Erasmus University Rotterdam
Erasmus School of Economics
Bachelor Thesis BSc^2 Econometrics and Economics

A comparison of the Markov and AdditiveHazard models for online advertisement attribution modelling

Lotte-Sophie Folbert (531284)

Supervisor:	dr. Kathrin Gruber
Second assessor:	Markus Mueller
Date final version:	2nd July 2023

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Abstract

Customers often encounter multiple advertisements before converting, which makes it difficult to analyse the contribution of each advertisement channel in the conversion. To solve this problem, many attribution models have been developed. In this paper, the Markov and AdditiveHazard attribution models are compared. The results indicate that the attribution scores of both models for this specific data are quite similar, whereas the AdditiveHazard model outperforms the Markov model when it comes to predictive performance. However, the robustness check indicated that the parameter estimation in the AdditiveHazard model is not robust, as it strongly depends on the chosen initial values. Furthermore, as the AdditiveHazard model is based on several assumptions, further research is required to improve the robustness of this model.

1 Introduction

Online advertising has been a trending marketing tool during the past years. In 2022, expenditures on digital advertisements reached €86 billion in Europe. In comparison, these expenditures amounted to only €18.9 billion euros in 2010 (Interactive Advertising Bureau Europe, 2023). These expenses and revenues are distributed over a wide range of online advertisement channels, such as paid search, display, video, and social media advertisements (Interactive Advertising Bureau Europe, 2023). Through these advertisements, customers can visit the company's website and eventually conclude a purchase transaction. A conversion often does not happen immediately after one advertisement interaction, but most customers see multiple advertisements before they convert (Li & Kannan, 2014). These advertisements influence customer behaviour in several ways. Firstly, they can lead to a carryover effect by influencing the customer to visit the company's website again through the same advertising channel. Secondly, these advertisements can motivate the customer to visit the website via a different online channel, which results in spillover effects (Li & Kannan, 2014). Because of the multiple online touchpoints in one consumer path[1] and their interactions, it becomes quite challenging to analyse the contribution of each advertisement channels to the potential conversion (Anderl et al., 2016). This process of allocating the value of a conversion to each of the marketing touchpoints across the online advertisement channels is called online advertisement attribution modelling (Moffett, Pilecki, and McAdams, 2014, as described in Kannan, Reinartz and Verhoef, 2016).

In the past, companies used heuristic measures to distribute the value of conversions over all channels in the path. One of these is the last-touch metric, which allocates the conversion completely to the last touchpoint in the consumer path (Singal, Besbes, Desir, Goyal & Iyengar, 2019). This metric ignores all the other advertisement impressions. Additionally, these rule-based approaches do not account for the timing and sequential characteristic of the interactions. The simplicity and inadequacy of these heuristics cause them to not fit reality so well (Li & Kannan, 2014; Shao & Li, 2011; Y. Zhang, Wei & Ren, 2014). To overcome this issue, several data-driven attribution models that use individual customer path data, have been developed. Some of these models include logistic regression models (Shao & Li, 2011), hierarchical models (Li & Kannan, 2014), mutually exciting point process models (Xu, Duan & Whinston, 2014), vector

[1] A consumer path, also called a customer journey, consists of all touchpoints with online marketing channels before a purchase decision is made (Anderl, Becker, Von Wangenheim & Schumann, 2016).

autoregression models (Kireyev, Pauwels & Gupta, 2016), hidden Markov models (Abhishek, Fader & Hosanagar, 2012), and neural networks (Du, Zhong, Nair, Cui & Shou, 2019). Many of these data-driven models have already proven to outperform the heuristic attribution measures (Abhishek et al., 2012; Berman, 2018; Ji & Wang, 2017; Li & Kannan, 2014; Singal et al., 2019; Xu et al., 2014; Y. Zhang et al., 2014).

Looking at the characteristics of these past models, some important features for accurate attribution modelling can be derived. First of all, an attribution model should reflect the sequential nature of consumer paths. Secondly, the potential interactions between the advertisement channels should be incorporated in the model (Li & Kannan, 2014). One approach that takes these aspects into account is the Markov model (Anderl et al., 2016; Dalessandro, Perlich, Stitelman & Provost, 2012; Singal et al., 2019). Besides incorporating these important characteristics, the model is also flexible and allows for multiple attribution scores to be computed with the model. Therefore, this model can be adjusted for and applied in various settings, as has already been shown by previous papers (Anderl et al., 2016; Dalessandro et al., 2012; Singal et al., 2019).

Based on these arguments, the Markov model seems to be an appropriate model for attribution modelling. However, there are also some potential important features of the consumer path data that are ignored. First of all, consumer data is always observed during a specified period. If no conversion is observed for a given consumer path, then it is assumed that this customer does not convert. However, a conversion can also happen after the observation period ended. This additional uncertainty is not incorporated in the Markov model. Secondly, even though the Markov model incorporates the sequential feature of a consumer path, it does not incorporate the timing. If an advertisement is shown only shortly before the conversion is made, then this advertisement is likely to have a much larger impact than if the advertisement was shown a long time ago. Therefore, considering a time dimension in attribution modelling may lead to more accurate predictions. To achieve this, it is important to look at the impact of the different advertisement channels. Some channels may have a strong impact when the advertisement is shown, but this influence can quickly disappear again, and for other channels it may be the other way around. Therefore, it is important to also incorporate some channel-specific characteristics, such as the impact strength and the time-decaying speed of the impact.

The above mentioned shortcomings of the Markov model are incorporated by the Additive-Hazard model (Y. Zhang et al., 2014). Even though these models focus on other features of the path data, a comparison between these two models has, to our knowledge, not been made yet. Therefore, to examine whether these differences also result in distinct model performances, this paper aims to answer the following research question: *how do the Markov and AdditiveHazard models compare with each other based on predictive performance and attribution allocation?*

The research is performed using a publicly available digital marketing dataset for an unspecified product, where customer paths were collected from 1 July 2018 until 31 July 2018. The results show that the attribution allocations for the AdditiveHazard and Markov model are similar for this data set. Additionally, the data-driven attribution metrics do not outperform the heuristic approaches. Regarding the predictive performance, the AdditiveHazard model outperforms the Markov model. The survival model has better area under the curve, F1 and precision values. However, the robustness check indicates that the parameter estimation in this model is not robust, as it strongly depends on the initial values.

This research is academically relevant, because it compares two attribution models, which have not been compared before. By highlighting the differences and similarities between the models, the research identifies the consequences of attribution modelling choices. In this way, we aim to contribute to the development of more accurate attribution models. Additionally, this paper is also socially relevant. Attribution metrics are used to analyse and optimise advertisement campaigns (Petersen et al., 2009; Shao & Li, 2011). Therefore, it is crucial to know which advertisement channels contributed to what extend in the conversions of the consumers. By comparing two attribution models and applying them to an empirical example, we show the strengths and weaknesses of both models, and in this way we contribute to the understanding of how both models work. By creating a better understanding of these models, it becomes easier for companies to correctly apply these methods, which in turn can contribute to making better attribution analyses and eventually better advertisement campaigns.

The rest of this paper is organised as follows. In Section 2, background information on the online advertising industry is provided, together with some previous works related to the AdditiveHazard and Markov model. Section 3 describes the data, followed by a description of the data used in Section 4. In Section 5, the attribution allocation, predictive performance and robustness checks are discussed. Finally, Section 6 contains the conclusion, which is followed by a discussion of the limitations in Section 7.

2 Literature Review

Attribution allows a company to determine the effectiveness of the different advertisement channels and this information can be used to optimize the allocation of the online advertising budget (Abhishek, Despotakis & Ravi, 2017). However, determining the efficiency of the advertisement channels can be complicated in reality. Companies often do not place advertisements themselves on websites, social media or search engines. Instead, there is a complete industry, the online advertising industry, which aims to connect advertisers to the consumers (Evans, 2009).

In the first part of this literature review, we elaborate on this online advertisement setting by describing how the online advertising industry works and addressing some economic issues that may arise. In the second part of this Section, the focus lies on the attribution models, as some previous applications of both the Markov and AdditiveHazard model are discussed and an overview of a few other attribution models is presented.

2.1 Online Advertising Industry

The online advertising industry is a two-sided market (Evans, 2008, 2009). In two-sided markets, one or several platforms, or intermediaries, facilitate interactions between two groups of end-users (Anderson & Gabszewicz, 2006; Rochet & Tirole, 2004). Two-sided markets are defined by the fact that it is mostly the price structure, instead of the price itself, that matters. Therefore, the platforms have to design the price structure in such a way to keep both sides on board (Rochet & Tirole, 2004). In the online advertising industry, the market consists of three groups of agents: 1) advertisers, 2) intermediaries, who facilitate the contact between the customers and advertisers, and 3) consumers (Evans, 2009). On one hand, the intermediaries sell knowledge about customers and advertising inventory to the advertisers. On the other hand, they offer content of various types such as information and entertainment to consumers, while they show them advertisements in exchange (Anderson & Gabszewicz, 2006; Evans, 2009). To keep both parties on board, it is important that intermediaries do not show too many advertisements to consumers, as the latter often think of advertisements as nuisance. Additionally, the benefits that the advertisers receive depend on the number of consumers on the other side of the market, hence not showing too many advertisers is also in interest of the advertisers themselves (Anderson & Gabszewicz, 2006).

The exact market structure in this industry depends on the type of online advertising. The markets for the various channels differ in the extent to which the intermediaries are integrated. Our focus will be on search and display advertising, as the online channels in our data fall into these categories. For search advertising, the intermediaries are mostly fully integrated, implying that there is only one company that interacts directly with both customers and advertisers. For display advertising this is not the case, as the intermediaries consist of several agents, such as publishers and software providers, who sometimes are also partially integrated (Evans, 2008, 2009). Both markets are explained in more detail below.

2.1.1 Search Advertising

Search advertising refers to the search engines that attract users to their websites by displaying organic search results, but also allocate a portion of the result page for paid advertising results (Evans, 2008). Search-based advertising platforms thus directly interact with consumers on their result pages and sell the paid search places straight to advertisers. Furthermore, they have the required technology to match advertisers to keywords and hence consumers. Therefore, all potential individual agents are combined into one fully integrated platform (Evans, 2009). Search engines want to maximise their revenue from selling their advertisement slots. Therefore, they need to identify the advertisers who are interested in a specific keyword, their potential revenue, and the relevance of those advertisements. The former is done using a keyword bidding system, where advertisers place a bid for the cost-per-click of their advertisement (Evans, 2008). Most search-based advertising platforms use a second-price auction with reserve price. This implies that the cost-per-click that an advertiser needs to pay equals the second highest bid, while there is a reserve price in place (Evans, 2008; Varian et al., 2006; Varian, 2007). However, it is not necessarily the case that the advertisement corresponding to the highest bid also ends up as the first paid search result. Search-based advertising platforms make higher profits when they put advertisements with lower cost-per-click in higher slots if these advertisements generate more clicks than high cost-per-click advertisements. Therefore, the search-ad platforms compute the expected click-through rate[2], to estimate the expected revenue for all bids. The platforms want to assign the highest slots on the result pages to the bids with the highest expected revenue. However, they also take the relevancy of the advertisements into account. Some advertisements are good at attracting consumers. However, if these advertisements are not relevant, then con-

[2] The click-through rate is the number of clicks on an advertisement, divided by the number of users who have seen the advertisement

sumers are dissatisfied because the advertisement was not helpful, and advertisers are unhappy because they paid for advertisements that do not result in conversions. In the end, the slot allocation is thus determined by the keyword bidding process, the expected revenue, and the advertisement relevancy (Evans, 2008).

The search advertising market is characterised by a few large competitors in every country or language group (Varian et al., 2006). This is also plausible, given some of the market characteristics. Firstly, online advertising is a scale intensive industry. As the purchase probability of most advertisements is relatively low, a large audience is needed to sell products. Therefore, new search engines encounter high fixed costs to reach a large enough audience. However, once a search engine entered the market, the marginal costs for placing new advertisements are relatively low. The switching costs for the advertisers are relatively low as well, as the competing search engines are only one click away. Because of these market characteristics, an important factor for the search platforms is learning-by-doing. The strong competition to bind advertisers requires the search engines to continually invest in their technologies and pricing strategies (Varian et al., 2006).

2.1.2 Display Advertising

Display advertising platforms also supply advertising inventory to advertisers, however, they attract consumers to their website with content other than search results (Evans, 2008). As mentioned before, this business is a bit more complicated than the search-based advertising market, as it exists of several intermediary businesses which are generally categorised into three groups (Evans, 2008, 2009). Firstly, publishers attract consumers to their websites and sell the resulting advertising inventory directly or indirectly to advertisers. Secondly, software providers handle the passage of ads from advertisers to the publishers advertising inventory. Thirdly, advertising networks are intermediaries between advertisers and publishers as they aggregate both supply and demand of advertising inventory. Some large publishers have integrated all these categories, but many publishers use third-party software and network providers (Evans, 2008).

In the display advertising market, the advertising inventory is sold via guaranteed and non-guaranteed selling (Choi, Mela, Balseiro & Leary, 2020). In the guaranteed selling channel, an advertiser and publisher negotiate a fixed price that depends on when, where and how the advertisement will be displayed. The price is often expressed as cost-per-mille, so the price per thousand impressions. As these contracts are drawn before the advertisement is displayed, both advertisers and publishers rely on expectations of the number of impressions and the user characteristics in the negotiation of the contract (Choi et al., 2020).

The non-guaranteed selling channel, also referred to as real-time bidding, involves the buying and selling of single advertisement impressions in real time, immediately after a user arrived on a website. Often, a second price auction is used in these settings (Wang, Zhang, Yuan et al., 2017). The ad inventory is not guaranteed for advertisers, as as only the advertisement of the winning bid is shown. However, the advantage of real-time bidding for advertisers is that they can buy impressions based on the user's past behavioural pattern (Choi et al., 2020).

For the two social media channels in our data, Facebook and Instagram, the sale of advertisement inventory is more similar to the auctions in search advertising. Meta Platforms, the company behind the two social media platforms, integrates the roles of publisher, advertising network and software provider to a large extent. Hence, the company has access to a lot of data which it uses to sell their advertisement inventory. Because of this, they use a bidding system similarly to the one used by large search-based advertising platforms, where they also take ad relevancy into account to align the advertisements with the user's interests (Choi et al., 2020).

2.2 Strategic Publisher Interactions

In the attribution modelling literature, it is often assumed that the various advertisement channels cooperatively try to make a customer path end with a conversion (Anderl et al., 2016; Singal et al., 2019; Y. Zhang et al., 2014). However, there are many different players in the industry who all try to maximise their own profit. Therefore, a potential strategic market environment in which several publishers compete with each other can also be considered (Abhishek et al., 2017; Berman, 2018). Depending on the type of payment and attribution schemes applied, economic issues such as adverse selection and free-riding may arise which affect the advertiser's profit. Now, we discuss some previous results related to how these problems arise and how different types of attribution and payment schemes are applied to solve them.

Berman (2018) considers an analytical model with an advertiser and two publishers to examine the impact of different payment schemes and attribution approaches on the decision of publishers to show advertisements and the profits of advertisers. The first scheme is the effort-based cost-per-mille contract (CPM), implying that publishers receive a payment for the number of advertisements they show. The second type of contract is performance based and called Cost Per Action (CPA), as publishers are compensated based on the observed output of the campaign. Without attribution, this latter scheme causes publishers to free-ride on the effort made by others and creates an opportunity for moral hazard. Because the individual effectiveness of each publisher cannot be determined, some publishers do not make any effort to publish advertisements to the right consumers, but they still benefit from the efforts made by the other publishers (Holmstrom, 1982). Because both publishers want the other to exert effort, but all parties are better off if they both put in more effort, this situation results in a prisoner's dilemma (Abhishek et al., 2017). As this free-riding behaviour does not happen under CPM schemes, these contracts result in higher advertiser profits than CPA schemes in scenarios without baseline conversion rates and attribution metrics (Berman, 2018).

However, CPM is not always preferred by the publishers. When publishers have market power, they are in the position to determine the payment scheme. If the publishers are extreme substitutes, they prefer a CPA scheme, as the possibility for free-riding is high. For example, if customers are extremely prone to advertising and a single advertisement is enough to influence them to convert, publishers showing any advertisement after the first one, receive an almost free commission. This also explains why search engines use cost-per-click prices. Search engines often arrive later in the customer path and hence they can partially free-ride on the impact of previous publisher advertisements.

Attribution models help to reduce the free-riding of publishers. Attribution creates a contest between the publishers to receive the credit for conversion, which forces the agents to increase their efforts. However, improvements in advertiser profits are not guaranteed when CPA with attribution is used, but it depends on the circumstances considered in the model. If there is no baseline conversion rate and the noise[3] is large enough, last-touch attribution improves the advertiser's profit compared to both the CPA and CPM schemes without attribution. However, CPA combined with the Shapley attribution scheme only improves the CPM profits if the publishers are strategic complements, and only improves the CPA with last-touch attribution for low levels of noise (Berman, 2018).

The above results from Berman (2018) hold in a one-dimensional setting. Abhishek et al. (2017) considers a two-stage model, where one publisher creates awareness and the other drives consumers towards conversion. In this model, they define f-contracts that linearly split the conversion credit between the two publishers, such that the first publisher receives $f \in [0, 1]$ and the other $1 - f$. This general contract definition contains many attribution models, such as linear-attribution ($f = 0$) and the Shapley value ($f = \frac{1}{2}$). It is shown that any f-contract suffers from free-riding in this model. To solve this issue, reinforcement contracts are introduced. These contracts not only compensate publishers for their effort, but they also penalize them for the effort made by the other publisher. A unique feature of this scheme is that it is able to assign negative credits to publishers who do not appear in a conversion path. In this way, publishers are incentivized to put in more effort, which solves the free-riding problem and results in higher profits for the advertiser.

Going back to the one-dimensional setting by, advertisers cannot distinguish whether a conversion was caused by advertising effects or customer characteristics like brand preference, when a baseline conversion rate is added to the model. Because publishers have more information about the customers reaching their sites, this private information causes adverse selection as publishers can target customers with higher baseline conversion rates to receive those credits (Abhishek et al., 2017; Berman, 2018). For example, when display publishers know which customers are most likely to convert, they flood them with advertisements (Abhishek et al., 2012). Under this information asymmetry, using CPA combined with last-touch attribution becomes inefficient, as many advertisements are shown to consumers with high baseline conversion rates. Applying the Shapley attribution scheme is then more profitable for a wide range of noise, as the Shapley value explicitly accounts for the baseline effect.

These results emphasize the importance of choosing an appropriate attribution model for advertisers, as it influences the advertiser's profit in multiple ways. An attribution metric is not only a tool to measure the efficiency of advertising channels, which determines the allocation of the advertising budget. Additionally, the chosen approach also influences the ability to create efficient campaigns and the resulting profits, as it affects the publishers' behaviour (Berman, 2018; Jordan, Mahdian, Vassilvitskii & Vee, 2011).

2.3 Previous Work

In the existing literature, Markov models are often used to model customer paths. To model the sequential nature of customer paths and in this way obtain insights on the interplay of advertise-

[3] Noise is defined as the uncertainty that publishers have about consumer's behaviour in terms of future visits and purchases (Berman, 2018).

ment channels, Anderl et al. (2016) formulate an attribution framework in which the customer paths are represented by first- and higher-order Markov walks. This approach is based on a data mining framework for aggregating user-level data, in which the customer paths are represented by random Markov walks in directed graphs. The results show that the higher-order Markov models outperform the first-order model in predictive performance and also provide a more detailed understanding of the interactions between channels (Anderl et al., 2016; Kakalejčík, Bucko, Resende & Ferencova, 2018).

Furthermore, Markov attribution model are also considered, because Markov chains enable the customer journey to be modeled in line with the widely used concept of a conversion funnel (Singal et al., 2019). In a conversion funnel, customers move from the state of being unaware of the product to eventually becoming interested and buying the product (Kotler & Armstrong, 2010). By setting the state space of the Markov chain in accordance with the stages of the funnel, state-specific attribution scores can be captured. This allows advertisement channels to have different effects dependent on the customer's level of product interest. Abhishek et al. (2012) show that advertisement channels indeed affect customers differently, based on their stage in the conversion funnel. For example, display advertisements have the most impact early on in the decision process. These results are obtained using a Hidden Markov Model, where the states of the conversion funnel are not directly observed in the data, but they can by inferred through observable customer actions, such as website visits.

The AdditiveHazard model is the first survival-based attribution model (Y. Zhang et al., 2014). Based on this model, Ji and Wang (2017) proposed the Additional Multi-Touch Attribution model, which also uses the additive hazard of the advertisements, but additionally takes the intrinsic conversion rate of the consumers and some contextual features, such as consumer preference, into account. These two aspects are also incorporated into Probabilistic Multi-Touch Attribution model, which uses the Weibull distribution to describe the time between the display of an advertisement and the eventual conversion and the hazard rate to compute the attribution scores (Ji, Wang & Zhang, 2016). However, the model does not directly measure the combined effect of the advertisements, as the conversion probability is formulated as the probability that at least one of the displayed advertisements has successfully influenced the consumer to convert.

Broadening our view to all online advertising-related literature, there are even more applications of survival-based models. Manchanda, Dubé, Goh and Chintagunta (2006) model the conversion time as a function of advertising exposure with a survival model, that controls for duration dependence with a piecewise exponential hazard function in which the advertising covariates have a proportional hazard formulation. Bolton (1998) also implement a proportional hazards model, but they use left-truncated data to model the length of the relationship between a customer and a continuous service provider.

Lastly, survival models are also useful for estimating the winning probability in a real-time bidding second price auction (Wang et al., 2017). As this is a sealed-bid auction, advertisers are only notified if they won the auction and they are otherwise not informed about the winning price. This implies that if the advertiser loses, it only knows that the winning bid is higher than its own bid, but the actual price is unobserved. When estimating the winning probability, not taking into account the lost auctions results in a biased estimate that overestimates the probability. If we regard the advertiser's bid price as the observation period, and the winning bid as the event of interest, a survival model can be applied, which takes into account both the won and lost auctions. This is done by Amin, Kearns, Key and Schwaighofer (2012) and W. Zhang, Zhou, Wang and Xu (2016), who use the Kaplan-Meier Product Limit estimator, a non-parametric maximum likelihood estimator of the data (Dabrowska, 1987), to estimate the winning price distribution. This approach leads to lower winning probabilities than if the lost auctions are disregarded (Wang et al., 2017).

3 Data

The data were obtained from a publicly available digital marketing dataset for an unspecified product[4] (Huyton, 2021). The dataset contains 586737 touchpoints, belonging to 240108 cookie-traced paths, and was collected from 1 until 31 July 2018. Each observation describes the visit of a customer to the product's website and includes information on the cookie-ID of the customer, timestamp, type of interaction (impression or conversion[5]), conversion value, and the advertisement channel through which the website was visited. The dataset considers five advertisement channels, namely Instagram, Facebook, Online Display, Online Video, and Paid Search. These advertisement channels can be divided into two types: 1) customer-initiated channels, where the marketing interaction is started by a (prospective) customer, 2) firm-initiated channels, which are set up by the company (Bhatnagar, De, Sen & Sinha, 2022; Bowman & Narayandas, 2001). From the five advertisement channels used in this research, all of them are classified as firm-initiated marketing channels, except for the Paid Search channel, which is initiated by the consumer who starts the interaction by using a certain search word. With these data, we are able to construct the customer journeys that describe the chronological click pattern across various online advertisement channels for each customer and the corresponding purchase behaviour. The dataset thus not only includes customer paths resulting in a conversion, but also paths that do not end with a purchase.

As mentioned before, the data was collected using cookies to identify consumer journeys. Therefore, a touchpoint is only included in the dataset if the consumer clicked on the advertisement and visited the company's website. This implies that only information about the consumers' actions is available, and not about the actions taken by the company. This can be considered as a limitation of the dataset, as we can only investigate the attribution of advertisements which caused a consumer action, a visit to the website, while other advertisements shown to the consumers, that did not result in a web visit, cannot be taken into account. Similarly, the dataset only contains conversions that directly result from an advertisement interaction: the customer clicks on an advertisement and immediately purchases a product. Other customer paths, where the conversion happens independently of an advertisement touchpoint, are not considered.

Besides this, there are some other limitations related to tying a cookie to a specific consumer (Flosi, FuLGoNi & VoLLMAN, 2013; Abraham, Meierhoefer & Lipsman, 2007; Kannan et al., 2016). On one hand, using cookies might overestimate the number of unique customers that

[4] We will refer to the corresponding producing firm as 'the company'.

[5] A conversion refers to a purchase transaction.

visited the website due to cookie deletion. If a cookie is deleted and this consumer visits the company's website again, then the server will count this visit as a completely new customer. Furthermore, cookie data does not take multi-device usage into account. If one customer visits the company's website from two different devices, then this is recorded as two separate customer paths, while they belong to one customer. On the other hand, cookies do not correct for device-sharing. If two customers visit the website on the same laptop, then this is recorded as one customer journey, while it belongs to two individual customers.

The dataset was cleaned in several ways to make the data more suitable for the models under consideration. Firstly, all paths that consisted of only one advertisement were removed, similar to Ji and Wang (2017) and Kakalejčík et al. (2018). These paths are removed under the assumption that consumers who only viewed one advertisement were already familiar with the product, and therefore the contribution to the potential conversion of these consumers would be unclear. Secondly, we removed all users who viewed more than 31 advertisements in July, at the suspicion of potential bots, similar to Singal et al. (2019). Finally, we removed all paths with multiple touchpoints on the same day.

Initially, the time unit would be set to days. This implied that there could be at most one touchpoint per day, as having multiple events happen at the same time causes problems for survival models, as the likelihood depends on the rank order of events (Hertz-Picciotto & Rockhill, 1997). However, changing the time unit from seconds to days substantially changed the data[6] and therefore the original time unit of the data (seconds) was used. In the end, we decided to still omit these paths with multiple touchpoints on one day, because of time constraints. The parameter estimation of the AdditiveHazard model is time-intensive, as in each iteration, multiple calculations for each path are required. After some test runs, it turned out that including these paths increased the computational time substantially. As multiple estimation runs are performed for this research and the available time was limited, it was not time-feasible to include these paths in the data.

After these data cleaning steps, the dataset still consists of 108582 touchpoints corresponding to 42622 paths. Some summary statistics of the cleaned dataset can be found in Table 3. As shown in Table 1, the conversion rate in this dataset is 7.38%, and the average path length is 2.55 advertisements. The average conversion has a value of 6.24. As information about the price of the product and the quantities purchased per conversion are not available, we set the price of the product to be constant at a value of one during the observation period. Differences in the conversion value thus arise due to different quantities being bought per conversion. Finally, Table 3 shows that most website visits happen due to advertisements on Facebook, while Online Video advertisements result in the least click-throughs.

[6] For example, time differences between two touchpoints at different days of 6 and 23 hours would both be rounded to one day.

Table 1: Descriptive statistics of the cleaned dataset.

	Cleaned dataset
Number of paths	42622
Number of ad clicks	108582
Average path length	2.55
Conversion rate (%)	7.38
Average conversion value	6.24
Facebook touchpoints (%)	33.21
Instagram touchpoints (%)	14.19
Online Display touchpoints (%)	15.29
Online Video touchpoints (%)	12.95
Paid Search touchpoints (%)	24.36

4 Methodology

The two main goals are the comparison of two attribution models, namely the first-order Markov and the AdditiveHazard model, and the comparison of their attribution scores with other existing attribution metrics. Therefore in this Section, we first introduce the Markov model and the AdditiveHazard model. Then, the predictive performance measures, namely the area under the ROC, and the precision, recall and f1-scores are discussed. Finally, the six attribution metrics, including three heuristic, two Markov-based and one survival analysis metric, are introduced.

4.1 Markov model

A Markov chain is a stochastic process, a collection of random variables, that can show the dependencies between sequels of observations of these random variables (Anderl et al., 2016; Ross, 2014; Styan & Smith Jr, 1964). We consider a discrete time first-order Markov chain, which implies that for each point $n \in Z_+$ in a customer journey, there is random variable X_n representing the n$^{\text{th}}$ state of the customer during the customer journey. The state space of the Markov model is similar to the set of states proposed by Anderl et al. (2016). Each of the five advertisement channels, *Facebook*, *Instagram*, *Online Display*, *Online Video*, and *Paid Search*, represents a state. Additionally, there is one indirectly observed state, namely *Start*, which corresponds to the starting point of the customer journey. Each path in the dataset starts with the first channel with which the customer had an interaction, but this *Start* state is added to indicate the start of the path. Finally, there are also two absorbing states, *Conversion*, and *Null*, for customer journeys that ended and did not end with a conversion respectively. Therefore, we can define the full state space as:

$$S = \{Start, Facebook, Instagram, OnlineDisplay, OnlineVideo, PaidSearch, Conversion, Null\}.$$

A first-order Markov chain assumes that the conditional distribution of the future state only depends on the present state and is independent of the past states. This implies that the one-step transition probabilities can be formulated as follows:

$$P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, ..., X_1 = i_1, X_0 = i_0) = \tag{1}$$

$$P(X_{n+1} = j | X_n = i) = p_{ij} \qquad \forall i, j \in S \tag{2}$$

$$0 <= pij <= 1 \qquad \forall i, j \in S \tag{3}$$

$$\sum_{j=1}^{N} p_{ij} = 1 \qquad \forall i \in S \tag{4}$$

The one-step transition probability p_{ij} between state $i, j \in S\{Start, Null, Conversion\}$ represents the probability that a contact with channel i is followed by a contact with channel j. The self-loop probabilities p_{ii} for the advertisement channel states is also non-zero, because it is possible that within a path, a customer sequentially interacts multiple times with the same channel. The probabilities $p_{Start,j}, j \in S$ represent the probabilities that the first channel with which the consumer interacted in the consumer path is channel j. It is not possible to go directly from *Start* to either *Null* or *Conversion* and hence $p_{Start,Null} = p_{Start,Conversion} = 0$. Finally, $p_{ij}, i \in \{Facebook, Instagram, OnlineDisplay, OnlineVideo, PaidSearch\}, j \in \{Null, Conversion\}$ represents the probability that after interaction i, the customer journey ends or leads to a subsequent conversion. Because *Null* and *Conversion* are absorbing states, their self-loop probabilities are equal to one.

4.2 Survival Analysis

This section introduces some basic concepts of survival models which are used to build the AdditiveHazard attribution model. Survival analysis involves the study of the time until an event of interest occurs (Clark, Bradburn, Love & Altman, 2003). It was originally developed within the biomedical research field to measure for example time to death, time between a response to treatment and recurrence, or the time between an infection and disease onset (?, ?). However in recent years, these statistical techniques have been applied to a wide field of studies, such as engineering, criminology, sociology and economics (?, ?). Time-to-event data needs special statistical techniques, because the event of interest, also often called the failure, does not necessarily take place within the observation period for all the participants of the study (?, ?; Clark et al., 2003). In case the failure does not happen at all or happens outside the observation period, the true time-to-event for the individual is unknown.

This phenomenon is often called right censoring. There are multiple types of censoring of which two are relevant for this marketing setting. Firstly, administrative censoring refers to the situation where the observation period ends before the failure has occurred. It is possible that a customer purchases the product after 31 July 2018, while it saw the advertisements during the observation period. For this customer, the exact survival time is unknown, because we could not observe when the purchase occurred. Secondly, if an individual cannot be observed any longer, this type of censoring is called loss to follow-up. If customers clear their cookie history or stop using the devices with which they started their customer journey, then it is not possible to longer track their cookie path, and hence their customer journey. We thus no longer can observe these customer paths, and hence it remains unknown if these paths ended with a conversion or not.

Because it is not possible to observe the type of the censoring in the data, all types are treated similarly in the survival models .

Besides right censoring, where the event of interest is not observed, there is also left censoring, which refers to the situation where the begin point is not observed. In this case, the complete survival time is unknown, because we do not know when the survival period began. In our setting, this involves paths that already started, before the start of the observation period on 1 July 2018. However, we assume that all observed paths started within the observation period, and hence that the first observed interaction is also the first channel in the customer path. Hence, we assume that there is no left censoring, but only right censoring in our data set.

Finally, there is one more assumption that needs to hold for the construction of the likelihood function. The survival times of the customers need to be independent of each other. As mentioned in Section 3, multiple paths could belong to one unique customer, which would imply that the observed paths are not independent of each other. This situation violates the independence assumption of the AdditiveHazard model. To limit the potential influence of such paths, paths consisting of only one interaction were deleted. Therefore, we now assume that each path belongs to one individual customer and hence all paths are independent of each other.

Survival data is characterized by five key functions. First, let T denote a non-negative, continuous variable, representing the time until the event-of-interest, the conversion, occurs. The time until the conversion occurs is referred to as the survival time. The first function is the survival function, which is defined as:

$$S(t) = P(T > t) \tag{5}$$

This function is non-increasing in time and $S(0) = 1$ and $S(\infty) = 0$. The survival function describes the probability that the conversion happens after some time point t. Therefore, the survival function focuses on the non-occurrence of the conversion.

The cumulative distribution function (cdf) is closely related to the survival function and represents the probability that the conversion happens before some specified time t. Furthermore, if the cdf is differentiable, the probability density function can also be defined.

$$F(t) = P(T \leq t) = 1 - P(T > t) = 1 - S(t) \tag{6}$$

$$f(t) = \frac{\delta F(t)}{\delta t} = -\frac{\delta S(t)}{\delta t} \tag{7}$$

The fourth function is the hazard function, the instantaneous potential per unit of time for the event-of-interest to occur given that the event has not happened up to time t (Clark et al., 2003). The hazard function is related to the other functions introduced above. The proof for the relation is omitted here, but can be found in Y. Zhang et al. (2014). The formal definition, where Δt represents a change in time, and the relation with the other functions are described as:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{8}$$

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)} = -\frac{\delta}{\delta t} log(S(t)) \tag{9}$$

Finally, there is also a cumulative hazard function, which is specified as the integral of the hazard function over time. The cumulative hazard function is also related to the survival function:

$$\Lambda(t) = \int_0^t \lambda(s)\delta s = -log(S(t)) \tag{10}$$

4.3 AdditiveHazard Model

As the basic survival analysis concepts have been introduced, these functions are used to develop the AdditiveHazard model, introduced by (Y. Zhang et al., 2014). As this model requires different notation than the Markov model, we first introduce the notation. The customers are denoted by $u \in \{1, ..., U\}$, and the advertising channels as $\{1, ..., 5\}$[7]. The path of a consumer u is in the form of $\{\{(\alpha_i^u, t_i^u)\}_{i=1}^{l_u}, X_u, T_u\}$. Within this path notation, α_i^u is the i$^{\text{th}}$ advertisement channel shown to customer u, t_i^u is the corresponding time point, and l_u represents the total number of touchpoints in the path. Additionally, X_u is a binary variable indicating the conversion result, with $X_u = 1$ indicating a conversion. Finally, T_u indicates the last timestamp of the path. If $X_u = 1$, then T_u indicates the time of conversion. If $X_u = 0$, then T_u is the last timestamp of the observation window. Note that the observation window is not necessarily the same as the observation period. As will be mentioned later on, the outcomes of the AdditiveHazard model depend on the specified time window T, $0 \leq T \leq 31$, as the time window can maximally be the observation period of 31 days.

With this notation, the AdditiveHazard model can be introduced. As the name already suggests, the model is based on the idea that the hazard function is additive for the clicking of online advertisements. If a customer clicks on an advertisement, the kernel function corresponding to this advertisement is added to the customer's hazard function. Hence, every time the customer clicks on an advertisement, the hazard function receives a boost and is shifted upwards, as shown in Appendix B. Each channel has an exponential kernel function with two parameters, $\beta_{a_i^u}$ and $\omega_{a_i^u}$, that model the influences of the advertisement of the conversion. The strength of the impact of advertisement a_i^u is reflected by $\beta_{a_i^u}$ and the time-decaying property by $\omega_{a_i^u}$. The definition of this additive hazard function is as follows:

$$\lambda_u(t) = \begin{cases} \sum_{t_i^u \leq t} \beta_{\alpha_i^u} \omega_{\alpha_i^u} e^{(-\omega_{\alpha_i^u}(t - t_i^u))} & if\, t \leq t_{l_u} \\ 0 & otherwise \end{cases} \tag{11}$$

As each channel has its own strength and time-decaying parameters, the model incorporates the unique effect of each advertisement channel on the conversion probability. Additionally, the time-decaying parameter takes into account that as time passes since the advertisement was seen, the influence of the advertisement decreases. This is time feature of the AdditiveHazard model is not considered by the Markov model and hence in this way, the survival model covers an additional dimension of the conversion setting.

In order to compute the conversion probabilities, the model parameters first need to be estimated which can be done by maximising the likelihood. Let $\Theta = \{\boldsymbol{\beta}\, \boldsymbol{\omega}\}$, with $\boldsymbol{\beta} = \{\beta_i\}$ and

[7] Each channel corresponds to one number: 1 - Facebook, 2 - Instagram, 3 - Online Display, 4 - Online Video, 5 - Paid Search

$\boldsymbol{\omega} = \{\omega_i\}$ and i being one of the five advertisement channels. Then, the log-likelihood function of the AdditiveHazard model for all paths is:

$$\mathcal{L}(\Theta) = \sum_{u=1}^{N} \left[\sum_{X_u=1} log(\sum_{i} \beta_{\alpha_i^u} \omega_{\alpha_i^u} exp(-\omega_{\alpha_i^u}(T_u - t_i^u))) - \sum_{i} \beta_{\alpha_i^u}(1 - exp(-\omega_{\alpha_i^u}(T_u - t_i^u))) \right] \tag{12}$$

In this case, the maximisation problem becomes:

$$\begin{aligned} &\max_{\Theta} \mathcal{L}(\Theta) \qquad s.t. \\ &\beta_i \geq 0 \qquad i = 1, ..., 5 \\ &\omega_i \geq 0 \qquad i = 1, ..., 5 \end{aligned} \tag{13}$$

In (Y. Zhang et al., 2014), an MM (minorise maximise) algorithm is suggested to maximise the likelihood and estimate the parameters. The goal of the MM algorithm is to simplify the optimisation problem by solving a simpler form of the problem, and in this way optimise the original problem (Hunter & Lange, 2004). An MM algorithm can simplify the optimisation in several ways. For this optimisation problem, it allows for the separation of the parameters. This can be useful for online attribution settings, as in cases with many different marketing channels, the number of parameters to be estimated can become substantial. Y. Zhang et al. (2014) simplify the likelihood function by constructing a lower bound $Q(\Theta|Theta^{(t)})$ for current estimation $\Theta^{(t)}$:

$$Q(\Theta|\Theta^{(t)} = \sum_{X_u=1} \sum_{i} p_i^u log(\frac{\beta_{\alpha_i^u} \omega_{\alpha_i^u} e^{(-\omega_{\alpha_i^u}(T_u - t_i^u))}}{p_i^u}) - \sum_{i} \beta_{\alpha_i^u}(1 - e^{(-\omega_{\alpha_i^u}(T_u - t_i^u))}) \tag{14}$$

The lower bound contains the term p_i^u, which is defined below. The interpretation of p_i^u is convenient for the determination of the attribution of the advertising channels and this will be elaborated in Section 4.4.3.

$$p_i^u = \begin{cases} \frac{\beta_{{\alpha_i}^u} \omega_{{\alpha_i}^u} e^{(-\omega_{{\alpha_i}^u}(T_u - t_i^u))}}{\sum_{i=1}^{l_u} \beta_{{\alpha_i}^u} \omega_{{\alpha_i}^u} e^{(-\omega_{{\alpha_i}^u}(T_u - t_i^u))}} & if \qquad X_u = 1 \\ 0 & if \qquad X_u = 0 \end{cases} \tag{15}$$

For the relation between the lower bound $Q(\Theta|\Theta^{(t)})$ and the likelihood function, there are some important theoretical properties mentioned by Hunter and Lange (2004) that hold according to Y. Zhang et al. (2014). If we let $\Theta^{(t+1)} = \max_{\Theta} Q(\Theta|\Theta^{(t)})$, then these properties are:

$$\mathcal{L}(\Theta) \geq Q(\Theta|\Theta^{(t)}) \quad \forall \Theta \tag{16}$$

$$\mathcal{L}(\Theta^{(t)}) = Q(\Theta^{(t)}|\Theta^{(t)}) \tag{17}$$

$$\mathcal{L}(\Theta^{(t+1)}) \geq Q(\Theta^{(t+1)}|\Theta^{(t)}) \geq Q(\Theta^{(t)}|\Theta^{(t)}) = \mathcal{L}(\Theta^{(t)}) \tag{18}$$

This shows that, theoretically, the likelihood should increase monotonically as the number of iterations increases, and hence Y. Zhang et al. (2014) state that it can be shown that the likelihood thus converges to a local optimal. As mentioned before, the usage of the lower bound allows for independent optimisation of the parameters. By taking the derivative of Q with respect to β_k and ω_k, $k \in \{1, ..., 5\}$, independent update equations for all parameters are obtained. This implies that during each iteration, all parameters are updated individually and independently, using the following formulas:

$$\beta_k = \frac{\sum_{u,i,X_u=1,\alpha_i^u=k} p_i^u}{\sum_{u,i,\alpha_i^u=k} 1 - e^{(-\omega_k^{(t)}(T_u - t_i^u))}} \qquad \mathit{for} \quad k \in \{1,2,3,4,5\} \tag{19}$$

$$\omega_k = \frac{\sum_{u,i,X_u=1,\alpha_i^u=k} p_i^u}{\sum_{u,i,\alpha_i^u=k} p_i^u(T_u - t_i^u) + \beta_k^{(t)}(T_u - t_i^u)e^{(-\omega_k^{(t)}(T_u - t_i^u))}} \qquad \mathit{for} \quad k \in \{1,2,3,4,5\} \tag{20}$$

Hence, during each iteration, we update the parameters using Equation 4.3 and Equation 4.3 in such a way that the lower bound is optimised. By doing this iteratively, the parameters eventually converge and these optimal values of the parameters can be used for further computations, such as the conversion probability or computing the attribution scores.

4.4 Attribution metrics

The attribution problem is analysed using six approaches. These metrics include three heuristic approaches, namely Last-touch attribution (LTA), First-touch attribution (FTA), and Linear attribution (LA). Besides these, there are also three data-driven approaches. Of these metrics, the Removal effect and Shapley value are used together with the Markov model, whereas the AdditiveHazard model has its own metric for attribution. Each of these attribution metrics is explained below.

4.4.1 Heuristic Attribution Metrics

Heuristic attribution approaches, and especially the last-touch metric, are widely used in practice (Berman, 2018; Li & Kannan, 2014). Even though they are easy to use (Singal et al., 2019), there are quite some limitations related to these methods as they are relatively simple and therefore may not fit the reality well (Y. Zhang et al., 2014). Firstly, these rule-based approaches ignore the interactions among the different touchpoints within the purchase funnel (Kannan et al., 2016). For example, last-touch attribution is based on the assumption that the consumer stochastically decides whether or not to buy the product, every time she sees an advertisement. If this assumption is violated, which often happens in practice, the last-touch attribution scheme can be inefficient (Jordan et al., 2011). Secondly, the heuristics only take paths that result in conversion into account, and they disregard the paths that do not lead to conversions (Petersen et al., 2009, as described in Li and Kannan, 2014). This implies that these metrics cannot make statements about the contribution of advertisements to the conversion decision, but only about the role of advertisements within a conversion. This is also described by Singal et al. (2019), who label these heuristics as "backward looking" and point out that they only split the purchase value after conversion.

Despite these limitations, we compare the data-driven attribution approaches to these methods to determine if these data-driven approaches result in different and potentially better attribution schemes. Comparing the heuristics, especially the last-touch attribution, to data-driven attribution methods is also common practice in academia, and previous research already showed that data-driven methods can improve these attribution schemes (Abhishek et al., 2012; Berman, 2018; Ji & Wang, 2017; Li & Kannan, 2014; Singal et al., 2019; Xu et al., 2014; Y. Zhang et al., 2014).

The first heuristic metric we consider is the last-touch attribution, which allocates the conversion completely to the last channel visited in the consumer path (Singal et al., 2019). Similarly, the second metric is the first-touch attribution and, as the name already suggests, it allocates the conversion to the first advertisement channel in the path (Li & Kannan, 2014). Both of the formulas are shown here, using the notation introduced in Section 4.3:

$$\pi_k^{first} = \sum_u \mathbb{1}\{X_u == 1 \ \& \ \alpha_1^u = k\} \quad for\ k \ \in \{1,2,3,4,5\} \tag{21}$$

$$\pi_k^{last} = \sum_u \mathbb{1}\{X_u == 1 \ \& \ \alpha_{l_u}^u = k\} \quad for\ k \ \in \{1,2,3,4,5\} \tag{22}$$

In these formulas, α_1^u and $\alpha_{l_u}^u$ represent the first and last advertisements in each consumer path respectively. Finally, we consider the linear attribution, where each advertisement encountered in a consumer path ending with a conversion receives an equal share for a conversion:

$$\pi_k^{linear} = \sum_{u,i,X_u==1} \frac{\mathbb{1}\{\alpha_i^u == k\}}{l_u} \quad for\ k \ \in \{1,2,3,4,5\} \tag{23}$$

Equation 4.4.1 also uses the notation introduced in Section 4.3. Note that this linear attribution model does not count the number of unique channels shown in the path, but it takes into account how often each channel is shown. In this way, this measure counts the number of advertisements shown in the path, even if some of these advertisements are shown via the same channel. Hence, it adjusts the allocation score per path for each channel by the number of times an advertisement is shown during the path.

4.4.2 Removal Effect

To allocate the attribution in the Markov model, the removal effect is used, which is the most commonly used metric for attribution (Singal et al., 2019). The removal effect reflects the change in eventual conversion probability, when the advertisements of a specific channel are removed from the path (Anderl et al., 2016; Archak, Mirrokni & Muthukrishnan, 2010; Singal et al., 2019). The computation is as follows:

$$\pi_k^{removal} = Visit(k) * EventualConversion(k) \quad k \in \{1,2,3,4,5\} \tag{24}$$

For each channel, *Visit(k)* is defined as the probability of passing an advertisement of channel k on a random walk beginning in the *Start* state, which we defined in Section 4.1. Additionally, *EventualConversion(k)* reflects the probability of sometime reaching conversion from state k (Archak et al., 2010). The removal effect for each channel can take values between zero and the total conversion rate (Anderl et al., 2016).

The removal effect incorporates a bit of counterfactual analysis, as it compares the situation with a specific advertisement channel to the fictional paths where advertisements of this channel are not shown, which is a desired property of attribution metrics (Anderl et al., 2016; Singal et al., 2019). Because this measure is forward-looking, the removal effect only needs information on the change in conversion probability when an advertisement channel is removed, without requiring knowledge on the previous touchpoints or how the paths in reality end (Singal et al., 2019). Even though the removal effect is widely used and has some attractive features, it has a serious issue with the total allocated value. Because the metric considers both immediate and eventual conversions, it is possible that one conversion is accounted for multiple times. In this way, the total allocation of the removal effect can be higher than the actual generated value (Singal et al., 2019).

4.4.3 Shapley Value

Besides the removal effect, we also use the Shapley Value (SV) as an attribution measure for the Markov model. The Shapley value originated in game theory and is used to assign credit to individual players in a cooperative game (Shapley et al., 1953). In this way, SV allocates the average marginal contribution of each advertising channel as its contribution to a conversion (Berman, 2018; Dalessandro et al., 2012). Translating this concept to our attribution setting, each advertisement channel can be viewed as a player in the cooperative game with the mutual goal of a conversion. We assume that each conversion generates a value of 1[8], the total value generated by the online advertising is the number of conversions. The credit assigned to each player is then the attribution allocation for the channel. Let $P = \{Facebook, Instagram, OnlineDisplay, OnlineVideo, PaidSearch\}$ be the set of players (channels) and $\mathcal{X} \subseteq P$ be a coalition of players. Then, the characteristic function $v(\cdot)$ maps coalition to the value (number of conversions) generated by a coalition. For the empty coalition $v(\varnothing)$, the value is set to 0, and for the coalition consisting of all players, $v(P) = 3145$, which corresponds to the number of conversions in the data set. The Shapley value for channel k is then defined as:

$$\pi_k^{SV} = \sum_{\mathcal{X} \subseteq P\{k\}} \omega_{|\mathcal{X}|} \times \{v(\mathcal{X} \cup \{k\}) - v(\mathcal{X})\} \tag{25}$$

with

[8] This assumption is made as the data on the conversion value is unreliable because it is not clear how it was computed.

$$\omega_{|\mathcal{X}|} = \frac{|\mathcal{X}|!(|P| - |X| - 1)!}{|P|!} \tag{26}$$

The Shapley value is theoretically attractive, because it is the only solution to a cooperative game with the following four properties (Fatima, Wooldridge & Jennings, 2008; Singal et al., 2019):

1. **Efficiency**: The allocations for all channels sum up to the total value generated by all channels together: $\sum_{k \in P} \pi_r^{SV} = v(P)$
2. **Symmetry**: Two players who have identical contributions to a coalition, should also have the same Shapley values: if there are two players, $k, k' \in P$, and $\forall \mathcal{X} \subseteq P \; \{k, k'\}$ it holds that $v(\mathcal{X} \cup \{k\}) = v(\mathcal{X} \cup \{k'\})$, then $\pi_r^{SV} = \pi_{r'}^{SV}$
3. **Linearity**: Let there be two characteristic functions, $v(\cdot)$ and $w(\cdot)$. Then $\forall k \in P$ and $\alpha \in \mathbb{R}$, it holds that $\pi_k^{SV}(v + w) = \pi_k^{SV}(v) + \pi_k^{SV}(w)$ and $\pi_k^{SV}(\alpha v) = \alpha \pi_k^{SV}(v)$
4. **Null Player**: For any player $k \in P$ that does not add any value to any coalition, the allocation is 0: $\forall \mathcal{X} \subseteq P\{k\}, v(\mathcal{X} \cup \{k\}) = v(\mathcal{X})$, it holds that $\pi_k^{SV} = 0$

Even though SV has appealing theoretical properties, there are also some small drawbacks. Firstly, the Shapley Value does not have a counterfactual feature, like the removal effect, so it does not focus on the added value of an advertisement compared to the base scenario where no advertisement was shown. Secondly, exact estimation of the SV is often computationally intractable, and hence the Shapley Value often needs to be approximated (Fatima et al., 2008). Luckily, because the number of players is not very large, the Shapley Value can still be computed using Equation 4.4.3. However, it is not necessary to compute this relatively complicated formula. Because we apply SV to a Markov model, it is also possible to use the unique-uniform attribution metric, which is equivalent to SV in such settings (Singal et al., 2019). The unique uniform metric is defined as follows:

$$\pi_k^{uu} = \sum_{u, X_u == 1} \frac{1}{\mathcal{U}(u)} \quad for\ k\ \in \{1, 2, 3, 4, 5\} \tag{27}$$

Hence, we sum over all consumers u whose paths ended with a conversion, $X_u == 1$. If a path contains at least one advertisement of channel k, then the allocated value is increased by the fraction, where $\mathcal{U}(u)$ returns the number of unique advertisement channels encountered in the path of consumer u.

4.4.4 AdditiveHazard Attribution

The AdditiveHazard model has its own measure to compute the allocation scores. To compute the contribution of each advertisement shown in the path to conversion, we use Equation 4.3. Using these contributions, we compute the contribution per advertisement channel in the conversion of a customer u:

$$p_k^u = \begin{cases} \frac{\sum_{i,\alpha_i^u==k} \beta_{\alpha_i^u} \omega_{\alpha_i^u} e^{(-\omega_{\alpha_i^u}(T_u - t_i^u))}}{\sum_{i=1}^{l_u} \beta_{\alpha_i^u} \omega_{\alpha_i^u} e^{(-\omega_{\alpha_i^u}(T_u - t_i^u))}} & if \qquad X_u = 1 \\ 0 & if \qquad X_u = 0 \end{cases} \tag{28}$$

Finally, we can combine those contributions per consumer to compute the overall attribution scores based on the AdditiveHazard model:

$$\pi_k^{AH} = \sum_u p_k^u \tag{29}$$

4.5 Performance Measurement

In general it is difficult to measure the validity of attribution models. The true attribution scores are unknown, and hence it is not possible to compare the modeled scores to the actual attribution scores with an error measurement. However, besides providing attribution scores, multi-touch attribution models should also be able to predict conversions. Therefore, the performance of the Markov and AdditiveHazard model is evaluated based on their predictive performance using three measures; the area under the receiver operating characteristic (ROC) curve and the precision-recall and F1 graphs.

The first measure is the area under the ROC curve (AUC). Based on the comparison between the predicted and actual conversion results, the ROC curve visualizes the prediction performance of a model by plotting the false positive rate on the horizontal axis, against the true positive rate on the vertical axis (Fawcett, 2006). However, comparing models based on a single figure or value is often preferred (Bradley, 1997), and therefore we reduce the information from the curve into one scalar value, the area under the curve (AUC). This value is computed by, as the name already suggests, calculating the area under the ROC curve. An advantage of this measure, in comparison to some other approaches such as the precision-recall graph, is that the ROC curve, and hence the AUC value, is invariant with respect to skewed class distribution and unequal classification error costs (Bradley, 1997; Fawcett, 2006). This is especially useful in the attribution setting, as the data consist of a relative small number of conversions compared to the total number of paths and the error costs are unknown.

A plot of a ROC curve also often contains a 45-degree line, which represents randomly classification, which corresponds to an AUC value of 0.5. A ROC curve below this diagonal line, or an AUC value below 0.5, performs worse than random guessing. Similarly, a ROC curve above the diagonal line has an AUC value above 0.5 and gives better predictions than a random classification would make (Fawcett, 2006). Therefore, the higher the AUC value, the better the model performance is.

Secondly, we will also look at the area under the precision-recall and F1 curves. Precision refers to the percentage of observations that are predicted to be positive and that are actually positive. Recall refers to the percentage of of observations that are positive and are also classified as positive (Forman et al., 2003). Finally, F1 is the weighted average of the precision and recall scores:

$$P = \frac{\sum_u (X_u^* == X)}{\sum_u 1} \tag{30}$$

$$R = \frac{\sum_u (X_u^* = 1 \ \& \ \hat{X}_u = 1)}{\sum_u X_u^* = 1} \tag{31}$$

$$F1 = \frac{2PR}{P+R} \tag{32}$$

Both the Markov and AdditiveHazard model predict the conversion with a probability, and they do not already assign a label (conversion or non-conversion) to a path. Therefore, whether a path is classified as conversion or not, depends on the probability threshold, above which a path is predicted to end with a conversion. Precision-recall graphs and F1-graphs take different thresholds into account, and plot the values for the scores for these different threshold values. Therefore, using the graphs instead of the actual scores solves the threshold issue. However, as a single score is preferred, we use a similar approach as with the ROC curve and compute the area under the precision-recall and F1 curves. These values are then used to compare the predictive performances of the models.

Compared to the AUC value, the precision, recall and F1 scores are affected by skewed data (Fawcett, 2006; Jeni, Cohn & De La Torre, 2013). However, as both models deal with the same data set, this does not affect the model performance for this specific data set. Furthermore, these measures are widely used to evaluate classification models ((Fawcett, 2006; Y. Zhang et al., 2014)), and therefore we consider them as well.

To compute these measures, we randomly split the data in a test and train group. The train group consists of 80% of the data, which is used to compute the transition matrix in the Markov model and the parameter estimates in the AdditiveHazard model. The test group contains the other 20% of the data points, for which both models make conversion predictions.

5 Results

5.1 Parameter Estimation

Before the attribution scores can be computed for the AdditiveHazard model, first the model parameters, two for each channel, have to be estimated. This is done in an iterative process using Equation 14 an 15. As a convergence rule, the difference between all the parameter estimates and the previous estimates needs to be smaller than $\delta = 0.01$. This is quite a large value for such a convergence rule, but with $\delta = 0.001$, no convergence was reached. Furthermore, the estimation seems to be quite heavily dependent on the initial values, which will be elaborated upon in Section 5.2. Furthermore, one of the key features of the AdditiveHazard model is that it takes into account that a conversion may still happen after the observation window T that needs to be defined. Here, $T = 31$ days, which corresponds to the observation period. In Section 5.2, other observation windows are considered for the parameter estimation and attribution as robustness checks. Table 5.1 shows the parameter estimation results. These results are obtained after an estimation process which started with initial values $(50, 40, 60, 40, 50)$ for β

Table 2: Parameter estimates for the AdditiveHazard model.

Channel	β	ω
Facebook	5.4750	4.6477e-9
Instagram	4.2665	5.9286e-9
Online Display	7.4848	2.8109e-9
Online Video	1.9461	1.7070e-8
Paid Search	4.9932	3.1449e-9

and $(5e-9, 5e-9, 5e-9, 5e-9, 5e-9)$ for ω and converged after around 1680 iterations in 12 hours. The strength of advertisement channel k is shown by β_k and the time-decaying property is given by ω_k. A higher β_k thus indicates that the advertisement channel is stronger, resulting in a larger upwards shift of the hazard function at the time point that the advertisement is shown. Note that the strength of the advertisement does not reflect the attribution value, but it indicates how much the hazard function is shifted upwards, which is associated with an increased probability of conversion. On the contrary, a higher value for ω_k implies that the impact of the advertisement on this channel diminishes faster. When it comes to channel strength, Online Display has the highest estimate of 7.4848, indicating the display advertisements have the most impact. Similarly, Online Display also has the smallest time-decaying speed of $2.8109e-9$, indicating that the effect of display advertisements decreases the slowest. The opposite holds for Online Video, which has both the lowest strength, $\beta_{OnlineVideo} = 1.9461$, and the highest time-decaying speed estimate of $1.7070e-8$. This indicates that the impact of Online Videos is the smallest, and also that this impact decreases the fastest out of the five advertisement channels. Similarly, Instagram has the second lowest strength estimate,$\beta = 4.2665$, and the second highest time-decaying speed estimate of $5.9286e-9$. Advertisements on Instagram are thus relatively weak, and their impact also decreases relatively fast.

Looking at Facebook and Paid Search, we see that both score relatively well both for strength and impact decrease. Facebook has the second highest strength, $\beta_{Facebook} = 5.4750$, and has the third lowest time-decaying speed, $\omega_{Facebook} = 4.6477e-9$. Finally, Paid Search has estimates $\beta_{PaidSearch} = 4.9932$ and $\omega_{PaidSearch} = 3.1449e-9$, indicating that Paid Search has the third highest advertisement strength and the second lowest time-decaying speed. This shows that Paid Search still has a relatively large impact when the advertisement is shown and this effect decreases relatively slowly, compared to the decline rate of the other channels.

5.2 Attribution

Figure 1: Relative attribution scores for the five advertisement channels.

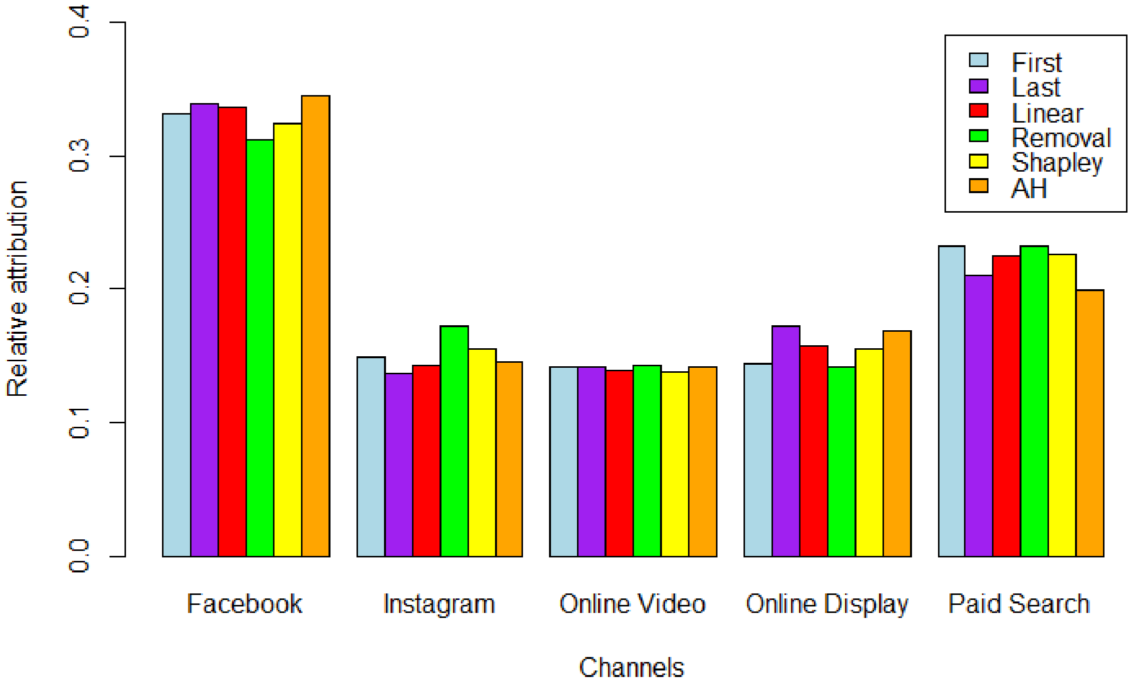


Figure 1 shows the percentage attributions of the six attribution metrics[9]. At first sight, the percentage attribution scores of the six methods seem to be quite similar. For all methods, Facebook receives the highest attribution score, which is a bit more than 30%. Similarly, all models allocate percentage attribution scores of around 20% to Paid Search, indicating that this channel receives the second most credit for conversions. The other three advertisement channels receive scores around the 15% and the exact order of these channels differs between the methods. For example, the Removal effect allocates the highest percentage to Instagram, followed by Online Video and Online Display respectively, while for the last touch metric, the percentages are ranked in the opposite way.

Even though the overall trends in allocations between the methods are quite similar, there are still some small differences that can be observed. The first touch approach allocates the highest percentage to Paid Search relative to the other metrics. This indicates that a relatively high share of the paths that end with a conversion, start with an interaction with a Paid Search advertisement. This might be an indication of pre-knowledge about the product. Consumers know the product already, they look for the product, are not convinced to buy it yet, and after seeing a few more advertisements, they are convinced to buy the product. Secondly, there are some interesting results for the last touch approach. On one hand, the heuristic gives a relatively high percentage score to the Online Display channel, indicating that this channel is

[9] Appendix D includes the corresponding attribution plots per method

often the last interaction before a conversion. On the other hand, a relatively low allocation is given to Paid Search. Both these outcomes are not in line as previous results indicate that heuristic approaches often underestimate firm-initiated channels, such as Online Display, and overvalue customer-initiated interactions which for example happen via Paid Search (Bowman & Narayandas, 2001; Wiesel, Pauwels & Arts, 2011). However, the Last Touch allocation for Instagram is in line with the fact that firm-initiated channels are often underestimated, because the percentage is the lowest compared to the other methods.

Thirdly, the allocation based on the Removal effect has two remarkable outcomes as well. Firstly, the allocation for Facebook is relatively low, whereas the allocation for Instagram is relatively high. This differs from the results found by (Singal et al., 2019), who found that the Removal effect scales with action intensity. In our data, most interactions are with Facebook advertisements, and hence the fact that the Removal effect allocates less to the channel with the highest interaction intensity compared to the other methods does not align with those other results.

The Shapley and linear allocations are quite similar, except for the allocations to Facebook and Instagram. For Facebook, the linear allocation is higher than for the Shapley value, whereas for Instagram, the Shapley allocation is higher than the linear attribution score. As the Shapley value is computed with the unique uniform definition introduced by Singal et al. (2019), the difference between these scores indicate that for paths ending with conversions, there are often multiple Facebook advertisements in one path, whereas for Instagram, this repeated interaction is less in such paths. This is also in line with the presence of the channels in the data set, as most interactions are with advertisements on Facebook (33.21%), whereas only (14.19%) of the interactions is with Instagram.

5.3 Robustness Checks

Shao and Li (2011) mention that the stability of the estimations is particularly important for attribution models, because they determine the performance metric of an advertising channel. As the AdditiveHazard model involves parameter estimations, it is important to check the robustness of the parameter estimates as this can help in validating the results. Therefore, we performed robustness checks for the AdditiveHazard model in two ways.

First of all, within the model, an observation window T needs to be chosen. In the main analysis, $T = 31$, which corresponds to the full observation period. However, to check for the robustness of this observation window, we also perform the AdditiveHazard model with $T = 26, 21, 16 and 11$ to see if and potentially how this changes the attribution results. The results are shown in Figure 5.3. For the observation windows of 26, 21 and 16 days, the relative attribution scores for most channels do not change a lot. However, for Online Video, the relative attribution score decreases as the observation window decreases. Contrarily, the relative attribution for Paid Search seems to increase slightly as the time window shrinks. When the window is only 11 days, the relative attribution scores change considerably. The attribution score for Facebook and Online Display are larger than for the other windows, and the scores for Instagram and Paid Search remain more or less the same. For Online Video, the drop in attribution score is the largest compared to the other periods. Note that these results express the relative attribution scores.

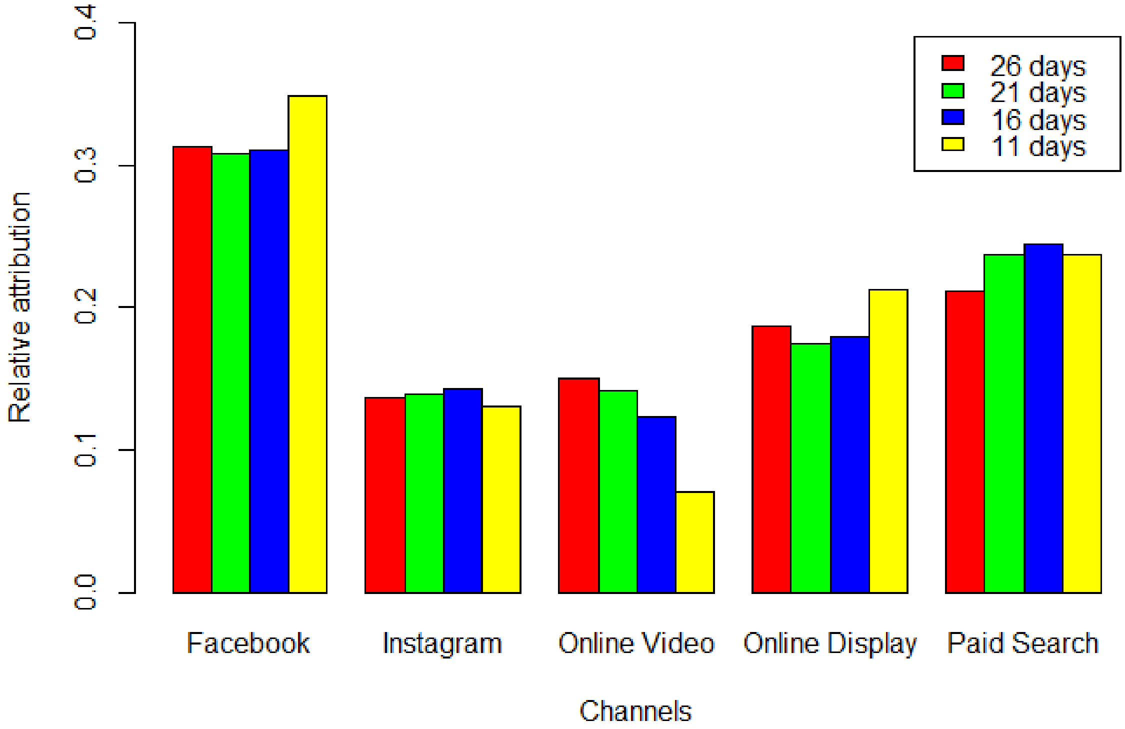


Figure 2: Relative attribution scores of the AdditiveHazard model for observation windows of 26, 21, 16 and 11 days.

The absolute attribution scores decreased as the time window shrank. As this period decreases, less conversions are observed and hence the total value to be allocated in the AdditiveHazard model is reduced as well.

Table 3: Parameter estimates of β and ω in the AdditiveHazard model for several initial values.

	$\beta : 0.5$ $\omega : 9e-10$		$\beta : 0.5$ $\omega : 5e-10$		$\beta : 0.5$ $\omega : 9e-09$		$\beta : 10$ $\omega : 9e-10$		$\beta : 10$ $\omega : 5e-10$		$\beta : 10$ $\omega : 9e-09$	
	β	ω	β	ω	β	ω	β	ω	β	ω	β	ω
Facebook	1.324	1.970e-08	1.635	1.586e-08	0.599	4.529e-08	6.278	4.054e-9	7.504	3.387e-09	2.209	1.167e-08
Instagram	1.433	1.797e-08	1.788	1.432e-08	0.625	4.283e-08	6.740	3.742e-09	8.170	3.083e-09	2.344	1.089e-08
Online Display	1.469	1.461e-08	1.857	1.150e-08	0.610	3.640e-08	6.848	3.075e-09	8.451	2.489e-09	2.344	9.076e-09
Online Video	0.938	3.641e-08	1.069	3.175e-08	0.540	6.601e-08	4.734	6.937e-09	5.103	6.430e-09	1.804	1.851e-08
Paid Search	1.120	1.333e-08	1.508	1.056e-08	0.509	3.247e-08	5.636	2.787e-09	6.901	2.274e-09	1.939	8.178e-09

Secondly, we checked for the robustness of the parameter estimates of β and ω. The results in Table 5.3 show that the estimates of the AdditiveHazard model depend strongly on the initial values. The β estimates for Facebook, for example, vary from 0.599 to 7.504 and from $4.054e-09$ to $4.529e-08$ for ω. Furthermore, for initial values $\beta = 0.5$ and $\omega = 9e-09$, the β estimates are considerably smaller than for the other starting values, except for Online Video. Even though the size of the estimates changes, the relative order remains almost the same. For example, for all initial values, except for the β estimate of 0.540 when $\beta = 0.5$ and $\omega = 9e-09$, Online Video has the lowest β estimate and the highest ω estimate. Similarly, the β estimate of Online Display is always the highest, apart from the value 0.610 for the initial values are $\beta = 0.5$ and

$\omega = 9e - 09$. Overall, the results in Table 5.3 clearly show that the parameter estimates vary as the initial value changes. Therefore, the parameter estimates are not robust, as only with specific initial values, the same estimates are obtained.

5.4 Predictive Performance

Even though it is not the main focus of attribution models, it is still important that these models can accurately make conversion predictions (Shao & Li, 2011). To evaluate the predictive performances of the Markov and AdditiveHazard models, the data set is randomly split into two groups. The training group contains 80% of the data and the 20% is in the test group. Based on the predictions for the test group, the areas under the ROC curve and precision, recall and F1 scores are calculated, as explained in Section 4.4.4. All results are computed in R with the *proc* package.

As shown by the ROC curves in Figure 5.4, the AdditiveHazard model seems to have a better predictive performance than the Markov model. The graph clearly shows that the AdditiveHazard curve lies further to the left of the 45-degree line than the Markov curve, indicating that the AdditiveHazard model provides better predictions, whereas the Markov model only performs slightly better than random predictions. Moreover, the Markov model has an area of 0.54 under the ROC, while the AdditiveHazard model has an area of 0.87 under the curve. This also shows that both methods perform better than random classification, which corresponds to a value of 0.50. However, the AUC value for the AdditiveHazard model is substantially larger than for the Markov model. Similar patterns can also be observed in Figure 5.4. These results imply that the AdditiveHazard model outperforms the Markov model when it comes to correctly predicting conversions, as the AdditiveHazard model less often classifies non-conversions as conversions and more often classifies conversions like that.

Similar results are obtained for the graphs shown in Figure 5.4. The left graph shows the Precision-Recall curves for both models, which are computed using the precision and recall scores for both models for several threshold values. The threshold values determine whether a path is labeled as a conversion or non-conversion. If the conversion probability is above the threshold, then it is considered as a conversion. The left graph in Figure 5.4 shows that the AdditiveHazard model has much higher precision scores. This indicates that among the paths predicted to convert, a higher percentage of these paths actually ends with a conversion for the AdditiveHazard model than for the Markov model. This is again an indication of the outperformance of the Markov model by the AdditiveHazard model. For the recall score, the Markov model also obtains relatively high values. However, those values on their own are not very informative for the analysis of the prediction performance. The recall score shows the percentage of observations that converts and is also predicted to convert. In the computation of the recall scores, a wide range of thresholds was considered. For relatively high threshold values, only the highest conversion probabilities are predicted as conversion. In this case, if these few observations are actually conversions, the recall score is high.

Finally, the right graph in Figure 5.4 shows the F1 curves. This graph also shows that the AdditiveHazard model outperforms the Markov model. For almost all threshold values, the AdditiveHazard model has a higher F1 score, which indicates a better predictive performance.

Because conversions are relatively sparse in the data, as the conversion rate is only 7.38%, it is quite challenging to accurately predict the conversions. However, the AdditiveHazard model achieves F1-scores above 0.5, which indicate a good prediction performance for this type of data (Y. Zhang et al., 2014). For the Markov model, the F1 score does not change much for different thresholds. This is the case, because the Markov conversion probabilities are very small. Hence for the larger thresholds, the number of predicted conversions does not change, because all probabilities are smaller. Still, even for the small thresholds, the F1-score of the Markov model is low indicating a poor predictive performance.

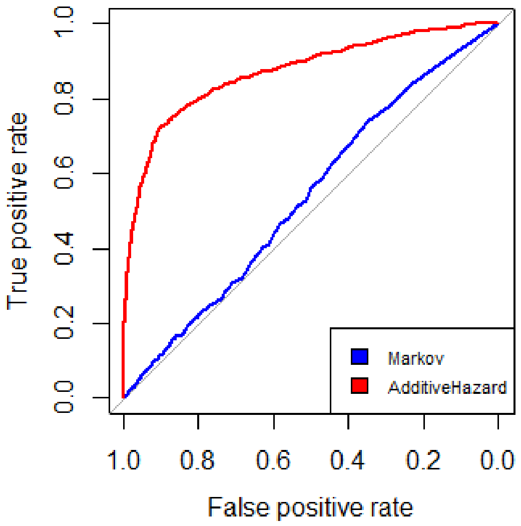


Figure 3: The ROC curves for the AdditiveHazard and Markov models.

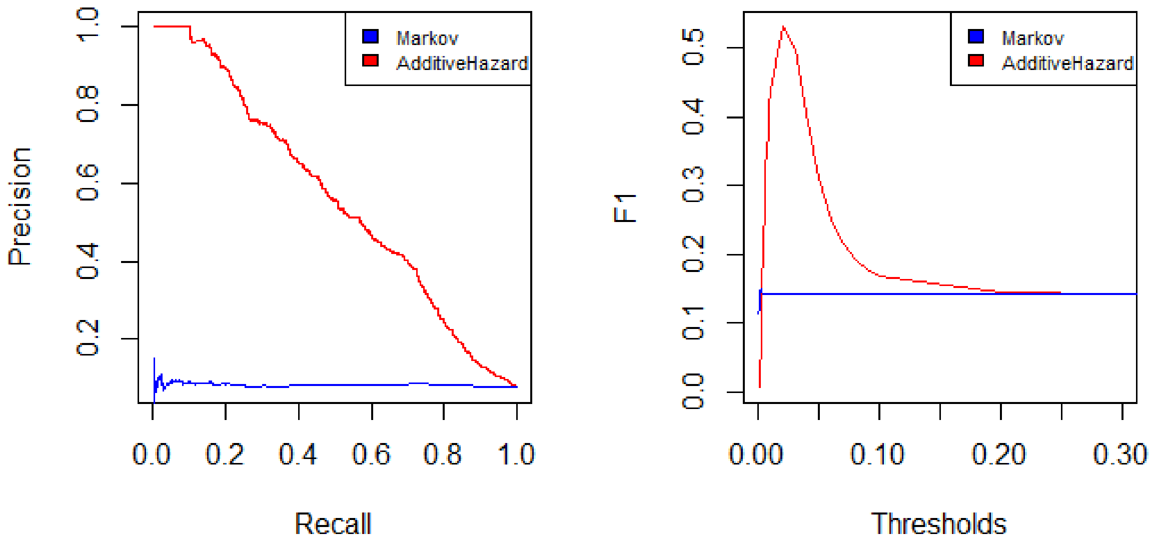


Figure 4: The Precision-Recall and F1 graphs for the AdditiveHazard and Markov models.

6 Conclusion

Over the past years, the usage of online advertising tools has heavily expanded (Interactive Advertising Bureau Europe, 2023). Before concluding an online purchase transaction, customers often encounter multiple advertisements (Li & Kannan, 2014). Because customers interact with multiple advertisements before they convert, it is challenging to determine the contribution of each advertisement channel towards a conversion (Anderl et al., 2016). In the past, companies used heuristic approaches to distribute the credits of conversions over the channels. However, these simple approaches turned out to sometimes inadequately reflect reality (Li & Kannan, 2014; Shao & Li, 2011; Y. Zhang et al., 2014). To improve these attribution approaches, several data-driven attribution models have been developed. One of these is the Markov model (Anderl et al., 2016; Dalessandro et al., 2012; Singal et al., 2019). This model incorporates the sequential nature of the customer journeys and the potential interactions between the channels. In addition, multiple attribution scores, such as the Removal effect and the Shapley value, can be computed with the Markov model. Another attribution model is the AdditiveHazard model (Y. Zhang et al., 2014), which incorporates some additional features compared to the Markov model. Firstly, besides the sequential nature, it also considers the timing of the interactions. Secondly, channel characteristics such as the impact strength and time-decaying speed of the impact are explicitly modeled. Finally, the model accounts for the potential right-censoring of the consumer paths. To investigate if these differences in model features also result in different model performances, this research aimed to answer the following research question: *how do the Markov and AdditiveHazard models compare with each other based on predictive performance and attribution allocation?*

The results showed that the six attribution metrics were quite similar. Facebook received the highest relative attribution scores for all methods followed by Paid Search. Instagram, Online Video, and Online Display all received scores of around 15%. The fact that the relative attribution scores for all approaches are quite similar, contradicts with previous findings of data-driven models outperforming the heuristic metrics (Abhishek et al., 2012; Berman, 2018; Ji & Wang, 2017; Li & Kannan, 2014; Singal et al., 2019; Xu et al., 2014; Y. Zhang et al., 2014). A potential explanation for this result is that the average path length is only 2.55 touch point, as shown in Table 3. The data-driven models cannot distinguish themselves from the heuristic approaches by incorporating the interactions between the channels, as each path does not contain much channels that could potentially interact. Comparing the Markov and AdditiveHazard models specifically, there are no substantial differences. However, the allocation of the Removal effect is relatively low for Facebook and high for Instagram, but this is due to the fact that these scores are scaled with action intensity (Singal et al., 2019). Therefore, we conclude that for this data set, the Markov and AdditiveHazard model have comparable attribution scores. Because the short average path length affected the results, it is not possible to generalise these results to other settings.

Moving on to the predictive performance, the results clearly indicated that the AdditiveHazard model outperforms the Markov model. Not only is the AUC value for the survival model much higher, 0.87 compared to 0.54 for the Markov model, the Prediction-Recall and F1-curves also show that the precision and F1 scores were better for the AdditiveHazard model. An explanation for these results is that the Markov model is not designed for prediction, while the AdditiveHazard model is. For the latter model, this conversion probability is computed based on the cumulative density function, whereas for the Markov model, no standard formula exists.

Finally, robustness checks were performed for the AdditiveHazard model. The attribution results were similar for the observation windows of 26, 21, and 16 days. However, when the window shrank to only 11 days, the attribution scores change. For Online Display, the attribution score decreases, whereas for Facebook and Online Display, they increase. The robustness check for the parameter estimates shows that the estimates depend strongly on the initial values. The reason for this is that the MM algorithm converges to a local optimum, of which the nearest value can change depending on the initial values of the algorithm (Y. Zhang et al., 2014). Based on these results, we conclude that the parameter estimates of the AdditiveHazard are not robust.

Overall, we can conclude that the attribution allocations of the Markov and AdditiveHazard model were similar for this empirical study. However, as these results are influenced by the short average path length, this conclusion cannot be generalised. In addition, the AdditiveHazard model has a better predictive performance than the Markov model. However, the parameter estimations of the AdditiveHazard model are not robust, as they depend strongly on the initial values.

7 Discussion

This research has both strengths and limitations. To begin with the former, the model investigates the AdditiveHazard model, which has not been investigated extensively in existing academic work. In this way, we elaborate on the potential of applying survival analysis methods to the attribution setting and broaden the knowledge over the different approaches that can be taken to solve the attribution problem. Furthermore, not only the attribution scores, but also the predictive performance of the models is taken into account. This provides a more general analysis of the attribution models.

However, this research also has several limitations. The main limitation of this research lies in the potential violation of the required assumptions for the AdditiveHazard model. We assumed that each path belongs to a unique customer, meaning that customers did not delete their cookies and clicked on another advertisement afterwards, and that they also did not use multiple devices. It is likely that this assumption was violated, as many consumers use multiple devices and internet users often clean their cookies (Abraham et al., 2007). Violation of this assumption implies that the paths are not independent of each other. Because the likelihood of the survival model depends on this assumption, this would affect the results of the AdditiveHazard model. Additionally, we assumed that the consumers did not click on an advertisement before the observation period started to prevent left-censoring. This assumption is potentially violated as well, as there is no reason why customers could not have seen the advertisement before 1 July. This violation also affects the survival Likelihood, as left censoring requires an adaption of this function. This also gives an interesting direction for a follow-up research. As it is likely that customers already interacted with advertisements before the start of an arbitrarily chosen observation period, incorporating the left censoring into the AdditiveHazard model could lead to even more realistic results.

Next, there are also some limitations related to the data. Firstly, the dataset only contains conversions that directly result from an advertisement interaction. This excludes many conversions that do not immediately follow upon an advertisement impression. Therefore, the results are not representative for conversions in general, as this feature affects the AdditiveHazard attribution scores. In this model, the final interaction receives a relatively high score, because the last advertisement is directly followed up by the conversion.

Moving on, both models assume that the different advertisement channels are cooperative, and together try to push the consumer in the direction of a conversion. However, as already addressed in Section 2.2, it is unclear if this assumption is appropriate (Singal et al., 2019). A violation of this assumption would affect the results, because if the advertisements are not cooperative, then the addition of the hazard kernels for the different channels becomes invalid, harming the validity of the results of the AdditiveHazard model.

Finally, the chosen optimisation procedure for the AdditiveHazard model is a limitation. The current procedure uses the update equations from the paper by Y. Zhang et al. (2014), which is based on an MM algorithm. This approach turned out to strongly dependent on the initial parameters, which harms the robustness of the estimates. Solving the AdditiveHazard optimisation problem with a different optimisation algorithm offers therefore a possibility for further research. One potential algorithm could be the Newton-Raphson algorithm, which generally requires fewer iterations than an MM algorithm. Additionally, it has a quadratic rate of convergence near the local optimum, compared to the linear convergence of the MM algorithm (Hunter & Lange, 2004).

References

Abhishek, V., Despotakis, S. & Ravi, R. (2017). Multi-channel attribution: The blind spot of online advertising. *Available at SSRN 2959778*.

Abhishek, V., Fader, P. & Hosanagar, K. (2012). Media exposure through the funnel: A model of multi-stage attribution. *Available at SSRN 2158421*.

Abraham, M., Meierhoefer, C. & Lipsman, A. (2007). The impact of cookie deletion on the accuracy of site-server and ad-server metrics: An empirical comscore study. *Retrieved October*, *14*, 2009.

Amin, K., Kearns, M., Key, P. & Schwaighofer, A. (2012). Budget optimization for sponsored search: Censored learning in mdps. *arXiv preprint arXiv:1210.4847*.

Anderl, E., Becker, I., Von Wangenheim, F. & Schumann, J. H. (2016). Mapping the customer journey: Lessons learned from graph-based online attribution modeling. *International Journal of Research in Marketing*, *33*(3), 457–474.

Anderson, S. P. & Gabszewicz, J. J. (2006). The media and advertising: a tale of two-sided markets. *Handbook of the Economics of Art and Culture*, *1*, 567–614.

Archak, N., Mirrokni, V. S. & Muthukrishnan, S. (2010). Mining advertiser-specific user behavior using adfactors. In *Proceedings of the 19th international conference on world wide web* (pp. 31–40).

Berman, R. (2018). Beyond the last touch: Attribution in online advertising. *Marketing Science*, *37*(5), 771–792.

Bhatnagar, A., De, P., Sen, A. & Sinha, A. P. (2022). Customer-initiated and firm-initiated online shopping visits under competition for attention: A conceptual model and empirical analysis. *Decision Support Systems*, *163*, 113844.

Bolton, R. N. (1998). A dynamic model of the duration of the customer's relationship with a continuous service provider: The role of satisfaction. *Marketing science*, *17*(1), 45–65.

Bowman, D. & Narayandas, D. (2001). Managing customer-initiated contacts with manufacturers: The impact on share of category requirements and word-of-mouth behavior. *Journal of marketing Research*, *38*(3), 281–297.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, *30*(7), 1145–1159.

Choi, H., Mela, C. F., Balseiro, S. R. & Leary, A. (2020). Online display advertising markets: A literature review and future directions. *Information Systems Research*, *31*(2), 556–575.

Clark, T. G., Bradburn, M. J., Love, S. B. & Altman, D. G. (2003). Survival analysis part i: basic concepts and first analyses. *British journal of cancer*, *89*(2), 232–238.

Dabrowska, D. M. (1987). Non-parametric regression with censored survival time data. *Scandinavian Journal of Statistics*, 181–197.

Dalessandro, B., Perlich, C., Stitelman, O. & Provost, F. (2012). Causally motivated attribution for online advertising. In *Proceedings of the sixth international workshop on data mining for online advertising and internet economy* (pp. 1–9).

Du, R., Zhong, Y., Nair, H., Cui, B. & Shou, R. (2019). Causally driven incremental multi touch attribution using a recurrent neural network. *arXiv preprint arXiv:1902.00215*.

Evans, D. S. (2008). The economics of the online advertising industry. *Review of network economics*, *7*(3).

Evans, D. S. (2009). The online advertising industry: Economics, evolution, and privacy. *Journal of economic perspectives*, *23*(3), 37–60.

Fatima, S. S., Wooldridge, M. & Jennings, N. R. (2008). A linear approximation method for the shapley value. *Artificial Intelligence*, *172*(14), 1673–1699.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, *27*(8), 861–874.

Flosi, S., FuLGoNi, G. & VoLLMAN, A. (2013). If an advertisement runs online and no one sees it, is it still an ad?: Empirical generalizations in digital advertising. *Journal of Advertising Research*, *53*(2), 192–199.

Forman, G. et al. (2003). An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.*, *3*(Mar), 1289–1305.

Hertz-Picciotto, I. & Rockhill, B. (1997). Validity and efficiency of approximation methods for tied survival times in cox regression. *Biometrics*, 1151–1156.

Holmstrom, B. (1982). Moral hazard in teams. *The Bell journal of economics*, 324–340.

Hunter, D. R. & Lange, K. (2004). A tutorial on mm algorithms. *The American Statistician*, *58*(1), 30–37.

Huyton, H. (2021). *Markov chain attribution data set.* `https://www.kaggle.com/code/hughhuyton/multitouch-attribution-modelling`.

Interactive Advertising Bureau Europe. (2023). *Iab europe adex benchmark 2022 study.* `https://iabeurope.eu/wp-content/uploads/2023/05/IAB-Europe`$_{A}dEx - Benchmark - 2022_{F}INAL - website.pdf$.

Jeni, L. A., Cohn, J. F. & De La Torre, F. (2013). Facing imbalanced data–recommendations for the use of performance metrics. In *2013 humaine association conference on affective computing and intelligent interaction* (pp. 245–251).

Ji, W. & Wang, X. (2017). Additional multi-touch attribution for online advertising. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 31).

Ji, W., Wang, X. & Zhang, D. (2016). A probabilistic multi-touch attribution model for online advertising. In *Proceedings of the 25th acm international on conference on information and knowledge management* (pp. 1373–1382).

Jordan, P., Mahdian, M., Vassilvitskii, S. & Vee, E. (2011). The multiple attribution problem in pay-per-conversion advertising. In *Algorithmic game theory: 4th international symposium, sagt 2011, amalfi, italy, october 17-19, 2011. proceedings 4* (pp. 31–43).

Kakalejčík, L., Bucko, J., Resende, P. & Ferencova, M. (2018). Multichannel marketing attribution using markov chains. *Journal of Applied Management and Investments*, *7*(1), 49–60.

Kannan, P., Reinartz, W. & Verhoef, P. C. (2016). *The path to purchase and attribution modeling: Introduction to special section* (Vol. 33) (No. 3). Elsevier.

Kireyev, P., Pauwels, K. & Gupta, S. (2016). Do display ads influence search? attribution and dynamics in online advertising. *International Journal of Research in Marketing*, *33*(3), 475–490.

Kotler, P. & Armstrong, G. M. (2010). *Principles of marketing.* Pearson Education India.

Li, H. & Kannan, P. (2014). Attributing conversions in a multichannel online marketing environment: An empirical model and a field experiment. *Journal of marketing research*, *51*(1), 40–56.

Manchanda, P., Dubé, J.-P., Goh, K. Y. & Chintagunta, P. K. (2006). The effect of banner advertising on internet purchasing. *Journal of Marketing Research*, *43*(1), 98–108.

Petersen, J. A., McAlister, L., Reibstein, D. J., Winer, R. S., Kumar, V. & Atkinson, G. (2009). Choosing the right metrics to maximize profitability and shareholder value. *Journal of Retailing*, *85*(1), 95–111.

Rochet, J.-C. & Tirole, J. (2004). *Defining two-sided markets* (Tech. Rep.). Citeseer.

Ross, S. M. (2014). *Introduction to probability models.* Academic press.

Shao, X. & Li, L. (2011). Data-driven multi-touch attribution models. In *Proceedings of the 17th acm sigkdd international conference on knowledge discovery and data mining* (pp. 258–264).

Shapley, L. S. et al. (1953). A value for n-person games.

Singal, R., Besbes, O., Desir, A., Goyal, V. & Iyengar, G. (2019). Shapley meets uniform: An axiomatic framework for attribution in online advertising. In *The world wide web conference* (pp. 1713–1723).

Styan, G. P. & Smith Jr, H. (1964). Markov chains applied to marketing. *Journal of Marketing Research*, *1*(1), 50–55.

Varian, H. R. (2007). Position auctions. *international Journal of industrial Organization*, *25*(6), 1163–1178.

Varian, H. R. et al. (2006). The economics of internet search. *Rivista di politica economica*, *96*(11/12), 8.

Wang, J., Zhang, W., Yuan, S. et al. (2017). Display advertising with real-time bidding (rtb) and behavioural targeting. *Foundations and Trends® in Information Retrieval*, *11*(4-5), 297–435.

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

Xu, L., Duan, J. A. & Whinston, A. (2014). Path to purchase: A mutually exciting point process model for online advertising and conversion. *Management Science*, *60*(6), 1392–1412.

Zhang, W., Zhou, T., Wang, J. & Xu, J. (2016). Bid-aware gradient descent for unbiased learning with censored data in display advertising. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 665–674).

Zhang, Y., Wei, Y. & Ren, J. (2014). Multi-touch attribution in online advertising with survival theory. In *2014 ieee international conference on data mining* (pp. 687–696).

A Proof of Conversion Prediction for AdditiveHazard Model

For the AdditiveHazard model, the probability of a conversion corresponds to the cumulative density function $F(t)$. In a survival model, it can be complicated to directly compute this probability and therefore we make use of the relation between $F(t)$ and the survival function $S(t)$:

$$F(t) = 1 - S(t) \tag{33}$$

Furthermore, the hazard function as defined in (Y. Zhang et al., 2014) and the relation between the hazard function and the survival function are needed, which are given below:

$$\lambda_u(t) = \begin{cases} \sum_{t_i^u \leq t} g_{\alpha_i^u}(t - t_i^u) & \text{if } t \leq t_{l_u} \\ 0 & otherwise \end{cases} \tag{34}$$

$$S(t) = exp(-\int_0^t \lambda(u)\delta u) \tag{35}$$

With these equations, we can write the formula for the conversion probability within time window T as:

$$\begin{aligned} F(T) &= 1 - S(T) \\ &= 1 - exp(-\int_0^T \lambda(t)\delta t) \\ &= 1 - exp(-\int_0^T \sum_{t_i^u \leq t} \beta_{\alpha_i^u}\omega_{\alpha_i^u} exp(-\omega_{\alpha_i^u}(t - t_i^u))dt) \\ &= 1 - exp(-\sum_i \int_0^T \beta_{\alpha_i^u}\omega_{\alpha_i^u} exp(-\omega_{\alpha_i^u}(t - t_i^u))dt) \\ &= 1 - exp(-\sum_i \beta_{\alpha_i^u}(1 - exp(-\omega_{\alpha_i^u}(T - t_i^u)))) \end{aligned} \tag{36}$$

B Examples of AdditiveHazard Hazard Function

Figure 5: Illustration the AdditiveHazard hazard function for a non-converted path.

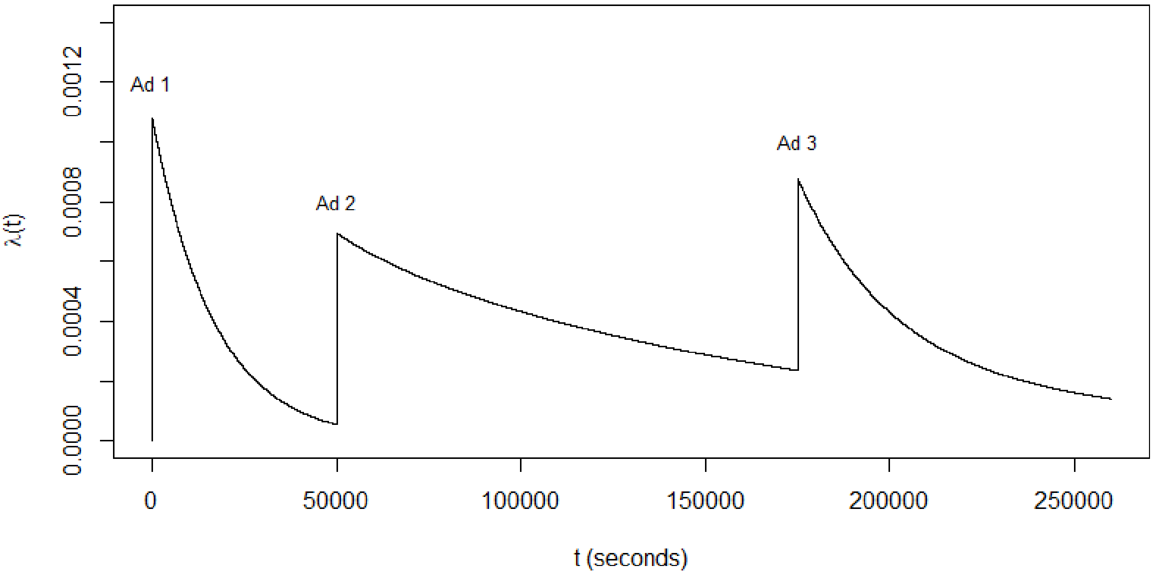
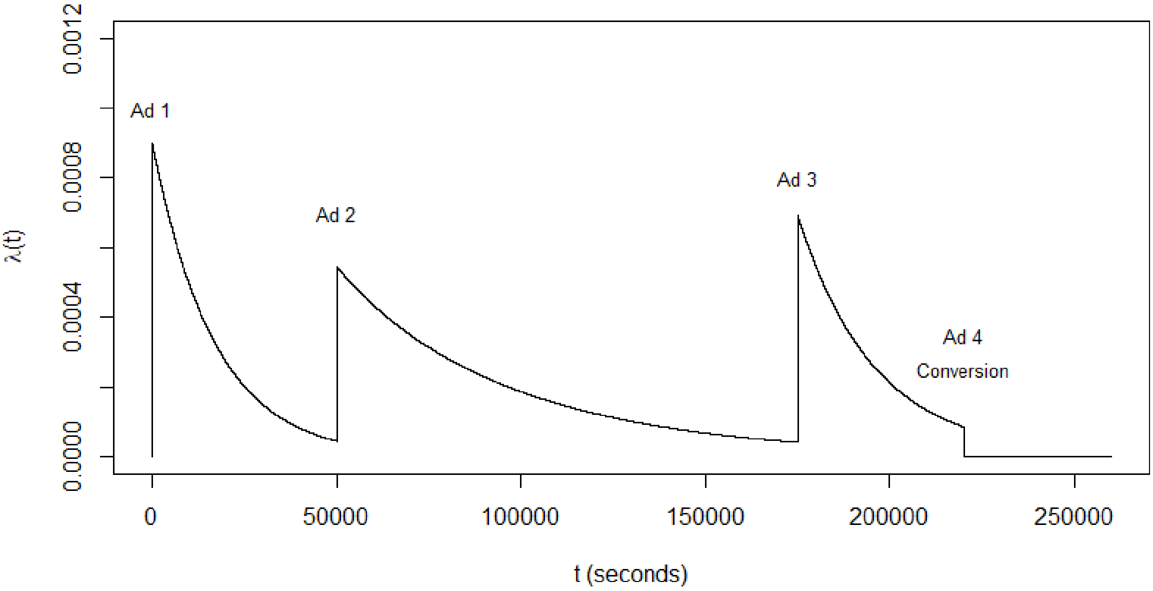


Figure 6: Illustration the AdditiveHazard hazard function for a converted path.



C Programming Code

Many different code files were used to obtain the results. The description of each code file together with the input and output files is provided below. Note that some code files cannot be performed without others being runned first, as some code requires output from other code. The code can be runned in the order that the documents are listed below.

1. Code - Data Cleaning
   description: cleans the data and computes some summary statistics
   input: t1.csv
   output: CleanedDataFrames.RData

2. Code - Linear, Last-touch and First-touch, Removal Effect Attribution
   description: Computes the linear, last-touch, first-touch and the Removal Effect metrics for the Markov model.
   input: CleanedDataFrames.RData
   output: No output file generated, but the results can be found in attributionresults.csv

3. Code - Changing data from R to Python
   description: Transforms the cleaned data into csv-files in accordance with the
   input requirements from the Python code
   input: CleanedDataFrames.RData
   output: FinalDataTest1.csv , FinalDataTest2.csv , Finaltest.csv , Finaltrain.csv

4. Code - Illustration of the AdditiveHazard hazard function
   description: creates illustrative graphs to show how the hazard function addition works in the AdditiveHazard model.
   input: no input file required
   output: no output file generated

5. Code FINAL Extension Seconds.ipynb
   description: computes the attribution score of the AdditiveHazard model
   input: FinalDataTest2.csv
   output: Beta.csv , Omega.csv (note that these are NOT the beta and omega files in the zip-file), attribution credits can be found in attributionresults.csv

6. Code FINAL Extension Seconds Initial Values RC.ipynb
   description: computes the attribution score of the AdditiveHazard model and is used for robustness checks of the initial values. The several initial values I used are provided in the code.
   input: FinalDataTest2.csv
   output: Beta.csv , Omega.csv (note that these are NOT the beta and omega files in the zip-file), and attribution results (but these are not stored in a file)

7. Code FINAL Extension Seconds Time Window RCipynb.ipynb
   description: computes the attribution score of the AdditiveHazard model for several

lengths of the time window. The time windows I used are 26,21,16,11 days.
input: FinalDataTest2
output: Beta.csv , Omega.csv (note that these are NOT the beta and omega files in the zip-file), and attribution results (but these are not stored in a file)

8. Code FINAL Extension Train.ipynb
description: computes the omega and beta parameter values based on the train data.
input: Finaltrain.csv
output: Beta.csv , Omega.csv

9. Code FINAL Extension Test.ipynb
description: computes the conversion prediction based on the AdditiveHazard model for the test data
input: Finaltest.csv , Beta.csv , Omega,csv
output: PredictionResult.csv

10. Code FINAL Linear, Last Touch, First Touch, Shapley.ipynb
description: computes the linear, last-touch, first-touch, unique-uniform and shapley attribution metrics for the Markov model input: FinalDataTest1.csv
output: attributionResult

11. Code - Creating graphs for the attribution scores
description: Creates several graphs that depict the attribution scores of the attribution metrics
input: attributionresults.csv
output: no output file generated

12. Code - Predictive performance of the Markov and AdditiveHazard models
description: computes several measures to evaluate the predictive performances of the Markov and AdditiveHazard models
input: CleanedDataFrames.RData , Finaltest.csv , PredictionResult.csv
output: no output file generated

D Attribution Scores Per Method

The relative attribution scores per attribution method are given in individual graphs.

Figure 7: Relative attribution scores for the five advertisement channels for the First touch attribution scheme.

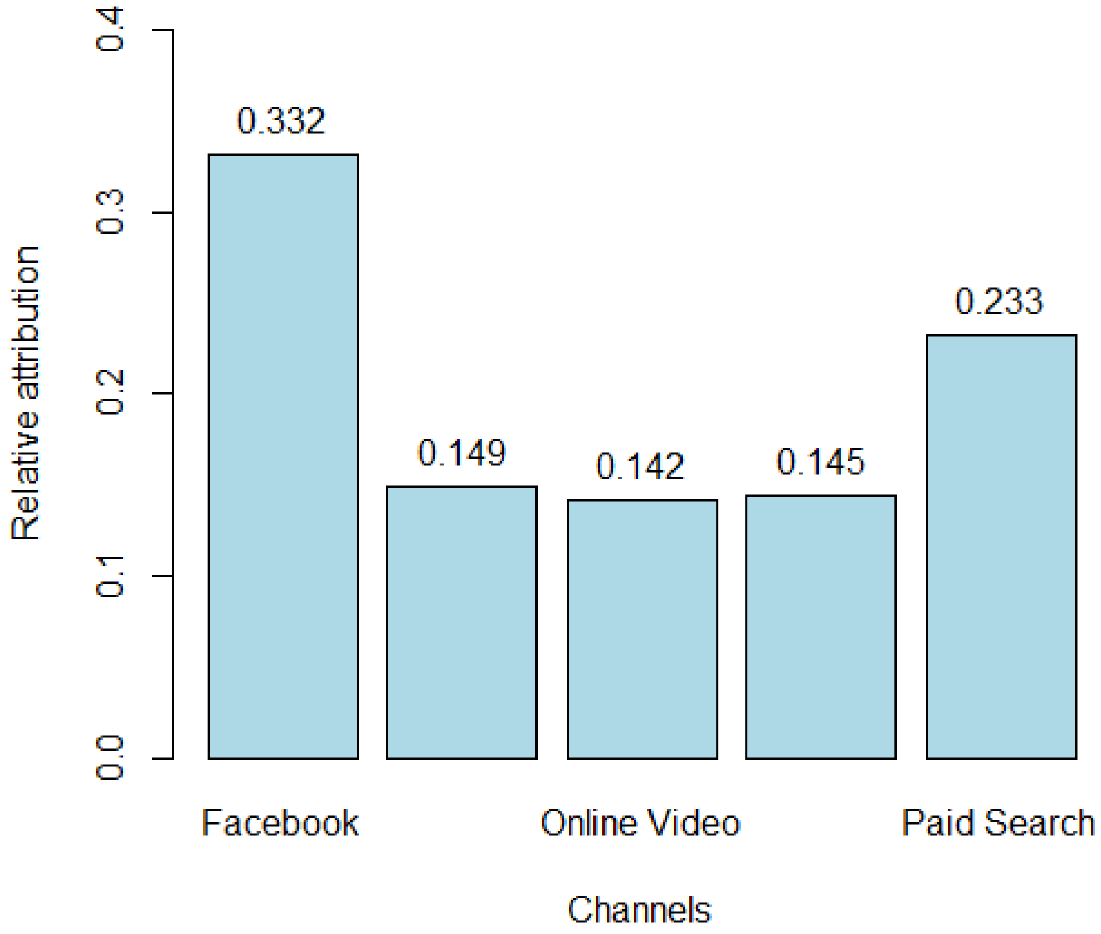


Figure 8: Relative attribution scores for the five advertisement channels for the Last touch attribution scheme.

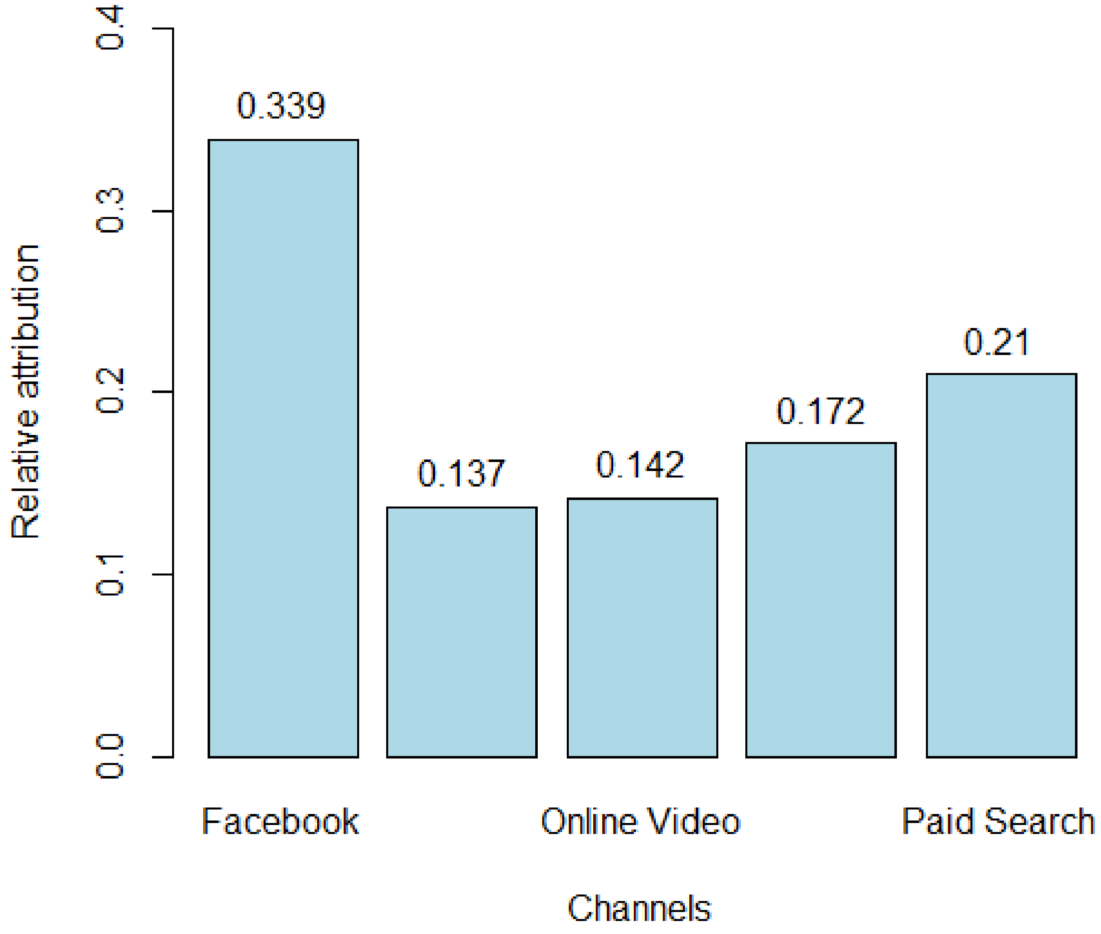


Figure 9: Relative attribution scores for the five advertisement channels for the Linear attribution scheme.

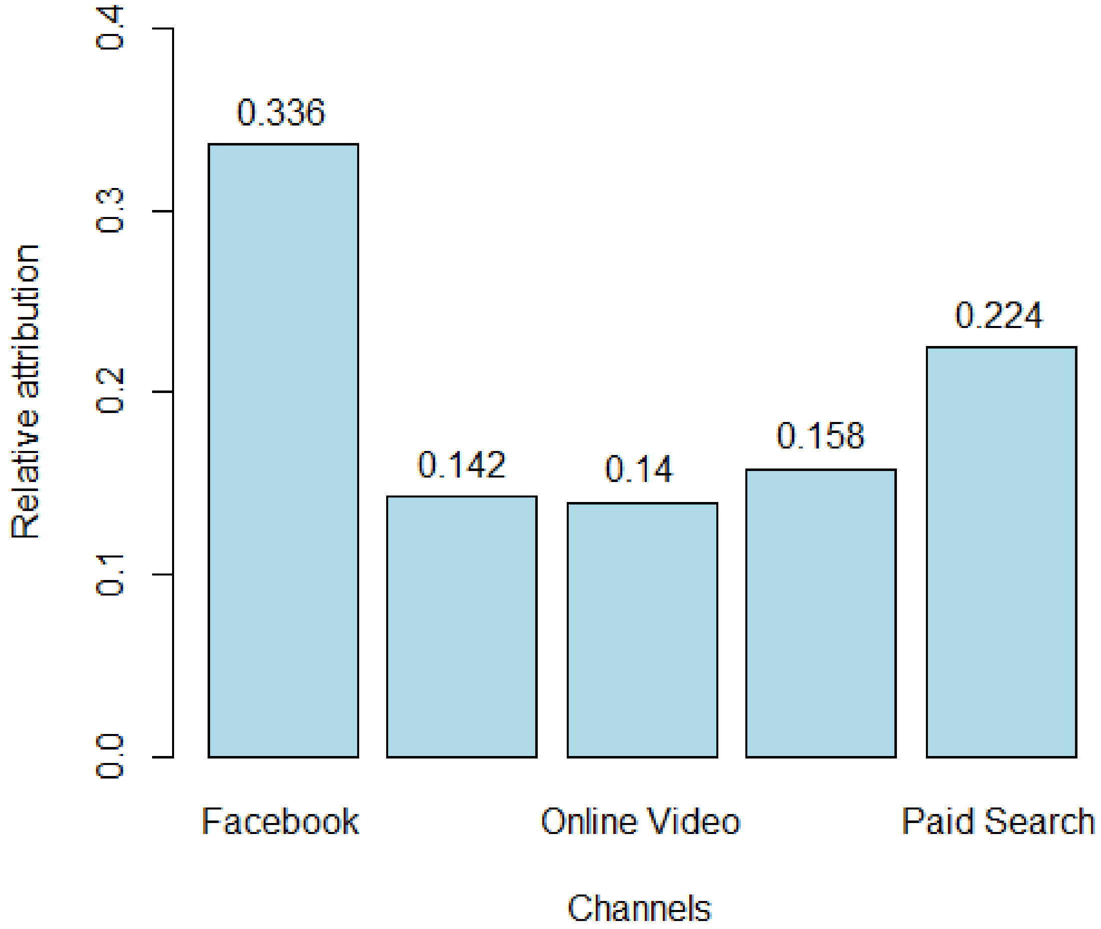


Figure 10: Relative attribution scores for the five advertisement channels for the Removal effect attribution scheme.

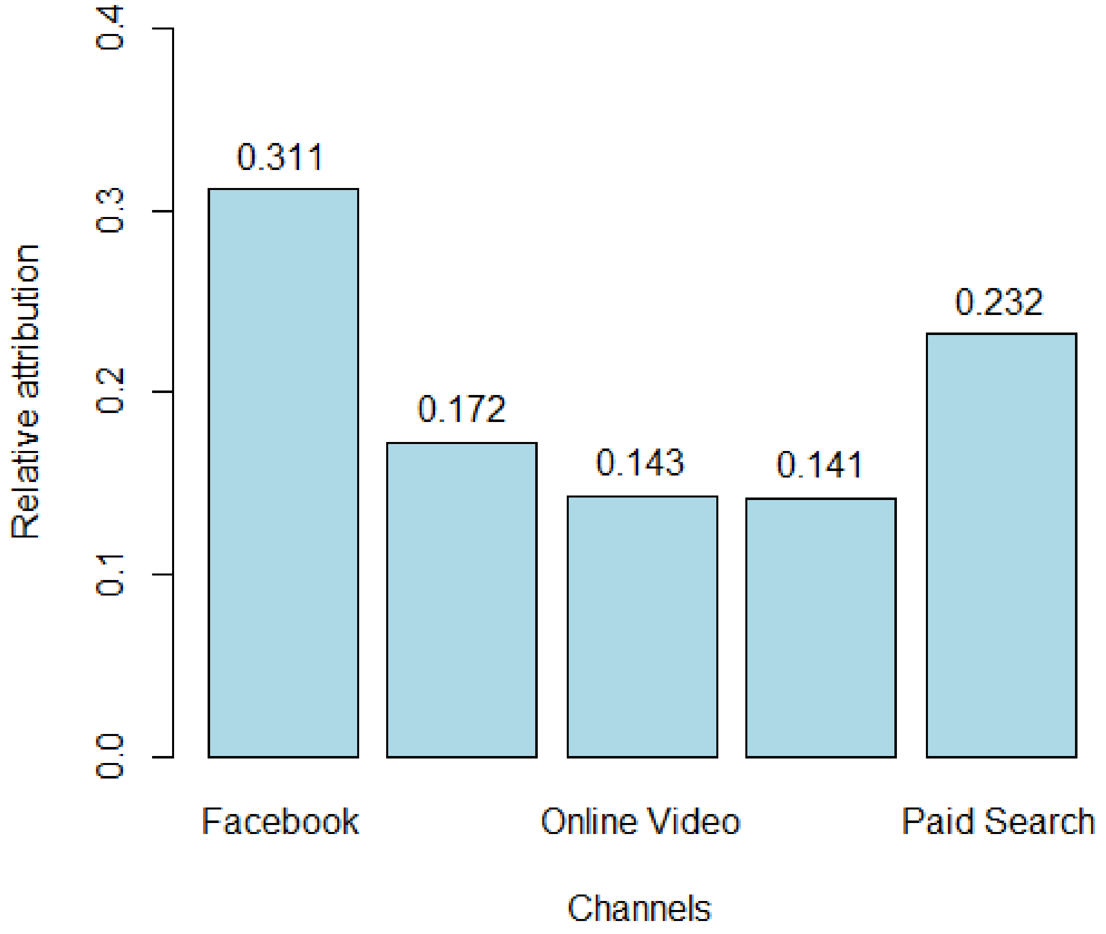


Figure 11: Relative attribution scores for the five advertisement channels for the Shapley attribution scheme.

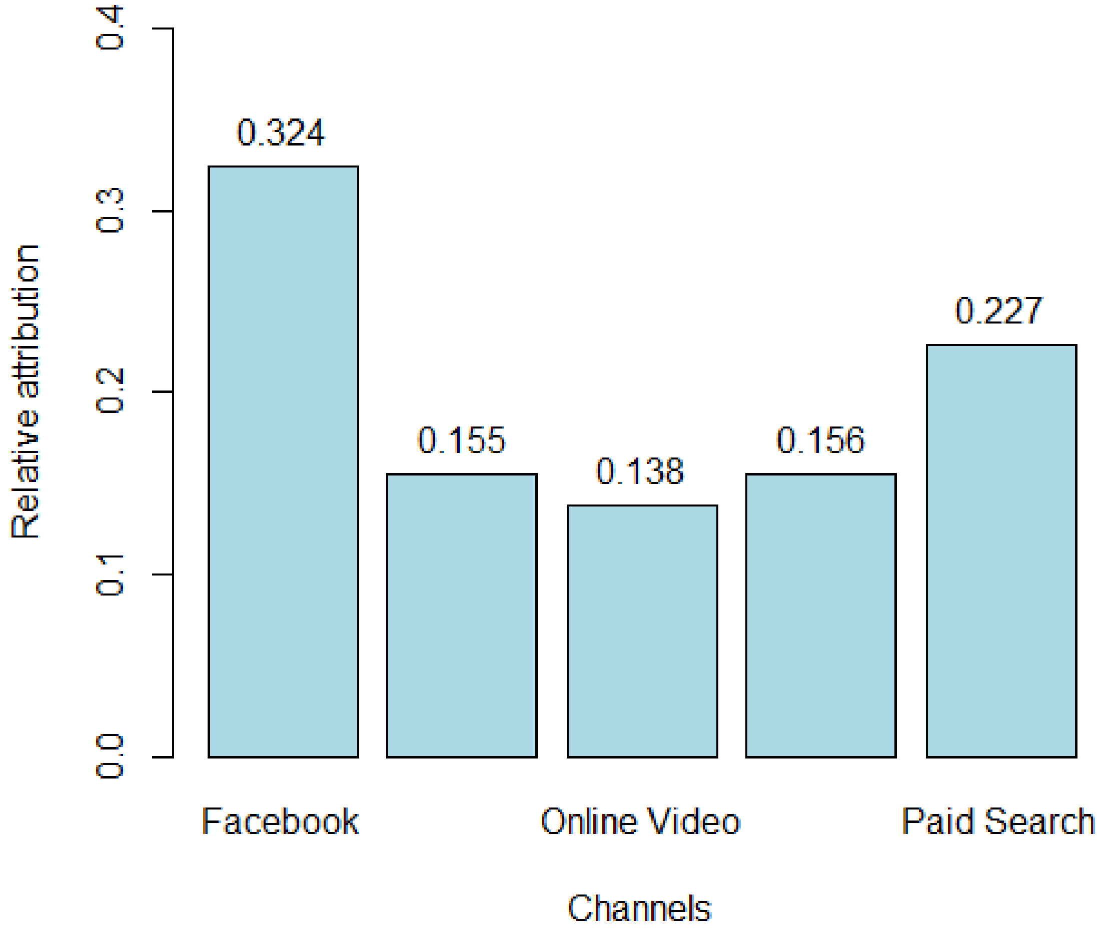
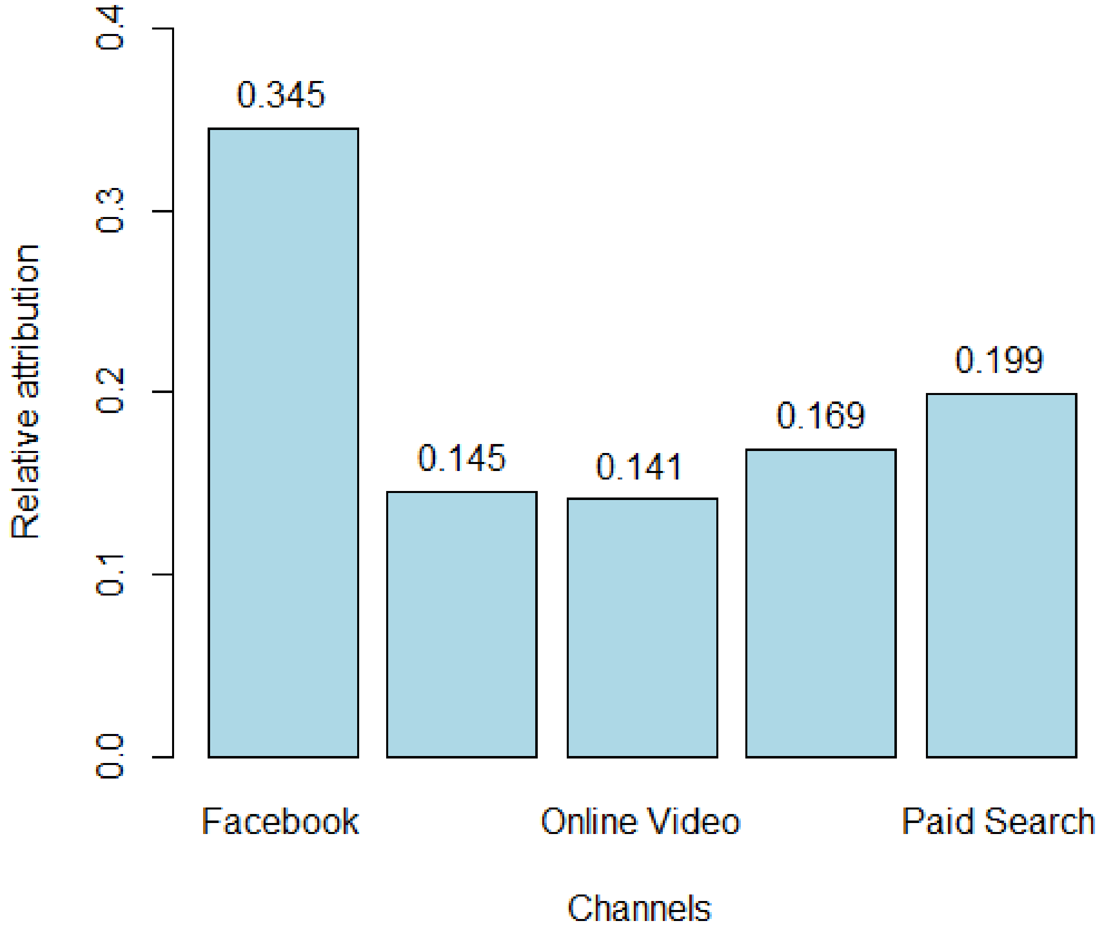


Figure 12: Relative attribution scores for the five advertisement channels for the AdditiveHazard touch attribution scheme.


Relative attribution
Facebook
Online Video
Paid Search
Channels