ClientComm Data Analysis

Gwen Rino and Eric Giannella

6/22/2018

Overview

We analyzed data from the Baltimore ClientComm project collected between (dates). One source dataset was composed of text messages between users (POs) and clients (people under supervision), and one source dataset recorded the outcome of supervision (success or failure) as reported in surveys of the POs. Because outcomes are at the user/client relationship level, variables were constructed to reflect all interactions within a relationship. For example, the variable max_user_msg_length records the length of the longest text message written by a user to a particular client.

We examined 11 variables for possible relationships to the outcome of supervision, and found 6 of them to show statistically significant correlations with supervision outcomes. While these correlations are statistically significant, they should be considered with caution. There are only 89 instances of supervision failure out of the 1200 observations, and the associations are not strong nor do they necessarily imply causality. However, they do suggest possible avenues for investigation in the development of new features for the ClientComm product.

Key Results

- Clients who use ClientComm to write even one message have a reduced chance of supervision failure.
- User messages that include future dates reduce the chance of supervision failure.
- Active users of the message scheduling feature have fewer supervision failures.

Analysis

We began by building a logistic regression model to investigate the impact of the number of messages sent by the user and by the client. We learned that clients who engage with the ClientComm system by sending any messages at all have better outcomes. The number of messages sent by the user is also signficantly associated with positive outcomes for the client, but the effect is confounded when the two variables are considered together, so we included only the number of client messages in the model.

Next we considered the impact of the length of messages written by the user, and found that while a longer median message length is only slightly associated with better outcomes, adding this variable improves the efficiency of the model.

Using the Python package TextBlob to quantify the polarity and subjectivity of the users' messages, we found that both of these qualities had an impact on client outcomes. Higher maximum polarity is positively associated with supervision success, and higher median subjectivity is negatively associated with supervision success. Both of these associations are statistically significant.

Next we examined the number of times that a user mentioned a future date in a message, and found that more mentions of future dates is significantly associated with better client outcomes.

We looked for an association between the amount of time from the receipt of a client's message to the user's response. This variable had substantial missing values and did not improve the model.

Finally, we examined the time between a user message's creation and its being sent (a proxy for a user's utilization of the message scheduling feature). We found that the feature is not used at all in about 70% of the observations. However, the ~9% of observations in which the median time interval is greater than 10 hours (heavy users of the message scheduling feature) are signficantly associated with better client outcomes.

The final model regresses the outcome variable supervision_failure against 6 variables: client_msg_count > 0 (clients who use ClientComm at all), median_user_msg_length, max_polarity, median_subjectivity, n_report_time (the number of times future dates are included in messages from the user), and median_schedule_time_hr > 10 (heavy users of the message schedule feature). There are no problematic correlations between any pair of these variables.

The table below compares the models with various combinations of the variables.

Table 1: ClientComm Model Results

	Dependent variable:					
	supervision_failure					
	(1)	(2)	(3)	(4)	(5)	(6)
user_msg_count	−0.039					
	(0.025)					
client_msg_count >0	−0.463*	−0.711***	−0.521**	−0.443*		−0.472*
	(0.238)	(0.239)	(0.248)	(0.250)		(0.251)
median_user_msg_length		−0.002	−0.003**	−0.003*	−0.004	−0.002*
		(0.001)	(0.001)	(0.001)	(0.003)	(0.001)
max_polarity			−1.436***	−1.100**	−1.503*	−1.153**
			(0.433)	(0.456)	(0.846)	(0.451)
median_subjectivity			1.370**	0.997*	1.652	0.878
			(0.547)	(0.573)	(1.020)	(0.568)
n_report_time				−0.134**	−0.079	−0.143**
				(0.067)	(0.082)	(0.068)
uc_max_response_time_hr					0.0004	
					(0.0004)	
median_schedule_time_hr >10						−1.076**
						(0.531)
Constant	−2.084***	−1.960***	−1.563***	−1.514***	−1.922***	−1.375***
	(0.178)	(0.294)	(0.325)	(0.322)	(0.651)	(0.329)
Observations	1,200	1,195	1,195	1,195	539	1,195
Log Likelihood	−312.445	−307.031	−301.424	−298.909	−106.034	−296.196
Akaike Inf. Crit.	630.889	620.061	612.847	609.818	224.068	606.392

Note: *p<0.1; **p<0.05; ***p<0.01

How Good Is the Model?

To get a sense of how accurate the final model is, we randomly divided the dataset into a training set (75% of the observations) and a test set (the remaining 25% of the observations). We fit the model to the training set and used the output to predict the probabilities of supervision success for each observation in the test set.

Because supervision failure is rare, there are only 21 occurrences of supervision failure among the 300 observations in the test set. With few instances of failure to detect, and with relatively weak associations between the variables and the outcomes, we must use a very low probability cutoff value (0.1) in order to correctly identify just 57% of the supervision failures and 82% of the supervision successes. This result is reflective of both the power and the limits of the model.

Next Steps

- Research TextBlob's polarity and subjectivity sentiment calculations in order to more clearly interpret the effect of these qualities on the model and their implications for product development.

- Conduct day of week/time of day analysis of user messages in order to refine our understanding of the key results concerning inclusion of future dates and the use of the message scheduling feature.