CZECH TECHNICAL UNIVERSITY IN PRAGUE
Faculty of Electrical Engineering
Department of Economics, Management and Humanities

Diploma thesis

Data Analytical Way to Identify an Appropriate Attribution Model for Digital Marketing

2018

Author: Bc. Matěj Matoulek

Supervisor: Jakub Novotný

ZADÁNÍ DIPLOMOVÉ PRÁCE

I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení: **Matoulek** Jméno: **Matěj** Osobní číslo: **406287**

Fakulta/ústav: **Fakulta elektrotechnická**

Zadávající katedra/ústav: **Katedra ekonomiky, manažerství a humanitních věd**

Studijní program: **Elektrotechnika, energetika a management**

Studijní obor: **Ekonomika a řízení elektrotechniky**

II. ÚDAJE K DIPLOMOVÉ PRÁCI

Název diplomové práce:

Data analytical way to identify an appropriate attribution model for digital marketing

Název diplomové práce anglicky:

Data analytical way to identify an appropriate attribution model for digital marketing

Pokyny pro vypracování:

- Description of attribution modelling in digital marketing (comparison of heuristic and probabilistic models)
- Traffic data description
- Traffic data analysis
- Models evaluation and conclusions for business application

Seznam doporučené literatury:

SHARMA, Himanshu. Attribution Modelling in Google Analytics and Beyond. Blurb, 2016. ISBN 1366694570, 9781366694577.
KAUSHIK, Avinash. Web analytics 2.0: the art of online accountability & science of customer centricity. Indianapolis, IN: Wiley, c2010. ISBN 0470529393.
BRODERSEN, Kay H., Fabian GALLUSSER, Jim KOEHLER, Nicolas REMY a Steven L. SCOTT. Inferring causal impact using Bayesian structural time-series models. The Annals of Applied Statistics. 2015, 9(1), 247-274. DOI: 10.1214/14-AOAS788. ISSN 1932-6157. Dostupné také z: http://projecteuclid.org/euclid.aoas/1430226092

Jméno a pracoviště vedoucí(ho) diplomové práce:

Jakub Novotný, Seznam.cz, a.s.

Jméno a pracoviště druhé(ho) vedoucí(ho) nebo konzultanta(ky) diplomové práce:

Datum zadání diplomové práce: **11.10.2017** Termín odevzdání diplomové práce: **25.05.2018**

Platnost zadání diplomové práce:
do konce letního semestru 2018/2019

Jakub Novotný podpis vedoucí(ho) práce	podpis vedoucí(ho) ústavu/katedry	prof. Ing. Pavel Ripka, CSc. podpis děkana(ky)

III. PŘEVZETÍ ZADÁNÍ

Diplomant bere na vědomí, že je povinen vypracovat diplomovou práci samostatně, bez cizí pomoci, s výjimkou poskytnutých konzultací.
Seznam použité literatury, jiných pramenů a jmen konzultantů je třeba uvést v diplomové práci.

.

Datum převzetí zadání Podpis studenta

Declaration

I hereby declare that this master's thesis is the product of my own independent work and that I have clearly stated all information sources used in the thesis according to Methodological Instruction No. 1/2009 – "On maintaining ethical principles when working on a university final project, CTU in Prague".

Prague, 22.05.2018

Matěj Matoulek

....

Acknowledgements

I would like to thank Jakub Novotný, my diploma thesis supervisor, for many useful comments and willingness to supervise my thesis in the first place, Seznam.cz, a.s. for providing me the data necessarily needed for this thesis, Eliška Šolcová for English proofreading and my family and close people for supporting me to finish this thesis and for their support during the whole time of my studies.

Abstract

This thesis is divided into two parts. The first part it theoretical, where the digital marketing environment is introduced, basic terms are explained, and attribution models are described. The overview of available attribution modelling approaches is focused mainly on data-driven models.

The second part focuses on the analysis of real historical data about online traffic of Zboží.cz. It describes the data, processing of the data, implementation of attribution model algorithms, possible difficulties, and conclusions drawn from the analysis.

The main goals of this thesis are to provide a complex overview of attribution models in digital marketing and to help traffic managers in Zboží.cz to make better managerial decisions about their online campaigns, mainly about marketing budget allocation.

Tato diplomová práce je rozdělena do dvou částí. První je teoretická, v níž je uvedeno prostředí digitálního marketingu, jsou vysvětleny základní pojmy a jsou popsány atribuční modely. Přehled dostupných atribučních modelů se zaměřuje především na data-driven modely.

Druhá část se soustředí na analýzu reálných historických dat o online návštěvnosti Zboží.cz. Popsána jsou v ní data, jejich zpracování, implementace algoritmů atribučních modelů, možné obtíže a závěry z analýzy.

Hlavními cíli této diplomové práce je poskytnout komplexní přehled o atribučních modelech v digitálním marketingu a pomoci traffic manažerům ve Zboží.cz k činění lepších manažerských rozhodnutí o jejich online kampaních, především o alokaci marketingových rozpočtů.

Keywords: attribution modelling, digital marketing, data analytics

Table of contents

1 Introduction of digital marketing

As the ecommerce market size [1] as well as advertising spends [2] are growing rapidly, question of advertising spends size comes into place. It is a nature of every market participant to allocate their advertising budgets efficiently. Although the data, the decision makers are working with, are exhaustive, they can suffer from a couple of technical issues such as ROPO (research online - purchase offline) effect, cookies-based measurement, fraud clicks and impressions, bot sites and many others.

Besides the technical issues, there are also conceptual challenges waiting for a solution. One of them is attribution modelling.

Advertising of ecommerce companies is nowadays relatively complex and consists of many different traffic sources. It means, that companies buy traffic from different advertising platforms (and therefore distribute their advertising spends). At the same time it is necessary to bear in mind, that visitors rarely finish the purchase during the first visit of the website. Instead, they perform a several phases described in theories about conversion funnel.

Depending on the efficiency of advertising strategy, some portion of visitors performs the conversion action. And here come the questions: Which of the so-called touchpoints caused the conversion action? Was it a combination of more particular sources? And how great is the value that was brought by this particular traffic source?

There are different approaches how to assess the value to the particular source and subsequently set the advertising budgets. To fully understand attribution modelling, it is necessary to build some theoretical background and that is going to be presented in the first part of the thesis. Main theoretical goals of the first part are to clarify which attribution model is the best one and how to choose the appropriate one.

How to implement such attribution model and how to solve real-world problems is going to be presented in the second part of the thesis. In this part, the data provided by the company Seznam.cz are going to be used. Seznam.cz is leading publisher company providing for its users several services such as freemail, news sites and content sites with many different topics and also price comparison website where users can compare prices of millions of products called Zboží.cz. More specifically, I am going to investigate conversion and related traffic data of ecommerce project Zboží.cz, which is a platform allowing visitors to compare goods online including the price and advertisers to promote their goods via cost per click model.

Zboží.cz is one of the projects running by the business company Seznam.cz introduced above. To maximise profits on this project, Zboži.cz traffic managers are trying to drive traffic in order to bring visitors to their site and potentially help advertisers to make more orders.

2 Attribution modelling in digital marketing

2.1 Definition of digital marketing

According to IAB (Interactive Advertising Bureau - institution trying to establish digital advertising standards and regulate the industry): "Digital advertising includes promotional advertisements and messages delivered through email, social media websites, online advertising on search engines, banner ads on mobile or Web sites and affiliates programs." [3]. Advertising is just a part of marketing, according to Dr. Philip Kotler's definition: "Marketing is the science and art of exploring, creating, and delivering value to satisfy the needs of a target market at a profit. Marketing identifies unfulfilled needs and desires. It defines, measures and quantifies the size of the identified market and the profit potential. It pinpoints which segments the company is capable of serving best and it designs and promotes the appropriate products and services." [4] digital marketing is much broader, than just advertising. In this thesis, digital marketing is understood with respect to Kotler's marketing definition as marketing using digital technology such as computers, mobile phones or other interactive devices in order to generate the value.

It is questionable, whether it makes sense to talk about marketing and digital marketing separately. As the whole industry professionalizes and takes advantage of technologies used in digital environment and the time spend with online media increases it is nearly impossible to leave out digital marketing from the whole marketing mix, regardless of the vertical.

2.1.1 Advantages

For online merchants, digital environment, respectively digital marketing, has two obvious major advantages. Firstly, there are technical possibilities which empower marketing techniques with exhaustive amount of data generated relatively easily, in comparison to traditional - offline environment.

This data could be used in marketing research as well as in marketing communication. Advertising platforms themselves provide wide range of tools which provide relevant and useful data. It resulted in abnormal attention of merchants who naturally starve for ways to increase their spends efficiency.

In practice, digital presentations, such as websites or mobile applications, are used in order to present products or services to visitors. For example, retailers use their websites as the online version of their offline stores and this parallel is often used to describe such a website.

In order to attract relevant visitors, data generated by users online activity is used. To attract them to the website, different so-called channels are used. The visitors interact with the presentation and, depending on many factors, part of the visitors sooner or later performs a conversion action.

This is not the end of the whole process, due to many statistics available [5], getting orders from returning visitors is significantly cheaper than acquiring new customers, and therefore it should be considered as a relevant source of visitors.

Digital marketing is not a closed environment. It is, of course, blended to offline environment where potential customers interact with products, brands or specific merchants. This aspect makes the key advantage of digital marketing - measurability - a little chaotic. However, this behavior is not likely going to disappear. What is more, connecting online and offline marketing is going to happen more and more often according to many sources [6] [7]. In this context, there are relevant terms like omnichannel, which is combination of online and offline marketing communication activities; O2O (Online to offline), which describes a situation when typical offline merchants try to catch attention of potential customer online; or ROPO effect, standing for Research Online, Purchase Offline.

2.1.2 History

In the next chapter, the history of digital marketing will be shortly outlined in order to introduce the readers to context and to point out how old the whole industry we are talking about is.

However, finding the first point in history, when we can talk about digital marketing, is hard, especially because it is not crystal clear what counts as digital marketing. There are several events that could be considered as the first appearance of digital marketing: inventing a radio, first usage of email, or invention of the first search engine [20]. But for the purposes of this thesis, the first usage of email spam, 3 May 1978 [21], is considered as the first digital marketing occurrence.

Another important moment in the digital marketing history is the purchase of the first banner advertisement. It was in 1994 and it had CTR of 44%! It was bought by AT&T on HotWired website [22] and it cost 30,000$ for 12 weeks placement [23]. Nowadays, it is much lower, average CTR in 2016 was 12% [23].

Company GoTo.com introduced the predecessor of the first PPC system in February 1998 [24]; in 2001 Yahoo acquired this company and started to use the system for Yahoo search engine.

Google released its AdWords PPC bidding platform in 2000, until that time advertising was sold based on CPM model through program called Premium Sponsorships [25].

Another chapter of digital advertising started to be written in 2004, when social network Facebook was founded [26]. In 2008, the first advertising system for Facebook was introduced.

In 2016, Google and Facebook generated advertising revenues of 106,27 billion US dollars in sum [2] [28]. Both companies received the highest advertising revenues in their history.

For comparison, print and digital advertising revenues of New York Times Media Group in 2016 was 0,58 billion US dollars [29].

2.1.3 Models of payment for digital advertising

There are several models of paying for digital advertising and the most frequently used varies across channels. There are only a few models, the most frequent ones, presented in this thesis.

Table 1 Payment models and their description [56] [57]

Model abbreviation	Description
CPM	Cost per mille, sometimes CPT (cost per thousand): Advertiser pays for every 1000 impressions. It is usually used for display (banner) advertising.
CPC	Cost per click: Advertiser pay for every click on the advertisement. It is usually used in search engine advertising or even display advertising.
CPA	Cost per action: Advertiser pay for some specific action provided by the publisher. This can include more complex interactions such as passing a new email in the newsletter database. Usually this is the model used in affiliate channel.
CPL	Cost per lead: Advertiser pays for the visitor that signalised that he is interested in the service or product and usually passed his contact information. This could be considered as a type of CPA.
PPP	Pay per post: Advertiser pays for publishing an article or a social media post promoting his business. This is rather rare payment model nowadays, however, it is still used.
CPMV	Cost per mile viewable: Advertiser pays for every 1000 impressions that were really displayed to a user.
CPV	Cost per view. Advertiser pays for every engaged viewer of a video ad. [67]

There is a discussion about what should be considered as an impression and how many impressions are actually never displayed to the visitor. This could be due to different reasons. First, web browser viewport is typically smaller than the whole webpage and advertisement could be displayed on a place, which is not visible for the user who does not scroll enough.

Another reason has more technical character. Some ad servers measure impression already in the moment a HTTP request to the ad server is fired. This does not necessarily mean that the advertisement was even loaded, because the user could hit the back button, close the browser window or lose the Internet connection in the meantime.

Additionally, there are problems with robotic traffic. Robots aka computers crawl the web and they fire the impression serving as well. Some estimates say, that volume of robotic traffic could be up to 60% [14]. This does not necessarily have to be fraud behavior. Robots such as Googlebot crawl the web in order to get the information and then use it for the search results. Besides that, there could be fraudster, behavior intending to increase the number of impressions and consequently the revenues for the publisher. For that reason, there are concepts as visible impressions [15] and CPMV payment model introduced earlier, which try to define the impression in the expected sense.

There are also obscurities about clicks. When advertiser pays for a click, he usually expects to receive this volume of traffic on his website. But because of the technical issues, unintended clicks, or due to the fact, that visitors simply change their mind in the meantime between the click and loading the page, the amount of clicks is usually higher than the volume of traffic received on the website. Number of sessions interacting only with the landing page and without any further interaction, referred to as a bounce rate, is sometimes used for measuring the level of engagement with the website and also refers to a quality and relevance of the campaign traffic source. However, it is important to bear in mind, that there are websites on which it is totally or nearly impossible to perform more than one interaction and then this metric is misleading.

Recently, Facebook reacted and updated its definition of click in its Audience Network [16]. In this case, advertiser is not charged, when user clicks on the advertisement and in less than 2 seconds gets back to the original Facebook page. It is considered to be an unintentional click.

Publishers usually sell their impressions or clicks in an auction model. It means, that in advertising platforms, marketers specify how they want to target their audience and set their maximum bid price, it is a maximum price they are willing to pay in a chosen payment model. When impression is fired somewhere on the web, real-time auction is issued and the highest bid buys the impression, respectively click. There are several positions in a search engine advertising, and the position is determined accordingly to the bid. Of course, there are publishers, that sell their traffic directly to one advertiser, but it becomes increasingly rare. Major platforms like Google AdWords, Facebook Ads, and other publishers using RTB (real-time bidding) use auctions.

Platform Google AdWords and similar search engine advertisement systems usually declare to sell clicks. However, factor described as advertisement quality contributes to the final result of the auction model. Auction model is based on CPM (cost per thousand impressions) which is recalculated value from CTR % metrics (number of impressions divided by number of clicks multiplied by hundred) This quality score consists of many factors, but one of them is CTR (click-through rate); the higher the CTR, the higher the quality score is. It is because the publishers do not want to display irrelevant advertisements in order to provide better user experience. But publisher's intents are not just noble. Low CTR means big portion of traffic, respectively impressions which are not paid for. Quality score influences the final price in the auction and therefore it is a little misleading to think about buying clicks. Publishers, in fact, sell impressions indirectly with one difference - advertiser is charged in the moment when the click is performed.

2.2 Web analytics

As it was written earlier, one of the key features of digital marketing is measurability of marketing activities efficiency. There are several software alternatives available, such as Google Analytics, Adobe Analytics, SiteCatalyst, or WebTrends. Google Analytics is one of the most popular from these, mostly because its basic version is for free.

In these tools, it is possible to track website (or mobile application) hits, web pages transitions, ecommerce events (such as conversions, including purchase revenue), and custom events (such as visitor's behavior on the page itself).

Tracking of events such as page hits or other is usually done with HTTP(S) requests. Such a request is fired in the moment a visitor performs an action in a browser (mobile application). This request is sent to a web analytics platform together with relevant parameters (such as browser information, operating system information, respectively revenue volume etc.). More of the technical details is not part of this thesis, as it is not its main subject.

Those web analytics platforms mostly rely on a technology called cookies. There are already other technologies able to identify the user based on probabilistic profiles of visitors or by user's account association (like in the case of Facebook) but cookies usually play a role in analytics systems anyway, despite the legal and other issues.

2.2.1 Technology of cookies

HTTP cookies are small text files stored in a web browser on user's device. Content of those files is sent with every request to the server which created them. Originally, it was used to store user-specific settings and to distinguish users across sessions.

2.2.1.1 Role in digital marketing analytics

Considering this application, the cookie file usually contains an identifier of a specific web browser. The identifier is sent to a server of web analytics platform or ads management (ad serving platform) and the server stores information about the browser behavior and the website interaction in order to be able to recognize the same user next time, evaluate his behavior, etc.

2.2.1.2 Issues with cookies

It is a common mistake, that people think about these cookies as identifiers of people. Due to the fact, that people nowadays typically use more than one device (sometimes even more browsers on the same device), and sometimes multiple users share one device, it is good to bear in mind that cookies identify browsers, not users.

Another problem which may occur is with the storage of cookies files. As the HTTP protocol can work without them, some browsers simply do not support cookies. If they do, they usually implement an option to delete cookies, for separate domain or all cookies. All of this imply problems with measuring, because it is not possible to identify the browser anymore.

In the EU, there is a law imposing an obligation to ask whether the user agrees with using cookies to analyse his behavior. Majority of webmasters interprets the law as so-called opt-out, which means, that user has the possibility to unsubscribe from using cookies, but, originally, it may have been intended to use the opposite principle.

2.2.2 A/B testing

A/B testing is a very often used concept of testing in online marketing. Typically, we want to compare performance of website funnels leading to a final conversion, two different color schemes or layouts of a page, two different advertisement pictures or texts.

The main idea is to collect information in order to be able to compare performance metrics. For instance, in the case of banners (advertisement pictures), it could be CTR (click-through rate).

To perform this test, two alternatives of the subject of testing are prepared and subsequently, the users are let to interact with them - part of the visitors with one alternative and the rest with the second alternative, simultaneously.

Using those alternatives simultaneously is crucial for avoiding problems with different conditions during the experiment. In the case the banner A is displayed 10000 times in the time from 8AM to 10AM and subsequently, banner B is displayed 10000 times from 10AM to 1PM, there is no way to be sure, whether there was some important factor influencing the number of clicks the two banners received, such as the different willingness to click in different daytimes.

Another important parameter of the test is to ensure that one user does not see both alternatives. This is usually done using the abovementioned cookies technology.

Support for such experiments is implemented in Google Analytics as well as in advertising platforms like Google AdWords or Facebook Ads.

2.3 Performance metrics

As it was written earlier, there is the possibility to track events of ecommerce activities quite well. Conversions which were generated on the website or in the application as well as their value are tracked. There are several issues, especially in the case of eshops. Sometimes it is not clear what is being tracked as a conversion value, because additional costs, such as shipping costs or VAT, are taken into consideration.

Usually, it is recommended to exclude the additional costs, but it depends on the merchant himself and the use of data. For a data clarity, it is best to track the costs separately in order to be able to operate with all the data later. This is possible only if the analytical software supports tracking cost components separately.

Another issue are different margins on different groups of goods. Small and mid-size merchants often do not track this and then it is necessary to bear this difference in mind when evaluating the performance of digital marketing.

To check how the investments into marketing channels perform, performance metrics are used. Obviously, the advertising investments (spends) are compared to generated revenues.

The simplest approach would be to identify which spends of the company could be clearly assigned as digital marketing spends and see how much of the company's revenue was generated online.

Then it depends on how far each merchant wants to go and how sophisticated metrics they want to use. Traditional ROI (Return on investment) is often the first choice, despite the fact that the timing of cash flows is not taken into account. It may not be necessarily a big problem, as marketing investments performance is often evaluated on a month, or multiple-month basis. Time does not need to play such an important role.

$$ROI = \frac{Digital\ marketing\ revenues}{Digital\ marketing\ spendings}$$

Equation 1 Return on investment

A bigger issue is the fact, that it is usually not the main parameter to be optimized. Generally, a total sum of revenues is the parameter to be interested in the most. Optimizing ROI can lead to reducing the investments rapidly and consequently shrinking the revenues volume simultaneously.

This issue leads to the approach that the focus is on maximizing the revenues, while keeping the ROI on sustainable level.

There are a few other names for ROI metric in the context of advertising. Firstly, it is ROAS (Return on advertising spend), which is basically the same as ROI, but as stated in the name, we are specifically talking about advertising investments. Sometimes we can see an inverse version of ROI in advertising systems or in the digital marketing community.

$$Revenue - spend\ ratio = \frac{Digital\ marketing\ spend}{Digital\ marketing\ revenues}$$

Equation 2 Revenue-spend ratio

There is a chance when acquiring a new client, that they will buy more in the future. Actually, it should be one of the main goals for merchants to increase the rate of returning customers. In this case, we should include the revenues in the equation. This could lead (especially in some verticals) to higher importance of a time parameter and metrics like NPV (Net present value) or IRR (Internal rate of return) would gain in importance, because they would reflect the reality better.

$$NPV = \sum_{t=1}^{T} \frac{C_t}{(1+r)^t} - C_0$$

Equation 3 Net present value: T := number of cash flows related; C_t := cashflow amount in time t; r := discount rate; C_0 := initial investment

$$0 = \sum_{t=1}^{T} \frac{C_t}{(1+IRR)^t} - C_0$$

Equation 4 Internal rate of return: T:= number of cash flows related; Ct := cashflow amount in time t; r := discount rate; C0 := initial investment

In this context, LTV/CAC (Lifetime value / customer acquisition cost) ratio should also be mentioned. Lifetime value is a metric calculating sum of revenues from one customer, normally based on historical data, while customer acquisition cost explains how much it costs to make a customer perform their first order. Overall, this metric says how much of customer’s revenues generated in the future we spend on acquiring the customer. This metric, however, does not take into account a time value of money.

So far, all of the metrics were calculated from aggregated data from all traffic channels of digital marketing. But it is not what is usually desired. To be able to better optimize the digital marketing strategy, all of the channels should be profitable. The data is usually available. Every digital marketing campaign could be labeled and therefore the interaction after which the conversion occurred could be determined. But is this really the right approach?

2.4 Digital marketing attribution

“Half the money I spend on advertising is wasted; the trouble is I don't know which half.”, this famous quote attributed to John Wanamaker is the essence of what digital marketing attribution is about. It basically tries to find out how big is the contribution of particular channels of digital marketing to occurrence of the conversion.

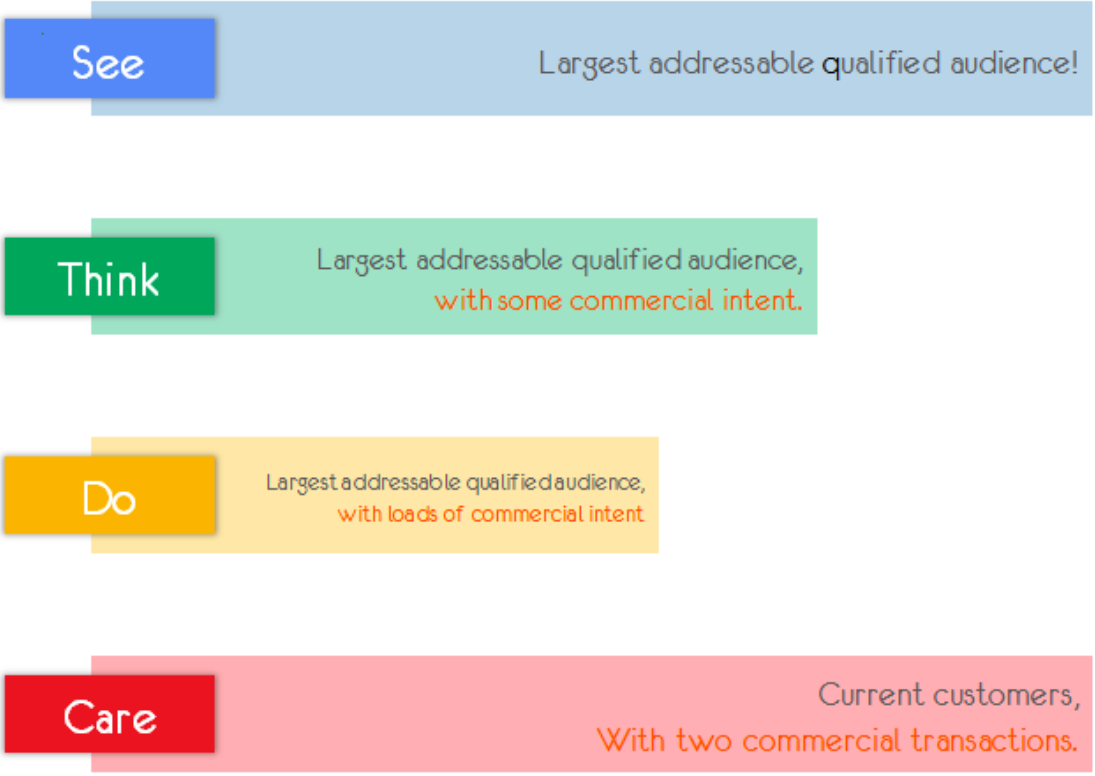
2.4.1 Multichannel attribution modelling

It is almost never the case that ecommerce merchant has just one source of traffic. Normally, there are multiple traffic sources and visitors are interacting with them. They undertake so-called customer journeys and visit merchant's website couple of times before they perform the conversion. Not only online channels play role in the customer journey, but also offline channels and non-media channels [17]. However, digital channels are way better traceable.

It is possible to observe patterns in the customer journeys and frameworks such as STDC (See-think-do-care) or AIDA (Awareness-interest-desire-action), which try to describe and formalize it.

2.4.1.1 See-think-do-care (STDC)

STDC is a framework by Avinash Kaushik [8] which addresses the problem of customer journeys and explains how to treat visitors, what to measure and how to build an appropriate strategy in every respective phase.


See
Largest addressable qualified audience!
Think
Largest addressable qualified audience,
with some commercial intent.
Do
Largest addressable qualified audience,
with loads of commercial intent
Care
Current customers,
With two commercial transactions.

Picture 1 See-think-do-care framework [8]

The "See" phase is the first contact with a potential customer. In this phase, there is an effort to reach the whole relevant audience and to let them know, that the company exists and what it does.

The "Think" phase already focuses on people that are trying to solve some problem and may want to buy something. One of the goals of the previous phase is to generate new members for the "Think" audience. An effort in this phase is to state the reasons why the company's solution is the appropriate one and to rank the best when considering other alternatives.

The "Do" phase focuses simply on performing the conversion. There needs to be high efficiency in terms of the checkout process and remove the last doubts about the company's goods or services.

The "Care" phase tries to make the current customers as satisfied as possible, monetize them and increase the lifetime value of customers.

2.4.1.2 Awareness-interest-desire-action (AIDA)

Framework AIDA probably created by Frank Hutchinson Dukesmith [9], sometimes called marketing funnel or sales funnel, is a widely used concept describing different phases that are undertaken before finishing the order. In the "Awareness" phase, a company should introduce the business and the product, in the "Interest" phase, it should attract the attention of a potential customer. With further interaction, in the phase called "Desire", a company should present the advantages of their service or product to make the potential customer interested in buying the service/product. The "Action" phase then focuses on finishing the order.

Originally, there is no phase focusing on a client retention in this concept. This problem could be explained by the argument that AIDA is a framework for acquisition of new clients. It may also be solved by adding the letter "R" as "Retention" at the end of the abbreviation [10].

This framework is not used in digital marketing only, but is well-known in other fields of business like sales too.

2.4.1.3 Summary

From both of the above mentioned frameworks it is clear that they expect several stages, phases or interactions with potential customer. Depending on the vertical in which a merchant operates, it shows up in online environment as well. If there are several steps necessary to perform the final action, the evaluation of performance of specific marketing channels should be treated with caution. If some channels tend to be at the beginning of a customer’s journey and therefore they do not bring that much of a completion of the goals, and at the same time the channel performance is judged based on its ability to finish the conversion, maybe there is a part of reality missing. The same applies to other scenarios where we attribute the whole value to one channel.

2.4.2 Terms

To prevent any misunderstandings, here are the definitions of some basic terms:

2.4.2.1 Conversion

Conversion is the action when a visitor converts into a customer, or at least into a prospect. This usually happens when the visitor performs an action on a web such as confirming an order or sending an online formular. As it was mentioned earlier, this action can have specified value, especially in the case of eshops. Normally, the information is about the revenue volume passed within the information about the conversion itself.

There are also so-called micro-conversions or the partial or secondary goals, that can be measured on a website or in an application. Typical example would be visiting the contact page. Goals like these have two reasons. First, there is no need to force the visitor to finish the goal which was set up in advance as primary. Every visitor has different behavior and habits. Somebody simply do not want to finish the order online but prefers to go to an offline store and finish the order there. Or there are people who do not want to fill in the form and rather call or write an email. Second, even if these micro-conversions are not completely clear, they can be used as a supporting criteria to judge the success on.

2.4.2.2 Channels of digital marketing

Channels were also already mentioned a few times. In digital marketing, this term describes a traffic source or group of traffic sources which are similar. It may be the same type of placement or the same type of advertisement format across different publishers. This work will use channels definitions (channel distribution) according to Google Analytics documentation [11].

Table 2 Channels definition by Google Analytics [11]

Channel	Description
Direct	When user enters the website address directly into web browser or uses a web browser bookmark.
Organic Search	When user clicks on the search result and it is not a paid result. Optimization of this channel is called SEO and it is often incorrectly used as a name for the whole channel.
Social	Interaction from social media.
Email	Clicks from email communication. Usually newsletters etc.
Affiliates	Interactions from a merchant partners who usually get paid for the promotion.
Referral	Not-paid clicks from external websites where link to the page was published.
Paid Search	Clicks from results of search engines that were paid. Often called PPC, which is referring to a model of payment usually used.
Other Advertising	Other advertising sources paid on different basis etc.
Display	Image advertisement published on external websites.

Due to historical reasons and established usage, there can be differences in channel names or even channel definitions among merchants. But the logic behind it usually follows the same pattern. However, some marketers also consider other criteria significant enough and separate the channel based on them. Good example is the type of device, from which the interaction was performed. In that case, channels for mobile, tablet and desktop can be seen, respectively combination of the device type and channels defined above.

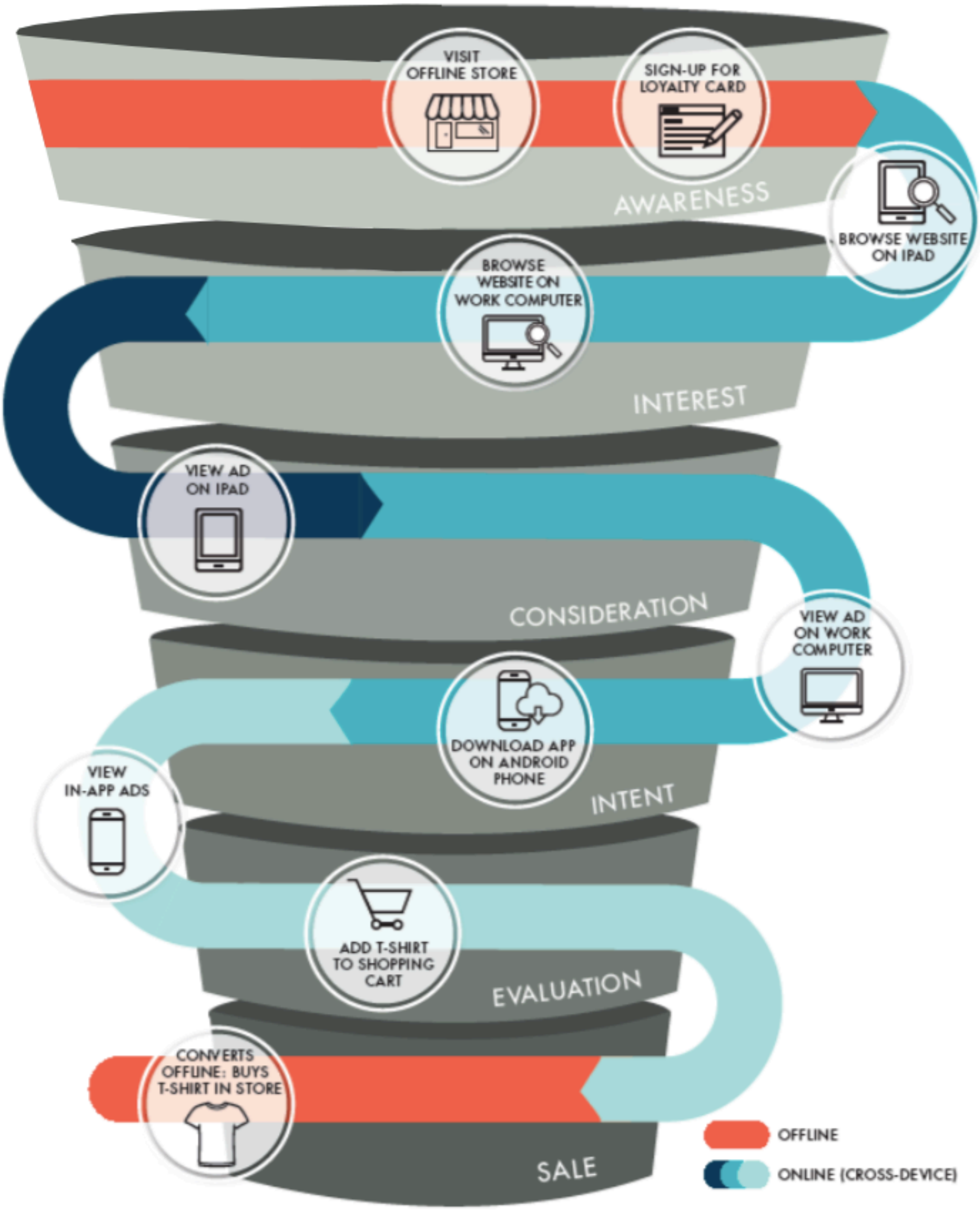
2.4.2.3 Touchpoints and customer journeys

Touchpoints are interactions of a user with a merchant in general. It could be either a visit of the site or just an impression.

Channels mentioned above are clearly click based. The interaction which is important for the purposes of this thesis is the visit of the website. According to some studies [12] [13], for some channels, it is not the visit what correlates with conversions the most, despite the fact that many marketers evaluate the campaigns based on the number of clicks. Some advertising platforms do not even provide the precise information about impressions.

Customer journeys are the sequences of interactions. A typical customer journey could be, for example, that a user saw an advertisement, recalled the company and entered the site by typing the address in the address bar of a web browser. The user researched the portfolio of products online and something distracted him, so he left the site. A few days later he got into the situation where he needed the product, but he had forgotten the name of the website. So he searched for it in the search engine, clicked on paid search result and finally finished the purchase.

This example finished with conversion, however, typically, this is the case just for low percentage points of the total traffic. A lot of conversion paths finish with no conversion at the end.



Picture 2 Conversion funnel showing multiple issues occuring [31]

This is related to the term "conversion window", which defines how long it takes from the interaction and the conversion. There is no universal or exact rule how long can man remember, that he interacted with an ad. This is up to each merchant to choose the right size of conversion window. Type of the interaction should be taken into consideration. Click interactions are considered to be more influential than impression interactions. Generally speaking, in retail conversion windows tend to be smaller than in other fields, but good pick of conversion window size is matter of the domain experience in the end. Normally, values from 7 to 90 days are considered as relevant.

2.4.3 **Attribution models**

According to IAB: "Attribution is the process of identifying a set of user actions (“events”) across screens and touch points that contribute in some manner to a desired outcome, and then assigning value to each of these events." [18]

Attribution models are then the prescriptions that assign the values for specific channels and eventually the process how to conduct these values.

2.4.3.1 Motivation

The right attribution model is the key for the right distribution of advertising budget. In order to allocate the budget appropriately, it need to be divided according to the contribution of the specific channel to the desired action. Digital marketing channels can be looked at as a portfolio which needs rebalancing in order to maximize revenues generated from it. It means, that revenues volume can increase without adding any additional advertising spend.

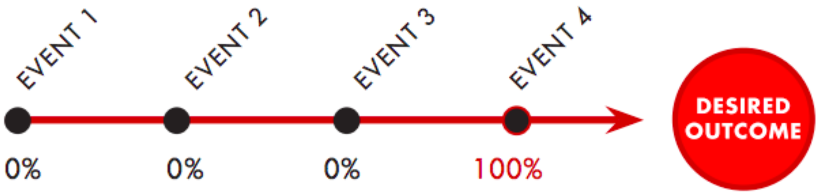
The whole ROI (ROAS) validity of specific channel completely depends on an attribution model. When we operate with an unreasonable attribution model, any optimization might be misleading.

2.4.3.2 Heuristic models

Heuristic models are based on assumptions and assign the exact value to the channels based on their position in customer journey. The basic concepts are attribution models with one only source of attribution, while the more complex ones are those, which assign the value to multiple sources.

2.4.3.2.1 Typical models

2.4.3.2.1.1 Single source attribution

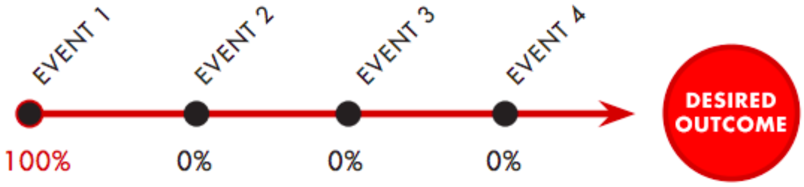


Picture 3 Last-click model [19]

There are two reasonable models with one source of attribution. First is so-called "last click". In this model, the whole value of conversion is assigned to the last source of traffic before the conversion happened. This means, that who uses this model believe, that the source on the end of the customer journey is the only one responsible for the conversion generation and conversely channels on the beginning of the customer journey are worth nothing.

There is also an alternative for this model called "last non-direct click" and it does the same, but if the last source is direct, or non paid search, or referral, then it assigns 100% of the value to the last paid source of traffic. The logic behind it is, that "direct" as a traffic source is not paid and will always be present. This could lead to an opinion that we should not assign any credit to it, because it is a by-product of the rest of the channels which are responsible for the fact, that a user remembered the brand of the eshop or service, respectively it's URL, and typed it into the address bar. This model is used in Google Analytics by default.

The problem is in the perception of a direct as a traffic source, because we feel that it is not a “source” in the original sense of the word. But if we start to think about a direct as an interaction, there could be reasons found why to assign a portion of credit to it. As we want to evaluate how influential the interactions from given source are, we also want to assign some portion of the overall performance to a direct, as it reflects so many factors, like special offer that invoked the direct interaction, offline campaigns, brand recognition, or overall website experience [34].



Picture 4 First-click model [19]

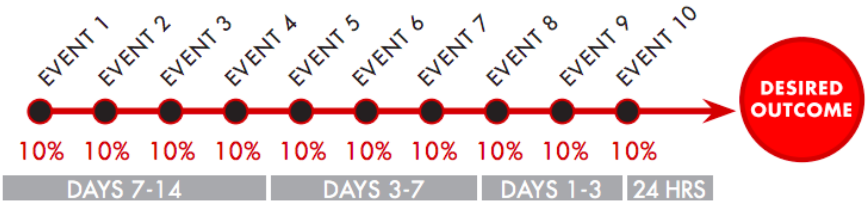
The second model with one source of attribution is called "first click" and it assigns 100% of the credit to the first source. This model assumes, that the only important thing is to let the customer know that the product or service exists.

Clearly, there is a big part of reality missing in the models, regarding the situation when users interact with more than one traffic source, which happens most of the time. Following models reflect this fact and attribute to multiple sources. If there is any uncertainty , whether to worry about attribution or the typical last-click (or single source in general) attribution model is enough, the following quote can be helpful:

"If a significant percent of your conversions have a greater than one path length, you have an attribution problem." [34]

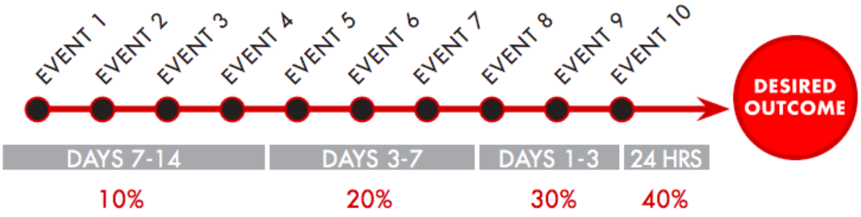
2.4.3.2.1.2 Multiple source attribution

"Multiple source attribution is the process of collecting and analyzing more than one advertising events contributing to an outcome. This type of measurement is based on the belief that all advertising events that occur within a users path—across channels, platforms, and formats—have a cumulative effect on consumer behavior when contributing to a desired outcome." [30]



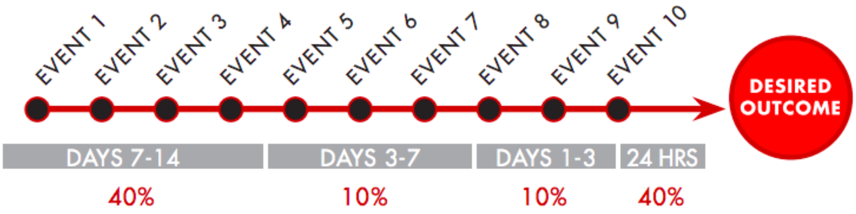
Picture 5 Equal-weighted model [19]

The simplest model is so called equal. Distribution between all traffic sources is very simple. Every source in the path receives equal portion of credit. This probably makes more sense than the models presented earlier and could be a good "hot-fix". However, if the budget should be distributed precisely, this model is probably not the best way. Logic behind this model is, that everything that caught attention of the user participated equally on the final action and it deserves equal portion of credit. Unfortunately, this model does not count with number and type of interactions delivered by each channel.



Picture 6 Time decay model [19]

The second alternative, usually referred to as "Time decay model", assigns the value increasingly with increasing time to conversion or interaction closer to conversion. This model assumes, that channels on the beginning of the customer journey have smaller influence, because users tend to forget things that happened in the past.



Picture 7 U-shape model [19]

The last of the classical models is called "U-shape". It assigns big portion to the first and the last point of conversion path and the rest is distributed among the interactions in the middle of the path.

2.4.3.2.2 Decision factors for heuristic model selection

There are several models described earlier. All of them have reasons to exist, but it may not be clear which one is the best.

In multi-touch situation, there can be barely a reason to use single click attribution model, because the channels on the conversion path contributed partially to the creation of the goal at the end.

However, the criteria can be the attitude towards growth and there can be prefered channels in the marketing strategy that begin the whole journey, because those channels that successfully introduce the product or service need to be supported. In this case the good beginning or introduction is crucial.

Conversely, if the growth is not the marketing priority and the focus is on short-term performance, the last-click model might be the best fit. Because optimizing by using the last-click model leads to decreasing of a budget for channels that are not directly prior to conversion. This leads to using channels that are attracting people in later phases of conversion funnel. This would be helpful point of view if we believe, that our product or service is comparatively good in the last phase, when people are making the final decision i.e. where to buy.

From this point of view, it would make sense to use the attribution model stressing the interactions in the middle of the conversion path. If it was clear that the interactions in the middle of the customer journey are crucial in the final decision, such a model would be used. This is applicable if there is a belief in strength in the phases “Think” of STDC framework or “Interest” and “Desire” phases in AIDA framework.

One of the main factors should be a basic understanding of the domain and statistics of current campaigns. From related reports it can be seen whether the channel tends to be the last interaction on the conversion path or whether it appears rather in earlier phases. So called time-lag report is useful for setting up correct conversion window length. It shows us how many conversions happen within the given time period. Conversion paths report shows us what the typical journeys which users undertake are. The user cannot be forced to undertake a specific path, but the the length of those paths can be guessed from the number of occurring paths and the place where the specific channels tend to appear. A good approach is to think about correct channel specification in order to have significant results for the division, but on the other hand, not to group channels that do not behave similarly [34].

The last-click attribution model is implemented in many advertising and analytical platforms. Some experts [34] recommend to switch to the time decay model in order not to be completely wrong.

“It should be completely obvious to you that this model is based on a specific client's business environment, my experience, and business priorities.” [34]

2.4.3.2.3 Critique

The main criticism of these models is, that the choice of attribution model is highly subjective. It is true, however, that using pre-defined or rule-based defined attribution models is extremely easy and it does not require any further data analysis or exhaustive data in the first place.

Using pre-defined model also creates advantages when comparing results with others. Most of the pre-defined models are well known and it is easy to explain what to use.

On the other hand, the choice of the heuristic attribution model is, in the best scenario, based on approximate rules conducted from domain specifics and from personal beliefs. When more accurate results are to be obtained, there are ways how to derive the attribution model from the data.

“If you spend more than $10 million on advertising/marketing, it might be well worth it for you to completely skip all the attribution analysis challenges and jump to media-mix modeling by leveraging controlled experiments.” [34]

Especially scientific community criticises heuristic models because of their non-exact base:

“The drawback of such rule-based models lies in the fact that the rules are not derived from the data but only based on simple intuition.” [55]

2.4.3.3 Data-driven models

Data-driven models are the ones in which the data analysis is performed first, before choosing an attribution model. This process should demonstrate the importance and value of the particular source in the overall context. As the process is based on related data, it should reveal the specific attribution model for the given business.

Even in data-driven models there are more methods. A few of them will be presented later in this chapter.

2.4.3.3.1 Data-driven model challenges

One of the key challenges is the overall data readiness for applying a data-driven attribution model [31]. There should be all the required data present, including user attributes, interaction information, or conversion data.

This is not necessarily as easy as it might seem. Any imprecisions in input data can destroy the whole result, because of GIGO (garbage in - garbage out) principle.

Firstly – if interactions about users are not tracked, but cookies are used instead, it might lead to biased results. For example, if a user performs some interactions on a specific device and browser and in the middle of his path to conversion he changes the device, or just the browser on the same device, in cookie-oriented analytical system it could be falsely interpreted as a part of another journey. Actually it is not and this conversion path will express another kind of behavior than the fully user-based information.

There are two different approaches of identifying a specific user across different devices.

A deterministic model is the one with high accuracy. Devices, browsers, and user identifiers, such as advertising IDs or cookie IDs, are paired with user’s login data. Login data is gathered from many sites and service providers and it is usually one of the most valuable data when it comes to cross-device attribution modelling and evaluation. Email login or login with other credentials are used in this case. It is widely used by companies like Facebook and Google.

The probabilistic method tries to establish a model which would be able to recognize the user across devices based on proxy signals such as IP address, web browser, geolocation, operating system, language used in the web browser, or web use behavior. For the identification of a user, a machine learning model, which uses identifiers mentioned above to predict the real user using the device, is used. Such a solution is implemented by companies like screen6, Roq.ad, or TapAd. The prediction accuracy is measured by two fundamental metrics - recall and coverage.

There is a big discussion about impression visibility concerning the quality of interaction data. What a big issue this is could be illustrated by a number of advertisement impressions that actually were not seen by anyone. This number varies according to a visibility definition and a person who conducted the particular research, but the average visibility it is between 31 and 56 percent [35], which is not an insignificant number. What plays an important role in interaction data quality is information about the channel and format itself. Different behavior can be observed when talking about different types of media, placement, advertisement format, and interaction engagement level (the above mentioned visibility, or whether the advertisement was clicked-through).

For example, bigger formats such as 1400 pixels wide branding format should get higher importance than 300x250 pixels banner even though both exhibit the same visibility time.

Consequences in data-driven attribution modelling would be dramatical. If the calculation includes the impression data as interactions but, in fact, there were none, the influence of the source can be easily overestimated.

Media Rating Council defined the viewable ad impression (impression with the potential to be seen) as follows: “A served ad impression can be classified as a viewable impression if the ad was contained in the viewable space of the browser window, on an in-focus browser tab, based on pre-established criteria such as the percent of ad pixels within the viewable space and the length of time the ad is in the viewable space of the browser.” [68], and added the pre-established criteria as follows: “Pixel Requirement: Greater than or equal to 50% of the pixels in the advertisement were on an in-focus browser tab on the viewable space of the browser page. Time Requirement: The time the pixel requirement is met was greater than or equal to one continuous second, post ad render.” And this is widely considered to be the industry standard [68].

The third source of issues is conversion data. The full range of conversion data should be operated with. A typical problem is with the data about offline conversions because of ROPO effect etc. The interactions with offline environment influence the online behavior and vice versa. In terms of offline data, online campaigns can invoke conversions offline and they should be tracked precisely by conversion paths. This is often not easy to implement and the ability of businesses to implement such a policy is related to an overall data readiness of the business [31].

The question which stands on top of all is whether there is enough data for performing the analysis and obtaining significant results. For the method used by Google Analytics, 400 conversions with path length higher than 2 interactions and 10 000 paths undertaken in last 28 days are required [36].

2.4.3.3.2 Research

Methods used for attribution model analysis originate in different fields of studies. Markov chains and Shapley value were originally part of the game theory [37] [19], Markov chains are used in computer science theory [38] and linear regression is used across the fields including environmental economics, medicine etc.

The goal of this thesis is not to present all of the possibilities, but only a few commonly used ones, especially in order to provide help for better understanding the analysis performed further.

As it was referred in [54], the variety of statistical models like logistic regression, simple probabilistic model, Bayesian inference, causally motivated methodology, mutually exciting point process, structural vector autoregression, Shapley value, or Markov chains were used. It was proposed by [54] to use Markov model as it satisfied all of the evaluation criteria.

During my research I discovered a paper [48], which could also be potentially relevant for attribution modelling.

2.4.3.3.3 Big data

A very often used term in the context of data-driven decisions is big data. Definition of this term is not uniform, and its aim is not just to set a threshold level of the greatness of big datasets that should be considered as big data, but it defines the structure needed and the processing of such data. One definition could be, that we can talk about big data as about: *“datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze”* [39]

The Internet produces and is able to store and process formerly unimaginable loads of data fast, easily, and unobtrusively. This, of course, applies to digital marketing environment. Every interaction with a web generates data which is usually stored somewhere. This applies to online purchases, filling in forms, open emails, web clicks, search queries, or even mouse movements [54].

2.4.3.3.4 Markov chains methodology

Markov chains have found its application in marketing already in 1964, when Styan and Smith wrote a paper investigating brand loyalty using Markov chains [40]; from that time several other studies were written [51] [52] [53].

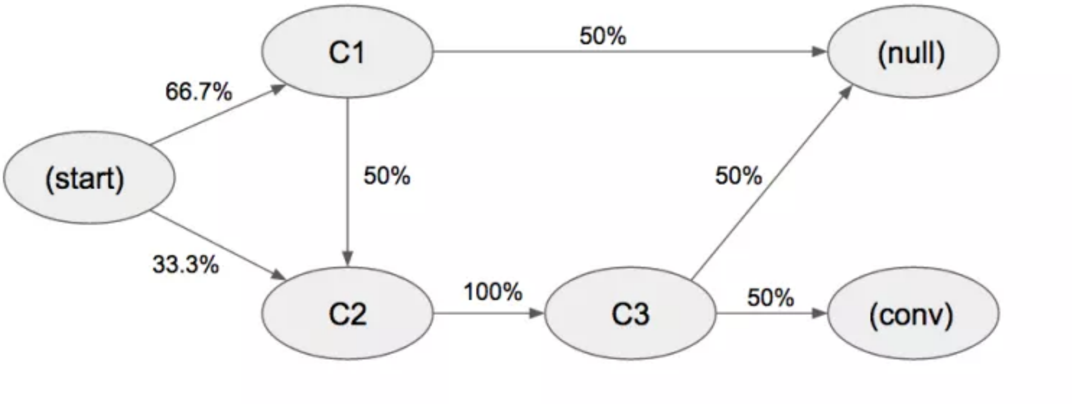
Besides that, it finds its application mainly in economics, finance or computer science, for example, its usage in the case of PageRank algorithm used to score and order web results in Google search engine.

Markov chains are mathematical probabilistic system, describing transitions from one state to another according to transition probability.

For multi-channel attribution modelling, states describe particular marketing channels and transitions describe paths that the user follows. Probabilities of transitions are calculated based on the underlying data.

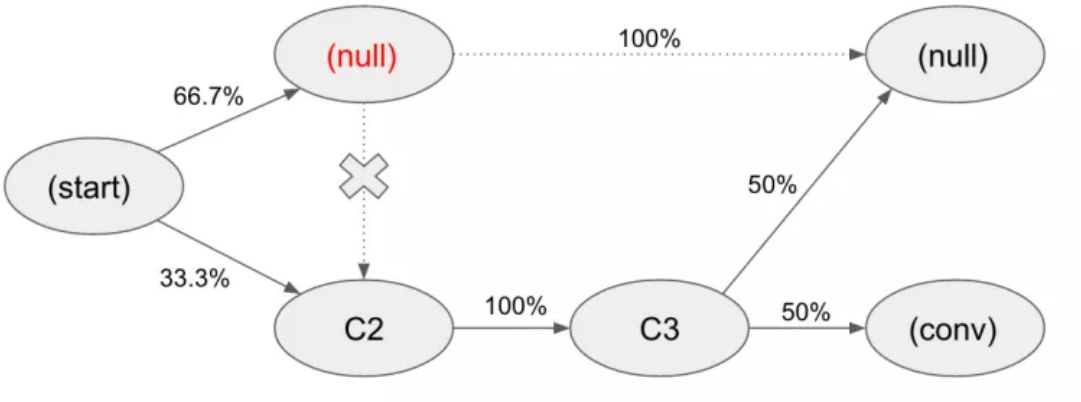
For the calculation purposes there are three more artificial states included - start state, conversion state, and null state. Null state describes the end of the path in which the user did not perform the desired action.

Markov property assumes that transition probability from one state to another depends strictly on the present state and not the one preceding it. In the case of web path analysis it means that if a user interacts with channel C, the next channel he uses depends strictly on C and not the channels that were used before C. This property is sometimes referred to as *memoryless*. While some researchers suggest it is non-problematic to use it for web usage [45], the prior research found out that customer journeys do not strictly follow this rule [41]. In this context, the prior property is referred to as first-order Markovian model and depending on how many states influence the transition we then talk about the nth-order Markovian model. However, first-order memoryless Markovian models are still used for attribution modelling because of their simplicity.



Picture 8 Markov chain graph for attribution modelling [43]

The technique used to calculate the importance of particular channels is called removal effect and it is quite straightforward. For every channel in the graph, the portion of conversion that would have been lost if the channel had not been used is calculated. This can be done by calculating the probability of conversion of the complete model – the sum of probabilities of all paths leading to conversion and subtracting the conversion probability of a model where this particular channel is replaced with another null state.



Picture 9 Removal effect in Markov chain [43]

Finally, to calculate the weights of channels, the portion of possibly lost conversions needs to be recalculated relatively to the sum of all possible losses. So for each channel, the individual portion of possible losses is divided by the sum of possible losses of all channels. Therefore these weights should make one all together. These weights then form the data-driven attribution model conducted on first-order Markov chain algorithm.

As the weights are relative, they can be interpreted as a portion of credit the advertiser should spend on the particular digital marketing channel of overall digital advertising budget.

There are several issues to deal with. Firstly, the conversion window length, because it is crucial to treat the data right. Secondly, first-time converting users and users that have already converted in the past should be distinguished because their behavior can be significantly different.

Using the nth-order Markovian model leads to higher computational complexity and therefore the real-world constraints are limited.

Other issues occur when channels are not tracked appropriately, however, it is not an issue of Markovian chain in particular [45].

2.4.3.3.5 Logistic regression methodology

Another approach was suggested by Shao & Li [46]. They proposed to use logistic regression in order to estimate channels attribution to the conversion. The model estimates whether the user performed the conversion and for that purpose predictors, representing the fact that the channel appeared in the conversion path are used.

In the same paper, a metric for evaluation was proposed [46]. The metric consists of two components – V-metric and A-metric. The V-metric is used for measuring variability of estimates across different estimations and it is calculated as the average of standard deviations of estimates of all coefficients used in the regressions.

The A-metric is used for measuring a model accuracy and it is calculated as the average of misclassification error rates of all model setups.

Both parts of bivariate V-A-metric are desired to be as small as possible. In attribution modelling, the accuracy of the overall prediction is just one part that concerns the marketers. As they want to distribute the credit to particular channels appropriately, they also care about unbiased estimators for every predictor. For that purpose, bagging idea was proposed by Shao & Li [46].

This bagging approach of logistic regression is in a way similar to machine learning algorithm called *Random Forest*. The main idea is to average results of estimation results in order to reduce bias created due to high correlation between the regressors [46].

The bagging process works as follows. Portion of regressors p_c and portion of observations p_s is randomly chosen and the estimates are recorded. This procedure is repeated M-times and the result of this bagged logistic regression is the average of the regressors. Authors recommend to use values around 0,5 for both p_c and p_s. For M, they used 1000 and they did not receive significantly better results by increasing this value.

2.4.3.3.6 Second-order probability methodology

In the above-mentioned paper [46], it was also proposed to use second-order probability model, even though lower accuracy was expected and experimentally verified.

This model calculates the probabilities of conversion given the exposure of particular channel. Subsequently, it calculates conditional probabilities given the exposure of pairs of channels in the customer journey.

The contribution of particular channels is then calculated with the following formula:

$$C(x_i) = p(y|x_i) + \frac{1}{2(N-1)} \sum_{j \neq i} \left\{ p(y|x_i, x_j) - p(y|x_i) - p(y|x_j) \right\}$$

Equation 5 Contribution of channel x_i: $p(y|x_i)$:= conditional probability of conversion given exposure to channel x_i; $p(y|x_i,x_j)$:= conditional probability of conversion given exposure to both - channel x_i and x_j; $p(y|x_j)$:= conditional probability of conversion given exposure to channel x_j; N := number of channels

Consequently, the contributions need to be normalized to express the weights of particular channels with the following formula [47]:

$$w(x_i) = \frac{C(x_i)}{\sum_{j=1}^{N} C(x_j)}$$

Equation 6 Weight of channel x_i: $C(x_i)$:= contribution of channel x_i; N := number of channels

Second-order probabilistic model was proposed because of high overlapping influence between channels, on the other hand, there is usually not enough data to use higher-order probabilities, even though it would lead to higher accuracy.

2.4.3.3.7 Causally motivated methodology

There was also a research focused on causally motivated attribution model. Very often cited work is the paper [47]. It is basically based on probabilistic approach introduced earlier. In contrast to the probabilistic approach, it adds a few strong assumptions to ensure causality.

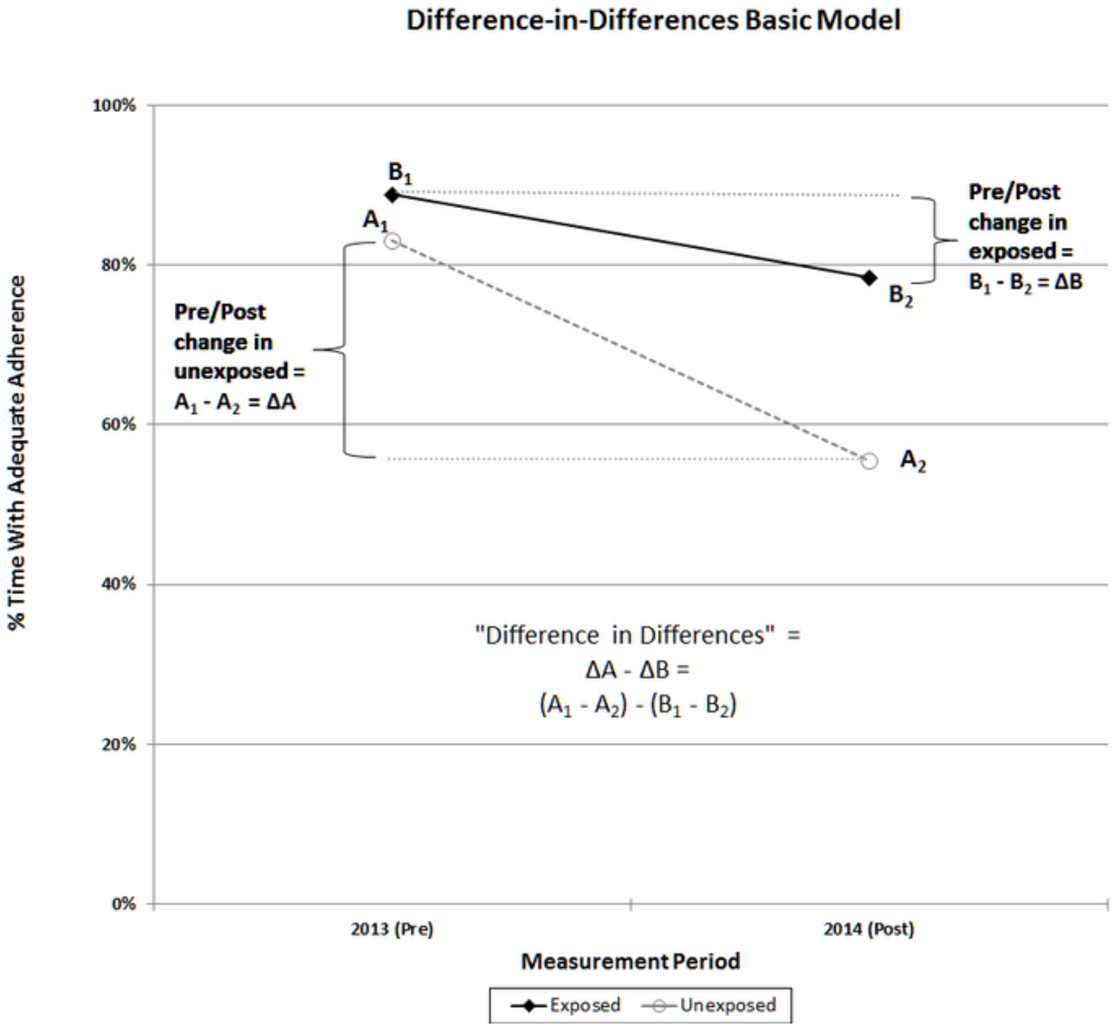
It starts by defining causally motivated attribution, but due to strong assumptions that: “the treatment precedes the outcome”, “any attribute that may affect both ad treatment and conversion outcome is observed and accounted for”, and “every user has some non-zero probability of receiving an ad treatment” it suggested to switch to the channel importance approach, which does not have such strong assumptions.

Particularly the assumption: “no unmeasured confounding”, could be violated, because most of advertising systems use their own logic to serve the ads in order to target users that are more likely to convert and this logic is always impossible to capture in the external model.

The assumption of “positivity” is likely to be violated in the data sample that is practically possible to cover, because it assumes that in the data set, there is at least one observation with positive number of conversions and one observation with zero conversions for every customer journey setup. As the number of possible customer journey can get quite high, it would be practically impossible to consider this assumption valid.

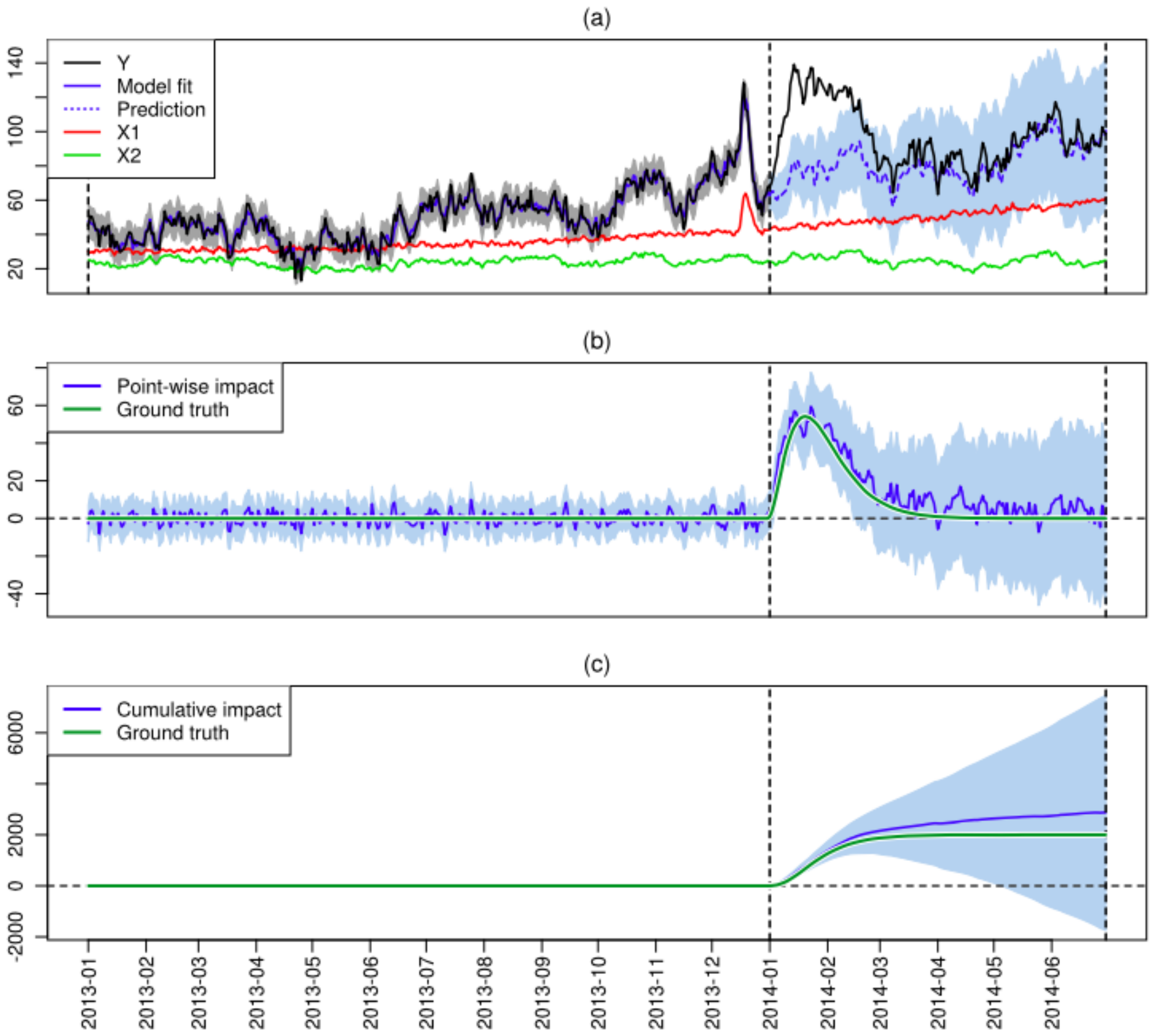
For that reason, channel importance was proposed to be estimated using the game theory approach, namely Shapley value [47].

Further research established a framework for estimating causal impact by adopting the well-known difference-in-differences approach and generalising it to time-series setting [48]. This paper got cited across different fields including economics and medicine [49].



Picture 10 Classic difference-in-differences model [50]

Basically, it predicts the time-series based on the data before the treatment, it finds a control that was predicting well the pre-treatment time-series and did not receive the treatment and prior knowledge of model parameters. Then this unobserved prediction is subtracted from the observed results after the treatment and it is considered to be the causal effect of the given campaign [48].



Picture 11 Causal effect Bayes estimation [48]

As this paper recommends to use at least three times longer period before treatment (campaign) for prediction in comparison to treatment length, it can get quite hard to use this technique for evaluation impact of every channel in the channel mix in order to establish data-driven attribution model. However, as this approach is able to estimate effect of one time-limited campaign, it is theoretically possible to use it to estimate (marginal) effects of each channel.

2.4.3.3.8 Shapley value methodology

Shapley value is an equation from game theory, namely cooperative games. Shapley value was first introduced by Llyod Shapley in 1953 [59] and it aims to assign the fair value of the contribution of the player in all coalitions which he participated in. Fairness is intuitively also one of the main goals of attribution model, so it makes Shapley value a good candidate for the attribution modelling approach.

As marketing channels can be considered for players and customer journey as coalitions, attribution modelling can be seen as a game theoretical problem.

Shapley value is defined as a payment equal to the average of marginal contributions of a player in all the coalitions.

In attribution modelling, values generated by the customer journeys are distributed, and can be measured in revenue or number of conversions.

Shapley value original equation assumes that all of the possible coalition values are defined. In fact, these values are known only for the customer journeys that have appeared in given time period for which the data are available.

$$\Psi_i(N,v) = \frac{1}{N!} \sum_{S \subseteq N \setminus \{i\}} |S|!(|N| - |S| - 1)![v(S \cup \{i\}) - v(S)]$$

Equation 7 Shapley value for channel i: N:= set of all possible customer journeys; S⊆N\{i}:= set of all possible customer journeys excluding those containing channel i; v(S):= value brought by customer journey S

The equation above basically means that all customer journeys with given channel are iterated through, the channel is removed from the customer journey and the likelihood of conversion for this modified customer journey is found. Then the difference between the original customer journey and the modified one is calculated and eventually multiplied by weight according to portion of revenues or conversions brought by this customer journey. All of the contributions are summed up and the final contribution of the channel is obtained.

2.4.3.4 Evaluation of attribution models

As it is clear from the section above, there are several ways of estimating the channel contribution to conversion generation even when only the data-driven approaches are considered. As the best one is wanted, certain criteria to judge those models on are needed. Commonly discussed criteria [47] [54] [60] are fairness, interpretability and data-driven basis.

There are several more criteria like computational efficiency or versatility. Investigating these properties is beyond the scope of this thesis.

However, all of the approximation methods presented in the section above are data-driven. Fairness is difficult to judge among these models and it is a matter of certain level of subjectivity, but in general should be ensured by dividing the conversions or their values according to channel's contribution. Interpretability for marketers is ensured by presenting clear set of rules or weights for particular channels.

Another quality to judge the model on is prediction accuracy. It aims to determine which of the compared models is able to distribute the advertising budget more efficiently because it is able to capture the reality better. Papers [54] [60] have done the comparison based on the other criteria.

2.4.3.4.1 Goodness of fit test

One of the methods that can be used for assessing the accuracy is goodness of fit tests. This method is used very often for regression models in the form of the well-known R^2 or adjusted R^2. These methods aim to describe how much of the variation in the data is captured (explained) by the model.

As the aim is to compare several different models which results were estimated by different approaches, it can be difficult to use such a specific metric. Even though there is goodness of fit metrics even for Markov chains [60], for example.

2.4.3.4.2 Training-test data split test

Another, commonly used, approach is to divide the dataset into two parts. The first part is used for training the model. With this data, the model is fitted and estimates are returned.

The second portion of the dataset is used for evaluating the model. The response value with the model can be predicted and compared to the actual observed value. As the result of this prediction is basically the likelihood of the conversion occurence, a certain threshold value for which the customer journey is considered to end with conversion is set and the likelihoods are translated into the prediction of conversion or non-conversion.

Then it can be seen how many wrong prediction the model performed. This is the classical misclassification error approach. As it is not important whether the misclassification was false negative or false positive (both would lead to wrongly distributed budget in the end), the misclassification metric is:

$$misclassification\ error\ = \frac{false\ negatives + false\ positives}{test\ dataset\ observations}$$

Equation 8 Misclassification error

2.4.3.4.3 Experiment

The third approach would be to perform a real-world experiment. Firstly, the model is estimated, the budget distribution is changed and then the change of overall revenues is observed.

2.4.4 Summary

In this part of the thesis the environment of digital advertising and marketing as a whole were shortly introduced. The advantages of digital environment for marketing were presented, and further it was outlined how measurement is performed and what its weaknesses and pitfalls are.

Next, the problem of attribution and the motivation of solving such a problem were introduced. Commonly known simplistic approaches were presented along with its critique.

Furthermore, data-driven approaches were shown along with the detail description of the selected ones from my research. Which properties are used for assessing the quality of the model were described and techniques for accuracy evaluation were proposed.

3 Attribution modelling in practice

In order to investigate attribution modelling practically, I managed to obtain data from the Czech company Seznam.cz, a.s. from its project Zboží.cz.

Seznam is a Czech Internet company operating many services on the Czech market. One of the services is the search engine seznam.cz which is a significant player on the search engine market. Its market share is roughly about 30%, depending on how it is measured [62, 63]. Other services like lide.cz (dating site), novinky.cz (news) or mapy.cz (online maps) are also in the portfolio of Seznam.cz. This results in the fact that Seznam has exhaustive information about web traffic, so it can analyze how the users interact within its services and how the customer journeys look like.

Zboží.cz is an online service comparing prices of goods across different eshops. To be present in the catalogue, a merchant needs to register and paste information about goods selling on its eshop. The information is usually transmitted via XML feed, where all necessary attributes of the goods are present including the name of the product, brand, identifier and description, and category according to Zboží.cz taxonomy system.

Zboží.cz is then able to categorize the goods from all merchants and compare them by price or by quality score. This score consists of availability of the product, eshop reviews, price, and other signals including the bid set by a merchant. Merchants can influence visibility of their products on Zboží.cz by setting a bid they are willing to invest to obtain a visitor to their eshop.

The bidding system simply provides the merchants with a possibility to be on higher positions on the list when sorted by the quality score. The higher bid the merchant sets, the higher he is displayed on the list and the more clicks and consequently traffic he obtains.

The visitor is already familiar with the price and with the attributes of the goods and therefore it is typically highly probable that he will purchase the goods in the eshop.

From the perspective of Zboží.cz, a click on the button leading to a redirect to merchants eshop is considered to be the conversion action, because in that moment Zboží.cz is entitled to charge the fee.

3.1 Motivation of Zboží.cz traffic data investigation

Zboží.cz business model relies on the merchant’s website button clicks from visitors. Therefore it is in the best interest of Zboží.cz to understand their visitors (customers) journeys and to know how different traffic sources are valuable in the process of generating eshop clicks. Even though from the perspective of Zboží.cz, some of the traffic sources are not paid because some of them belong to the same company, the advertising slots could be used for another purpose and therefore the costs could be expressed as opportunity costs.

For example, the impression from news portal Novinky.cz could be sold via one of the RTB systems or there can be displayed advertisement for a different service of Seznam.cz.

The final revenue depends on the total volume of visitors of Zboží.cz website and the user experience with the website at the same time, because it influences the likelihood of the visitor returning when another shopping situation arises and also the likelihood of choosing the product and shop and make the final step which is redirect to the merchant.

From the other perspective, the total revenue of Zboží.cz depends on the number of merchants promoting on Zboží.cz and their level of satisfaction with the quality of the traffic they receive. The quality, measured as likelihood or readiness to buy the product on merchant's website, is crucial for the merchants and it influences the amount they are willing to offer in the form of bid for the Zboží.cz traffic. The way the merchants evaluate the performance of the traffic is also important. For example, which attribution model they use. Luckily for Zboží.cz, even if the merchants tend to use the simplistic, but widely used, last-click attribution model, traffic from Zboží.cz relatively often seems like well-performing. That is because of the nature of price comparison services traffic source – they tend to appear at the end of a customer journey and therefore perform well when evaluated by the last-click model.

Zboží.cz wants to optimize its traffic mix in order to maximize their profit and that is the reason for traffic data analysis. We will explore the data in order to understand the customer journeys better and then we will estimate the channel importance.

3.2 Data description

The dataset that will be used for this analysis consists of two parts for each of the two time periods. The first part consists of the information about Zboží.cz visits. From this dataset, we are able to see which users interacted with the website and performed a visit i.e. clicked on some advertisement, typed the URL of Zboží.cz in the address bar etc. These interactions are classified by the type of interaction. It can be either a general interaction with the website i.e. page visit, or even a click on the button that takes the user to the merchant's eshop.

The second part consists of the data about interactions with advertisements promoting Zboží.cz on the websites of Seznam.cz services portfolio. There are two types of interactions – impressions and clicks in the dataset.

In both datasets, there is a unique identifier that identifies users across Seznam.cz websites based on cookie, timestamp, and additional information about the interaction, such as URL address where the interaction was performed or to which URL should the user be redirected in case of click interactions. Based on these addresses, there is also a category of the source according to Seznam.cz categorization.

After a few discussions, I obtained this pair of datasets for two time periods. First – from the beginning of November until the end of December 2016, and second – from the beginning of July until the end of August 2017. These two time periods were chosen in order to have more information about user behaviour during the summer when the activity of users on the

Internet and their willingness to buy intuitively tend to be lower and the period around Christmas which is told to be the main season for the majority of mainstream merchants.

The length of the period was set to two months because, according to common beliefs, most of the ecommerce verticals have the majority of customer journeys 30 days long maximum. The length of the journey and how it is treated will be discussed later.

The datasets obtained from Zboží.cz take a form of CSV (comma separated values) files. This is a widely used format for tabular data, because it is easily readable by humans. On the other hand, there is no precise standard what they should look like, which makes them harder to use. Common problems are strings handling and how to distinguish between strings and other value formats, separator use or text encoding.

Overall package of four files takes up about 6,4 GB of disk space and consists of 28 million of rows.

3.3 Data treatment

I obtained the data from Seznam.cz in a compressed form, total of 504,3 MB, so there was no problem to download it via the Internet. When I first tried to open the uncompressed data I experienced the first big data problems. Microsoft Excel for Mac 2011 was unable to load the full contents of the files and loaded only about one million of rows. Text file editor Sublime Text 2 was unable to load the file and froze.

Then I figured out that it will be easier to use terminal commands like *head*, *tail*, *cut* or *wc* in order to get an idea what the data look like. These commands are way faster and they are sufficient in many cases.

Then I tried to use statistical program R and the environment RStudio to open the CSV files. It takes from about five minutes to thirty seconds to open a file consisting of ten million rows, depending on the method used for the opening. R loads the entire dataset into RAM [64] and it could be a problem on machines with smaller memory available.

In the end, I decided to load the data in PostgreSQL relational database. It stores the data on hard drive and loads them into memory on demand. It allows me to query the database for specific subsets of data, load it into memory and then work with this separate piece of data efficiently.

3.4 Analysis setup

I am performing the analysis on my laptop computer. It is not a scientific machine optimized for the purpose of handling big data or performing statistical analysis. It is Macbook Air Mid 2011 with 1,8 GHz Intel Core i7 with 4 GB 1333 Mhz RAM and 250 GB of SSD hard drive.

In the end, I decided to use combination of PostgreSQL database which is open-source, free to use database and Python, which is multi-purpose programming language often used for statistical analysis, machine learning and related areas. For those purposes many packages are used. For handling tabular data I used Pandas package, for producing plots I used Matplotlib library.

Some of the basic data transformation is better to perform in terminal programs like cut, head, wc, sed or awk [64]. These programs are less intuitive, however very efficient and fast. This is a very desirable property when working with large amount of data.

For some preprocessing operations, terminal program csvkit [65] was used.

3.5 Data sanitization

I figured the CSV files contained unescaped quotes inside quoted string values. To indicate that the value is a character string, double quotes are usually used. If the string contains double quote, it could lead to a problem, as the programs are not able to determine where the string value ends.

For this purpose, so-called escaping is used. It means that if the string itself contains double quote (or corresponding character indicating string value), it is preceded with another special character, which gives the double quote its usual, non-special, meaning. There is no standard on how to do it. Some programs use double quote for escaping the double quote, others use backslash character for such purpose.

Therefore I have written a Python script which adds one double quote before every double quote inside a string value using regular expressions. I have also written a similar script for replacing the missing values, which were encoded as “\N” in the original files I obtained, with empty strings, which is prefered by PostgreSQL within CSV format [66] [A1].

This finally enabled me to insert all the rows in the database.

3.6 Data overview

I imported the four files in four database tables with names “sluzby” for data about campaigns on Seznam.cz websites portfolio and “zbozi” for data about interactions directly on Zboží.cz website. These names are followed by “16”, respectively “17” to distinguish between data for time period in year 2016 and 2017. After the first transformations, which were undertaken in order to clean and normalize the data to be able to import them to the database (importing scripts are a part of appendices [A2]), I ended up with tables that will be the basis for my analysis.

I was considering decomposing the tables into more tables, for example for categorical variables, but it is not necessary for this purpose.

3.7 Description of tables “zbozi”

Tables “zbozi16” and “zbozi17” contain the data about interactions directly on Zboží.cz website for the time periods from the beginning of November until the end of December 2016 and from the beginning of July until the end of August 2017 respectively.

These tables contain 11 columns of original data and one artificial column with unique identification number of the interaction. The rows represent particular interactions.

Table 3 Description of “zbozi” tables columns

Column name	Description	Example
id	The unique integer identifier for each interaction with the Zboží.cz website in the given time period	3304
cookieid	This identifies the cookie that was sent with the HTTP request. This will be used as a user identifier, despite the fact that it is not absolutely precise due to reasons described above.	-1000800877469163709
counter	The integer number of the visit	2
datetime	The timestamp of when the interaction occurred	2016-12-01 12:43:14.51
interaction	There are two types of interactions: *“impress”* – a user request for one of the Zboží.cz’s pages, *“click-to-shop”* – a user click on the button that redirects him to a merchant’s eshop	impress
url	The URL of the page where the interaction took place	https://www.zbozi.cz/hledani/?q=vlo%C5%BEky%20do%20ly%C5%BEa%C5%99sk%C3%BDch%20bot&strana=6
sourcecategory	The category of the traffic source according to Zboží.cz‘s classification. The full list of categories can be found in [A3].	Seznamácké weby/Proženy
firsturl	The URL of the first interaction with the Zboží.cz website	https://www.zbozi.cz/hledani/?q=%C5%BEidle%20k%20psac%C3%ADmu%20stolu%20pro%20d%C4%9Bti#utm_source=search.seznam.cz&utm_medium=hint&utm_content=items-opesBB&utm_term=%C5%BEidle%20k%20psac%C3%ADmu%20stolu%20pro%20d%C4%9Bti
referrer	If the first interaction of the visit was initiated by a click from another website, the referrer stores its URL.	http://m.facebook.com
product	If the interaction is with a product page, then it stores the name of the product.	IR kamera VERIA I515-C700
category	If the interaction is with a product page, then it also stores the name of the category of the product.	IP kamery
fee	If the interaction is of the ‘click-to-shop’ type, this column stores the integer amount of fee charged to the merchant in hundreds of Czech crowns.	1672

Table 4 Descriptive statistics about columns in tables “zbozi” and comparison of 2017 and 2016 dataset

Characteristic	**Dataset 2016**	**Dataset 2017**
Number of cookies identifiers	314346	399206
Number of cookies that performed at least one click to shop	133474	166449
Average fee per click to shop	227	268
Number of traffic source categories	23 (including NULL)	24 (including NULL)
Average number of visits per cookie	1,448	1,338
Average conversion rate	20,82%	20,87%

3.8 Description of tables “sluzby”

Tables “sluzby16” and “sluzby17” contain the data about advertisements on websites of Seznam.cz. The data set contains mainly the information about displayed advertisements and in less than 1% the information about clicks on the advertisements for the time periods from the beginning of November until the end of December 2016 and from the beginning of July until the end of August 2017 respectively.

These tables contain 7 columns of original data and one artificial column with unique identification number of the interaction. The rows represent particular interactions.

Table 5 Description of "sluzby" tables columns

Column name	Description	Example
id	The unique integer identifier for each interaction with the Zboží.cz website in the given time period	3304
cookieid	This identifies the cookie that was sent with the HTTP request. This will be used as a user identifier, despite the fact that it is not absolutely precise due to reasons described above.	-1000800877469163709
service	The name of the website where the interaction took place. This column takes values of: *"fulltext"* *"firmy"* *"hp"* *"novinky"* *"sbazar"* *"prozeny"* *"sobrazky"*	firmy
interaction	This describes whether the advertisement of Zboží.cz was displayed or clicked. This column takes values of: *"mousedown"* *"impress"*	impress
datetime	The timestamp of when the interaction occurred	2016-12-01 12:43:14.51
url	The URL of the page where the interaction took place	https://www.firmy.cz/detail/349736-schorov-obecni-urad-schorov.html#utm_source=search.seznam.cz&utm_medium=hint&utm_term=obc%20scho%C5%99ov&utm_content=search
href	If the interaction is *"mousedown"*, this column holds the URL to which the user is redirected	https://www.zbozi.cz/obleceni-a-moda/?utm_source=HP_Seznam&utm_medium=seznam_sluzeb
query	If the service is *"fulltext"*, this column holds the search query the user inserted in the search field	Pandora

Table 6 Descriptive statistics about columns in tables “sluzby” and comparison of 2017 and 2016 dataset

Characteristic	Dataset 2016	Dataset 2017
Number of cookies identifiers	34104	30356
Number of cookies that performed at least one click on advertisement	23471	19265
Number of impressions	9923514	9947878
Average number of impressions per cookie	291	328
Average click through rate	0,77%	0,52%
Number of websites present in the sample	7	7
Average visits per cookie	1,448	1,338

The datasets “zbozi16” and “sluzby16” share 33847 cookie identifiers. It means that 33847 cookies performed at least one visit of the Zboží.cz website were also present on a different Seznam.cz service and an advertisement promoting Zboží.cz was displayed to them. The number for “zbozi17” and “sluzby17” is 30094 shared cookies.

3.9 Data selection for analysis

For the purpose of the analysis for attribution model, the data need to be transformed into so-called customer journeys consisting of interactions of a visitor and Zboží.cz. This interaction can take two forms. The first one is displaying some of the Zboží.cz webpages, from the perspective of advertising systems this type of interaction is often called the click. This can be misleading as the user who clicks on the link does not necessarily need to reach and display the webpage or in the case of “direct” channel it’s not actually click, as the user perform this interaction by typing the address in the address bar in the web browser.

The second form of interaction is displaying an advertisement promoting Zboží.cz. Regrettably, I do not have the data about visibility of these so-called impressions. As mentioned above, there is a big discussion about the topic of visibility. The main point is that the impressions that were not properly shown to the user can hardly influence his consumer behavior. As I do not have any further information about the visibility, I need to make a simplifying assumption that all of them were visible enough to be influential.

Both of these types of interaction stored in the tables “zbozi16” and “sluzby16” (“zbozi17” and “sluzby17” respectively) I merged into the tables “mergedinter” and inserted columns “cookieid”, “datetime”, “interaction”, “sourcecategory” and “fee” into columns of the same name. Columns “cookieid”, “datetime” and “interaction” from the table “sluzby” were inserted into the same columns of the table “mergedinter” and the column “service” was inserted into the “sourcecategory” column.

A new column “advertisement” was created which indicates whether this interaction was a displayed advertisement or an interaction with the site (this column distinguishes between the table from which the row originated).

Basic website interactions are encoded as “impress” string of characters, which is the same as encoding of a displayed advertisement promoting Zboží.cz. For clarity, I changed the “impress” string to “hit” for the interactions of the website, which is also a commonly used term for the visit of a single webpage of the website.

To construct the customer journeys representation I needed to define the interaction touchpoints from the data.

Firstly, there is a problem concerning incomplete information about impressions and even website hits. As some of the traffic sources (advertisement platforms) do not offer the option to track advertisement impressions, merchants are only able to track website hits. On the other hand, if there is a significant number of impressions prior to the website hit for some of the sources, the results will be biased. Luckily, most of the traffic originates on the Seznam.cz’s websites portfolio and I have the data concerning the impressions. Big part of the data is direct – it means that there were no impressions from this source prior to the visit; and big part of the traffic originates from search engines – for those, there is usually significantly higher click through rate and therefore the effect of the impressions is not so important. The reason also is that advertisements in search engines are purely textual and therefore the brand awareness potential (usually considered to be influenced by advertisement impressions) is not so significant.

The difference between the number of impressions and hit interactions will be also reduced by merging the interactions of the same source that are following each other in single interaction. It means that if there are several impressions with the same value in the column service, then it will be represented as a single interaction.

This reduction is done by removing all the rows that represent the same traffic source which are following each other, except for the last one.

Secondly, there is an inconsistency between values in the “service” column from the original table “sluzby” and the “sourcecategory” column from the original table “zbozi”. From the investigation of interaction patterns and consultation with traffic specialist from Seznam I figured out a few relations of these two categorizations and I unified them in order to avoid multiple interaction from the same source in a row.

Table 7 Unification of the traffic sources

Original category	**Possible interactions**	**Replaced value of category**
firmy	click	firmy
prozeny	click, impress	prozeny
sobrazky	click, impress	sobrazky
novinky	click, impress	novinky
sbazar	click, impress	sbazar
null	click, impress	null
Seznamácké weby/Link	click, impress	hp
Seznamácké weby/Novinky	click, impress	novinky
PPC/Facebook	click, impress	social
Seznamácké weby/Proženy	click, impress	prozeny
PPC/Sklik	click	ppc-sklik
Hinty/Product cards	click, impress	fulltext
Social (FB)	click	social
Seznamácké weby/Ostatní služby SZN	click, impress	sostatni
Bannery (selfpromo)	click, impress	sostatni
Vlastní návštěvy	click	direct
Direct	click	direct
hp	click, impress	hp
PPC/AdWords	click	ppc-adwords
SEO/Seznam	click, impress	seo-seznam
Seznamácké weby/Obrázky	click, impress	sobrazky
Seznamácké weby/Sbazar	click, impress	sbazar
Hinty/PI	click, impress	fulltext
SEO/Global (Google a ostatní)	click	seo-global
Seznamácké weby/Sport	click, impress	ssport
fulltext	click, impress	fulltext

Then I iterated through all the cookies and transformed the dataset of interactions into a dataset of journeys. The beginning of the journey is either the first interaction in the dataset or the first interaction of the cookie after a conversion. Because of a conversion character, which is only the click on the merchant's button, I grouped together all the conversions within one hour from the first click-to-shop interaction and I made a sum of their value, which is a fee paid by an advertiser.

I also needed to filter out the “hit” interactions as they are not marketing channels interactions. However, these interactions were useful to detect the preceding “click” interaction, which cannot be reliably tracked due to technical reasons mentioned above. The rule is that if there is a “hit” interaction which is not preceded by a “mousedown” interaction, the first “hit” interaction should be used to express the “click” interaction and the rest of the “hit” interactions should be dropped.

Then I constructed the journeys in a usual format. The channel is represented by the unified source name and interaction type (either “click” or “impress”) and these two parts are divided by a slash sign. The journeys is constructed as a chain of channels divided by “greater than” sign.

I was working with big amount of data. To process the dataset for one time period takes about 10 hours on my computer. At this point, I was investigating parallelization possibilities, but due to the implementation difficulty I decided to perform the algorithm as it was.

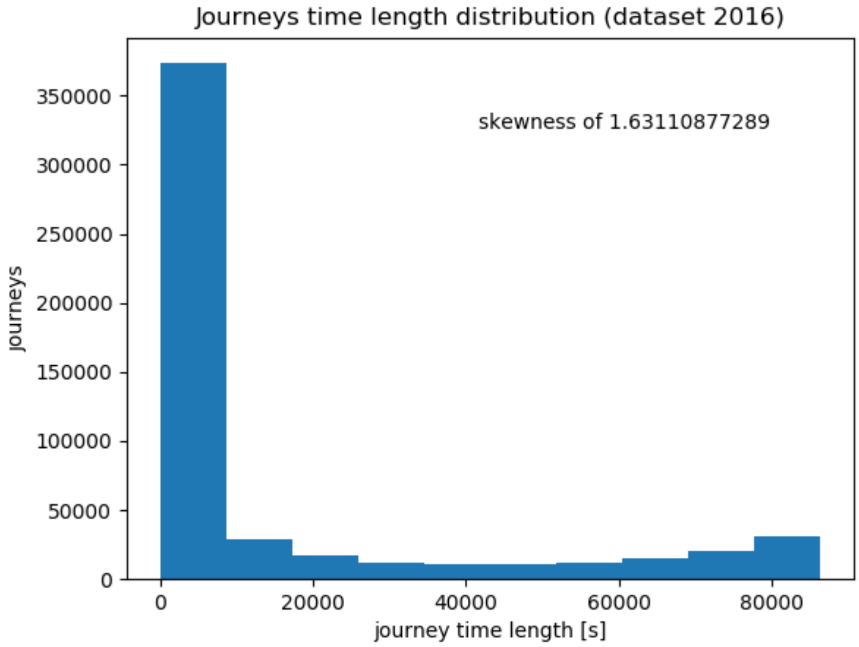
3.10 Customer journeys processing

I decided to implement Shapley value attribution algorithm, as it is one of the industry’s mostly used approach. However, even when using this algorithm, there are multiple ways of implementing it.

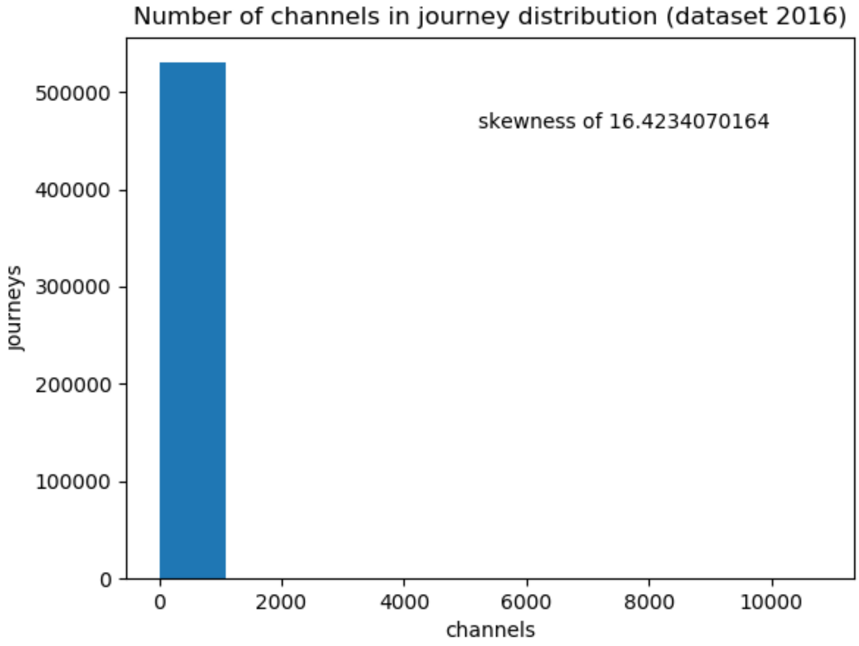
First of all, the computational complexity of this algorithm is in #P-complete class, which means that its computational complexity is at least polynomial. In practice, it makes the algorithm unusable on big datasets if the number of channels is more than 15-20 [69]. However, there is also an approximation algorithm, whose computational complexity is linear and uses a randomization approach [70].

In order to make the computation feasible, I used the algorithm variant that operates only on the present data and not on all the theoretically possible combinations. This allows me to reduce the complexity, even though it had to be done by reducing the data that the algorithm used for the computation.

Firstly, since the original data are time intervals, it is not possible to determine whether we have broken the customer journey in the middle. It biases the results, and to eliminate this bias I tried to filter out the journeys of extreme duration by keeping the 90% of the journeys around average. However, as it turned out, the journey duration distribution is positively skewed so much, that 5% percentile has a value of 0 seconds of duration. It basically means that I was able to filter out only the extremely long journeys and I was unable to identify the journeys broken shortly before converting. It turned out that it is quite frequent that cookie converts after only one interaction.



Picture 12 Journey time length distribution for the dataset from 2016



Picture 13 Number of channels in journey distribution for the dataset from 2016

From the positively skewed charts above, it is clear that the majority of the journeys tends to be shorter than the average length. I leveraged this fact by dropping the journeys that occured in the data only once. This reduces the number of unique journeys, which are used for the computation, from 49720 unique journeys to 3606 (7,3%); but the sum of fees captured by this selection of data dropped from 1417369 CZK to 1275536 CZK (90%). As a big number of channels in the journey increases the computational complexity of Shapley value disproportionately, it seems like a reasonable tradeoff of precision and computational time.

Another property of Shapley value algorithm is to decide what should be used as a payoff function. Some experts consider the conversion rate of the particular journey to be the best payoff function [69]. In my point of view, it depend on the precise purpose of attribution model. If it aims to divide the conversions generated in the given time period, the amount of conversions or their value would be a better measure. On the other hand, if we want to compare the performance of the channels, it is not fair to judge the channels performance based on the conversions volume, because it can be highly influenced by the budget spend on the channels present in the journey. For the fair budget allocation in the future, I consider the conversion rate to be a better measure. If we also want to take into account the monetary performance, I would consider the conversion rates multiplied by the average conversion value a better measure. This ensures that channels that contribute to better paying customers are rewarded for such benefit. It is quite easy to change the payoff function in the Shapley value implementation.

After selecting the data for the computation as specified above, it takes only about few second to compute the Shapley values on my computer.

As a second attribution model, I decided to use widely criticised but also widely used last-interaction model in order to be able to compare the differences of weights.

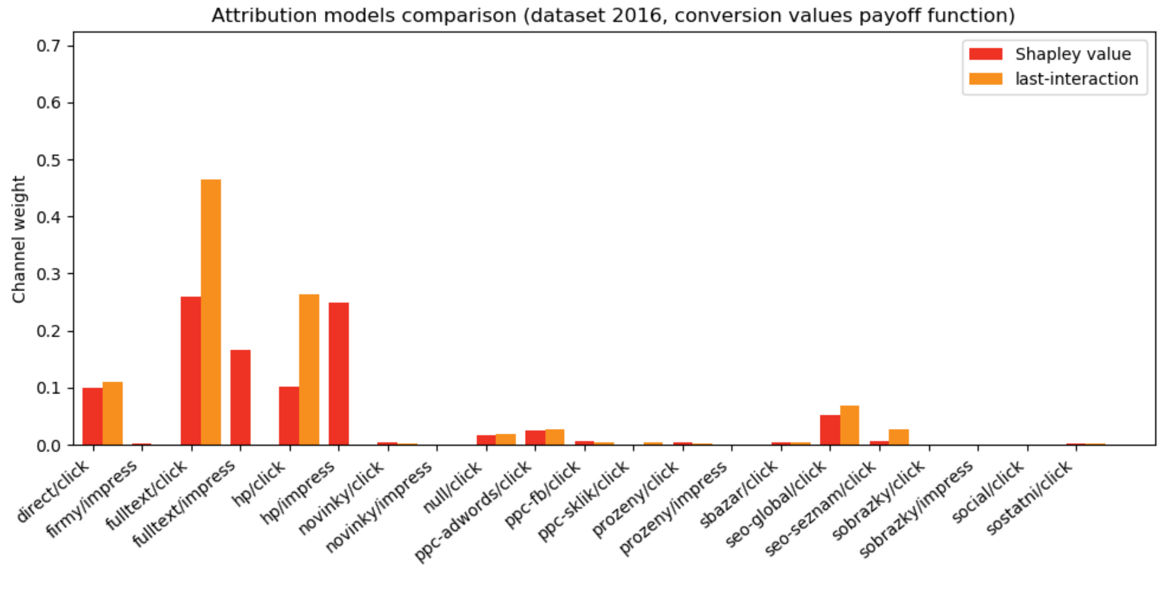
In order to preserve the same conditions as were set for the Shapley value algorithm, I filtered out the long and rare journeys as described above. However, in this case it did not cause any problem to include them as this approach is significantly faster to compute. In order to group the data for last-interaction attribution, I searched for the last interaction in the journey and I made the sum of conversion numbers and values.

3.11 Comparison of attribution models results

Firstly, I will present the results of Shapley value and last-interaction model with payoff function based on the contribution to overall conversion values. In my opinion, this payoff function is more suitable for the situations when we want to assign the proportion of conversion values to the marketing channels ex-post. Then it would be a good idea to consult these results with information about spending on the particular channels.

Table 8 Attribution model comparison for dataset from 2016 and conversion values payoff function

Channel	**Shapley value attribution**	**Last-interaction attribution**	**Difference**
direct/click	9,92%	10,97%	-1,05%
firmy/impress	0,13%	0,00%	0,13%
fulltext/click	25,87%	46,46%	-20,60%
fulltext/impress	16,68%	0,00%	16,67%
hp/click	10,22%	26,34%	-16,12%
hp/impress	24,88%	0,01%	24,87%
novinky/click	0,38%	0,23%	0,16%
novinky/impress	0,06%	0,00%	0,05%
null/click	1,68%	1,95%	-0,27%
ppc-adwords/click	2,61%	2,64%	-0,02%
ppc-fb/click	0,58%	0,37%	0,20%
ppc-sklik/click	0,00%	0,53%	-0,53%
prozeny/click	0,36%	0,17%	0,19%
prozeny/impress	0,11%	0,00%	0,11%
sbazar/click	0,45%	0,36%	0,09%
seo-global/click	5,22%	6,77%	-1,55%
seo-seznam/click	0,62%	2,81%	-2,19%
sobrazky/click	0,00%	0,02%	-0,02%
sobrazky/impress	0,00%	0,00%	0,00%
social/click	0,00%	0,06%	-0,06%
sostatni/click	0,23%	0,29%	-0,06%


Attribution models comparison (dataset 2016, conversion values payoff function)

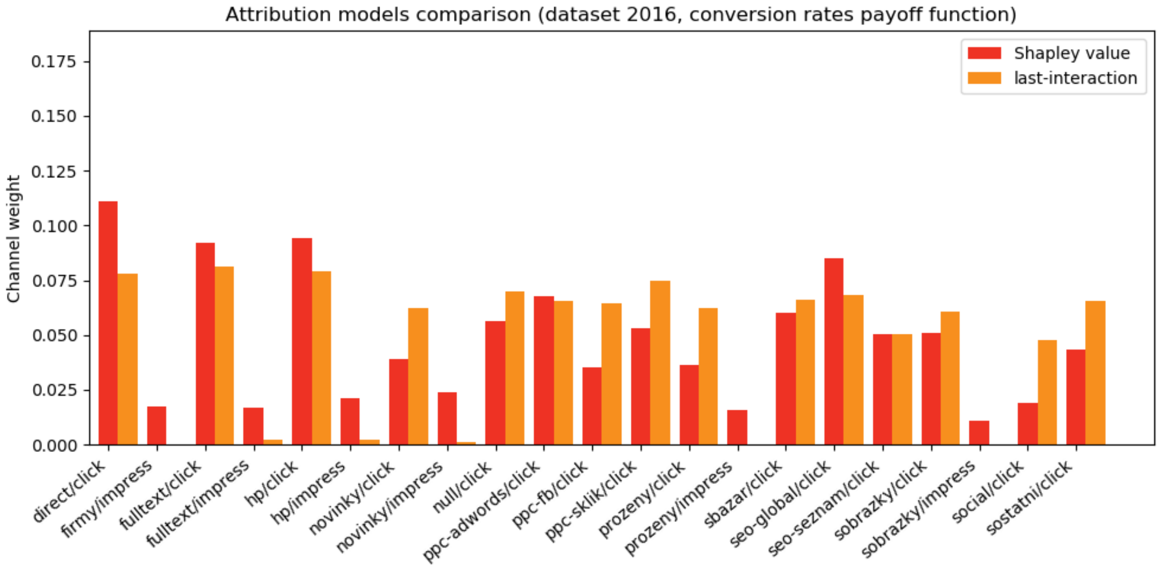
Picture 14 Attribution models comparison for dataset from 2016 and conversion values payoff function

From the results above, it is clearly visible that Shapley value assigns bigger weight to impression channels. However, this is caused mainly by the fact that the Shapley value is a multi-touch attribution model in contrast to the last-interaction model. By its definition, it is necessary to enter the website (in other words to click) to be able to perform the conversion. It means that click interactions appear at the end of the customer journey and therefore the last-interaction model assigns the credit mainly to “click” channels. I would recommend the Shapley value with conversion values payoff function to use for assigning the conversion values for the corresponding time period and for evaluating the efficiency with respect to costs for the traffic acquisition.

The results, if we switch to use the conversion rate as a payoff function, look as follows.

Table 9 Attribution model comparison for dataset from 2016 and conversion rates payoff function

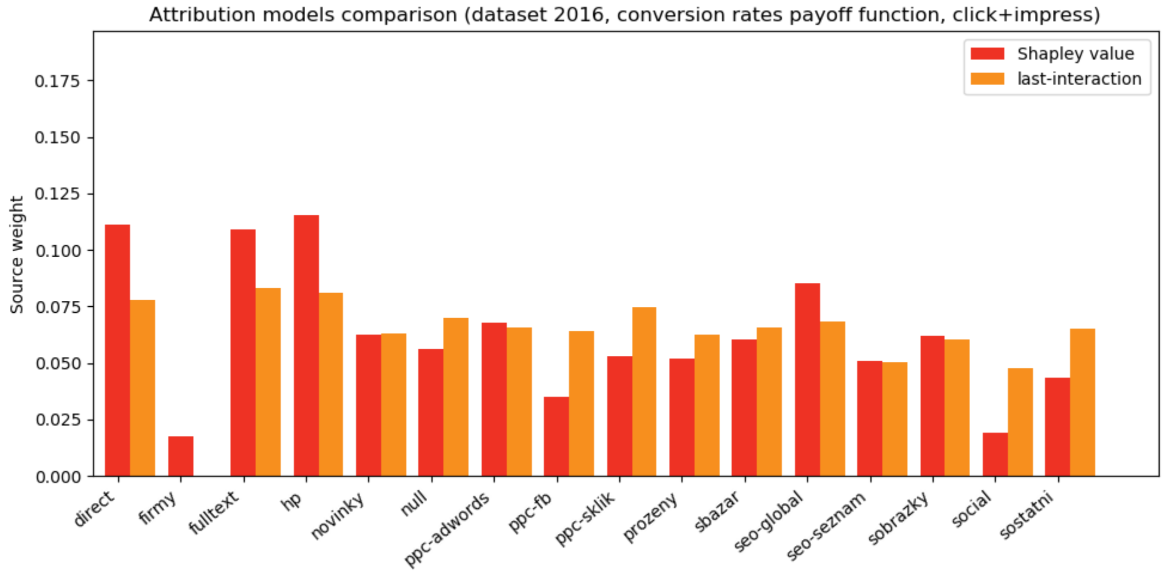
Channel	Shapley value attribution	Last-interaction attribution	Difference
direct/click	11,10%	7,78%	3,31%
firmy/impress	1,73%	0,00%	1,73%
fulltext/click	9,23%	8,10%	1,13%
fulltext/impress	1,66%	0,23%	1,43%
hp/click	9,44%	7,89%	1,55%
hp/impress	2,11%	0,22%	1,89%
novinky/click	3,91%	6,22%	-2,31%
novinky/impress	2,36%	0,09%	2,27%
null/click	5,64%	6,99%	-1,35%
ppc-adwords/click	6,75%	6,55%	0,21%
ppc-fb/click	3,50%	6,43%	-2,93%
ppc-sklik/click	5,29%	7,45%	-2,16%
prozeny/click	3,64%	6,26%	-2,61%
prozeny/impress	1,56%	0,00%	1,56%
sbazar/click	6,01%	6,58%	-0,57%
seo-global/click	8,53%	6,80%	1,73%
seo-seznam/click	5,06%	5,05%	0,01%
sobrazky/click	5,10%	6,05%	-0,95%
sobrazky/impress	1,10%	0,00%	1,10%
social/click	1,90%	4,77%	-2,87%
sostatni/click	4,36%	6,53%	-2,17%



Picture 15 Attribution models comparison for dataset from 2016 and conversion rates payoff function

When conversion rates are used as payoff function, we can observe much more equally distributed values across the channels, and the difference between models is also much lower. However, the impression channel's relative underestimation persists. More equally distributed weights are caused by the fact that when using conversion rates as payoff function, results are not influenced by the effect of unequally distributed media budget and the amount of interactions that are issued.

As the last-interaction tends to assign the credit to click interaction, the chart above should not be used for comparison of the attribution models. Therefore I made a sum of the weight of click and impression interaction of corresponding channels in order to see the difference better.



Picture 16 Attribution models comparison for the dataset from 2016 and conversion rates payoff function with summed up click and impression interactions

From this chart, it is possible to draw the commonly proclaimed conclusion, which is that the last-interaction attribution model tends to underestimate channels that tend to be at the beginning of the customer journey and conversely overestimate channels that tend to be at the end of the journey. To channels like "hp", or "fulltext" (as in this setting it receives a lot of impression interactions), the last-interaction assigns lower value than Shapley value model. In contrast to "ppc-sklik", "seo-seznam", "ppc-adwords", "ppc-fb" (as it is probably used primarily for retargeting) to which the last-interaction model assigns lower or similar value as Shapley value model.

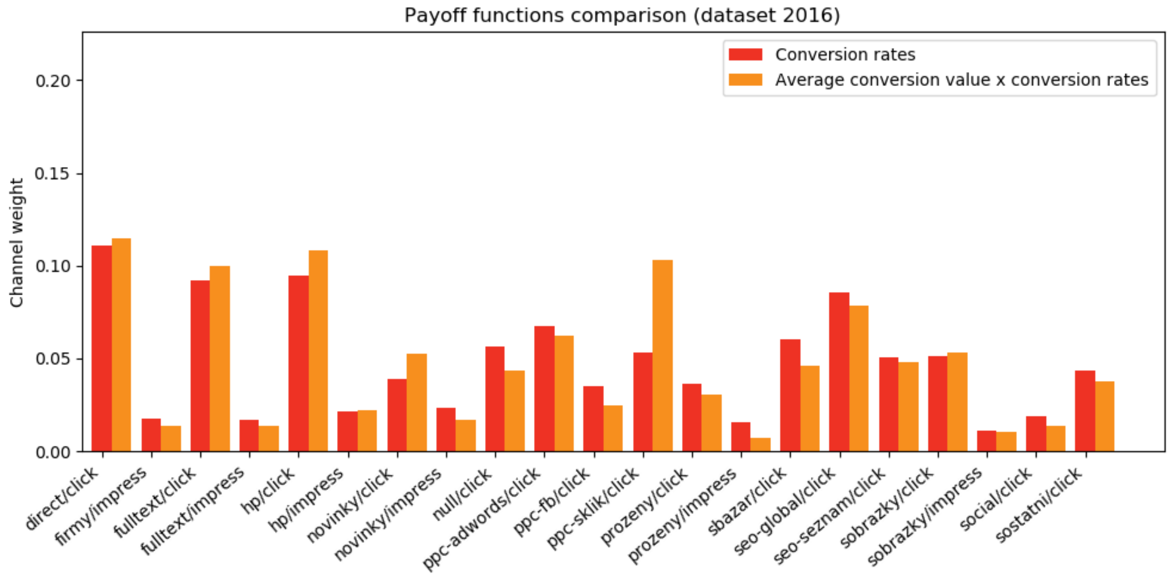
I am surprised by the results for "direct" and "social" channel. Direct channel is usually considered to be the typical channel for the end of the journey, but the Shapley value model is assigning significantly higher value to "direct" than the last-interaction model here. It might be caused by the fact that "direct" is not only a channel which is used for finishing the conversion after initiating the journey from other traffic source. In the case of Zboží.cz and Seznam.cz, well known and long-term operating online entities, it is possible that users initiate the whole process by typing the address directly into the address bar, or at least it is easy for them to restart the purchase process by typing in the address. Generally speaking, for such a well-known brand, it is not rare to have "direct" as a channel also on the other position than at the end of the journey, and it causes that it is also highly valued by the multi-touch Shapley value.

To speculate on the relatively big difference in models' evaluation for "social" channel, I would expect that "social" value will be higher with multi-touch attribution model, because interactions with organic social messages are usually not initiated by the user, but rather they depend on whether something new was posted on the page or whether algorithm of that particular social network decides to display the message. In this case, however, the last-interaction model valued "social" channel higher which might be caused by the fact that the social network has enough data, that it displays Zboží.cz's organic messages at the moment when the user has already begun the journey, or by the fact that it is not tracked properly and this channel contains also a part of paid social traffic that, I assume, is mainly retargeting.

From the above presented data, I can conclude that the median value of the rate between click and impression channel for the channels that have both variants in the model presented is about 4,5. It means that the click interactions are typically 4,5 times more valuable than impression interactions. This is a number that I would expect to be significantly higher. From the practise, I experience that the rates between prices for click and impression interactions are about 300-1200, meaning that the clicks are usually at least 300 times more expensive than impressions. These numbers are only rough estimates, as it is the median or the average value of not precisely specified conditions. However, the difference is of more than two orders.

I suspect that it has something to do with a high correlation of impression and click interactions of corresponding channels. As it applies to many channels that an impression should precede a click, it would lead to splitting the attribution between the impression and the click channel variant. It could be a subject of further discussion.

I can also compare the data for a simple conversion rate payoff function and a conversion rate multiplied by the average conversion value, which would, in my opinion and under given circumstances, reflect the best which channels are the most valuable.



Picture 17 Payoff functions comparison for the dataset from 2016 and for conversion rates and the average conversion value multiplied by conversion rate payoff function

From the chart above, it is visible that payoff function, which takes into account the average conversion value, leads to assessing bigger value to “ppc-sklik” channel. This might be caused by the fact that this traffic source is easily influenceable and it is probably optimized to acquire mainly traffic for categories that have bigger average conversion value. I would recommend this model to use for a future budget allocation optimization, despite its drawbacks, which will be described in the last chapter.

4 Conclusions

In this chapter I would like to summarize the results of the practical part of this thesis, outline the possible pitfalls and areas of further investigation, the value of the analysis and summarize my personal takeaways.

I implemented two attribution models for the purposes of comparison. One is simplistic, but still very often used last-interaction attribution model. This model assigns the conversion and its value to the last channel in the customer journey. Second one is much more complex, multi touch, data-driven Shapley value attribution model. This model distributes the conversions and their value to multiple marketing channels that are along the customer journey. Moreover, the weight it assigns to all of the channels is calculated from the rest of the journeys to ensure that the portion is fair. This approach has its roots in a game theory and is used in the data-driven attribution model of Google Analytics.

It turns out that it is hard to compare the results of these two models, because the last-interaction attribution model assigns virtually no value to impression interactions (channels). Therefore I summed up the values for every traffic source and it turned out that sources that tend to be on the beginning of the customer journey are being underestimated by the last-interaction attribution model.

I also computed the results for different payoff functions. The first and commonly used payoff function assigns part of the overall conversions values to particular channels. This distribution approach is influenced by the number of occurrences of the channel, which consequently means that channels that have more interactions are valued more. As the frequency of the channel in the journey can be influenced by the budget spent on that particular channel, this payoff function makes sense only in the case when we want to distribute the conversion values of the corresponding time period to particular channels and compare it with the medium costs. However, this is valid use-case and it can help analytical team in Zboží.cz to make more precise reports of their marketing activities.

However, in order to be able to make better budget allocation decisions, I consider more meaningful to use payoff function, which uses a conversion rate contribution weighted by the average conversion value. This reflects how interesting the users reached by the channel from the business perspective are. As this approach is not burdened by the previous unequal budget allocation, it suits well the decision for the future budget allocation. By leveraging those results combined with costs information, Zboží.cz traffic managers can optimize the budget allocation in order to maximize the revenues from a merchant advertising on Zboží.cz.

I am aware of the fact that average conversion value is also influenced by the channels optimization, namely some marketing channels are easier to optimize in order to acquire users interested in the goods with higher conversion values (click fees). It is important to bear in mind that they then appear as better performing in the attribution analysis. It is also important to be aware of the nature of these marketing sources and to know that the traffic acquisition is not perfectly independent.

The insights from the analysis would be even more powerful if it also used the data about costs expanded on the marketing channels. However, to join the outputs of this analysis and cost data is still feasible.

There is also a big variability of how to construct the customer journeys for marketing channels interaction data, which I have never heard of. From this variability also arise my concerns about the proper attribution value assessment in the case of impression and click interactions of the corresponding traffic sources. I assume that these interactions are highly correlated, which influences the results. To verify the presented results, the correlation should be investigated and new interaction grouping rules should be eventually proposed. However, this is beyond the scope of this thesis.

For me personally, there are several takeaways. First of all, working with huge amount of data gave me a lesson I can use later. It does not matter whether it is using command line tools to preprocess the CSV files, because it is incredibly fast in comparison with other tools, data sanitization, using SQL database, basic notion of parallelization possibilities in Python and its disadvantages, or the real consequences of computational complexity of a big amount of data.

Another takeaway is that articles about data science usually discuss the cutting edge technologies like machine learning, deep learning, neural networks, or artificial intelligence. However, most of the work often can be done by basic data manipulation, selection, filtering, grouping, and combining. It is something that is not so interesting for an article reader, but in reality it is a significant part of a data specialist. In the practical part of this thesis, it actually took the vast majority of time to preprocess, sanitize, and transform the data. Implementation of the attribution algorithm itselfs took unproportionally smaller part of the workload. When I was going through the literature, I have not read much about this part of work, but from my experience, the final result highly depends on how the data was treated in the early phases.

I believe that working with Python, SQL, data analysis library Pandas and the plotting library Matplotlib will be very beneficial in my future career. Working with digital marketing traffic data and data analysis libraries enabled me to utilize my previously gained information technology knowledge from the bachelor part of my studies.

I strongly hope that outputs of this thesis will be helpful for Zboží.cz and I will be happy to cooperate with them further to explore more insights from the available data.

I would like to further work with, refactor, document, and generalize my algorithms in order to enable more digital marketers to make smarter decisions based on the collected raw data, as I did not find any comprehensive Python library that would help with this task.

Reference list

[1]Statista Inc., "Global retail e-commerce market size 2014-2021", Statista, 2017. [Online]. Available: https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/. [Accessed: 05- Dec- 2017].

[2]Statista Inc., "Google: ad revenue 2001-2016", Statista, 2017. [Online]. Available: https://www.statista.com/statistics/266249/advertising-revenue-of-google/. [Accessed: 05- Dec- 2017].

[3]IAB Europe, "IAB Europe - Interactive Advertisiting Europe", Iabeurope.eu, 2017. [Online]. Available: https://www.iabeurope.eu/. [Accessed: 05- Dec- 2017].

[4]Kotler Marketing Group, Inc., "Dr. Philip Kotler Answers Your Questions on Marketing", Kotler Marketing Group, 2017. [Online]. Available: http://www.kotlermarketing.com/phil_questions.shtml. [Accessed: 05- Dec- 2017].

[5]G. Charlton, "Companies more focused on acquisition than retention: stats", Econsultancy, 2013. [Online]. Available: https://econsultancy.com/blog/63321-companies-more-focused-on-acquisition-than-retention-stats. [Accessed: 05- Dec- 2017].

[6]L. Burgess, "8 Expert Predictions for 2017’s Top Digital Marketing Trends", Fresh Egg, 2017. [Online]. Available: https://www.freshegg.co.uk/blog/digital-trends/8-expert-predictions-for-2017-s-top-digital-marketing-trends. [Accessed: 05- Dec- 2017].

[7]D. Hendricks, "3 Advertising Trends That Are Revolutionizing Offline Marketing", Inc., 2017. [Online]. Available: https://www.inc.com/drew-hendricks/3-advertising-trends-that-are-revolutionizing-offline-marketing.html. [Accessed: 05- Dec- 2017].

[8]A. Kaushik, "See, Think, Do, Care Winning Combo: Content +Marketing +Measurement!", Ocamm's Razor, 2015. [Online]. Available: https://www.kaushik.net/avinash/see-think-do-care-win-content-marketing-measurement/. [Accessed: 05- Dec- 2017].

[9]R. Dragon, "Who Created AIDA? | DragonSearch", Dragon SEARCH, 2011. [Online]. Available: https://www.dragonsearch.com/blog/who-created-aida/. [Accessed: 05- Dec- 2017].

[10]D. Rowles, "Using the AIDAR purchasing funnel model", Smart Insights, 2013. [Online]. Available: https://www.smartinsights.com/customer-relationship-management/social-crm/aidar-model/. [Accessed: 05- Dec- 2017].

[11]Google LLC, "Default channel definitions", Analytics Help. [Online]. Available: https://support.google.com/analytics/answer/3297892. [Accessed: 05- Dec- 2017].

[12]comScore, Inc., "For Display Ads, Being Seen Matters More than Being Clicked", comScore, 2012. [Online]. Available: http://www.comscore.com/Insights/Press-Releases/2012/4/For-Display-Ads-Being-Seen-Matters-More-than-Being-Clicked. [Accessed: 05- Dec- 2017].

[13]Visual IQ Inc., "A Click Is Not A Click Is Not A Click", Visual IQ, 2011. [Online]. Available: https://www.visualiq.com/resources/marketing-attribution-newsletter-articles/a-click-is-not-a-click-is-not-a-click. [Accessed: 05- Dec- 2017].

[14]I. Zeifman, "Bot Traffic Report 2016", IMPERVA INCAPSULA, 2017. [Online]. Available: https://www.incapsula.com/blog/bot-traffic-report-2016.html. [Accessed: 05- Dec- 2017].

[15]Google LLC, "Bid on viewable impressions using viewable CPM", AdWords Help, 2017. [Online]. Available: https://support.google.com/adwords/answer/3499086?hl=en. [Accessed: 05- Dec- 2017].

[16]B. Vogel, "Reducing Unintentional Clicks", Facebook Audience Network, 2017. [Online]. Available: https://www.facebook.com/audiencenetwork/news-and-insights/reducing-unintentional-clicks. [Accessed: 05- Dec- 2017].

[17]D. Kehrer, "The Truth About Cross-Channel Attribution In Marketing", Forbes, 2014. [Online]. Available: https://www.forbes.com/sites/forbesinsights/2014/12/02/cross-channel-attribution/. [Accessed: 05- Dec- 2017].

[18]Interactive Advertising Bureau, "IAB Attribution Hub", IAB Attribution Hub. [Online]. Available: https://www.iab.com/guidelines/iab-attribution-hub/. [Accessed: 05- Dec- 2017].

[19]IAB - Interactive Advertising Bureau, DIGITAL ATTRIBUTION METHODOLOGIES. IAB - Interactive Advertising Bureau, 2017.

[20]J. Lincoln, "What Is Digital Marketing? (Full History) + More", Ignite Visibility. [Online]. Available: https://ignitevisibility.com/what-is-digital-marketing/. [Accessed: 05- Dec- 2017].

[21]H. Tschabitscher, "When the 1st Spam Email Was Sent and What it Advertised", Lifewire, 2016. [Online]. Available: https://www.lifewire.com/when-was-the-first-spam-email-sent-1171212. [Accessed: 05- Dec- 2017].

[22]A. LaFrance, "The First-Ever Banner Ad on the Web", The Atlantic, 2017. [Online]. Available: https://www.theatlantic.com/technology/archive/2017/04/the-first-ever-banner-ad-on-the-web/523728/. [Accessed: 05- Dec- 2017].

[23]N. Leonette, "Advertising Industry Current State & Pain Points", JMC Brands, 2017. [Online]. Available: https://jmcbrands.com/blog/advertising-current-state-pain/. [Accessed: 05- Dec- 2017].

[24]J. Pelline, "Pay-for-placement gets another shot", CNET News, 1998. [Online]. Available: https://archive.is/20121208124417/http://news.com.com/Pay-for-placement+gets+another+shot/2100-1023_3-208309.html. [Accessed: 05- Dec- 2017].

[25]G. Marvin, "Google AdWords Turns 15: A Look Back At The Origins Of A $60 Billion Business", Search Engine Land, 2015. [Online]. Available: https://searchengineland.com/google-adwords-turns-15-a-look-back-at-the-origins-of-a-60-billion-business-234579. [Accessed: 05- Dec- 2017].

[26]N. Carlson, "At Last -- The Full Story Of How Facebook Was Founded", Business Insider, 2010. [Online]. Available: http://www.businessinsider.com/how-facebook-was-founded-2010-3. [Accessed: 05- Dec- 2017].

[27]A. Oberoi, "The History of Online Advertising", AdPushup Blog, 2013. [Online]. Available: https://www.adpushup.com/blog/the-history-of-online-advertising/. [Accessed: 05- Dec- 2017].

[28]Statista Inc., "Facebook ad revenue 2009-2016", Statista, 2017. [Online]. Available: https://www.statista.com/statistics/271258/facebooks-advertising-revenue-worldwide/. [Accessed: 05- Dec- 2017].

[29]O. Malik, "How is The New York Times Really Doing?", Om Malik, 2017. [Online]. Available: https://om.co/2017/02/20/how-is-the-new-york-times-really-doing/. [Accessed: 05- Dec- 2017].

[30]IAB - Interactive Advertising Bureau, DIGITAL ATTRIBUTION PRIMER 2.0. 2016.

[31]IAB - Interactive Advertising Bureau, MULTI-TOUCH ATTRIBUTION (MTA) IMPLEMENTATION AND EVALUATION PRIMER. 2017.

[32]Google LLC, "Choose an attribution model that best fits your needs", AdWords Help. [Online]. Available: https://support.google.com/adwords/answer/7002714?hl=en-GB. [Accessed: 05- Dec- 2017].

[33]Google LLC, "About conversion windows", AdWords Help. [Online]. Available: https://support.google.com/adwords/answer/3123169?co=ADWORDS.IsAWNCustomer%3Dfalse&hl=en. [Accessed: 05- Dec- 2017].

[34]A. Kaushik, "Multi-Channel Attribution Modeling: The Good, Bad and Ugly Models", Ocamm's Razor, 2013. [Online]. Available: https://www.kaushik.net/avinash/multi-channel-attribution-modeling-good-bad-ugly-models/. [Accessed: 05- Dec- 2017].

[35]G. Marvin, "Google's Report That 56% Of Ads Aren't Seen Isn't Shocking & Here's Why", Marketing Land, 2014. [Online]. Available: https://marketingland.com/googles-report-56-percent-ads-arent-seen-isnt-shocking-heres-110433. [Accessed: 05- Dec- 2017].

[36]Google LLC, "About Data-Driven Attribution", Analytics Help. [Online]. Available: https://support.google.com/analytics/answer/3264076?hl=en. [Accessed: 05- Dec- 2017].

[37]Investopedia LLC, "Shapley Value", Investopedia. [Online]. Available: https://www.investopedia.com/terms/s/shapley-value.asp. [Accessed: 05- Dec- 2017].

[38]J. Keeley, "What Are Markov Chains? 5 Nifty Real World Uses", MakeUseOf, 2016. [Online]. Available: http://www.makeuseof.com/tag/markov-chains-5-nifty-real-world-uses/. [Accessed: 05- Dec- 2017].

[39]G. Press, "12 Big Data Definitions: What's Yours?", Forbes, 2014. [Online]. Available: https://www.forbes.com/sites/gilpress/2014/09/03/12-big-data-definitions-whats-yours/#ad37fa613ae8. [Accessed: 05- Dec- 2017].

[40]G. P. H. Styan and H. Smith, “Markov Chains Applied to Marketing,” Journal of Marketing Research, vol. 1, no. 1, p. 50, Feb. 1964.

[41]F. Chierichetti, R. Kumar, P. Raghavan, and T. Sarlos, “Are web users really Markovian?,” in Proceedings of the 21st international conference on World Wide Web - WWW ’12, 2012.

[42]A. L. Montgomery, S. Li, K. Srinivasan, and J. C. Liechty, “Modeling Online Browsing and Path Analysis Using Clickstream Data,” Marketing Science, vol. 23, no. 4, pp. 579–595, Nov. 2004.

[43]S. Bryl', "Marketing Multi-Channel Attribution model with R (part 1: Markov chains concept)", AnalyzeCore by Sergey Bryl', 2016. [Online]. Available: https://analyzecore.com/2016/08/03/attribution-model-r-part-1/. [Accessed: 05- Dec- 2017].

[44]S. Bryl', "Marketing Multi-Channel Attribution model with R (part 2: practical issues)", AnalyzeCore by Sergey Bryl', 2017. [Online]. Available: https://analyzecore.com/2017/05/31/marketing-multi-channel-attribution-model-r-part-2-practical-issues/. [Accessed: 05- Dec- 2017].

[45]Zhao Li and J. Tian, “Testing the suitability of Markov chains as Web usage models,” in Proceedings 27th Annual International Computer Software and Applications Conference. COMPAC 2003.

[46]X. Shao and L. Li, “Data-driven multi-touch attribution models,” in Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’11, 2011.

[47]B. Dalessandro, C. Perlich, O. Stitelman, and F. Provost, “Causally motivated attribution for online advertising,” in Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy - ADKDD ’12, 2012.

[48]K. H. Brodersen, F. Gallusser, J. Koehler, N. Remy, and S. L. Scott, “Inferring causal impact using Bayesian structural time-series models,” The Annals of Applied Statistics, vol. 9, no. 1, pp. 247–274, Mar. 2015.

[49]Google LLC, "Google Scholar: Citation of 'Inferring causal impact using Bayesian structural time-series models'", Google Scholar. [Online]. Available: https://scholar.google.cz/scholar?cites=4836725887667944732. [Accessed: 05- Dec- 2017].

[50]E. Warton, M. Parker and A. Karter, How D-I-D you do that? Basic Difference-in-Differences Models in SAS®. Oakland, p. 3.

[51]P. E. Pfeifer and R. L. Carraway, “Modeling customer relationships as Markov chains,” Journal of Interactive Marketing, vol. 14, no. 2, pp. 43–55, Jan. 2000.

[52]A. Prasad, V. Mahajan, and B. Bronnenberg, “Advertising versus pay-per-view in electronic media,” International Journal of Research in Marketing, vol. 20, no. 1, pp. 13–30, Mar. 2003.

[53]H. Che and P. B. (Seethu) Seetharaman, “‘Speed of Replacement’: Modeling Brand Loyalty Using Last-Move Data,” Journal of Marketing Research, vol. 46, no. 4, pp. 494–505, Aug. 2009.

[54]Anderl, E. (2017). Three Essays on Analyzing and Managing Online Consumer Behavior. [ebook] Wirtschaftswissenschaftlichen Fakultät der Universität Passau. Available at: https://d-nb.info/1064147364/34 [Accessed 9 Dec. 2017].

[55]Y. Zhang, Y. Wei, and J. Ren, “Multi-touch Attribution in Online Advertising with Survival Theory,” in 2014 IEEE International Conference on Data Mining, 2014.

[56]K. Yamaguchi, "Pay Per What? Choosing Pricing Models In Digital Advertising", Marketing Land, 2014. [Online]. Available: https://marketingland.com/pay-per-pricing-models-digital-advertising-97913. [Accessed: 09- Dec- 2017]

[57]J. Friedman, "Blogging for dollars raises questions of online ethics", Los Angeles Times, 2007. [Online]. Available: http://articles.latimes.com/2007/mar/09/business/fi-bloggers9. [Accessed: 09- Dec- 2017]

[58]L. Shapley, A value for n-person games. 1953 [Online]. Available: http://www.library.fa.ru/files/Roth2.pdf. [Accessed: 09- Dec- 2017]

[59]L. S. Shapley, “17. A Value for n-Person Games,” in Contributions to the Theory of Games (AM-28), Volume II, Princeton University Press.

[60]C. H. W. Jayawardane, S. K. Halgamuge, and U. Kayande, “Attributing Conversion Credit in an Online Environment: An Analysis and Classification,” in 2015 3rd International Symposium on Computational and Business Intelligence (ISCBI), 2015.

[61]M. S. Bartlett, “The frequency goodness of fit test for probability chains,” Mathematical Proceedings of the Cambridge Philosophical Society, vol. 47, no. 1, p. 86, Jan. 1951.

[62]J. Vidim, "Infografika: Podíl vyhledávačů Google a Seznam na českém internetu #2", eVisions.cz, 2017. [Online]. Available: https://www.evisions.cz/blog-2017-11-23-infografika-podil-vyhledavacu-google-a-seznam-na-ceskem-internetu-2/. [Accessed: 18- Dec- 2017]

[63]M. Domes, "Google vs Seznam: Jaký je podíl vyhledávačů v roce 2016", Martin Domes, 2017. [Online]. Available: http://www.martindomes.cz/google-vs-seznam-jaky-je-podil-vyhledavacu-v-roce-2016/. [Accessed: 18- Dec- 2017]

[64]S. Pradeep and P. Moy, "Handling large datasets in R", Rpubs.com, 2015. [Online]. Available: https://rpubs.com/msundar/large_data_analysis. [Accessed: 24- Dec- 2017]

[65]C. Groskopf, "csvkit 1.0.2 documentation", Read the Docs, 2016. [Online]. Available: http://csvkit.readthedocs.io/en/1.0.2/. [Accessed: 26- Dec- 2017]

[66]The PostgreSQL Global Development Group, "COPY", PostgreSQL 9.2.24 Documentation. [Online]. Available: https://www.postgresql.org/docs/9.2/static/sql-copy.html. [Accessed: 26- Dec- 2017]

[67]Google LLC, "Cost-per-view (CPV): Definition", AdWords Help. [Online]. Available: https://support.google.com/adwords/answer/2382888?hl=en. [Accessed: 25- Mar- 2018]

[68]"MRC Viewable Ad Impression Measurement Guidelines", Mediaratingcouncil.org, 2014. [Online]. Available: http://www.mediaratingcouncil.org/063014%20Viewable%20Ad%20Impression%20Guideline_Final.pdf. [Accessed: 20- Apr- 2018]

[69]T. Paulsen, "Attribution Theory: The Two Best Models for Algorithmic Marketing Attribution – Implemented in Apache Spark and R", The Data Feed Toolbox, 2017. [Online]. Available: http://datafeedtoolbox.com/attribution-theory-the-two-best-models-for-algorithmic-marketing-attribution-implemented-in-apache-spark-and-r/. [Accessed: 15- May- 2018]

[70]S. S. Fatima, M. Wooldridge, and N. R. Jennings, “A linear approximation method for the Shapley value,” Artificial Intelligence, vol. 172, no. 14, pp. 1673–1699, Sep. 2008 [Online]. Available: http://dx.doi.org/10.1016/j.artint.2008.05.003. [Accessed: 15- May- 2018]

Appendices

[A1] Python CSV sanitization scripts

[A2] Database importing scripts

[A3] List of Zboží.cz traffic source categories

[A4] SQL queries used for the original data transformation

[A5] Python script for transforming the interaction data into customer journeys

[A6] Python script for computing the Shapley value

[A7] Python script for computing the last-interaction attribution

Appendices can be found on the CD attached.

List of abbreviations

Abbreviation	Meaning	Brief explanation
AIDA	awareness-interest-desire-action	marketing funnel model
CAC	customer acquisition cost	advertising performance metric
CPM	cost per mille	digital advertising payment model
CPMV	cost per mille viewable	digital advertising payment model
CSV	comma separated values	file format
CTR	click through rate	advertising performance metric
CZK	Czech Koruna	Czech national currency
GB	gigabyte	unit of memory size
GIGO	garbage in - garbage out	general principle
HTTP	hypertext transfer protocol	basic protocol used for websites
IAB	Internet Advertising Bureau	advertising business organization
LTV	lifetime value	customer classification metric
MB	megabyte	unit of memory size
NPV	net present value	investment performance metric
O2O	online to offline	online and offline world interaction effect
PPC	pay per click	digital advertising payment model
RAM	random access memory	kind of computer memory
ROAS	return on advertising spend	investment performance metric
ROI	return on investment	investment performance metric
ROPO	research online - purchase offline	online and offline world interaction effect
RTB	real time bidding	advertising technology
SEO	search engine optimization	techniques used to appear higher in the organic search results
SQL	structured query language	database language
STDC	see-think-do-care	marketing funnel model
URL	uniform resource locator	address of a online website or file
VAT	value added tax	tax kind
XML	extensible markup language	file format

List of tables

List of pictures

List of equations