-parametric models

	Dependent variable:			
	Duration of Campaign in Days			
	Cox PH	stratified Cox PH	Log-Normal	Weibull
	(1)	(2)	(3)	(4)
log(goalOfCampaign)	-0.830***		1.306***	1.182***
	p = 0.000		p = 0.000	p = 0.000
startedDecember	0.150	0.142	-0.486**	-0.222
	p = 0.274	p = 0.302	p = 0.030	p = 0.254
categoryOther	-0.087	-0.241**	-0.098	0.073
	p = 0.473	p = 0.049	p = 0.618	p = 0.671
categoryEvents	0.387***	0.302**	-0.695***	-0.587***
	p = 0.004	p = 0.025	p = 0.002	p = 0.003
categoryFamily	-0.542***	-0.900***	0.556**	0.738***
	p = 0.002	p = 0.00000	p = 0.031	p = 0.003
categoryMedical	0.501***	-0.017	-1.125***	-0.796***
	p = 0.005	p = 0.921	p = 0.00004	p = 0.002
categoryNewlyWeds	-0.265*	-0.319**	0.162	0.317
	p = 0.070	p = 0.030	p = 0.484	p = 0.124
categorySports	-0.154	-0.249	0.136	0.135
	p = 0.349	p = 0.131	p = 0.600	p = 0.561
categoryTravel	0.159	0.084	-0.547**	-0.269
	p = 0.290	p = 0.578	p = 0.028	p = 0.207
log(SocialMediaShares)	0.370***	0.257***	-0.677***	-0.511***
	p = 0.000	p = 0.000	p = 0.000	p = 0.000
MostlyPositive	0.239***	0.227***	-0.359***	-0.359***
	p = 0.001	p = 0.002	p = 0.002	p = 0.0004
Constant			-0.308	-0.035
			p = 0.392	p = 0.919
Observations	8,464	8,464	8,464	8,464
R^2	0.126	0.030		
Log Likelihood	-10,879.040	-10,142.800	-9,371.415	-9,339.881
chi^2 (df = 11)			1,119.482***	1,167.365***
LR Test	1,139.144*** (df = 11)	259.275*** (df = 10)		
Score (Logrank) Test	898.809*** (df = 11)	261.686*** (df = 10)		
Note:				*p<0.1; **p<0.05; ***p<0.01