uc3m | Universidad **Carlos III** de Madrid

Master Degree in Statistics for Data Science

2019-2020

Master Thesis

"Data Driven Attribution Models In Digital Marketing"

Sergio Sierra Garrido

1st Ana Arribas Gil

Madrid, 2020

SUMMARY

Digital marketing has suffered an incredible growth in the last decade. Companies advertising their products target online users with specific ads and customized marketing campaigns. This has allowed them to increase the return of investment over traditional marketing campaigns, that include for instance TV or radio ads. The main benefit of digital marketing is that it can be tracked and measured through the use of cookies stored on the browsers of the users. Mathematical models have been developed with the aim of attributing the sales to the different marketing channels used to reach the users. In this work we present the models currently used by most of the advertisers, that are based on simple rules. We also present three data driven models that introduce more granularity, and that have been developed to improve the shortcomings of the simple rule based models. Finally, we apply the models to a dataset and compare the results, with the aim of understanding and characterizing the behavior and properties of the different models.

Keywords: Attribution model. Digital marketing. Customer journey. User. Conversion. Advertising channel. Ad. Campaign.

DEDICATION

This work is dedicated to my parents Curro and Sonia for helping and supporting me during the college years, and to my girlfriend Clara whose encouragement helped me earn the master's degree.

Another special mention goes to Fernando and Juan, who introduced me to the digital marketing world when I worked at Constellation Consulting Iberica. The realization of this work has been possible thanks to the knowledge they shared with me.

Finally, I would like to thank the professors who took part in the Master in Statistics for Data Science for their dedication and effort, and specially to Ana for guiding me throughout this project.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Traditional marketing campaigns have been developed on platforms such us TV, radio, newspaper ads, etc. The recent development in technology and internet has allowed companies to extend the channels in which ads are advertised, emerging what is called digital marketing. New channels and ad formats have appeared. Some examples are search engine marketing (SEM) where Google and its browser is the best example by showing to the users ads at the top and at the bottom of the browsing page. Social media marketing, that is growing quite fast with the expansion of platforms such as Facebook, Instagram or Twitter. Display ads are commonly known by banners and they pop up on web pages such as online newspapers and on mobile apps. Email marketing consists on emailing publicity, etc.

The real impact that traditional marketing campaigns have on users can not be directly measured, making it one of the biggest drawbacks of traditional marketing. For instance, a car brand cannot count exactly the number of cars it has sold after making a TV campaign advertising a new car. This implies budget to be allocated on general ideas and historical campaign's performance, making it difficult to find its optimal distribution. On the other hand, digital marketing allows companies to target users with specific ads which can also be tracked. The customer journey is the path a user does engaging with the company before the final conversion (or non conversion) occurs, and usually companies know it through the use of cookies stored on the web browser of the user. An example of a customer journey could be: {Display ad → Display ad → Email ad → Display ad → SEM ad → Conversion}. Following the previous example, let us suppose that the previous car brand decides to advertise the new car through two digital marketing channels, by creating display and SEM campaigns. When the campaigns start, users are hit many times with display ads and finally some of them decide to google the car, activating the SEM ad. The company knows precisely which and how many ads have been shown to them and at which time. This information can be used to optimize the budget allocation in the channels and campaigns in order to increase the return of investment.

In the last decade, data driven attribution models have been developed aimed to determine which channels have influenced and how the final decision of the users given their customer journey. However, industry still resists to widely implement them and the problem is usually addressed by simple rules, such us attributing all the conversion effort to the last touch point, to the first one, or to all of them evenly. This approach has some disadvantages, since some channels tend to be at the beginning of the customer journey such as prospecting display campaigns, and other usually appear later as SEM. Thus, last touch models assign more credit to SEM ads when display ads may have had a large influence on the final decision. Based on this information, one could think of reducing considerably the budget allocated on display campaigns in order to allocate it on SEM campaigns, but it could lead to a lower number of users hit by the display ads, reducing at the end the number of sells. Similar results yield other simple rule based attribution models.

In this paper we will present the simple models that are widely used, together with the data driven models developed recently that address the problem through more complex formulations, but that provide better results in terms of attributing the conversions to the channels. Later, we will present the results of these models applied to a dataset to finally obtain the conclusions. The aim of the project is to understand and characterize the behavior and properties of the data driven models, and how they perform with respect to the simple models when applying them to the dataset.

2. RELATED WORKS

One of the first works was developed by Shao and Li [1] and they proposed a Bagged Logistic Regression model. The main idea is to use the customer journeys to predict the final decision of the users, that is, whether they make a conversion or not. Then, the coefficients of the model are used to compute the contribution of the channels. The advantage of the bagging idea is the reduction of the variability of the estimated coefficients. The predictors of the model are whether the channels appear on the path or not. They also proposed a simple probabilistic model that relies on the proportion of times the channels appear on the customer journey of users who converted over the ones who did not. Dalessandro, Perlich, Stitelman, *et al.* [2] proposed a causal framework that accounts the individual effect that each ad has on the final decision. The attribution of each channel is then computed through the aggregated effect of each ad. Geyik, Saxena, and Dasdan [3] used the Probabilistic model of Shao and Li to formulate a budget allocation problem given the information provided by the attribution model. They showed that using the attribution model at sub-campaign levels instead of campaign levels leads to a better allocation of the budget. Since a campaign may be formed of different sub-campaigns with different ads, it would mean to apply the model at a deeper level. Zhang, Wei, and Ren [4] presented an AdditiveHazard model based on survival theory, that attributes at channel level taking into account both the temporal variable and the additive effect that ads have on users. Ji and Wang [5] continued the work of Zhang, Wei, and Ren by proposing an Additional Multi-Touch attribution model, that goes deeper into the effects of the ad exposure. They consider both the conversion rate and the delay to the conversion, outperforming the previous models. Ren, Fang, Zhang, *et al.* [6] developed a Dual-Attention Recurrent Neural Network model that assigns credit to individual touch points of the customer journey, to later aggregate them at the campaign and channel levels.

3. ATTRIBUTION MODELS

So far we have explained that data driven attribution models introduce useful information from customer journeys, and that companies can benefit from them to optimize the budget allocation of digital marketing campaigns based on the results of the models and previous campaigns. Traditionally, platforms and marketers have used models based on simple rules. Usually, they assign the credit of the conversions to the adds and channels based on the position of these adds in the customer journey. First, we are going to present these simple models and we will use them to explain the behavior of the channels of the dataset, that will be presented later. Then, we will present the data driven models and we will compare among them and them with respect to the simple models, that will be used as a benchmark.

3.1. Simple rule based attribution models

The main reason these models are widely used is that they are included in the Google platform to analyze the campaigns [1], and most advertisers use Google to promote their products.

3.1.1. Last interaction

The Last interaction model assigns all the credit of a conversion to the last ad of the customer journey. Thus, channels that usually appear at the end of the path such as SEM have a higher rating. These channels are characterized because users are the ones who search for the product. This happens when they are somehow prepared to buy.

3.1.2. First interaction

The First interaction model assigns all the credit of a conversion to the first ad that hits a user. In this case it is common that a brand tries to find new clients by placing ads to users that may be interested in the product advertised. Usually, display campaigns are used with this purpose, and marketers launch these campaigns using demand-side platforms (DSP) due to the vast information these systems have of the users.

3.1.3. Linear

The Linear attribution model assigns equal credit to each touchpoint in the path of a user who has converted. Thus, it does not benefit channels that usually appear at the beginning

[1] https://support.google.com/analytics/answer/1662518

or at the end of the path but to all of them. However, there are channels that hit users with many ads before the conversion, and others that just appear once or twice in the path. For instance, users can be hit with more than ten or twenty display ads before they decide to buy a product. On the other side, channels as SEM or email are less frequent and just one or two interactions are necessary before users take a decision. Thus, display channels would benefit from this model. We will denote this model by Linear Proportional, since the contribution is proportional to the number of appearances.

There is an alternative model that assigns equal credit to each channel of the path of a user who makes a conversion. In this case, it does not matter how many times an add or channel appears, but just if it appears or not. The effect is that channels such as display do no benefit from hitting users many times. We will denote this alternative by Linear Same.

3.1.4. Position based

The Position based attribution model, also called Bathtub, assigns 40% credit to the first ad, another 40% credit to the last ad and the remaining 20% is distributed evenly among the other ads. From the attribution point of view, the behavior is a mixture of the previous models, giving more weight to both first and last interaction models.

3.1.5. Time decay

The Time Decay attribution model assigns more credit to the touchpoints that are closer in time with the conversion. Therefore, sometimes it has a similar effect to the last touch model. SEM ads are assigned in general more credit in detriment of others as display ads.

3.1.6. Position decay

The Position Decay attribution model assigns more credit to the touchpoints that are closer in the path to the conversion, but does not take into account the time. It is not used in practice but it will help us to understand better the behavior of the channels of the dataset.

3.2. Data driven attribution models

Before defining the models, let us introduce the notation we will follow. Let us denote any user as u and any channel as Ch. Moreover, we denote the set of users as $\{1, ..., U\}$, the set of channels as $\{1, ..., n\}$ and the set of ads shown to a user u as $\{a_i^u\}_{i=1}^{l_u}$, where i stands for each ad in the customer journey and l_u is the length of the path. The set of times at which the ads are displayed is denoted as $\{t_i^u\}_{i=1}^{l_u}$. The conversion of a user u is specified as $C_u = 1$ ($C_u = 0$ if no conversion occurs).

3.2.1. Probabilistic

The Simple Probabilistic model was introduced by Shao and Li [1]. It is based on the probability of a user converting given the channels that appear in the path. For that, they compute the first and second order probabilities of conversion for each channel and each pair of channels respectively. Mathematically, it is expressed as:

$$P(C = 1 \mid Ch_i) = \frac{N_{\text{conv}}(Ch_i)}{N_{\text{conv}}(Ch_i) + N_{\text{no conv}}(Ch_i)}$$

$$P\left(C = 1 \mid Ch_i, Ch_j\right) = \frac{N_{\text{conv}}\left(Ch_i, Ch_j\right)}{N_{\text{conv}}\left(Ch_i, Ch_j\right) + N_{\text{no conv}}\left(Ch_i, Ch_j\right)}$$

where Ch_i and Ch_j are any channels, N_{conv} and $N_{\text{no conv}}$ are the number of conversions and no conversions in which channels Ch_i and Ch_j are in the path, and C is the binary response denoting if a user makes a conversion or not. Finally, the contribution of each channel is computed as:

$$c(Ch_i) = P(C = 1 \mid Ch_i) + \frac{1}{2(n-1)} \sum_{i}^{n} \sum_{j, i \neq j}^{n} \left[P\left(C = 1 \mid Ch_i, Ch_j\right) - P(C = 1 \mid Ch_i) - P\left(C = 1 \mid Ch_j\right)\right]$$

where n is the number of channels.

As opposed to the previous models, the Probabilistic model uses the information of both the users who converted and the users who did not. However, it does not use the temporal distribution of the ads.

3.2.2. Logistic regression

The logistic regression model was introduced by Shao and Li [1] and they use a bagged logistic regression to predict the conversion of the users given their customer journey. The bagging idea is introduced for the sake of improving the accuracy. For a user u, the model is defined as:

$$\log \frac{p_u}{1 - p_u} = \beta_0 + \sum_{i}^{n} \beta_i x_i, \quad u = 1, ..., U$$

where p is the probability that user u will convert given the channels that appear in its path. Thus, predictors x_i are binary.

Following the bagging idea, we randomly sample a subset of users. We use the *l2* norm as penalty term in the loss function. Thus, the optimization problem to solve at each iteration is:

$$\min_{\boldsymbol{\beta},r} \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 + \lambda \sum_{u=1}^{U} \log\left[\exp\left(-C_u\left(\boldsymbol{x_u}^t\boldsymbol{\beta}\right) + r\right) + 1\right]$$

where λ is the regularization parameter. Then, we create a data partition for training and testing. This allow us to find the optimal value for λ and for the probability threshold that better divides between conversions and no conversions. Finally, we compute the contribution of the channels as the mean value of the coefficients obtained.

As was the case in the Probabilistic model, the Logistic Regression model uses the information of both the users who converted and the users who did not. However, it does not use the temporal distribution of the ads.

3.2.3. AdditiveHazard

The AdditiveHazard model was proposed by Zhang, Wei, and Ren [4] and is based on survival models. They use the survival function to model the time that goes between the first ad is displayed to a user and end of the time window, that can be either the time of the conversion or a pre-specified time interval if no conversion occurs. The survival function is defined as:

$$S(t) = P(T > t)$$

where the value of the function $S(t)$ is the probability that the time of the conversion T is greater than some time t. From there, they define the hazard function as the instantaneous rate of occurrence of an event:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t \mid T > t)}{\Delta t}$$

which is simplified after some operations. Substituting in the survival function we obtain:

$$\lambda(t) = -\frac{dS(t)}{S(t)}$$

Finally, the survival function is computed as:

$$S(t) = \exp\left(-\int_0^t \lambda(\tau)\, d\tau\right)$$

Thus, from the previous expression and the definition of the survival function, we get that the hazard function measures the influence that an ad makes on a user. They use an exponential function to model that effect. For users whose path do not lead to a conversion, the end of the time window is set to 15 days after the last ad is displayed, and that time is denoted as T_u for user u. Following the previous notation, the hazard function with exponential kernel for user u is expressed as:

$$\lambda_u(t) = \begin{cases} \sum_{i=1}^{l_u} \beta_{a_i^u} \omega_{a_i^u} \exp\left[-\omega_{a_i^u}\left(t - t_i^u\right)\right] & \text{if} \quad 0 \leq t \leq t_{l_u} \\ 0 & \text{otherwise} \end{cases}$$

where $\beta_{a_i^u}$ measures how much ad i in the path of user u influences user u, and $\omega_{a_i^u}$ measures how long the effect of the ad lasts on the user. From the previous definition, we obtain two variables to be optimized for each channel. The set of parameters $\Theta = \{\boldsymbol{\beta}, \boldsymbol{\omega}\}$ is estimated through maximum likelihood, with the log-likelihood expressed as:

$$\begin{aligned} \mathcal{L}(\Theta) &= \sum_{u=1}^{U} \log\left[\left(S(T_u)\lambda(T_u)\right)^{C_u} S(T_u)^{1-C_u}\right] \\ &= \sum_{u=1, C_u=1}^{U} \log\left[\sum_{i=1}^{l_u} \beta_{a_i^u}\omega_{a_i^u} \exp\left(-\omega_{a_i^u}(T_u - t_i^u)\right)\right] \\ &- \sum_{u=1}^{U}\sum_{i}^{l_u} \beta_{a_i^u}\left[1 - \exp\left(-\omega_{a_i^u}(T_u - t_i^u)\right)\right] \end{aligned}$$

Therefore, the optimization problem to be solved is:

$$\begin{aligned} \max_{\Theta} \quad & \mathcal{L}(\Theta) \\ \text{s.t.} \quad & \beta_i \geq 0, \quad i = 1, ..., n \\ & \omega_i \geq 0, \quad i = 1, ..., n \end{aligned}$$

As they point out, the problem can be addressed easier through the Majorize-Minimazation algorithm [7], because the estimates can be updated independently. Further explanations can be found in [4]. The equations to optimize the coefficients are:

$$\beta_k = \frac{\sum_{u=1}^{U}\sum_{i=1, C_u=1, a_i^u=k}^{l_u} p_i^u}{\sum_{u=1}^{U}\sum_{i=1, a_i^u=k}^{l_u} 1 - \exp\left[-\omega_k\left(T_u - t_i^u\right)\right]}$$

$$\omega_k = \frac{\sum_{u=1}^{U}\sum_{i=1, C_u=1, a_i^u=k}^{l_u} p_i^u}{\sum_{u=1}^{U}\sum_{i=1, a_i^u=k}^{l_u} p_i^u\left(T_u - t_i^u\right) + \beta_k\left(T_u - t_i^u\right)\exp\left[-\omega_k\left(T_u - t_i^u\right)\right]}$$

where p_i^u is defined as the contribution of ad i to the final decision of user u:

$$p_i^u = \begin{cases} \frac{\beta_{a_i^u}\omega_{a_i^u}\exp\left[-\omega_{a_i^u}\left(T_u - t_i^u\right)\right]}{\sum_{i=1}^{l_u}\beta_{a_i^u}\omega_{a_i^u}\exp\left[-\omega_{a_i^u}\left(T_u - t_i^u\right)\right]} & \text{if} \quad C_u = 1 \\ 0 & \text{if} \quad C_u = 0 \end{cases}$$

Thus, the contribution of the channels to the conversion of the users is computed through the sum of the contribution of each of their ads.

The AdditiveHazard model uses the information of both the users who converted and the users who did not. It also uses the temporal distribution of the ads, since the model measures not only the strength of the impact of the ads, but also their time decaying effect on the users.

4. DATASETS AND DATA PRE-PROCESSING

4.1. Dataset description

We use the R package *ChannelAttribution* [2] which includes a synthetic dataset with 10 thousand rows and 4 fields:

- Path: the customer journey of each user.
- Total conversions: the number of users who converted with the given path.
- Total conversion value: the value of the conversion.
- Total null: the number of users who did not convert with the given path.

From the description we have that for each path there are usually some users who converted and others who did not. Thus, we can generate a new dataset from it with the individual path of each user, so users are no longer aggregated on the same path. To avoid having more than one user with the same journey, we are going to take just one user from each path and sample if it ends in a conversion or not. The probability of conversion is computed as the proportion of the number of conversions over the number of users (conversions plus non conversions). That makes a dataset with over 10 thousand users. There are 12 channels hitting the users, which have been mapped to integer values for the sake of reducing the computational cost. The mapping to integer values is as follows:

$$\begin{Bmatrix} \text{eta} \rightarrow 0 & \text{iota} \rightarrow 1 & \text{alpha} \rightarrow 2 & \text{beta} \rightarrow 3 \\ \text{theta} \rightarrow 4 & \text{lambda} \rightarrow 5 & \text{kappa} \rightarrow 6 & \text{zeta} \rightarrow 7 \\ \text{epsilon} \rightarrow 8 & \text{gamma} \rightarrow 9 & \text{delta} \rightarrow 10 & \text{mi} \rightarrow 11 \end{Bmatrix}$$

Since the dataset does not contain the time distribution of the ads, we have chosen to simulate it based on *Criteo* dataset, which is described below.

4.2. Criteo dataset description

Criteo AI Lab is a digital marketing company that has published a dataset [3] for attribution modeling purposes. The dataset was published in 2017 by Diemert Eustache, Meynet Julien, Galland, and Lefortier [8] together with a paper in which they show how attribution models can increase the efficiency of bidding in display advertising. It contains live traffic data generated from its campaigns during a period of 30 days. That makes up to a total of

[2] https://cran.r-project.org/web/packages/ChannelAttribution/index.html

[3] http://ailab.criteo.com/criteo-attribution-modeling-bidding-dataset/

16 million ads impressions over 700 campaigns. Each line corresponds to an impression displayed to a user and it contains information such as the timestamp of the impression, the channel and the user. This information has been used to generate the timestamp of each ad in the previous dataset.

We choose to use *ChannelAttribution* dataset instead of *Criteo* dataset because users in the second one are hit usually by just one or at most two channels. Nowadays it is rare to think of big companies using just two channels to advertise their products. Also, when users are hit by two channels, usually their ads are not mixed in the path. Moreover, other authors have used *Criteo* dataset with some of the models we are going to use, so we are testing them with new data.

4.3. Event timestamp generation

Generating the timestamp of an ad is a difficult task and requires us to make some assumptions. The main idea is to use *Criteo* dataset to generate the time that goes between two consecutive ads for all the ads and users of *ChannelAttribution* dataset. Let's denote this time period as interarrival time. That is, if for a given channel and user we know the interarrival time distribution from which each ad has been displayed, then we can sample new timestamps for other ads from that distribution.

To address the problem and given that *Criteo* dataset is huge, we have ordered it by users and selected the first 200 thousand ads, to later compute the interarrival times of each channel. Taking the channels with more ads displayed and plotting a histogram of the interarrival times of their ads, we realize that most of these distributions may be fitted to an exponential or a gamma distribution. Figure 4.1 shows an example with three of the selected channels, and Figure 4.2 shows the kernel density estimation of their interarrival times, together with a fitted gamma distribution, for the sake of comparing the shape of both distributions. Thus, we can carry out a goodness of fit test to check if any of these distributions fit the data. For that, we have used the Kolmogorov-Smirnov test (KS) with a significant level of 0.01. The test does not make any assumption about the distribution of the data. The null hypothesis H_0 is that the interarrival times of a given channel are consistent with the distribution we are comparing against, whilst the alternative hypothesis H_1 is that they are not consistent with that distribution. However, we can not compute the parameters of the distribution and obtain the critical values of the test, since they would not be valid. Instead, we rely on Monte Carlo (MC) simulation to compute the critical values. The process is as follows:

1. Run the KS test for each of the chosen channels of *Criteo* dataset, with H_0 establishing that the interarrival times are distributed according to the distribution we are testing against. The parameters are estimated through Maximum Likelihood.
2. Simulate from that distribution a number of n MC samples with the number of

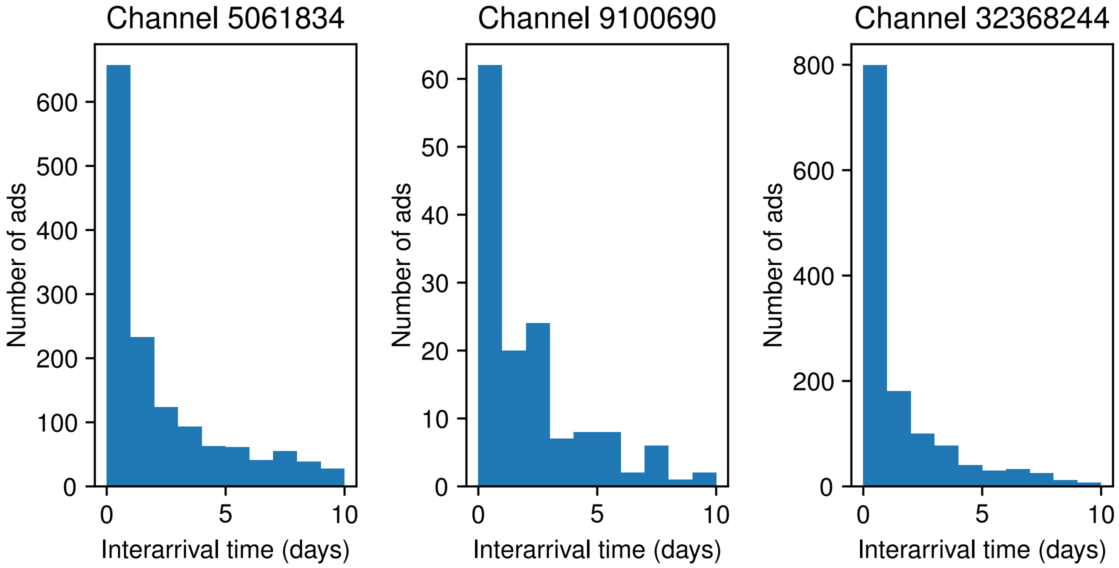

Fig. 4.1. Histograms of the interarrival times of three channels of *Criteo* dataset. The channels IDs refer to the original ones in *Criteo* dataset and are kept for the sake of identification and reproducibility.

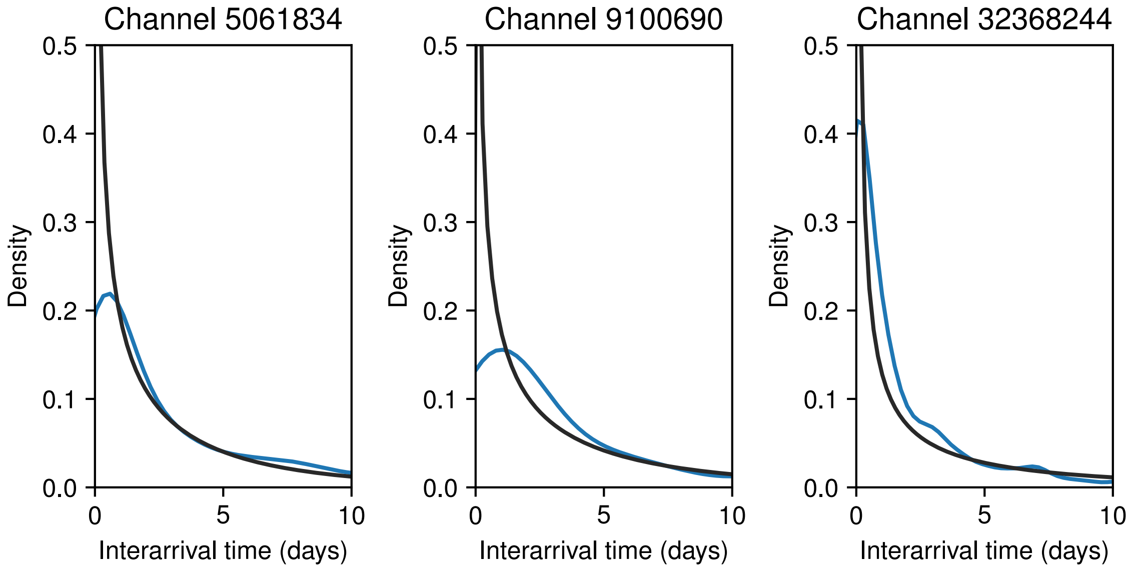

Fig. 4.2. Kernel density estimation of the interarrival times of three channels of *Criteo* dataset. The black line is the probability mass function of a gamma distribution fitted to the data, and the blue line is the kernel density estimation of the interarrival times. The channels IDs refer to the original ones in *Criteo* dataset and are kept for the sake of identification and reproducibility.

observations equal to the number of interarrival time values that the channel has. We have chosen $n = 1000$ simulations.

3. For each simulation we run the KS test with the same H_0 and H_1, and we keep the test value.

4. The critical value or p-value of the KS test is approximated by the fraction of test statistics of the simulated samples that are greater than the test statistic obtained from the channel.

After running the test on the 50 channels with more ads displayed, we do not reject the null hypothesis in most of them when testing against a gamma distribution. Thus, we can use the fitted parameters of 12 of them to sample new values and generate the timestamp of the ads of *ChannelAtribution* dataset. Table 4.1 includes the list of channels taken from *Criteo* dataset, together with the approximated p-value of the KS test and shape and scale parameters of the fitted gamma distribution. The selected channels are the ones that have more ads displayed and for whose null hypothesis is not rejected. Finally, the parameters of each of these channels are used in *ChannelAtribution* (mapping the channels of both datasets as shown in Table 4.1) and interarrival times are sampled randomly with the gamma distribution. We choose to sample from the gamma distribution instead of directly sampling from the empirical interarrival time distribution of *Criteo* dataset because it allows us to better understand and characterize the distributions of the time between ads.

For a given channel, we are generating all its interarrival times from a gamma distribution with the parameters fixed. Thus, we are assuming that interarrival times with different positions in the path can be drawn from the same probability function. That is, we need to check that the time elapsing between the first and the second ad, the time elapsing between the second and the third ad, etc, follow the same distribution. That is due to the fact that users in *Criteo* dataset are reached by at most two channels and in many cases by one channel, and thus we are checking if the distributions of the interarrival times with different position in the path of these users are the same. For that, we rely again on the KS test. This time the null hypothesis H_0 is that the distribution of the interarrival times with different position in the path is consistent with the distribution of all the interarrival times. The alternative hypothesis H_1 is that both distributions are not consistent with each other. As an example, if we had a customer journey in *Criteo* dataset such as $\{\mathrm{Ch}_0 \rightarrow \mathrm{Ch}_0 \rightarrow \mathrm{Ch}_1 \rightarrow \mathrm{Ch}_0 \rightarrow \text{No Conversion}\}$, we would be testing if the distributions of the times elapsing between the first and the second ad and between the third and the fourth ad, are consistent with the distribution of all the elapsing times that reach Ch_0 (which does not take into account the position in the path), for all the users reached by Ch_0. This time we do not need to follow the previous procedure. We just have to compute the critical values of the KS test given the positioned interarrival times and the parameters shown in table 4.1 for each channel, and we set a significance level of 0.01. The main issue we find here is that not all the channels have a large number of users with

Criteo channel	P-value	Shape parameter	Scale parameter	*ChannelAtribution* channel
23644447	0.342	0.598	3.759	0
16491630	0.438	0.297	6.503	1
73327	0.268	0.318	4.943	2
5061834	0.148	0.441	6.138	3
21016759	0.152	0.533	5.528	4
9100690	0.170	0.422	7.862	5
19602309	0.274	0.406	7.004	6
17710659	0.232	0.598	4.546	7
30535894	0.130	0.432	6.194	8
8980571	0.142	0.413	5.356	9
32368244	0.860	0.230	13.636	10
14235907	0.206	0.418	5.673	11

Table 4.1. Channels selected of *Criteo* dataset. P-value column refers to the approximated p-value of the KS test against the gamma distribution, and shape and scale columns are the parameters of the distribution. ChannelAttribution column is the mapping of the channels from *Criteo* dataset to *ChannelAtributtion* dataset.

four or more ads displayed. There are channels that usually appear just two or three times in the user paths, so we are going to skip these channels from the test and assume that they behave similarly to the ones tested. We test the first three interarrival times, that in almost all the cases correspond to the first four consecutive ads, because since we explained before, ads belonging to the same channel are usually together in the path. The results are shown in table 4.2, and we do not reject the null hypothesis in most of the cases. Thus, we assume that the interarrival times of a given channel can be sampled from the same distribution, no matter the position of the ads in the path. Figure 4.3 shows an example of the distribution of the first three interarrival times of one of the channels selected.

Criteo channel	1^{st} to 2^{nd}	2^{nd} to 3^{rd}	3^{rd} to 4^{th}
5061834	0.075	0.032	0.023
19602309	0.762	0.010	0.085
30535894	0.005	0.046	0.030
8980571	0.496	0.033	0.023
32368244	0.000	0.003	0.000
14235907	0.064	0.463	0.017

Table 4.2. P-value obtained when comparing the first three interarrival times of some *Criteo* Dataset channels against the gamma distribution specified with the parameters shown in table 4.1

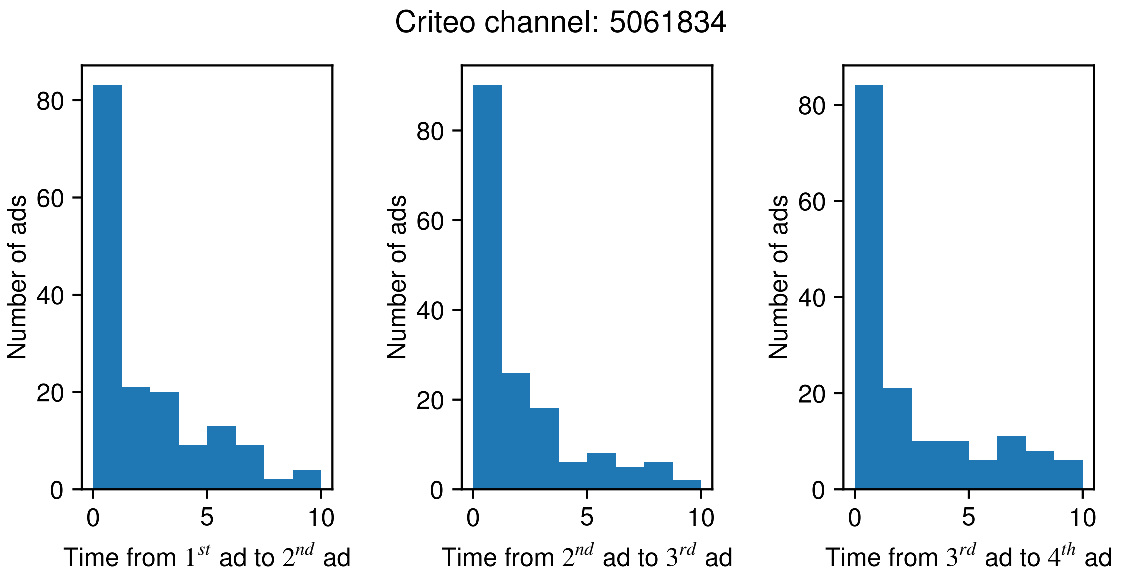

Fig. 4.3. Histograms of the individual interarrival times (days) of channel 50618334 of *Criteo* dataset.

4.4. Remarks on the process of timestamp generation

So far, we have covered how to generate the time distribution of the ads in *ChannelAtribution* dataset relying on those of *Criteo* dataset. Through the process we have done some assumptions, but we should consider a few more:

- The results of the KS test presented in table 4.1 indicate that gamma distributions are consistent with the time distribution of the interarrival times of the ads. However, it is unlikely to be completely true. A channel (or campaign) can only display ads to the users when they are surfing the internet, so actually there are time slots in which ads are displayed. For instance, online traffic is usually lower during the night than during the day making it impossible to display ads to many users who are offline during that period.

- If we look closely at the *Criteo* dataset, we realize that most of the users have seen ads that belong just to one channel, in some cases to two channels, and very rarely to more than two channels. That is the main reason why we have computed the interarrival times with both ads belonging to the same channel. It is frequent to find consecutive ads from different channels in *ChannelAtribution* dataset, so when generating their time distribution we assume that it can be generated from the gamma distributions shown in table 4.1.

- The previous remark explains why Shao and Li [1] conclude that the second order probabilistic model provides sufficient granularity. They use *Criteo* dataset and since it is very unlikely to find a user reached by more than two channels, then, adding a third order term to the model does not result in a better performance. Nonetheless, the current trend is to hit users through many channels.

5. RESULTS

In this section we are going to analyze the results obtained by applying the previous models to *ChannelAtributtion* dataset. First, we will cover the simple rule based models and use them to describe the behavior of the channels. Later, we will compare the data driven models, and we will use the simple models as a benchmark.

Since there are twelve channels, we are going to select the six of them with more ads displayed, because there are three channels that have displayed just a few ads and reached even less users. For example, two of them appear no more than twenty times, which causes some models as Logistic Regression and Probabilistic to return odd results.

For the sake of comparing among the models, we will normalize the contribution of the channels between 0 and 1, so that the sum of the contribution of all the models is 1.

All the code written to obtain the results and to process the data is available from Github [4].

5.1. Simple models

Figure 5.1 contains the contribution of the channels. We can see how some models have a similar behavior as Bathtub and Last, Bathtub and First, and Linear Proportional and Linear Same. We have to remind that these models take into account only the users who converted. Following with the description of the channels we have:

- Channel 0: appears with more frequency at the end of the path. Last, Pos. Decay and Bathtub models give it fairly credit. However, Linear and First also give it credit, which implies that ads from channel 0 are the most frequent in the path of users that convert.
- Channel 1: has a behavior similar to channel 0, although in this case it hit users with more frequency at the beginning rather than at the end.
- Channel 2: appears practically at the end of the journey. All the models assign it less credit compared to channels 0 and 1, so it hit users with a lower frequency. This can be checked by looking at both the Linear Proportional and Linear same models.
- Channel 3: behaves similar to channel 1, but as channel 2 it has a lower frequency. From the Linear Same model attribution, we infer that usually appears with just a few channels in the path.

[4] https://github.com/sesiga/attribution_models_for_digital_marketing

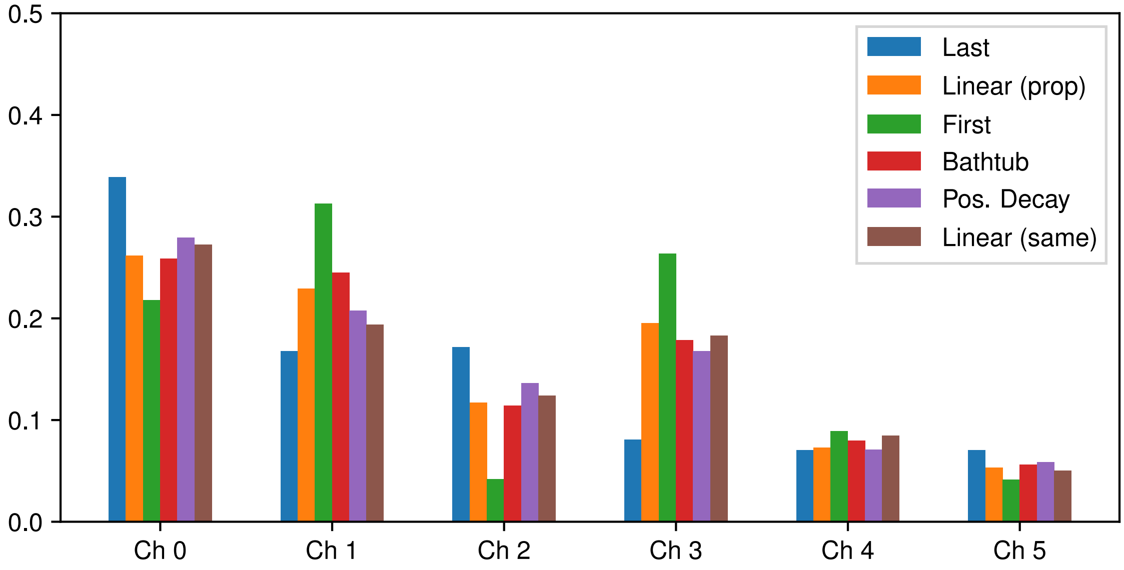

Fig. 5.1. Contribution of the simple ruled based models applied to the six channels with more ads displayed. The contribution of the models is normalized between 0 and 1, so that the sum of all the contributions is 1.

- Channel 4: it has a similar credit over all the models but attribution is low. Hence, channel 4 appears indifferently in any position in the journey. Since the credit is low, it does not appear in many users who finally end in a conversion.
- Channel 5: has a low attribution with the tendency of hitting users at the end of the journey. Thus, as channel 4, channel 5 is not that common in users who convert as the previous channels.

5.2. Data Driven Models

Figure 5.2 contains the contribution of the channels with the data driven models together with the most used simple models, as benchmark to compare with. The time decay model is presented here, since none of the previous models take into account the temporal variable.

Probabilistic and Logistic Regression models have a similar behavior. The credit given to all the channels is quite similar. However, in comparison channels 0, 1 and 2 receive more credit from the rest of the models, while channels 3, 4 and 5 have a similar attribution. The main reason is that channels 0, 1 and 2 hit more users than other channels and they hit them with more ads. Since Probabilistic and Logistic Regression models do not take into account how many ads of the same channel have been displayed to a user but just if the channel appears in the customer journey, these models do not give extra credit to channels 0, 1 and 2 for hitting users many times. On the other hand, AdditiveHazard model is more similar to Time Decay model and Last touch model. It does give more

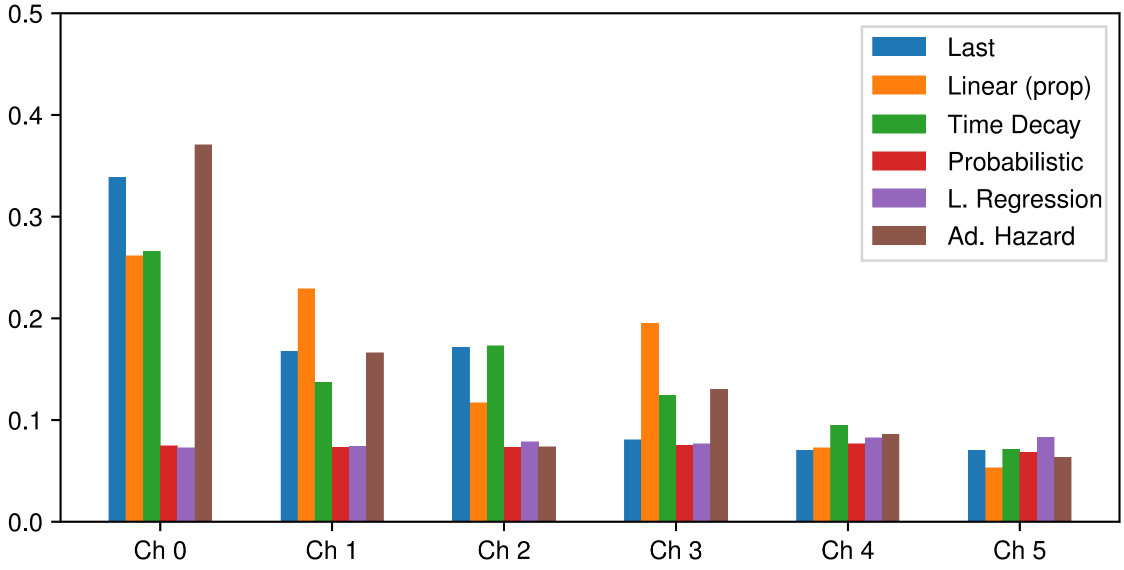

Fig. 5.2. Contribution of the data driven models applied to the six channels with more ads displayed. The contribution of the models is normalized between 0 and 1, so that the sum of all the contributions is 1.

credit to channels and user with many ads because each ad has an additive effect on the final probability of conversion.

From the results of the Probabilistic model, we find that channels 0, 1 and 2 do not outperform other channels in relative terms. That is, the proportion of times that for instance channel 0 (or 1 or 2) appears in the path of users who end with a conversion with respect to those who end with no conversion is similar to channels 3, 4 and 5. Thus, when channel 0 (or 1 or 2) is present in the path of a user, the probability of converting of this user is similar to the probability of converting if instead of channel 0 (or 1 or 2), channel 3 (or 4 or 5) was present. However, since channel 0 appears many more times than channel 3, the AdditiveHazard model gives it more credit.

6. CONCLUSIONS AND FUTURE WORK

We have presented a comparison of different attribution models that address the problem of channel attribution in digital marketing. First, we presented the simple rule based models widely used in the industry, and later we presented three data driven models. We used a synthetic dataset to compare the models and we saw that the simple ruled based models only take into account users who make a conversion, skipping in consequence most of the users and ads. The result was that these models do not assign correctly the contribution of the channels. Later, we presented three data driven models: Probabilistic, Logistic Regression and AdditiveHazard. We saw that the Probabilistic and the Logistic Regression models have a similar behavior with this dataset. Both models only take into account which channels appear in the path of all the users, the ones who make a conversion and the ones who do not. However, these models do not use the temporal distribution of the ads. The AdditiveHazard model is based on survival theory and measures the influence of the impact of each ad through the survival function. It considers both the strength of the impact and how long the effect lasts on the users. We observed that the attribution of this model is different to the previous ones, and it is more similar to the attribution of the time decaying model.

There are some interesting directions in which this project can be continued. Other data driven models have been developed recently [6] and could be analyzed through this dataset. Moreover, it would be interesting to analyze these models with a real dataset different from *Criteo* dataset, since most of the works use it, and we explained previously that it contains users usually reached by just one or two channels. Nowadays it is common to make campaigns on different channels such as SEM, social media, email and display, with the last one having usually more than one sub-campaign.

BIBLIOGRAPHY

[1] X. Shao and L. Li, “Data-driven multi-touch attribution models,” in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 258–264.

[2] B. Dalessandro, C. Perlich, O. Stitelman, and F. Provost, “Causally motivated attribution for online advertising,” in *Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy*, 2012, pp. 1–9.

[3] S. C. Geyik, A. Saxena, and A. Dasdan, “Multi-touch attribution based budget allocation in online advertising,” in *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*, 2014, pp. 1–9.

[4] Y. Zhang, Y. Wei, and J. Ren, “Multi-touch attribution in online advertising with survival theory,” in *2014 IEEE International Conference on Data Mining*, IEEE, 2014, pp. 687–696.

[5] W. Ji and X. Wang, “Additional multi-touch attribution for online advertising,” in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[6] K. Ren *et al.*, “Learning multi-touch conversion attribution with dual-attention mechanisms for online advertising,” in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2018, pp. 1433–1442.

[7] D. R. Hunter and K. Lange, “A tutorial on mm algorithms,” *The American Statistician*, vol. 58, no. 1, pp. 30–37, 2004.

[8] Diemert Eustache, Meynet Julien, P. Galland, and D. Lefortier, “Attribution modeling increases efficiency of bidding in display advertising,” in *Proceedings of the AdKDD and TargetAd Workshop, KDD, Halifax, NS, Canada, August, 14, 2017*, ACM, 2017, To appear.