Survival Analysis of Questions Posted on the iFixit Answers Forum

Lisa Oshita[a], Anthony Pileggi, Shannon Pileggi

Department of Statistics, California Polytechnic State University

[a]Frost Research Fellow, recipient of the Frost Undergraduate Student Research Award

Overview

➤ iFixit's online question and answer forum, *Answers*, features over 120,000 user-asked questions related specifically to device repair. Analysis of question response times can reveal factors that affect how quickly questions receive answers, which can lead to suggestions for how users can ask better questions to minimize response times and for how forum design can be improved.

➤ Objective Develop a Cox proportional hazards model to predict the survival probability (probability that a question remains unanswered beyond a certain time t), of questions on the forum, with the goal of identifying variables significantly associated with response time.

Data

Data analyzed contained 7,760 questions posted from April 2017 to July 2017. 63.8% received an answer. Shortest response time was 0.5 hours; longest was 2,159 hours (90 days). Figure 1 shows the distribution of response times for questions analyzed. Original variables in the data include: device name and category, question title, text, tags, new user status, post date and answer date. Fourteen variables, capturing textual and user information, were derived from the data.

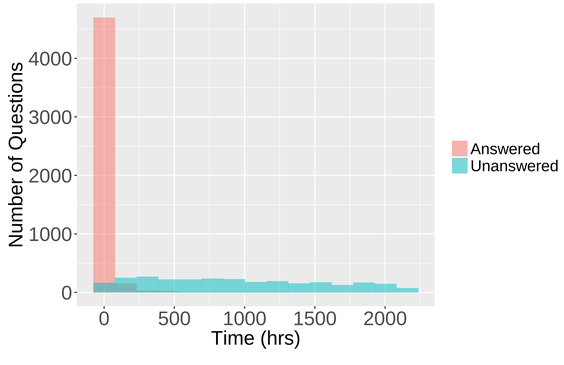

Figure 1: Distribution of response times

Methods

➤ Univariate Analysis Used to identify which variables, as well as the optimal form of continuous variables (if transformations and restricted cubic splines should be applied), to include in the model.

➤ Five Fold Cross-Validation In each iteration, the model was built on the training set and used to predict hazard ratios on the test set.

➤ Assessing Performance Predicted hazard ratios were entered into separate Cox models as the single quantitative predictor with response times as the survival time. Resulting Nagelkerke's R^2, concordance statistic, Somers' Dxy, partial likelihood ratio (PLR) and p-value, were averaged over each iteration and assessed as performance indicators.

➤ Final Model The model was fit to the full data and proportional hazards assumption was assessed.

Model Building Results

Univariate analysis determined that all categorical predictors should be retained. Table 1 shows the form of continuous predictors included in the final model.

Predictor	Knots
$\sqrt{\text{average tag frequency}}$	0
$\log(\text{average tag length} + 1)$	4
$\sqrt{\text{line breaks / text length}}$	5
$\sqrt{\text{line breaks / text length}}$	3
$\log(\text{text length})$	5

Table 1: Form of continuous predictors included in the final model (knots indicates the number of knots included in restricted cubic splines

Table 2 displays the average performance metrics for models fit to test and training sets in cross-validation, as well as metrics for the model fit to the full data. Results for training, test, and the full data did not change significantly, indicating that the model was not over fit.

Final Model Results

Final model statistics

➤ PLR = 1265.29 (p-value <0.0001)

➤ $R^2 = 0.15$

➤ Somers' $Dxy = 0.27$

Significant Predictors Device category, if the question was posted on a weekend or a weekday, whether or not the question's title contained at least one word considered frequently-used among unanswered questions, whether or not the question's title ends in a question mark.

	HR	LR	p-value	R^2	Dxy	C
Training	2.03	937.39	<0.0001	0.14	0.27	0.63
Test	1.99	220.83	<0.0001	0.14	0.26	0.63
Full	2.03	1165.03	<0.0001	0.14	0.28	0.63

Table 2: Performance metrics (HR: Hazard Ratio, LR: Partial Likelihood Ratio, C: Concordance)

Assessing the proportional hazards assumption indicated that several levels of the device categorization variable violated the assumption.

Interpreting Hazard

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. Table 3 contains interpretations of select hazard coefficients of the final model.

Suggestions

Results from this analysis lead to suggestions for changes in CQA design. Many users incorrectly specified device names and tags (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").

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.

Table 3: Hazard coefficient interpretations

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 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 by restricting 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 time, as well as create a set of consistent questions for improved analysis.

Acknowledgements

This research was supported by the Bill and Linda Frost fund of the California Polytechnic State University of San Luis Obispo. We also thank iFixit for providing access to the CQA data.