Survival Analysis of Questions Posted on the iFixit Answers Forum

Lisa Oshita[a], Anthony Pileggi[b], Shannon Pileggi[a]

[a]Department of Statistics, California Polytechnic State University, CA, 93407; [b]iFixit, 1330 Monterey St, San Luis Obispo, CA 93401, United States

ARTICLE HISTORY
Compiled September 18, 2018

ABSTRACT

iFixit's community-driven online question and answer forum (CQA), *Answers*, provides a platform for thousands of users to ask and answer device repair-related questions. This paper presents a survival analysis of the time until a question posted on iFixit's *Answers* forum receives it's first response. A Cox proportional hazards model was developed to identify variables associated with response time. Though several predictors were significant, the model's predictive accuracy was low ($R^2 = 0.15$). Significant predictors included the device category of the question (questions pertaining to Apple products received answers faster than others (HR = 2.54, 95% CI = (2.32, 2.79)), factors related to the question's title (e.g., if it was phrased as a question (HR = 1.29, 95% CI = (1.21, 1.38))), and the day of the week the question was posted (questions posted over the weekend received answers slower than those posted on a weekday (HR = 0.87, 95% CI = (0.82, 0.93))). Future studies can investigate whether factors identified as significant in this analysis can be generalized to other CQAs.

KEYWORDS
Cox regression; CQA forum; Kaplan-Meier method; response time; survival analysis; survival probability;

1. Introduction

Founded in 2003, iFixit teaches thousands of users around the globe how to fix their own broken devices. This company provides over 40,000 free online repair guides and hosts a CQA, the *Answers* forum, featuring questions pertaining to over 11,000 devices with over 130,000 user-posted solutions. One critical key feature of CQAs like *Answers* is question response times. Fast response times have been shown to be associated with positive user experience and increased web traffic—factors that are extremely valuable to the reputation and longevity of the CQA [18]. Analysis of response times can reveal factors that affect how quickly questions receive answers, which can lead to suggestions for how users can ask questions to minimize response times and for how forum design can be improved.

However, analysis and prediction of response times on CQAs have not been thoroughly investigated. There is need for further analysis in this area, as the majority of existing research focuses on assessing and predicting question or answer quality. This paper presents a survival analysis on the time until a question receives its first

CONTACT Lisa Oshita Email: l.oshita@yahoo.com

answer on iFixit's *Answers* forum, in order to determine factors significantly related to response time and to predict survival probabilities of questions.

1.1. Related Work

With the recent increase in the popularity and use of CQAs, these data-rich platforms have been the subject of a multitude of studies. Regarding analysis of response times, [2] developed a classification model using both textual and non-textual features to estimate response times for questions posted on *Stack Overflow.* [13] used parametric methods and proposed models to predict response times based on exponential distributions. [1] instead focused exclusively on unanswered questions and created a model to predict how long a question would remain unanswered. However, all research mentioned restricted analysis to either questions for which response times were observed or to questions that remained unanswered. The research presented in this paper uses survival analysis, a method not yet applied to CQAs, to allow for analysis of both answered and unanswered questions in a unified framework.

The majority of existing research has been focused on predicting question and answer quality. [21] developed a classification algorithm to assess the quality of posts on *Nabble.com* using primarily textual features of questions. [4] also determined that textual features (e.g. word count) was the most accurate predictor of Wikipedia article quality (Wikipedia articles can represent the same kind of user-generated content featured on CQAs). [7] instead found non-textual features, like revision and comment count or the number of points a user has, were the most useful indicators of quality. A number of other studies have also developed classification algorithms using textual and non-textual features with the similar goal of predicting question or answer quality [16, 17, 20, 22]. The present analysis seeks to utilize both textual and non-textual features of questions on iFixit's *Answers* forum to investigate if factors that indicate a question is of high quality also lead to faster response times.

2. Materials

The data analyzed contained 8,025 questions posted from April 8, 2017 to July 7, 2017 (date of data download). Variables in the data included: device name, device category, question title, text, tags, new user status, date and time the question was posted, and date and time the first answer was received. Variables derived can be found in Table 1. The Appendix contains details on variable derivation, as well as an example of an answered and unanswered question in Table 5.

3. Methods

Questions analyzed were restricted to those posted in English. Time until event was defined as the time since posting until a question received its first answer. For questions that did not receive an answer by the download date, time until event values were defined as the time since posting to the time the data was downloaded. Such questions were considered right-censored, meaning exact answer times for these questions are greater than the observed download time (questions may receive answers after the download date) [12].

Survival was defined as the event that a question did not receive an answer beyond a certain time, t. Estimates of survival probability were generated with the Kaplan-Meier method, which adjusts to the presence of right-censored observations, or unanswered questions [3]. From these estimates, survival curves were constructed to examine overall survival experience. Mean and median survival times were also computed.

As the probability distribution for response times is unknown, a non-parametric Cox proportional hazards model was built to predict question survival probability [14]. To construct the model, the full data was partitioned into five separate sets [19]. Univariate analysis, performed on one set, was used to identify variables to include in the final predictive Cox model [9]. For this process, each predictor was entered into separate Cox models and evidence of association with response times was assessed. Those with partial likelihood ratio p-values of less than 0.001 were included in the final model. Transformations and restricted cubic splines were also explored through a similar process. The combination of which transformation, square root or log-transformed, and restricted cubic spline with three, four, or five knots, were selected for each continuous predictor based on the AIC statistic [11]. All significant categorical and continuous predictors, in final form, were included in the model for five-fold cross-validation.

In each iteration of cross-validation, the full model was built on the training set and used to compute predicted hazard ratios on the corresponding test set. To assess performance, predicted hazard ratios were entered into separate Cox models as the single quantitative predictor with response times as survival time. The resulting Nagelkerke's R^2 statistic, concordance statistic, Somers' Dxy, partial likelihood ratio and p-value, were assessed as performance indicators [5]. The partial likelihood ratio is a ratio of the log partial likelihood function evaluated at the parameter estimates (equivalent to a goodness of fit measure of the model with predictors) and the log partial likelihood function for the null model without predictors and only the baseline hazard function (equivalent to a goodness of fit measure for the null model). The p-value for this statistic was calculated from the χ^2 distribution [15]. Concordance statistics and Somers' Dxy are a measure of the model's discriminative ability. Concordance is defined as the probability that for any two randomly chosen questions, the question with the shorter response time also has the higher predicted hazard. Concordance statistics close to 1 indicate high discriminative ability, while statistics close to 0.5 indicate discordance, or random predictions. Somers' Dxy is the difference between the model's concordance statistic and discordance, 0.5. A value of 0 indicates random predictions, while a value of 1 indicates perfect predictions [11]. These metrics were computed for each iteration on every training and test set. Averages across training and test sets were evaluated. The final model was then fit to the full data and the same metrics were computed and compared.

Correlations between scaled Schoenfeld residuals, differences between observed and expected predictor values for questions that received answers, and a function of time, were examined to assess the proportional hazards assumption that the effect of predictors on hazard does not depend on time [8]. Significant correlations would indicate a violation of this assumption.

All computation was executed under `R` version 3.4.3. Kaplan-Meier estimates of survival probability were computed with the `Surv` and `survfit` functions of the `Survival` package in R. To compute predicted survival probabilities using the final model, the `predictSurvProb` function was used from the `pec` package. This function uses the Cox model and a vector of times to compute predictions. Data used for this analysis was published on Mendeley Data (DOI: 10.17632/shfkj785yb.1). Code for variable setup and analysis is available upon request.

4. Results

Of 8,025 questions in the full data, 7,760 were in English (97% of the full data). Of those questions, 4,951 (63.8%) received an answer by the download date. The shortest response time was 0.5 hours. The longest was 2,159 hours (90 days). Figure 1 shows the distribution of response times for all questions analyzed.

Figure 2 shows the Kaplan-Meier estimated survival probability curve for all questions in the data. The curve indicates that if a question does not receive an answer soon after it is posted, the likelihood of it receiving an answer in the future is low. The Kaplan-Meier estimated mean survival time, or the average time until a question received its first answer was 776 hours, or 32.3 days. The median survival time, the time at which 50% of the questions in the data received an answer, was 9.2 hours.

Each training set contained 6,208 questions; each test set contained 1,552 questions. Univariate analysis performed on one training set indicated that all categorical predictors were significant with partial likelihood ratio p-values of less than 0.001. Continuous predictors, excluding device length and average tag length, also resulted in p-values of less than 0.001. Applying either square root or log transformations to all continuous predictors resulted in p-values of less than 0.001. The following transformations were applied to continuous variables: square root transformation of average tag frequency, log plus one transformation of average tag length, log plus one transformation of device name length, square root transformation of line break to text length ratio, and log transformation of text length.

Table 2 shows the results of determining the optimal number of knots for each continuous predictor. Those marked with an asterisk (*) were included in the final model.

Average performance metrics for test and training sets are in Table 3. Partial likelihood ratios and p-values indicate that the model as a whole is significantly associated with response time. However, R^2 statistics and discrimination indexes are considerably low. Metrics for the final model's performance on the full data, also found in Table 3, are consistent with metrics found in cross-validation and indicate low predictive accuracy. Results for training, test, and the full data did not change significantly, indicating that the model was not overfit.

Assessing the proportional hazards assumption indicated that several levels of the device categorization variable, specifically Apple Product, Camera, Game Console, and Other, violated the assumption.

The final model trained on the full data set was significant with a partial likelihood ratio of 1265.29 (p-value <0.0001). Its R^2 statistic was 0.15, and Somers' Dxy was 0.27.

The following are interpretations of select hazard coefficients in Table 4. Hazard is approximately equivalent to the conditional probability that a question will receive an answer within the next moment in time, given that it has not already received an answer.

- The estimated hazard of receiving an answer is 154% higher (95% CI (132%, 179%)) for questions pertaining to Apple products than the hazard for questions about Android and Other phones, controlling for all other predictors.

- The estimated hazard of receiving an answer is 25% lower (95% CI (19%, 29%)) for questions with titles that contain at least one word considered to be frequently-used among unanswered questions, than the hazard for questions with titles that do not, controlling for all other predictors.
- The estimated hazard of receiving an answer for questions posted on a weekend is 13% lower (95% CI (7%, 18%)) than the hazard for questions posted on a weekday, controlling for all other predictors.

5. Discussion

Results and coefficients of the final model reveal how users can ask questions in order to decrease response time. Findings suggest that users should phrase the title in the form of a question, use correct grammar (e.g. punctuation and capitalization), include specific and concise tags, and post the question on a weekday.

The model's weak predictive accuracy, which may be explained by limitations in the data, lead to suggestions for changes in CQA design. Many users incorrectly specified the device name (e.g. a user asking a question about a Turtle Beach Ear Force XO ONE headset defined the device name as, "Turtle Beach Ear Force Xmy grandson chewed through the wire while he was playing it's brand-new is there anyway I can have it fixed0 One"), or did not include a device name. Many users also incorrectly used the tagging system by including ambiguous and lengthy tags like "someone sat on it :(" or "help me please!!!!!" (tags are generally a couple key words that allow the answering community to quickly discern the question's topic). It is likely that these inconsistencies contributed to the final model's low predictive power. This, along with the results of the final model, reveal some of the ways the CQA can be structured to potentially decrease response times. As findings indicate that questions with correctly defined tags and device names may lead to quicker response times, the CQA can provide a stricter framework for asking questions. Allowing users to enter any device name or tag leaves room for user error, shown by the examples described above. Instead, the CQA can restrict the tags or devices users can include to a drop-down list. The CQA can also include more tips to guide users asking questions. Implementing a more structured framework for asking questions can help users create understandable and clear questions and in turn decrease response times, as well as create a set of consistent questions for improved analysis.

Acknowledgements

We would like to thank iFixit for providing access to the CQA data.

Disclosure Statement

We wish to confirm that there are no known conflicts of interest associated with this publication.

Funding

This research was supported by the Bill and Linda Frost fund of the California Polytechnic State University of San Luis Obispo. This sponsor was not involved in any aspect of this paper's analysis.

References

[1] M. Asaduzzaman, A.S. Mashiyat, C.K. Roy, and K.A. Schneider, *Answering questions about unanswered questions of Stack Overflow*, 2013 10th Working Conference on Mining Software Repositories (MSR) (2013), pp. 97–100. Available at http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6624015.

[2] V. Bhat, A. Gokhale, R. Jadhav, R. Pudipeddi, and L. Akoglu, *Min (e) d Your Tags : Analysis of Question Response Time in StackOverflow*, IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (2014), pp. 328–335.

[3] J.M. Bland and D.G. Altman, *Statistics Notes: Survival probabilities (the Kaplan-Meier method)*, Bmj 317 (1998), pp. 1572–1580. Available at http://www.bmj.com/cgi/doi/10.1136/bmj.317.7172.1572.

[4] J.E. Blumenstock, *Size matters: word count as a measure of quality on wikipedia*, in *Proceedings of the 17th International Conference on World Wide Web*. ACM, 2008, pp. 1095–1096. Available at http://portal.acm.org/citation.cfm?id=1367673.

[5] H.C. Chen, R.L. Kodell, K.F. Cheng, and J.J. Chen, *Assessment of performance of survival prediction models for cancer prognosis*, BMC Medical Research Methodology 12 (2012), p. 102. Available at https://doi.org/10.1186/1471-2288-12-102.

[6] D. Correa and A. Sureka, *Fit or unfit: Analysis and prediction of 'closed questions' on Stack Overflow*, Proceedings of the first ACM Conference on Online Social Networks (2013), pp. 201–212. Available at http://dl.acm.org/citation.cfm?doid=2512938.2512954.

[7] H. Fu, S. Wu, and S. Oh, *Evaluating answer quality across knowledge domains: Using textual and non-textual features in social Q&A*, Proceedings of the Association for Information Science and Technology 52 (2015), pp. 1–5.

[8] P. Grambsch and T. Therneau, *Proportional hazards tests and diagnostics based on weighted residuals*, Biometrika 81 (1994), pp. 515–526. Available at https://doi.org/10.1093/biomet/81.3.515.

[9] K.E. Hammermeister, T.A. DeRouen, and H.T. Dodge, *Variables predictive of survival in patients with coronary disease. Selection by univariate and multivariate analyses from the clinical, electrocardiographic, exercise, arteriographic, and quantitative angiographic evaluations.*, American Heart Association, Inc. 59 (1979), pp. 421–430. Available at http://dx.doi.org/10.1161/01.CIR.59.3.421.

[10] F.M. Harper, D. Raban, S. Rafaeli, and J.A. Konstan, *Predictors of Answer Quality in Online Q&A Sites*, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (2008), pp. 865–874. Available at http://dl.acm.org/citation.cfm?id=1357054.1357191.

[11] F.E. Harrell, *Regression Modeling Strategies*, Vol. 64, 2015.

[12] D. Kleinbaum and M. Klein, *Survival Analysis: A Self-Learning Text, Third Edition (Statistics for Biology and Health)*, in *Biometrical Journal*, M. Gail, K. Krickeberg, J. Samet, A. Tsiatis, and W. Wong, eds., 3rd ed., chap. 1, Springer-Verlag New York, 2011, p. 715.

[13] J. Mahmud, J. Chen, and J. Nichols, *When Will You Answer This? Estimating Response Time in Twitter*, in *Seventh International AAAI Conference on Weblogs and Social Media.* 2013, pp. 697–700.

[14] D. Moore, *Applied Survival Analysis Using R*, 1st ed., chap. 5, March, Springer International Publishing, 2010, p. 226.

[15] D. Oakes, *Survival Times: Aspects of Partial Likelihood*, International Statistical Review / Revue Internationale de Statistique 49 (1981), pp. 235–252. Available at http://www.jstor.org/stable/1402606.

[16] L. Ponzanelli, A. Mocci, A. Bacchelli, and M. Lanza, *Understanding and classifying the quality of technical forum questions*, Proceedings - International Conference on Quality Software (2014), pp. 343–352.

[17] S. Ravi, B. Pang, V. Rastagori, and R. Kumar, *Great Question ! Question Quality in Community Q & A*, International AAAI Conference on Weblogs and Social Media (2014), pp. 426–435.

[18] A. Rechavi and S. Rafaeli, *Not all is gold that glitters response time & satisfaction rates in Yahoo! answers*, Proceedings - 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, PASSAT/SocialCom 2011 (2011), pp. 904–909.

[19] J.D. Rodríguez, A. Pérez, and J.A. Lozano, *Sensitivity analysis of kappa-fold cross validation in prediction error estimation*, IEEE transactions on pattern analysis and machine intelligence 32 (2010), pp. 569–575.

[20] H. Toba, Z.y. Ming, M. Adriani, and T.s. Chua, *Discovering high quality answers in community question answering archives using a hierarchy of classifiers*, Information Sciences 261 (2014), pp. 101–115. Available at http://dx.doi.org/10.1016/j.ins.2013.10.030.

[21] M. Weimer, I. Gurevych, and M. Mühlhäuser, *Automatically assessing the post quality in online discussions on software*, Proceedings of the ACL (2007), pp. 125–128. Available at http://atlas.tk.informatik.tu-darmstadt.de/Publications/2007/acl2007.pdf.

[22] Y. Yao, H. Tong, T. Xie, L. Akoglu, F. Xu, and J. Lu, *Detecting high-quality posts in community question answering sites*, INFORMATION SCIENCES 302 (2015), pp. 70–82. Available at http://dx.doi.org/10.1016/j.ins.2014.12.038.

6. Appendices

Appendix A. Posted questions

Table 5 provides observed and derived variable values from an example of an answered question (regarding the iPhone 6) and unanswered question (regarding android tablet). All fields in device, title, text, and tags are shown as the users entered it.

Appendix B. Variable Derivation

B.1. Average tag frequency

This variable was created to investigate if including popular or widely-used tags in a question can lead to faster response times, based off of findings from [2]. This was computed as the average frequency, or proportion of times a question's tags appear in all of the data set. Questions without tags were assigned a value of 0. If a question has a higher average, then at least one of its tags are considered frequently-used.

B.2. Average tag length

This variable was created to investigate if questions with correctly defined tags have faster response times than questions that do not include tags or incorrectly use the tagging system. It is hypothesized that questions with ambiguous or lengthy tags make it difficult for the answering community to discern the question's topic, and thus have slower response times or remain unanswered.

B.3. Device categorization

Original device categories defined by iFixit included: Apparel, Appliance, Camera, Car and Truck, Computer Hardware, Electronics, Game Console, Household, Mac, Media Player, PC, Phone, Skills, Tablet, Vehicle. Device titles were parsed and certain categories were combined or separated to create a new device categorization variable to use in the Cox model. Final categories created included: Android/Other Phone, Apple Product, Camera, Electronics, Game Console, Home, Other, PC, Tablet, Vehicle. Under the original device categorization, 1,954 questions (25.2% of 7,760) were categorized incorrectly and indicated an NA for its category. Missing values were a result of users creating questions for devices not already in the iFixit's database, or from the user incorrectly defining the device name. For questions with missing categories, key words were searched for in device titles to re-categorize accordingly. All other questions with ambiguous device titles were categorized as Other.

B.4. Device name length

This variable was created to investigate if questions with correctly defined device names have faster response times than questions that do not include a name or are incorrectly defined. Similar to the average tag length variable, it is hypothesized that questions with incorrectly defined device names (often lengthy) make it difficult for the answering community to discern the question's topic, and thus have slower response times.

B.5. If the question's text contains end punctuation

This variable was created to investigate if sentences with no end punctuation take longer to receive an answer.

B.6. If the question's title contains terms considered frequently-used among answered/unanswered questions

These variables were created to investigate if questions pertaining to popular topics receive answers faster than questions that do not. Similarly, another variable was created to investigate if questions pertaining to unpopular topics receive answers slower than questions that do not. This variable was created based off of methods from [6] and [17].

To create the variables, the data was separated between answered and unanswered questions. For the data frame containing answered questions, text mining techniques were used to create a list of every word within each questions' titles and the proportion of times those words occurred among all answered questions' titles, defined as the frequency. The same was performed for the data frame containing unanswered questions. "Frequently-used" words in answered questions were defined as those that appeared in 1% or more of all answered questions' titles, and appeared in more answered questions than in unanswered questions. To determine the latter, a ratio of frequencies was assessed. The ratio was calculated as the proportion of times a word occurred among answered questions, to the proportion of times a word occurred in unanswered questions. As an example, if “cracked” appeared in 2% of answered questions and 0.1% of unanswered questions, it would be considered “frequently-used” among answered questions as it occurs in more than 1% of answered questions’ titles and occurs 20 times more in answered question than in unanswered questions (ratio = 0.02/0.001 = 20). Similarly, “frequently-used” words in unanswered questions must occur in 1% or more of all unanswered question’s titles and occur in more unanswered questions than answered questions. As there was some overlap between “frequently-used” words in each list and levels of the device categorization variable, every word that matched a device name was removed from the lists. Table 6 contains the resulting terms for answered questions (111 words) and for unanswered questions (32 words).

B.7. If the question’s title ends in a question mark

This variable was created to investigate if questions with titles in the form of questions receive answers faster than those otherwise phrased [2].

B.8. If the user made an effort to solve the problem prior to asking the question

This variable indicates true if the user included words in the question’s text that indicate the user made prior effort to find the solution before posting. Words searched for include: tried, searched, researched, tested, replaced, used, checked, investigated, considered, measured, attempted, inspected, fitted. This variable was created based off of findings and methods from [2] and [10].

B.9. Line break to text length ratio

This variable was created to investigate if questions that include line breaks in the text are easier to read than questions that do not include any, and thus have shorter response times. Questions with a ratio of 0 tend to include either short, ambiguous texts (e.g. ”Start button not working.”) or long and lengthy texts, and tend to remain unanswered or have slower response times. Questions with a ratio of greater than 0 tend to have faster response times.