EMAC 2015 LEUVEN (BELGIUM) MAY 26-29 2015

Purchase Conversions and Attribution Modeling in Online Advertising: An Empirical Investigation

Author: TAHIR NISAR - Email: t.m.nisar@soton.ac.uk

University: SOUTHAMPTON UNIVERSITY BUSINESS SCHOOL

Track: Modelling and Marketing Analytics

Co-author(s): Man Yeung (University of Southampton)

Access to this paper is restricted to registered delegates of the EMAC 2015 Conference.

EMAC EUROPEAN MARKETING ACADEMY

Purchase Conversions and Attribution Modeling in Online Advertising: An Empirical Investigation

Abstract

In a purchase funnel, a consumer may interact with an assortment of ad platforms ranging from display ads, paid search and organic search to social media and email. In this study, we consider attribution models that can be applied to assign sales credit to these and other online channels. Using an online firm's conversion data, we investigate the commonly used the last-click attribution model and compare its results to a cooperative game theory based (Shapley Value) attribution model. Our findings show that individual rewards vary significantly for different online channels under these two models. We also compute contributions of the various estimated factors using the Shapley Value regression approach in order to decompose a consumer funnel by regressed sources. Our empirical research provides insights into the complexity of attribution modeling.

Keywords: Purchase funnel; Attribution modeling; Last-click; Shapley Value

Conference Track: Modeling and Marketing Analytics

1. Introduction

Digital advertising campaigns are often launched across multiple channels, a selection of which may include search, display ads, social media, mobile, video, and email. These channels assist consumers to make purchase decisions, or sign up to a service being advertised, as they are exposed to advertisement impressions. To gauge the effectiveness of such advertising campaigns, it will be necessary to know which media channels or advertising formats have contributed to a purchase conversion. This is a process known as attribution. A better understanding of this process or assigning conversion credit to the various relevant channels can serve a number of research and industry purposes. For example, marketing managers may use such attribution models to interpret the influence of advertisements on consumer behavior and optimize their advertising campaigns.

In this paper we first examine the last-click attribution model and then consider a cooperative game theory (Shapley Value) based attribution approach as a statistical model for online businesses (Osborne and Rubinstein, 1994). The Shapley Value model assesses the contributions of a set of factors whose sum accounts for the purchase conversion. In our context, the approach yields an exact additive decomposition of any touch points into its contributory factors. Using an online firm's purchase conversion data, the study sheds light on how these attribution models can be used to better measure advertising performance. As the effect of changing attribution models for different online channels has been largely unstudied, an analysis of these models will allow conclusions to be made on whether an advertising format's revenues significantly differ between the models. To facilitate our analysis, we compare the performance of display advertising with other online sales channels. We first provide a brief literature survey to identify the challenges of attribution modeling in online advertising markets. Our empirical results about the outcomes of different attribution models are presented in the next section. The following section describes our findings on Shapley Value regression model. The study then progresses to consider implications for different online sale channels and attribution. These are summarized in the last section.

2. Attribution in online advertising: A literature survey

There is a small but rapidly growing body of literature that examines the entire clickstream history of individual consumers in terms of whether visits to different ad formats have positive effects that accumulate toward a purchase (e.g., learning about a product that the shopper intends to buy. See Wiesel, Pauwels and Arts, 2011). This strategy of modeling the purchases as a result of the accumulative effects of all previous interactions largely focuses on how non-purchase activities (e.g., advertisement clicks, website visits) affect the probability of purchasing. Their concern with the non-purchase activities means that they cannot directly deal with the question of attributing credit for conversion to each individual ad format. Relatedly, Xu, Duan and Whinston (2014) study the specific "exciting effects" between advertisement clicks (i.e. how the occurrence of an earlier advertisement click affects the probability of occurrence of subsequent advertisement clicks). Li and Kannan (2014) use a probit-based consideration and nested logit formulation for visit and purchase to attribute conversions. These and other predictive models have (Li et al., 2010) generally focused on the classification accuracy and, more importantly, they do not pay enough attention to the stability issue of the variable contribution estimate.

2.1. Shapley Value-based attribution model

In digital advertising, multi-channel attribution is one of the most important problems, especially as a wide variety of media are involved. In recent years, researchers have made efforts to develop a true data-driven methodology to account for the influence of each user interaction to the final user decision. Shao and Li (2011) have developed a probabilistic model based on a combination of first and second-order conditional probabilities. There are two steps involved in generating the probabilistic model:

Step 1. First compute the empirical probability of the main factors,

$$P(y|x_i)= \frac{N_{positive(x_i)}}{N_{positive(x_i)}+ N_{negative(x_i)}} \quad (1)$$

and the pair-wise conditional probabilities

$$P(y|x_i, x_j)= \frac{N_{positive(x_i,\ x_j)}}{N_{positive(x_i, x_j)}+ N_{negative(x_i, x_j)}}, \quad (2)$$

for $i \neq j$. A conversion event (purchase or sign-up) is denoted as y which is a binary outcome variable, and $x_i, i = 1,...,p$, denote p different advertising channels. $N_{positive}(x_i)$ and $N_{negative}(x_i)$ denote the number of positive or negative users exposed to channel i, respectively, and $N_{positive}(x_i, x_j)$ and $N_{negative}(x_i, x_j)$ denote the number of positive or negative users ex- posed to both channels i and j.

Step 2. The contribution of channel i is then computed at each positive user level as:

$$C(x_i) = p(y|x_i) + \frac{1}{2N_{j\neq i}} \sum_{j\neq i} \{P(y|x_i, x_j) - P(y|x_i) - P(y|x_j)\}, \quad (3)$$

where $N_{j\neq i}$ denotes the total number of j's not equal to i. In this case it equals to N-1, or the total number of channels minus one (the channel i itself) for a particular user. An advantage of using this estimation is that it includes the second-order interaction terms in the probability model. As there is significant overlap between the influences of different touch points due to the user's exposure to multiple media channels, the model fully estimates the empirical probability with the second-order interactions. Another important assumption is that the net effect of the second-order interaction goes evenly to each of the two factors involved. Dalessandro, et al. (2012) show that, after rescaling, this probability model is equivalent to their Shapely Value formulation under certain simplifying assumptions.

3. Data description

We utilize logs from a large-scale online sales platform to first identify where different online channels feature in the customer journey. In total, 996,708 transactions are included in the analysis, with total revenue of $158,519,417, at an average order value of $159.04. Our conversion data span 104 weeks from January 1, 2012 – February 28, 2014. Currently, the firm we investigated attributes revenue generated through online transactions to its various paid marketing tools on a last-click basis. In our data, we have information about the following digital channels: display ad, organic search, paid search, price comparison sites, email, retargeting, and social media.

4. Attribution models: An empirical investigation

Our specific hypotheses relate to examining the financial importance of display advertising channel under the current last-click model; and the effects of moving to Shapley Value-based attribution model. We test the hypothesis that, as being a convertor, display advertising generates more revenue under the last-click model than Shapley Value-based attribution model. In addition, we compute contributions of the various estimated factors using the Shapley Value regression approach so as to decompose a consumer funnel by regressed sources. The approach has the merit of computing the weighted marginal contributions of an estimated conversion source in various coalitions of conversion sources. These weighted contributions exactly sum up to the considered channel impact measure.

4.1. The last-click model

Current industry practice indicates that the majority of online sales are attributed on a "last ad" or "last-click" model. The model attributes all conversions to the last referring impression within a customer journey, which means it is the final interaction that matters from a marketing perspective (Li and Kannan, 2014). The contribution of display ads and the other online marketing tools to online revenue are presented in Table 1. It can be seen that using the current last-click method, display ads generate 18.42% of total online revenue. The highest revenue generating online marketing tool is that of organic search, bringing 63.45%. Social media contributes the least with the current model, at 0.02%. The mean order value for display ads offer insight into this as it is higher than any other of the marketing tools at $159.04. We conduct two-sample t-test comparing average order value of display ads to the rest of online marketing tools. It examines if there is any significant difference between the means of the average order values for display ads against the rest of the online marketing tools. The T statistic of 21.22 is greater than the two-tail critical value of 1.96 and therefore indicating (with a 95% confidence level) there is significant difference between the average order values. Furthermore, the p-value of 3.13E-98 is considerably lower than 0.05.

Table 1: Different online marketing tools and revenue generated under last-click.

Tool	Revenue (%)	Orders (%)	Average Order Value (in dollars)
Display ads	18.42	13.41	159.04
Organic Search	63.45	68.71	106.11
Paid Search	10.92	10.83	115.80

Price Comparison	2.15	1.81	136.16
Email	0.86	0.89	111.84
Retargeting	1.22	1.28	109.93
Social Media	0.02	0.06	48.11
Other	2.95	3.01	112.65

4.2. The Shapley Value-based attribution model

The Shapley Value methodology was developed in a cooperative game setting, and has been applied from measuring systemic risk in a macroeconomic environment to inequality indices (Osborne and Rubinstein, 1994). In a typical Shapley Value cooperative game, a group of players generates a shared "value" (e.g. wealth, cost) for a group as a whole. The Shapley Value of a player in a game is calculated as his expected marginal contribution over the set of all permutations on the set of players. The Shapley Value of an advertising medium is its expected marginal contribution over all possible sets of the interacting channels. We have noted these assumptions in the formulation in Section 2.1, and use it to calculate the percentage of value allocated to each given channel.

Figure 1 shows the effects on revenue attribution for the online marketing tools using the Shapley Value-based model. Our results show that display ads represent 14.34% of the revenue generated, down on the 18.42% revenue accumulated under the last click model, whereas organic search registers only a small increase from 63.45% to 64.17%. Social media and email record the largest changes in value percentage, as reflected in their revenue generation contributions of 2.14% and 2.58%, respectively. There is also a sizeable increase in paid search, increasing from 10.92% under the last-click model to 12.85% under the probability model. We conduct two-sample t-test comparing last click and probability based display ad rewards. The T statistic of 28.43 is greater than the two-tail critical value of 1.96 and therefore indicating (with a 95% confidence level) there is significant difference between the average display advertising return. It could therefore be concluded that the Shapley Value based attribution model on average attributes lower revenue to display ads. This is also supported by Table 2 that shows that display ads are allocated 24.86% lower revenue under the Shapley Value-based attribution model.

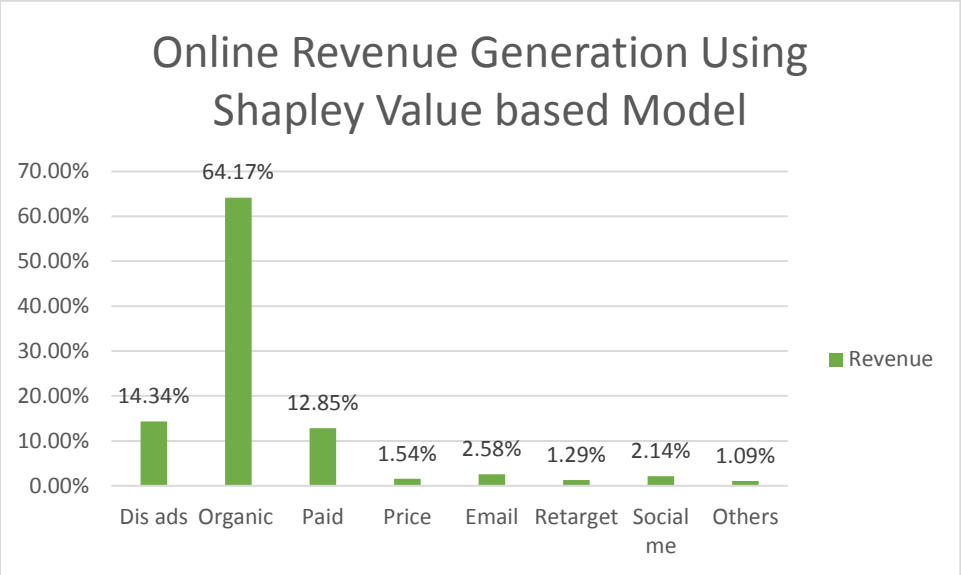

Figure 1. Shapley Value-based attribution modeling

Table 2: Display ads revenue from Shapley Value-based attribution compared to last-click.

Tool	Display Ads Revenue (%)	Last-click Difference (%)	Increase/decrease from Last-click (%)
Last-click	29.73	n/a	n/a
Shapley Value	22.57	7.16	-24.86

5. The regression results

In the preceding section, we have examined the challenge of attributing credit in a multi-channel online sales environment. In this section, we determine the exact contributions and statistical significance of each explanatory variable to the variance of the dependent variable of a regression. The Shapley Value regression method provides a systematic way of quantifying the different contributions of the explanatory variables to the goodness of fit of a regression. As the Shapley Value uses the marginal contributions of a variable from all sequences, two highly correlated variables are expected to have similar Shapley contributions because they will be low or high depending on whether the variables that this variable is correlated with, are already included or not. One can then be confident that the approach takes the potential correlation amongst regressors into account, where the contribution of each attribute is measured by the improvement in R-square. In Table 3, the Shapley Value approach is developed to derive the exact contributions of the various explanatory variables of a linear regression to its R-Square. It shows the two decompositions (Shapley and Nested-Shapley) along with a 95% level confidence interval for each component. As expected, organic search makes the largest contribution to the explanation of purchase conversion: it accounts for about 55% of the R-Square. Second in importance is display advertising, while paid search is at number three in importance. Both retargeting and price comparison sites explain equally well the variation in purchase conversions, although their contributions do not significantly differ from each other. More significantly, the regression captures the important role that email and social media now play in a purchase funnel. They explain 2.43% and 2.19% variations in the regression model, respectively.

Table 3 Contributions of the purchase funnel medial channels to the R-Square Total (RSquare)

	Shapley		Nested-Shapley	
Display Ads	0.35784 (0.122, 0.149)	17.34%	0.12593 (0.121, 0.147)	16.89%
Organic Search	0.23746 (0.025, 0.033)	55.27%	0.24268 (0.020, 0.034)	51.66%
Paid Search	0.034765 (0.027, 0.13)	13.44%	0.02315 (0.024, 0.11)	11.34%
Price Comparison	0.0276 (0.15, 0.16)	3.56%	0.01573 (0.12, 0.11)	6.68%
Email	0.036218 (0.117, 0.113)	2.43%	0.02714 (0.112, 0.109)	2.54%
Retargeting	0.01783 (0.24, 0.226)	2.85%	0.01549 (0.23, 0.221)	2.73%
Social Media	0.00016	2.19%	0.00016	4.89%

	(0.121, 0.137)		(0.121, 0.128)	
Other	0.02786	2.92%	0.01843	3.27
	(0.015, 0.013)		(0.016, 0.014)	
Total (*R*-Square)	0.1947	100%	0.1947	100%

95% confidence intervals are reported in parentheses.

6. Conclusion

We examined the hypothesis that, as being a convertor, display advertising generates more revenue under the last-click model than Shapley Value-based attribution model. The results showed that the last-click model generated the most revenue for display ads. The revenue attributed to display advertising under the last-click model was 29.73% of the total revenue - higher than the Shapley Value's 22.57%. Shapley Value simplifies the analysis in such a way that advertisers can assign values to individual advertising channels in accordance with their contributions to the generation of a shared value. When multiple channels, such as search, display, mobile, social and email are involved in a purchase conversion, Shapley Value method allows all these channels to get their fair reward for making the sales transaction possible. As the model is stable and relatively easy to interpret, advertisers can develop a clear strategy to optimize their resource allocations among multiple advertising channels. Our Shapley Value-based regression results further demonstrate the efficacy of adopting this approach.

References

Dalessandro, Brian, Ori Stitelman, Claudia Perlich, & Foster Provost (2012). Causally motivated attribution for online advertising, ADDKDD'12, 1-9.

Li, Hongshuang (Alice) & P.K. Kannan (2014). Attributing Conversions in a Multichannel Online Marketing Environment: An Empirical Model and a Field Experiment, *Journal of Marketing Research*, 51(1): 40-56.

Li, W., X. Wang, R. Zhang, Y. Cui, J. Mao, & R. Jin (2010). Exploitation and exploration in a performance based contextual advertising system. In Proceedings of the Fifteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Osborne M. & A. Rubinstein (1994). A course in game theory. The MIT press.

Shao, X., & L. Li. (2011). *Data driven multi-touch attribution models*, KDD'11, August 21-24 2011, San Diego, California, USA

Wiesel, T., K. Pauwels, & J. Arts (2011). Practice prize paper—Marketing's profit impact: Quantifying online and off-line funnel Progression, *Marketing Science*, *30*(4), 604-611.

Xu, L., J. Duan, & A. Whinston (2014). Path to purchase: A mutually exciting point process model for online advertising and conversion, *Management Science*, forthcoming.