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. It provides a place for people of all backgrounds and levels of expertise to provide input on a certain question. 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 2 receive around 40 million visits every month. These online question and answer forums are clearly becoming popular sources of information.

The CQA analyzed in this paper is iFixit's Answers forum. Founded in 2003 by two engineering students at the California Polytechnic State University in San Luis Obispo, iFixit now helps thousands of people repair their broken devices everyday. The company provides over 30,000 free online repair guides with pictures and step-by-step instructions, and sells the specialized tools and parts needed for such repairs—tools like iPhone 7 battery adhesive strips or Nexus 9 LCD digitizers. The main goal of this company is to enable people to extend the lifetime of their own devices, effectively saving them money and reducing electronic waste.

As the goal of this company is to provide people with the knowledge necessary to repair their devices, an important part of this company is their CQA, called Answers. This platform features questions related specifically to device repair, with over 9,000 device topics and over 100,000 solutions. As many rely on this forum and the community of people who provide answers, for help with their own device repairs, it's important that these users receive timely answers. Fast response times will drive up user engagement and generate more traffic, which in turn is great for the reputation and longevity of the Answers forum. As such, investigation of response times would be both beneficial and informative. Analysis can reveal factors within the forum and from elements of the questions that affect how quickly questions are answered. Suggestions for how users can ask questions to minimize answer times and for how the forum design can be improved, can be derived from such analysis.

However, to the best of our knowledge, analysis and prediction of answer times on CQAs have not been thoroughly investigated. A majority of the existing research focuses on assessing and predicting question and answer quality. As such, 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. It attempts to determine the factors that are significantly related to answer time, and create a cox proportional hazards model that can accurately 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

The data analyzed in this paper contained 7,760 questions posted from April 8, 2017 to July 7, 2017, the date the data was downloaded. We worked exclusively with questions in English. Of the questions in the data set, 4,951, or 63.8%, recieved an answer. Questions that remained unanswered were considered right-censored. The time until event variable used in survival analysis methods was defined as the time until a question posted on the forum receives an answer. Comments posted on the question are not considered answers, and an answer does not have to be accepted as the "chosen solution" to be considered the answer. We are only considering the time until the question receives its first answer. The time until answer value for right-censored questions, questions that did not receive an answer, is defined as the time from when the question was posted, to the time the data was downloaded. Figure a shows the distribution of these answer times for all questions in the data set.

The following statistics were calculated from Kaplan-Meier estimates of questions survival probability. In this analysis, survival is defined as the event that a question remains unanswered beyond a certain time, t. As Kaplan-Meier estimates adjust for the presence of censored observations, these statistics are adjusted for the presence of unanswered questions. For questions in this data set, 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 A shows the Kaplan-Meier Curve for all questions in the data. The curve indicates that around 100 hours after 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.

Variables created:

The data contained information about each question's title, text, and device, from which we created the following variables to use in the Cox regression model.

The time until a question receives its first answer was calculated as the time, in hours, from the time the question was posted to the time the first answer was received. If the question did not receive an answer, the time until answer value was calculated as the time from when the question was posted, to the time the data was downloaded, July 7, 2017 at 2:41 PM PDT. The censoring variable indicates 1 if the question received an answer, and 0 if the question remained unanswered.

Categorical Variables:

- Device category the question pertains to. Categories include: Apple Products, AndroidOther Phone, PC, Tablet, Electronics, Camera, Vehicle, Game Console, Home, Other. Table B shows the percentage of questions within each device category.
- Whether or not the question's title contains at least one word that is considered "frequently used" among answered questions.

- Whether or not the question's title contains at least one word that is considered "frequently used" among unanswered question.
- Whether or not the question's title ends in a question mark.
- Whether or not the question's title begins with a "Wh" word (e.g. "What", "Where").
- 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 asker edited or added information to the questions text sometime after posting it.
- Whether or not the asker included a greeting (e.g. "Hello", "Greetings") in the question's text.
- Whether or not the asker used polite language in the question's text (e.g., "Thank you", "please"). (don't like the way this is worded)
- Whether or not the asker made an effort to solve the problem on their own, prior to asking the question.
- The day of the week the question was posted.
- The time of day the question was posted. Times include: Morning, Afternoon, Evening, and Night.

Numeric Variables:

- The average frequency, or proportion of times a tag appeared in all of the data, for all of a question's user-defined tags. Questions without tags were assigned a value of 0 for this variable.
- 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 newlines to the number of characters in the question's text.

5 Methods

Univariate analyses for each variable of interest were performed on one of the training data sets. Variables with a p-value of less than 0.01 were entered into the final model. As the goal of this model is prediction, it was decided that all variables would be kept in the model to maintain predictive power.

5-fold cross-validation was used to assess predictive performance. Training data sets contained 6,208 questions. Test data sets contained 1,552 questions. For each iteration, the model fit to the training data was used to calculate predicted hazard ratios from the corresponding test data. To assess prediction performance, those predicted hazard ratios were entered into a separate Cox regression model as the single predictor with the question's answer times as the survival time. The resulting hazard ratio, R-square statistic, partial likelihood ratio and p-value, Somers' Dxy and Concordance statistics for that model were used as indicators of the prediction model's performance. Significant statistics for this model would indicate that the predicted hazard ratio is signficantly associated with survival time and that our modle performs well. The concept behind these metrics came from **?**. These metrics were also computed for predictions on the training data. For each of the five iterations, there was no signficant difference between the metrics computed on the training data prediction, and the metrics computed ont he test data. Table C shows the average metrics for training and test data predictions. The metrics evaluated were the average of all metrics computed on the test data. Although these metrics are not considerably high, they do not vary signicantly from training to test data sets. (should I explain more about Somers' Dxy and Concordance?)

6 Results and Discussion

The final model included all variables described above, along with the following transformations:

- Stratification on the time of day the question was posted.
- 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.

Assessing the indepenedence between the model's Schoenfeld residuals and time indicated that the "Apple Product" category of the device variable violated the proportional hazards assumption. Other variables at risk for violation included the Camera and Game Console categories of devices, and the number of characters in the user-defined device name. As "Apple Product" was the only major violation of this assumption, and stratifying on this variable substantially decreased predictive power, these variables were left as is. Visually assessing martingale residuals for each quantitative predictor indicated that functional form for each appeared to be adequate. Assessing the deviance residuals indicated that a considerable amount of questions were not predicted well by the model. Similarily, assessing

the score residuals showed that some questions might have an influence on the fit of the model. However, fitting the model without questions identified by score and deviance residuals to be either poorly predicted or highly influential, resulted in worse predictive performance. Hence, all questions were left in the data set.

The partial likelihood ratio test statistic was 1307.15 with a p-value of 0. The R-squared value for this model is 0.1555. Significant predictors with a p-value of less than 0.01 are: the Apple product, Camera, Game Console, Home, PC, and Vehicle categories of the device variable, whether or not the question's text contained at least one term considered "frequently used" among answered questions, whether or not question's title ended in question mark, whether or not the asker edited the question's text after posting, whether or not the asker made effort prior to posting the question to find their own solution, the average frequency of a question's tags, and average number of characters for a question's tags. Include some interpretations of the hazard ratio?

While the model is significant overall, it's predictive accuracy is low. Table blah shows metrics for the model fit to the full data set. One possible explanation for this is the high number of categorical variables included in the model. Including more quantitative variables may improve predictive accuracy. 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.

7 Conclusion

8 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. $\text{Contain}_{a}nsweredandcontain_{u}nasnweredarelogicalvariables, indicatingtrueifaquestion'stitlecontaineda$

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.