Time-Dependent Cox Model for Revenue Attribution

Fan Yang*[1]

[1]School of Statistics, University of Minnesota, Minneapolis, MN 55455.

Abstract
In marketing analytics, revenue attribution is an important aspect. Most retailers have multiple channels for both ordering and marketing, for example, email order and catalog order for order channel; promotion email and paid search for marketing channel and so on. Retailers want to correctly identify the sources of orders that come in back to the various marketing channels that drove them, in order to realize a more efficient way of marketing. This paper proposes a time-dependent Cox model built to analyze the relationship between the number of contacts made through a diversity of marketing channels and the number of purchases via many order channels. The method is applied to data from a retailer company. Weights of each marketing channel which measure how important a marketing channel is, are given finally for different order channels using the model. Numerical results show that the proposed method is very useful for revenue attribution and provides a good way to estimate weights for marketing channels.

Key Words: Time-dependent Cox model ; revenue attribution

1. Background

Stores and retailers usually have multiple marketing channels to attract customers to purchase. Especially in this information era, besides traditional ways of promotion like promotion mail, many online marketing channels emerged, including email, website affiliates, mobile apps, paid search on search engines and so on. Retailers invest a large amount of money and time in doing a variety of promotion activities. They want to better understand which marketing channels are bringing them the largest number of orders, the second largest and so on. With a good estimate of the relevance of each marketing channel to the final order, they can achieve a more efficient way of allocating their money and time by investing more in marketing channels that are highly effective and less for those that are not as effective.

Besides the effect of marketing promotions, retailers also believe there is a portion of the orders coming from "organic demand". This is the number of orders they can get even without doing any marketing activities. Retailers are also very interested in estimating the "organic demand".

Not only marketing has multiple channels, retailers usually also have multiple order channels. For example, phone order, mail order, web order and so on. They want to understand how much effect their marketing activities have on the number of orders they get from each order channel. Hereto we have fully explained the so-called "Revenue Attribution". In one word, through revenue attribution we try to correctly identify the sources of orders that come in through order channels back to the marketing channels that drove them. Also, identifying the importance of "organic demand" is of interest. What retailers need are weights of contribution to orders from each order channel for all marketing channels and "organic demand".

*yang2732@umn.edu

2. Review of Existing methods

Currently there are some methods in the industry, but all are fairly intuition-based and not able to model all factors that may be relevant to customers' purchase behavior. These methods are summarized below.

2.1 Last-Touch Attribution

This is the current de-facto industry standard according to [1]. It gives full credit to the last marketing channel that touches the customer. The contribution of all other events are ignored. Though very simple to use, this clearly cannot reflect the relevance of all marketing activities.

2.2 Equal Attribution

This method gives equal credit to every marketing channel preceding an order. Though also easy to carry out, it doesn't take into account the potential different effect of every marketing channel.It is not a good approach either.

2.3 Arbitrary Parameters

People in the industry want to weight first marketing touch , last marketing touch, and intermediate touches disproportionately to adjust for the different effects of different events. But currently there is no good way to assign the weights and they are just assigned arbitrarily. So this approach is not a clear answer.

3. Brief Review of Time-dependent Cox Model

Survival analysis studies the time to event of interest. Among all models in survival analysis, the Cox model, also known as the proportional hazard model, is a semiparametric model and assumes that for the ith individual or observation

$$h(t) = h_0(t)\text{exp}(\beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}) \tag{1}$$

where t is time, $h(t)$ is the hazard rate function, $h_0(t)$ is the baseline hazard rate function, $(X_{i1}, X_{i2} \cdots X_{ip})^T$ are the covariates for individual i and $\beta_1, \beta_2, \cdots, \beta_p$ are the corresponding coefficients. A basic interpretation for this model is that, fixing all other covariates, if we increase one covariate say X_k by 1, then the hazard rate will be multiplied by $\exp(\beta_k)$. This is the reason why Cox model is also called proportional hazard model.

A variant of the basic Cox model is the time-dependent Cox model which assumes

$$h(t) = h_0(t)\text{exp}(\beta_1 X_{i1}(t) + \beta_2 X_{i2}(t) + \cdots + \beta_p X_{ip}(t)) \tag{2}$$

As can be seen from the above formula, the difference between the basic Cox model and time-dependent Cox model is that the covariates of the latter are allowed to change with time.

4. Apply Cox Model to Revenue Attribution: Idea and Details

Consider a fixed time period where the data is given. After the marketing events in the fixed period, customers either purchase or not purchase. This can be taken as an analogy of "alive" or "dead" in survival analysis. Customers who purchased are "dead", with the event happened. Customers who didn't purchase are censored and still "alive".

Retailers typically have data about all marketing events and orders information for each user during a fixed period, including the dates of the events.Thus we can do analysis from a time-to-event perspective. Time-dependent Cox model can be used to incorporate the date information of all events. Basically the count of each type of marketing event at a given date will be used as the covariate. Each time a new marketing event occurs, a new time interval is set up in the time-dependent Cox model, with the count of teh relevant marketing event increasing by 1.

Marketing events after the last purchase are simply deleted because they don't contribute to any purchase. If a customer purchased more than once in the study time period, events between two purchases are regarded as new customers. This is done to avoid using recurrent events Cox model, which will make this model much more complicated.

To account for the fact that a marketing event should have less and less effect on the customer's behaviour as time goes on, the counts of a marketing event is decayed according to length of time it has passed. Specifically, the following formula is used

$$count' = count * \alpha^{\Delta d/7} \tag{3}$$

where Δd is the day difference between $count'$ and $count$, and α is a decaying factor from 0 to 1. The count of an event during oen time interval is fixed. But for the next time interval, it will decay according to (3) using the stop time of each interval. By marketing people's experience, for promotion mail, α is set 0.8, and for all other marketing channels, α is set as 0.5. In other words, for example after a week, the effective count of a promotion mail will decrease to 0.5 from 1.

History orders are treated as fixed covariates which doesn't vary by time, as there is no exact date information. And this term is added to estimate "organic demand".

To assign a weight to each marketing channel, only main effects are included in the model without interactions or higher order terms.

5. Application to dataset

The data is from a store that sells men's clothes. Marketing channels include promotion mail, referring sites, direct load, email open, email click, paid-search, FPM, linkshare, natural search and other-web-channel. And order channels include Retail ,Catalog, Outlet,Store Web and Web.

The data includes all customers who had purchased from any order channel during the study period and a 20% sample of customers who didn't purchase in the study period but has history orders. So weights are used to adjust this 20% sampling of nonresponders in the model. There are 674,069 customers in the study in total, with 8,917,143 records with each record for one event.

The retailer recorded each customer's marketing touch point and order information including date and type of event in 07/01/2012∼ 08/31/2012 for study purpose. Customers' order history in 01/01/2012∼ 06/30/2012 is also available.

A part of the data is shown in table 1. The first column is customer id, and the 2nd to 6th columns are history orders. indicator_date is the date of the event and indicator_desc is the description of the event.

	customer_id	web_orders	catalog_orders	retail_orders	outlet_orders	store_web_orders	responder	transaction_date	indicator_date	indicator_desc
30	957	2	0	2	0	0	1	2012-08-26	2012-08-14	Email Open
31	957	2	0	2	0	0	1	2012-08-26	2012-08-14	Email Open
32	957	2	0	2	0	0	1	2012-08-26	2012-08-23	Email Open
33	957	2	0	2	0	0	1	2012-08-26	2012-08-23	Email Open
34	957	2	0	2	0	0	1	2012-08-26	2012-08-23	Email Open
35	957	2	0	2	0	0	1	2012-08-26	2012-08-23	Email Open
36	957	2	0	2	0	0	1	2012-08-26	2012-08-23	Email Open
37	957	2	0	2	0	0	1	2012-08-26	2012-08-23	Email Open
38	957	2	0	2	0	0	1	2012-08-29	2012-08-29	Channel Retail
39	957	2	0	2	0	0	1	2012-08-29	2012-08-23	Email Click
40	957	2	0	2	0	0	1	2012-08-29	2012-08-23	Email Click
41	957	2	0	2	0	0	1	2012-08-29	2012-08-23	Email Click
42	957	2	0	2	0	0	1	2012-08-29	2012-08-23	Email Click
43	957	2	0	2	0	0	1	2012-08-29	2012-08-23	Email Click
44	957	2	0	2	0	0	1	2012-08-29	2012-08-23	Email Click
45	957	2	0	2	0	0	1	2012-08-29	2012-08-25	Email Open
46	957	2	0	2	0	0	1	2012-08-29	2012-08-18	Email Open
47	957	2	0	2	0	0	1	2012-08-29	2012-08-17	Email Open
48	957	2	0	2	0	0	1	2012-08-29	2012-08-15	Email Open
49	957	2	0	2	0	0	1	2012-08-29	2012-08-23	Email Open
50	957	2	0	2	0	0	1	2012-08-29	2012-08-23	Email Open
51	957	2	0	2	0	0	1	2012-08-29	2012-08-23	Email Open
52	957	2	0	2	0	0	1	2012-08-29	2012-08-23	Email Open
53	957	2	0	2	0	0	1	2012-08-29	2012-08-23	Email Open
54	957	2	0	2	0	0	1	2012-08-29	2012-08-23	Email Open
55	396250	1	0	1	0	0	0	2012-08-01	2012-07-11	Promotion Mail
56	396250	1	0	1	0	0	0	2012-08-01	2012-07-11	Promotion Mail
57	714286	0	0	1	0	0	0	2012-08-01	2012-07-27	Promotion Mail
58	1485523	0	0	2	0	0	0	2012-08-01	2012-07-27	Promotion Mail

Table 1: Original data (part)

SAS is used for implementation. Example code is included in the Appendix. Variable selection was done to delete insignificant variables and reduce the model to a significant submodel. After doing variable selection, we obtain the final coefficient estimate of each marketing channel. To further derive the weights, mean count of each marketing channel is also calculated. Then the final weights are set to be proportional to the product of coefficient for each marketing channel and its mean count. The same is done for history orders, which can give us the estimate of organic demand.

The final results for all five order channels are shown in the pie chart in Figure 1. For example, for retail channel the three most important factors are history retail order, promotion mail and email, with 69.9%, 23% and 5.7% relevance to a purchase respectively. This is very reasonable from the point of view of the marketing people. So the model employed is doing fine.

6. Summary

Time-dependent Cox model fits into the problem of revenue attribution nicely. It models the response variable of responding to marketing events or not, and utilize the time(date) information of marketing events. This model gives reasonable results for how important each marketing channel is in attracting customers to purchase. The effect of organic demand is also studied in the model as a fixed variable. Time-dependent Cox model is a very useful and powerful model to apply for revenue attribution.

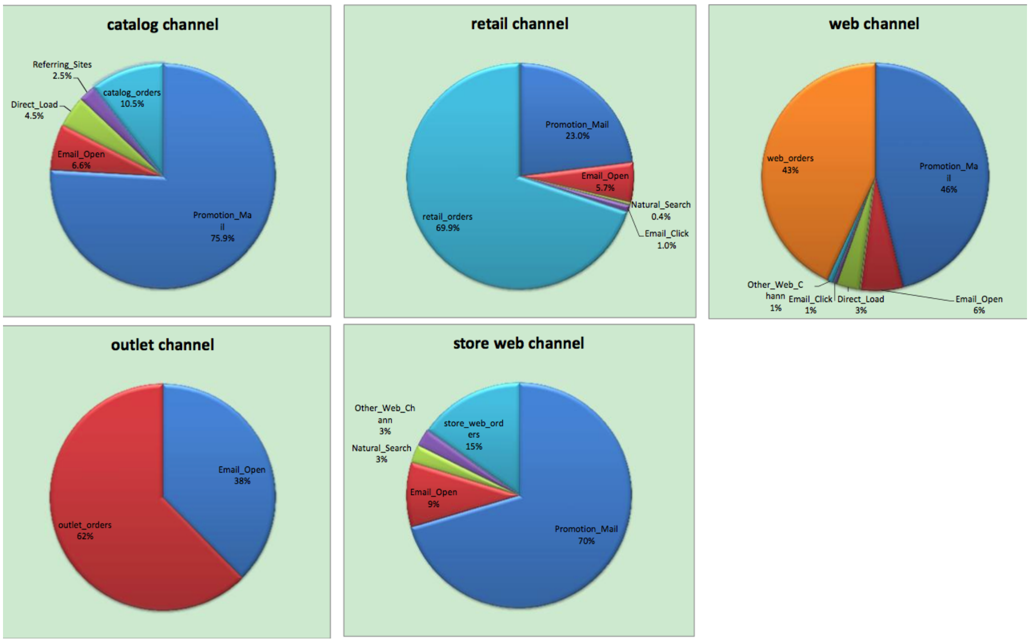

Figure 1: Pie chart for weights

Appendix: SAS Code

An example of code for the catalog order channel model is

```
proc phreg data=mylib.catalog4_08_05 ;
  model (start,stop) * status(0) = Promotion_mail
  Email_open Direct_load Email_click Natural_search
  Referring_sites other_web_chann PaidSearch_FPM
  web_orders retail_orders catalog_orders
  outlet_orders store_web_orders;
  weight weights;
run;
```

where `weights=1` for responders(people who buy) and `weights=5` for nonresponders.

References

[1] Datasong. Revenue attribution methodology: A marketer's guide. 2013.

[2] Terry M. Therneau. Multiple and correlated events slides. 2009.

[3] Masaaki Tsujitani, Yusuke Tanaka, and Masato Sakon. Survival data analysis with time-dependent covariates using generalized additive models. *Computational and Mathematical Methods in Medicine*, 2012.

[4] Upstream. Retail marketing analytics. 2012.