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

➤ 7,760 questions from April 2017 to July 2017.
➤ 63.8% received an answer
➤ Shortest response time: 0.5 hours
➤ Longest response time: 2,159 hours (90 days)

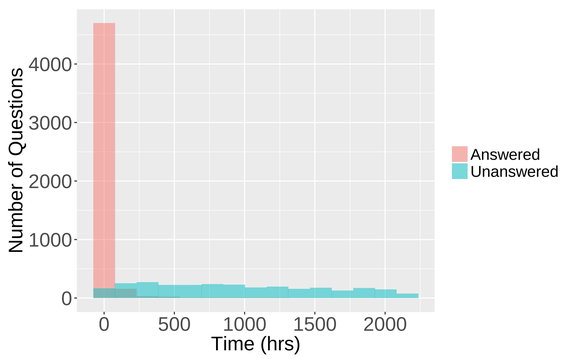

Figure 1: Distribution of response times

Original variables include: device, question title, text, tags, new user status, post date and answer date. Fourteen variables, capturing textual and user information, were derived.

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.

➤ Assess Performance

Predicted hazard ratios were entered into separate Cox models as the single predictor. Resulting R^2, concordance, Somers' *Dxy*, partial likelihood ratio (PLR) and p-value, were averaged over each iteration and assessed.

➤ Final Model

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

Cross-Validation Results

Table 1 displays average performance metrics achieved in cross-validation, as well as for the model fit to the full data.

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

Final Model Statistics

➤ PLR = 1265.29 (p-value <0.0001)
➤ $R^2 = 0.15$
➤ Somers' $Dxy = 0.27$
➤ Several levels of the device categorization variable violated the proportional hazards assumption

Interpreting Hazard

Controlling for all other predictors, the estimated hazard of receiving an answer is:

➤ 154% higher (95% CI (132%, 179%)) for Apple products vs. Android/other phones
➤ 13% lower (95% CI (7%, 18%)) for questions posted on the weekend vs. weekday

Forum Design Suggestions

➤ Rather than allowing users to enter any tag or device name, restrict options to a drop-down list
➤ Include more tips to guide users when asking questions

Final Model Coefficients

Variable		Coefficient (SE)	p-value
Device Category	Apple Product	0.93 (0.048)	<0.0001
	Camera	-0.32 (0.090)	
	Electronics	-0.01 (0.078)	
	Game Console	0.24 (0.083)	
	Home	0.34 (0.070)	
	Other	-0.13 (0.056)	
	PC	0.28 (0.060)	
	Tablet	-0.16 (0.081)	
	Vehicle	0.40 (0.069)	
	Android/Other Phone (reference)	—	
Weekend		-0.13 (0.033)	<0.0001
Text contains end punctuation		0.03 (0.050)	0.613
Text is in all lower case		-0.18 (0.064)	0.006
Title contains terms considered frequently-used among answered questions		0.05 (0.042)	0.260
Title contains terms considered frequently-used among unanswered questions		-0.28 (0.034)	<0.0001
Title ends in a question mark		0.26 (0.033)	<0.0001
User edited or added to the question's text after posting it		0.30 (0.086)	0.001
User was a member for less than one day before posting		-0.11 (0.036)	0.003
User made an effort to solve the problem prior to asking the question		-0.07 (0.036)	0.045
Square root of the average tag frequency		2.23 (0.720)	0.002

Table 2: Coefficients for predictors in the final model. (Continuous predictors fit with restricted cubic splines are not shown)

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.