Rough Draft

Lisa Oshita

1 Introduction

Community-driven online question and answer forums (CQA) are becoming widley 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 liek these 2 recieve around 40 million visits every month. These online question and answer forums are clearnly becoming popular sources of information.

The CQA analyzed in this paper is iFixit's Answers forum. iFixit is a company founded in 2003 by two engineering students of the California Polytechnic State University in San Luis Obispo. What began as a small business running out of the college dorms, is now a company that helps thousands of people fix their broken devices everyday, providing over 30,000 free repair guides with pictures and step-by-step instructions. iFixit also sells the specialized tools and parts used in their repair guides, specialized tools like an iPhone 7 batter adhesive strip or a Nexus 9 LCD screen and digitizer. The main goal of this company is to teach people how to fix their own devices, and in extending the lifetime of their own devices, they save money and cut down on electronic waste. As such, this company helps thousands of people fix their devices everyday.

Along with repair guides and tool sales, another important component of this company is their online question and answer forum called Answers. This platform features questions related to device repair, featuring over 9,000 devices and over 100,000 solutions and answers. Questions on this forum range from broken devices like a jammed zipper on a Patagonia jacket, to a shattered iPhone 6 plus screen. This forum is an integral part in teaching people how to fix their own devices. If people run into problems, they can always trust that they have a large community to turn to for help. Thus it's important, for both the individuals as well as iFixit, that askers get prompt answers. Fast response times will drive up user engagement and generate more traffic, which in turn is great for the reputation and longevity of Answers and iFixit. As such, investigation of response times would be both beneficial and informative. Analysis can reveal factors within the forum that affect how quickly questions get answers.

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 forums has not been investigated by many researchers. A majority of the research focuses on assessing and predicting the 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 survival probability of a question.

2 Related Work

In regards to investigation of forum response times, (Bhat et al., 2014) 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.

(Mamykina et al., 2011) 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.

(Asaduzzaman et al., 2013) 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.

3 Materials

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

The following statistics were calculated from Kaplan-Meier estimates of survival probabilities. In this analysis, survival probability is defined as the probability that a question remains unanswered after a certain time t. For questions in this data set, the mean survival time, or the average time until questions recieve answers is 775.75 hours, or 32.32 days.

The median survival time, the time at which 50% of the questions in the data recieved an answer is 9.16 hours. By 0.88 hours, 52.8 minutes, 25% of the questions received an answer. By 683.19 hrs, 28.47 days, 64% of the questions have been answered. Figure b shows the Kaplan-Meier curve for the questions in the data. The curve indicates that around 100 hours, the probability that a question recieves an answer after that time hovers around 0.35. The survival probability never reaches 0 due to the presence 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 recieves its first answer, as described above, was calculated as the time from the date the question was posted to the date the first answer was received. If the question did not receive an answer, the time until answer value was calculated as the time from the date the question was posted, to the date the data was downloaded, July 7, 2017 at 2:41 PM PDT. The censoring variable required for a cox regression model indicates 1 if the question received an answer, and 0 if the question remained unanswered by the time the data set was downloaded.

New_category

This variable grouped questions into categories based off of the type of device it pertains to. The data contained the original category variable as defined by iFixit. This category variable contained NAs (how many), a result of the asker creating a question for a device not already in the website's data base. These missing values were coded to "Other" in the new categorization variable. 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. Table A shows the proportion of questions within each of these device categories.

Contain_answered and contain_unanswered

These variables were created based on the intuition that certain topics might be more popular among the answering community, and that questions concerning these topics might recieve an answer faster than questions that don't. It was also based on the flip side of this idea, that certain topics might not be popular, or that some topics might be too narrow or specific and would be common among questions with long answer times or those that remained unanswered. To investigate this intuition, contain$_answeredandcontain_unansweredwerecreated.$

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. These lists contained every word within the question's titles and the frequency, or number of times they occured in their respective data set. For each of these lists, rules were applied to determine if a word could be considered "frequently use". For frequently used words in answered questions, that word would have to appear in more than 1% of the titles, and would have to appear in answered questions more than it appeared in unanswered questions. To determine the latter part, a ratio of proportions was calculated. 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 unanswered questions, to be considered "frequently" used, 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. Contain$_answeredandcontain_unasnweredarelogicalvariables, indicatingtrueifaquestion'stitlecontaineda$

Title$_questionmarkandtitle_beginwh$

These two variables were created to determine if stating the title in the form of a question is associated with faster answer times. Title$_questionmarkisalogicalvariableindicatingtrueifthequestion'sti$

Text$_contain_punct$

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

Text$_all_lower$

A logical variable indicating true if the text of the question is in all lower case.

Update

A logical variable indicating true if the asker updated the question (edited or added information about the problem) sometime after posting it.

Greeting

Logical variable indicating true if the asker included a greeting (e.g. Hello, Hi) as the first word of the question's text.

Gratitude (change the name of this variable)

Logical variable indicating true if the asker used polite language in the question's text (e.g. Thank you, please, appreciate).

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 (Bhat et al., 2014).

Weekday

The day of the week the question was posted.

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

The frequency or score of a tag is defined as the proportion of times that tag appears in the entire data set. Avg_tag_score is the average of all of a questions tag scores or frequencies. This variable was created to investigate if questions that include popular tags, or tags that are frequently used and thus have a high "score", also have faster answer times. This variable was created based off of findings from (Bhat et al., 2014).

Text_length

The number of characters in the question's text.

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's hypothesized that users who use the tagging system correctly, as in the latter example, recieve answers quicker than the user in the former.

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. long text and also include newlines are generally easier to read than questions that dont include any newlines.

The final model

The variables described above were entered into the final cox regression model with these variable transformations:

Stratification on ampm: prior analysis of initial cox regression models indicated that this variable violated the proportional hazard assumption. Stratification resolved this violation.

Square root of avg_tag_score: Viewing the distribution of answer times against the original avg_tag_score variable indicated heavy right skewing. Taking the square root of this variable decreases some of this skewing.

Quadratic polynomial on text$_length$

Restricted cubic splines were included for device$_length, avg_tag_length, andnewline_ratio.Themethodforfitti$

4 Methods

Univariate analyses for each of the variables of interest was carried out 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. After building the model on one of the training data sets, assumptions and residuals were checked.

The proportional hazards assumption was assessed by using the cox.zph function of the survival package in R. This test is based on the correlation between scaled Schoenfeld residuals and a function of time. Signficant p-values of this test indicate a possible violation of the proportional hazard assumption. For our model, the "Apple Product" level of the device category variable had the lowest p-value of 0.01. The next lowest p-values hovered around 0.04, for variables like device$_length, andtheCameraandGameConsolelevelsofcategory.AsAppleProductw$

Assessing the martingale residuals for each quantitative predictor indicated that the functional form appears to be adequate.

Assessing the deviance residuals shows that a considerable amount of questions have residuals with an absolute value of well over 2.5. However, removing those questions from the data set and and refitting the model with cross-validation techniques resulted in worse predictive accuracy.

Assessing the score residuals for each predictor indicated that some questions might a strong influence on the fit of the model. However, similar to the case with the deviance residuals, refitting the model without those certain questions with cross-validation techniques resulted in worse predictive performance.

After determining that the model did not make any extreme violations, a 5 fold cross-validation scheme was used to assess predictive performance. The data set was split into 5 training and test data sets. Each training data set contained 6,208 questions. Test data sets contained 1,552 questions. The model was built on each of the training data sets, and tested on the corresponding test data sets. To assess the prediction performance of this model, the model and its coefficients built from the training data sets, were used to calculate the predicted hazard ratios from the test data sets. Predictions were then entered into a separate cox regression model with the answer times as the survival time value. Metrics calculated calculated and assessed were the hazard ratio, along with the R2 (explain what R2 is in this context, how it's calculated), partial loglikelihood statistic and p-value. The idea for these metrics came from (Chen et al., 2012). Somers Dxy along with a concordance statistic were also calculated as performance metrics. (explain what Somers Dxy and concordance is). This process of building the model on a training data set, and then using it to predict on the test data set and calculating performance metrics from the result was repeated for each cross-validation fold. The metrics were also computed for predictions on the training data itself. For each iteration of cross-validation, there was no signifcant difference between the metrics on the training data and the metrics on the test data. The metrics taken into consideration was the average of all of the metrics for the test data. Table b shows the average performance metrics for the training data sets as well as for the test data sets. Although the metrics are not considerably high, they do not vary signficantly from training to test data sets.

After cross-validation, the model was fit to the full data set. Using the model fit to the full data set, predicted hazard ratios were extracted and the same metrics as described above were calculated. Table c shows those metrics. These metrics were not signifcantly different from the metrics obtained in cross-validation.

5 Results and Discussion

The final model gave an AIC value of 69945.32. The partial loglikelihood ratio test statistic was 1307.15 with a p-value of 0. The R-squared value for this model is 0.1555. Variables that proved to be significant predictors with a p-value of less than 0.01 are: new$_{c}ategory(appleproduct, camera, gameconsole, home, pc, andvehical), contain_{u}nanswered, title_{q}uesitor$

While the model itself overall is significant, the metrics for the models predictive accuracy are not impressive. This indicates that perhaps a different model might perform better as a model to predict the survival time for questions. Or, a model with more quantitative variables might preform better as well.

References

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

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

Chen, H.-C., Kodell, R. L., Cheng, K. F., and Chen, J. J. (2012), “Assessment of performance of survival prediction models for cancer prognosis,” *BMC Medical Research Methodology*, .

Harrell, F. E. (2015), *Regression Modeling Strategies*, Vol. 64.
URL: *papers://55069ee6-504c-4f60-bfa9-053c4dcabb39/Paper/p398%5Cnhttp://link.springer.com/10.1*
3-319-19425-7%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/21531065%5Cnhttp://linkinghub.elsevier.com

Mamykina, L., Manoim, B., Mittal, M., Hripcsak, G., and Hartmann, B. (2011), “Design Lessons from the Fastest Q&A Site in the West,” *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI ’11*, p. 2857.
URL: *http://dl.acm.org/citation.cfm?doid=1978942.1979366*