Rough Draft

Lisa Oshita

1 Abstract

Community-driven online question and answer forums (CQA) are becoming increasingly valuable sources of information. These platforms house an expansive amount of crowd-sourced knowledge in the form of thousands of questions and answers posted everyday. There are forums that cover a broad range of topics, like *Yahoo! Answers*, and forums focused on specific topics, like computer programming-focused *Stack Overflow*. An example of the latter is iFixit's *Answers* forum. The *Answers* forum features questions asked by users specifically related to device repair, which are answered by both repair experts and everyday users, furthering iFixit's mission to enable individuals to repair their own devices. Thus, it is important that questions receive timely answers. This paper presents a survival analysis on the time until a question receives its first answer. We developed a Cox proportional hazards model to predict the failure probability of a question, or the probability that a question receives an answer before a certain time. Though we identified signifcant predictors, the predictive accuracy was low ($R^2 = 0.15$). Our findings indicate that the most important predictors were the device category of the question (questions pertaining to Apple products received answers faster than others (HR = 2.56, 95% CI = (2.31, 2.82))) and factors related to the question's title (e.g., whether or not it was phrased as a question (HR = 1.31, 95% CI = (1.22, 1.41))). Future studies could investigate if the factors identified as signifcant in our study can be generalized to other CQAs.

2 Introduction

Community-driven online question and answer forums (CQA) are becoming widely used sources of information. These online platforms allow for anyone to ask or answer a question, regardless of their background or expertise level. One popular CQA is Yahoo! Answers, featuring questions of all topics and expertise levels. On the other hand of the spectrum is Stack Overflow, the CQA that features strictly computer programming question and answers. Forums like these receive around 40 million visits every month.

The CQA analyzed in this paper is iFixit's *Answers* forum. Founded in 2003, iFixit helps thousands of people repair their broken devices everyday, by providing over 30,000 free on-

line repair guides and by selling the specialized tools and parts needed for such repairs. The mission of iFixit is to enable people to extend the lifetime of their own devices, effectively saving them money and reducing electronic waste.

As not all possible repairs are covered in the published repair guides, and users may have questions related to those guides, iFixit's *Answers* forum is another important resource. This platform features questions related to over 9,000 devices with over 100,000 solutions. Questions range from broken devices like jammed zippers to shattered iPhone screens. As many rely on this forum for help with repairs, it is important that users receive timely answers. Fast response times will enhance user engagement and generate more web traffic, which is valuable to the reputation and longevity of the *Answers* forum. Analysis of answer times can reveal factors that affect how quickly questions are answered, which can lead to suggestions for how users can ask better questions to minimize answer times, and for how the forum design can be improved.

However, analysis and prediction of answer *times* on CQAs have not been thoroughly investigated. A majority of existing research focuses on assessing and predicting question and answer quality. Therefore, there is need for further analysis of response times in these forums. This paper presents a survival analysis on the time until a question is answered on iFixit's *Answers* forum, in order to determine factors significantly related to answer time and to predict the "failure" probability of a question.

3 Related Work

In regards to investigation of forum response times, (**?**) developed a classification model to analyze response times of questions posted on *Stack Overflow*, and found that tag-based features like the number of tags included or the number of subscribers a certain tag has, were the best predictors of answer time.

(**?**) found that the swift answer times of *Stack Overflow's* community is a result of the reputation system and the strict emphasis on factual and informative questions and answers, rather than discussion-based.

(**?**) analyzed unanswered questions on *Stack Overflow* to determine common characteristics and found that questions that went unanswered shared certain characteristics in that they were too short and vague, or utilized the tagging system incorrectly.

4 Materials

good question ex: id 408124, 406786 (answered) bad question ex: id 410310, 405882 (unanswered)

The data analyzed contained 8,025 questions posted from April 8, 2017 (10:14 PM) to July

7, 2017 (9:28 PM) (the date the data was downloaded). Variables in the data included: device name and category, title, text, tags, whether or not the user was a member of iFixit's site for less than one day before the question was posted, date the question was posted, date the first answer was received. Variables derived:

Categorical Variables:

- Device category the question pertains to. Categories include: Apple Products, AndroidOther Phone, PC, Tablet, Electronics, Camera, Vehicle, Game Console, Home, Other.
- Whether or not the question's title contains at least one word that is considered "frequently used" among answered questions. See appendix for a complete list of these terms.
- Whether or not the question's title contains at least one word that is considered "frequently used" among unanswered question. See appendix for a complete list of these terms.
- Whether or not the question's title ends in a question mark.
- Whether or not the question's text contains any end punctuation marks (. ? !).
- Whether or not the question's text is in all lower case.
- Whether or not the user edited or added information to the question's text sometime after posting it.
- Whether or not the user made an effort to solve the problem on their own, prior to asking the question.
- The day of the week the question was posted.

Numeric Variables:

- The average tag "score" for all of a question's tags. Score is defined as the proportion of times a tag appears in all of the data. Questions without tags were assigned an average tag score of 0.
- The average number of characters in each question's tags.
- The number of characters in the question's text.
- The number of characters in the user-defined device name.
- The ratio of the number of line breaks to the number of characters in the question's text.

5 Methods

Questions analyzed were restricted to those posted in English (97% of the full data). The time until event variable used in survival analysis is defined as the time since posting until a question receives 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. These questions were considered right censored, meaning that the exact answer times for these questions are higher than the recorded times (questions can still receive answers after the download date).

Survival was defined as the event that a question remains unanswered beyond a certain time, t. Estimates of survival probability were generated with the Kaplan-Meier method (cite). This method adjusts to the presence of right-censored, or unanswered questions. Estimates factor in the number of questions that have not been answered on each ordered interval of complete answer times. The estimates of survival probablity are the products of all probabilities up to that interval's time (fix this explanation). From these estimates, survival curves were constructed to examine the survival experience of questions. Mean, median and other percentiles of survival times were also generated.

Five-fold cross-validation was used for model building. The full data was split into five training and five test sets. Each training set contained 6,208 questions, and each test set contained 1,552 questions. Univariate analysis, performed on one of the training sets, was used to identify variables to include in the full Cox proportional hazards model. Predictors were entered into separate Cox models and the strength of association between those predictors and answer times were assessed. Those with partial likelihood ratio test p-values of less than 0.01 were included in the full model.

In each fold of cross-validation, the full model, with variables found to be signficant in univariate analysis, was built on one of the training sets and used to generate predicted hazard ratios on the corresponding test sets. To assess prediction performance, the predicted hazard ratios were entered into separate Cox models as the single quantitative predictor with answer times as the survival time. The resulting estimated hazard ratio, R-square statistic, partial likelihood ratio and p-value were assessed as indicators of the full model's performance (cite). Signficant statistics would indicate high predictive accuracy. Concordance statistics and Somers' Dxy were also assessed to measure the model's predictive discrimination (cite). Concordance is defined as the probability that for any two randomly chosen questions, the question with the shorter survival 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) (cite). These metrics were computed for each training and test set. The average of metrics across test sets was evaluated.

To select the final model, the model with the lowest AIC statistic and most consistent performance metrics, in that the values did not change substantially from training to test data sets, was selected.

To assess model adequacy, model residuals were examined. Martingale residuals plots were evaluated for each continuous predictor to assess the functional form. Martingale residuals are the differences between the observed number of answers received by question i, and the expected number of answers experienced by question i under the model. Deviance residuals, which are transformed martingale residuals with a mean of 0 and roughly normal distribution, were assessed to check for outliers, or questions with outcomes that are poorly predicted by the model. Score residuals, which are a measure of the approximate change in the parameter estimates that would occur if the ith question were removed from the sample, were assessed to acheck for influential questions, or questions that highly effect the parameter estimates.

After assessing model residuals, splines and other transformations were applied to variables. Restricted cubic splines were applied to continuous predictors using the method outline in **?**. As our data was fairly large (¿100), splines of 5 knots were initially fit to each continuous predictor. A higher number of knots were assigned to predictors with lower p-values in the model, those thought to be more important. The optimal number of knots for each spline was determined by the model with the lowest AIC statistic. Other transformations (i.e. square root and polynomial terms) were also applied to variables.

An assumption upon which Cox regression models are built, is the assumption of porportional hazards, or that the effect of predictors on hazard doesn't depend on time. To assess this assumption, the correlations between scaled Schoenfeld resiudals and a function of time were evaluated. Schoenfeld residuals are the differences between observed and expected predictor values for each question with a complete answer time (each question that received an answer). Signficant results from this test would indicate that a predictor has violated this assumption.

6 Results

Of 7,760 questions, 63.8% received an answer by the download date. Figure a shows the distribution of answer times for all questions, grouped by whether or not the question was answered. The following statistics were calculated from Kaplan-Meier estiamtes of survival probability. The mean survival time, or the average time until a question receives its first answer is 775.75 hours, or 32.32 days. The median survival time, the time at which 50% of the questions in the data received an answer is 9.16 hours. Table 1 provides more percentiles of survival time. Figure B shows the Kaplan-Meier curve for all questions in the data. The curve indicates that at around 100 hours after a question is posted, the probability that a question receives an answer after that time hovers around 0.35. The survival probability never reaches 0 due to the large amount of unanswered questions.

After performing univariate analysis, the predictors with partial likelihood ratio test p-values of less than 0.01 were entered into the final model. Results for univariate analysis can be seen in table 2 (results in pvalues dataframe). Each variable was entered into the Cox regression model.

The metrics for each iteration of cross validation are in table 3. Although not impressively high, the metrics did not vary substantially from test to training data, indicating that the model does not overfit.

The final model selected contained variables defined above as well as the following transformations:

- Square root of the average frequency of a question's tags.
- Quadratic polynomial on the length of the question's text.
- Restricted cubic spline with 5 knots on the number of characters in the device name.
- Restricted cubic spline with 4 knots on the average number of characters in a question's tags.
- Restricted cubic spline with 4 knots on the ratio of newlines to the number of characters in a question's text.

Results for this model can be found in table z.

Assessing residuals for model adequacy indicated that assumptions were met. Score and deviance residuals showed that there were a few questions that were outliers or highly influenced the fit of the model. However, refitting the model without those questions resulted in worse predictive accuracy. Assessing the proportional hazards assumption indicated that the "Apple Products" level of device category was the only probable variable that violated the assumption, but as stratifying on device category substantially reduced predictive accuracy, it was left as is.

7 Discussion

(Discuss importance of findings, not repeat them. A combined results + discussion section is often appropriate) The data we analyzed for this model presented some limitations. Currently, askers have a considerable amount of freedom in the way they ask a question. As a result many have incorrectly specified various input fields when asking their question. For example, a question about a faulty Android tablet screen included as a tag "someone sat on it :(". Generally, tags are not more than a couple key words that describe the topic of the question. Many askers on this forum included lengthy and somewhat ambiguous tags in their question. Another issue was the incorrect naming of the device the asker's question was about. A user asking a question about their Turtle Beach Ear Force X device included as the name of their device "Turtle Beach Ear Force Xmy grandson chewed through the wire while we was playing it's brand-new is there anyway I can have it fixedO One". The inconsistencies in this data made it somewhat difficult to analyze, and possibly contributes to the model's low predictive accuracy.

8 Conclusion

This study developed a Cox regression model to predict the probability that a question posted on iFixit's *Answers* forum receives an answer before a certain time. Predictors found to be signficant in the model included: device category, whether or not the question contained words considered to be "frequently-used" among unanswered questions, whether or not the title ends in a questionmark, whether or not the user updated the question's text after the initial posting, whether or not the user indicated that he or she made an effort prior to answer their question prior to posting it (how should I write out that different knots of the splines were signficant). While overall the model was signficant, its predictive performance was considerably low.

9 Acknowledgement

This research is supported by the Bill and Linda Frost fund.

10 Appendix

device

The data contained the original category variable as defined by iFixit. This category variable contained NAs (include how many), a result of the asker creating a question for a device not already on the website's data base. Questions that made the device clear were categorized accordingly. However there were a number of questions that did not explicitly declare the device their question pertained too. These questions were categorized as "Other". All Apple products (e.g., iPhones, iMacs, Apple watches) were given their own category. After pulling out iPhones from the original "Phones" category, the remaining phones were categorized as "AndroidOther Phone". Appliance and Household categories of the original variable were merged into "Home". "Car and Truck" were added to the "Vehicle". Lastly, any category that contained less than 2% of the questions were categorized as "Other". All other categories remained the same as the original category. The final categories in the new categorization variable were Apple Product, Android/Other Phone, PC, Tablet, Electronics, Camera, Vehicle, Game Console, Home, and Other.

contain answered and unanswered

These variables were created based on the intuition that certain topics might be more popular, and even unpopular, among the answering community, and that questions concerning these topics might recieve an answer faster, or slower, than questions that don't.

To create these variables, the data was separated between answered and unanswered questions. Using text mining techniques, two lists were created-one for answered questions and another for unanswered questions–containing every word within the question's titles

and the frequency or the number of time they occured among answered and unanswered questions. For "frequently-used" words in answered questions, that word would have to appear in more than 1% of answered question's titles, and would have to appear in answered questions more than it appeared in unanswered questions. To determine the latter, a ratio of proportions was assessed. This ratio was calculated as the proportion of times a word occured among answered questions, to the proportion of times a word occured in unanswered questions. As an example, if "cracked" appeared in 2% of answered questions and in 0.1% of unanswered questions, it would be considered frequently used among answered questions as it occurs in more than 1% of answered question's titles and occurs 20 times more in answered questions and in unanswered questions (ratio = 0.02/0.001 = 20). As for "frequently-used" words in unanswered questions, they must occur in 1% or more of the question's titles and occur in more unanswered questions than unanswered questions (ratio ¡ 1). Since there was some overlap between the words in each list and the device categories, every word that matched a device name was removed from the list. The resulting list for answered questions was 111 words, and 32 words in uanswered questions. $\text{Contain}_{a}nsweredandcontain_{u}nasnweredarelogicalvariables, indicatingtrueifaquestion'stitlecontaineda$

111 terms found to be "frequently-used" among answered questions (ordered from highest frequency): screen, turn, working, replacement, power, work, replace, charging, charge, button, touch, black, turning, broken, start, stuck, new, lcd, upgrade, problem, change, port, replaced, card, open, boot, replacing, remove, reset, back, drive, error, cable, ssd, cracked, hard, one, dropped, logic, lines, white, keeps, pro, dead, now, front, damage, switch, parts, glass, still, charger, issue, sim, turns, digitizer, just, mode, model, backlight, usb, stopped, logo, starting, know, unresponsive, password, factory, call, use, damaged, find, sensor, possible, fixed, side, galaxy, data, ipod, problems, issue, slow, system, connector, without, overheating, code, ram, air, microphone, please, red, much, plugged, getting, booting, left, way, buy, plus, time, loose, lock, coming, got, shuts, says, install, key, door, top

32 terms found to be "frequently-used" among unanswered questions (ordered from highest frequency): sound, light, wifi, speaker, connect, picture, stay, noise, bluetooth, isnt, apps, going, rear, question, play, stop, service, take, hear, lights, showing, network, volume, come, keep, connection, flashing, shut, print, blue, buttons, edit

Title_questionmark and title_beginwh

These two variables were created to determine if stating the title in the form of a question is associated with faster answer times. $\text{Title}_{q}uestionmarkisalogicalvariableindicatingtrueifthequestion'sti$

Text_contain_punct

A logical variable indicating true if the questions text contains any end punctuation marks (. ? !). This variable was created to investigate if run-on sentences, or sentences with no end punctuation, take longer to receive an answer.

Prior_effort

Logical variable indicating true if the asker used words that indicate prior effort or research

was done before asking the question (e.g. "tried", "attempted", "tested"). This variable was created based off of findings from **?**.

Ampm (change name)

The time of day the question was posted. If the question was posted between 5am and 12pm, it was categorized as "morning". If it was posted between 12pm and 5pm, it was categorized as "afternoon", between 5pm and 8am- "evening", and 8pm and 5 am, "night"

Avg_tag_score

This variable was created to investigate the idea that some tags are more popular, or widely used than other, and that including such tags might increase the likelihood of that question recieving an answer. This variable is the average frequency, or proportion of times a question's tags appear in all of the data set. If a question has a higher average, than at least one of it's tags are frequently used. This variable was created based off of findings from **?**.

Device_length

The number of characters in the user-defined device variable. This variable was included to capture when a user incorrectly inputs the device name. For example, the device variable for question 390271 is, Turtle Beach Ear Force Xmy grandson chewed through the wire while we was playing it's brand-new is there anyway I can have it fixedO One and is 136 characters long. 11 questions in this data set also didnt include any device. This variable was created based on the intuition that users who incorrectly define their devices make it harder for answerers to discern the topic of their question, and thus have longer answer times or might not recieve an answer.

Avg_tag_length

The average number of characters in each question's user-defined tags. If a question did not include a tag, this variable was set to 0. This variable, like the device-length variable, was created to capture when a user correctly or incorrectly used the tagging system. For example, question 390989 includes tag: i need to repair the headset because i can not find the bluetoot and is 64 characters long, while question 410254 includes tags "sound", "sound driver" and "speaker" and has an average tag length of 8 characters. It is hypothesized that users who use the tagging system correctly, as with the latter user, receive answers quicker than the former user.

Newline_ratio

This ratio of the number of newlines to the number of characters in the question's text. This variable was based on the intuition that question's that include newlines in the text are easier to read and thus have faster answer times. Questions with long text that also include newlines are generally easier to read than questions that dont include any.