

Need for methodologies to optimize time-alignment of predictor variables with target variables in building models for Quality of Experience as a function of network behavior

Yana Kane-Esrig

Comcast

yana_kane-esrig@cable.comcast.com

Abstract

At Comcast, we strive to continuously improve the quality of experience (QoE) for our customers.

As part of this effort, we conduct data collection and analysis projects to build quantitative models that predict various behaviors that express a customer's problems with QoE. The candidate predictors are metrics that measure the health of the various portions of the network involved in serving the customer and/or the quality of service delivered by the network to the customer.

A specific type of studies presents a challenge that would be a good topic of discussion and that may serve as a starting point for research and/or building open source tools. In these studies each value of the target variable has a time stamp. For example, a customer does or does not call the service center a given time to report an issue with their service.

In these studies it is challenging to align in time the corresponding values of the target variable and the various predictors. There may be lags between the time when an adverse event occurs (for example, the customer experiences degraded service) and the time when the customer calls the service center. These lags are not fixed, but may vary from customer to customer, or even within a single customer's behavior.

Moreover, there may be the effect of "being sensitized by accumulated grievances": the customer may be unlikely to call the first time they experience a service degradation, but may be likely to call the second or the third time they experience a degradation. The impact of this sensitization effect may be dependent on the lengths of the gaps of time between episodes of service degradation. Another time-dependency may be the duration of the degradation and whether it happens during peak usage or at an off-time.

It would be useful to develop a methodology and tools that would optimize the process of fitting the models that explicitly incorporate and estimate the time lag and “sensitization by accumulation of grievances” aspects of the relationship between the customer’s behavioral expression of QoE and the various predictors. It would also be useful to have visualization tools for such models.

Paper

At Comcast, we strive to continuously improve the quality of experience (QoE) for our customers.

As part of this effort, we conduct data collection and analysis projects. In these projects we aim to build quantitative models that predict various behaviors that express a customer’s problems with QoE. The candidate predictors are metrics that measure the health of the various portions of the network involved in serving that customer and/or the quality of service delivered by the network to that customer.

The target variable in such a study may be self-reported by the customer (for example, an opinion about service quality expressed in a survey), or it may be the observed behavior of the customer (for example, the customer placing or not placing a call to the service center).

Various characteristics of the customer, such as the service tier and the devices that the customer uses, are included as covariates that may have interactions with the predictors of interest.

These predictive models are typically being built to serve one of two goals:

1. Identifying customers whose QoE is unsatisfactory or is deteriorating, so that Comcast can proactively address the problem
2. Understanding how predictive various network health metrics and/or Quality of Service metrics are of QoE problems to help us prioritize which aspects of the network to examine further for potential causal link with QoE problems.

A specific type of studies presents a challenge that would be a good topic of discussion and that may serve as a starting point for research and/or building open source tools. In these studies each value of the target variable has a time stamp. Typically, we take snapshots of the behavior of individual customers in some population over a set of windows of time and code their behavior as a binary target variable (for example, the customer placing or not placing a call to the service center) or a multinomial variable (for example, the satisfaction score on a 5 point scale the customer gives the reliability of our service).

In these studies it is challenging to align in time the corresponding values of the target variable and the various predictors. There may be lags between the time when an adverse event occurs (for example, the customer experiences degraded service) and the time when the customer calls the service center. These lags are not fixed, but may vary from customer to customer or even within a single customers’ behavior.

Moreover, there may be the effect of “being sensitized by accumulated grievances”: the customer may be unlikely to call the first time they experience a service degradation of a certain severity or duration, but may be much more likely to call the second or the third time they experience a degradation, even if the subsequent degradations are less severe or prolonged than the first one. The impact of this

sensitization effect is likely to be dependent on the lengths of the gaps of time between the episodes of service degradation.

Another time-dependency may be the duration of the degradation and whether it happens during peak usage or at an off-time.

It is possible to deal with these issues by the usual model-refining approach of fitting a large number of models and comparing their performance on the training data set, then selecting a small number to compete in the model validation stages. However, these time lag, duration, and grievance accumulation considerations (combined with the usual potential for interactions among predictors) can make the process of data collection and model selection particularly laborious and time consuming.

It would be useful to develop a methodology and tools (for example, R packages) that would optimize the process of fitting the models that explicitly incorporate and estimate the time lag and “sensitization by accumulation of grievances” aspects of the relationship between the customer’s behavioral expression of QoE and the various predictors. It would also be useful to have visualization tools for such models.

It would be especially helpful to add such capabilities as extensions to the widely used existing methodologies for which there are diagnostic tools and implementations in open source software, such as ordinary and logistic regression, and partition trees.