ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Bachelor Thesis BSc^2 in Econometrics and Economics

Direct Traffic Imputation Using MICE for Data-Driven Multi-touch Attribution Modelling

Name student: **Evgeny Astapov**

Student ID number: **471394**

Supervisor: **dr. Kathrin Gruber**

Second assessor: **dr. Phyllis Wan**

July 4, 2021

Abstract

We live in a world where the amount of data collected by companies is ever-increasing. At the same time, the concerns about privacy are rising. This Thesis aims to identify a "direct traffic effect" in attribution modelling data, a case where the source of a visit from a user is obfuscated from the marketer. This traffic is first simulated as missing data under a set of arbitrary assumptions. Secondly, the missing data points are imputed using Missing Imputation by Chained Equations. Lastly, probabilistic, logistic and bagged logistic attribution models are fitted on both the original and imputed data. It is found that despite 18% of the users being affected by the missing data, the probabilistic model on the imputed data yields coefficients only differing by 2.1% in absolute value from the "true coefficients". The (bagged) logistic models' coefficients differ by 12% on average. Further, simply treating the simulated missing data as a separate "direct" channel results in a -6.6% downward bias of each of the original channels' contribution. This could lead to a lower calculated ROI for each channel, and wrong business decisions to be made.

Keywords— attribution modelling, data-driven attribution, missing data imputation, direct traffic, bagged logistic regression, probabilistic attribution, logistic attribution

Contents

1 Introduction

The problem of analyzing the effectiveness of marketing channels has been bothering marketeers for decades (Bharti, 2020). With the advent of online shopping and the ability to separately observe each customer's interactions with the brand, marketers have gained the ability to assign each conversion credit to a specific channel (Yadagiri, Saini, & Sinha, 2015). This is known as the marketing attribution problem, and has quickly gained importance among researchers and marketers alike. In fact, the Marketing Science Institute has identified the problem of marketing attribution, or more specifically, attributing outcomes to specific causal factors, as one of its research priorities for years 2020-2022 (MSI, 2020).

At the same time, the concerns about internet privacy are constantly on the rise. Since 2018, the General Data Protection Regulation (GDPR) requires companies to inform their users from the European Union about how the data generated by the user is collected and stored, and collect the users' explicit consent to do so (Nouwens, Liccardi, Veale, Karger, & Kagal, 2020). Tech giants are also starting to take a stand for their own users' privacy. For instance, Apple's recent iOS 14.5 update, released in 2020, which updates the operating system running on the vast majority of iPhones, has prevented third parties from "tracking" the user activity without explicit consent from the user (Sokol & Zhu, 2021). Various sources report that as much as 95% of the users didn't give consent to this tracking after being prompted (Reichert, 2021). In the most recent Worldwide Developer Conference 2021, Apple also introduced a feature called "Private Relay", which encrypts all outgoing traffic from Safari, a browser that is installed by default on all devices running iOS (Wodinsky, 2021). Clearly, as time goes on, and users become more aware of how their data is collected and used, it will only become harder to collect, store and use data for marketing purposes.

In line with the recent developments in privacy and limitations of user tracking, this Thesis aims to develop a way of imputing uncertain user activity for the purpose of attribution modelling. Namely, this paper deals with the specific problem of "direct" traffic, which is a channel frequently used in web analytics as a bucket for (among other things) unknown traffic sources. This effect is investigated by firstly simulating its patterns in the underlying data, and using Multiple Imputation by Chained Equations to impute versions of the data without missing values. Then, the performance of this method will be assessed by estimating probabilistic and (bagged) logistic attribution models developed by Shao and Li (2011), on both the original and imputed data, and the differences between the results will be compared.

Thus, this Thesis will investigate the question, "**To what extent can direct traffic be recovered from marketing attribution data using Multiple Imputation by Chained**

Equations, and what effect does (not) imputing direct traffic have on various attribution models?"

Using the models introduced by Shao and Li (2011), it was found that for the probabilistic model, if the channel attributions are averaged from all imputations, these only differ by 2.1% in absolute value from the underlying full dataset. For the (bagged) logistic models, the resulting coefficients differ by 12% on average. Further, if the "direct effect" is not imputed, but simply treated as a separate channel, this would result in a -6.6% bias on the overall contributions of all the other channels, which could lead in decreased ROI per channel (on paper) and wrong business decisions to be made.

The aim of this Thesis with regards to scientific relevance is to encourage research into a slightly new area of attribution modelling. While most previous research on attribution modelling has been steadily focused on building the most accurate attribution model possible, this paper takes a different approach. Namely, to incorporate the uncertainty of underlying user data into the attribution model. The motivation for this follows from the aforementioned: while data collection methods are getting more complex, the rules on protection of user data are getting more stringent. The problem of uncertainty is bound to only rise over time as more and more governments and companies take user privacy more seriously.

This is linked to the societal relevance of this paper. Namely, much of the research in attribution modelling is bound to influence real-world products that offer computation of attribution models as a product to their users, as was the case with the research of Shao and Li (2011). Much of this software can be used by business owners, junior marketers, and other parties with limited statistical knowledge. It's important to inform those users about the fact that their results are only as accurate as the underlying data, and provide them with the tools to compute (more) accurate models when faced with large amounts of uncertain data.

This Thesis will be structured as follows. Section 2 will introduce the topic of attribution modelling, direct traffic and missing data imputation. Section 3 will describe the data used for this research. Section 4 will provide the methodology for this Thesis, by describing the probabilistic, logistic and bagged approach of attribution modelling of Shao and Li (2011) and building on the theory for missing data imputation. Section 5 will describe the results of implementing the aforementioned models of Shao and Li (2011), as well as the imputation of direct traffic. Finally, Section 6 will present a discussion of the results, and the findings will be presented in Section 7 along with suggestions for further research.

2 Literature Review

In this Section, firstly, the customer journey is introduced, and the ways this journey can be modelled are presented. Next, the reader is eased into the concept of attribution modelling, along with its corresponding terminology. The existing research on attribution models is briefly covered. Next, the concept of direct traffic is introduced. Finally, direct traffic is tied to the concept of missing data by introducing the definition of missing data, its categories, and approaches to dealing with missing data. Lastly, the preferred method of dealing with missing data, Multiple Imputation by Chained Equations, is briefly covered, and the way this method is used to simulate the "direct traffic effect" is described.

2.1 Customer journey

There is a close link between attribution and consumer decision making (Mizerski, Golden, & Kernan, 1979), thus it is important to define the consumer decision making model and identify how the customer journey is related to attribution modelling prior to diving deep into the specifics of attribution modelling.

Lemon and Verhoef (2016) define the customer journey as a multidimensional construct, as all the stages of a process that customers go through that makes up their overall experience.

Many models exist to describe the customer journey from both a marketing and psychological point of view. For instance, a popular model is to consider the customer journey as one consisting of distinct stages: problem recognition, search for alternatives, purchase of the product, and post-purchase evaluation (Neslin et al., 2006). Another famous model that is often used to describe the customer journey from the point of view of advertising is the AIDA model (Lemon & Verhoef, 2016) - an acronym that describes the stages a customer goes through before purchasing a product. It stands for how the product can initially capture a user's Attention, cause Interest in the consumer's mind, create a Desire for the product, and finally, cause Action to purchase said product.

2.2 Modern marketing tactics for the customer journey

Referring to the user journey, modern marketing tactics have been developed to target different parts of the AIDA funnel. For instance, the upper funnel (i.e., awareness and interest) are frequently targeted by display, social media, and search engine marketing; while the lower funnel (desire and action) use more involved methods, like content marketing and email marketing (Gruber, 2020). All these channels make up a "marketing mix", further referred to below, which can be roughly split into 3 categories: owned channels (owned and operated by the brand, such as their website), paid channels (paid advertising such as display ads and paid search advertising),

and earned channels (the likes of word of mouth and shares on social media) (Lovett & Staelin, 2016). The "direct traffic effect", introduced later in this Section, largely affects the channels the brand cannot control directly, i.e. earned channels.

2.3 Attribution modelling

2.3.1 Background definitions

When discussing the user journey in context of attribution modelling, there are a number of definitions that should be kept in mind.

Firstly, when considering the customer journey within attribution modelling, one usually talks about the user's (conversion) path. This path contains the interactions that user has made with separate digital marketing channels, that may or may not have lead to a conversion (Matthews, 2016).

Each of these interactions are often referred to as "touchpoints", getting their name from being synonymous to every time the customers "touch" any part of a product, service, brand or organization across multiple channels at various points in time (Matthews, 2016).

A conversion is an action taken by the user that lead them to purchase a product or sign up for a service that is offered by a company (Abhishek, Fader, & Hosanagar, 2012).

Finally, this leads to the definition of a channel, which is simply a medium or contact point through which a firm and a customer interact (Neslin et al., 2006).

2.3.2 Definition of attribution modelling

According to Bharti (2020), the term "attribution modelling" (also known as multi-channel attribution) has been on a steady rise in popularity online since the early 2010's. Yadagiri et al. (2015) define attribution modelling as a two-step process of crediting a conversion to a specific marketing channel. In the first step, the value of exposure to different marketing channels is estimated. In the second step, the exposure effects from the previous steps are used to assign a numerical "contribution", or attribution, of each channel.

The problem of attribution is far from being new: in the days of traditional advertising mediums such as television and radio, advertisers have used marketing-mix models, i.e., aggregate data on interactions across multiple marketing activities (Naik, Raman, & Winer, 2005), to estimate the effects of separate marketing efforts (Abhishek, Despotakis, & Ravi, 2017). However, the ease of data collection in the era of online advertising allows for a more user-centric approach: as instead of aggregate data, marketeers have access to individual-level data on interactions, conversions and demographics of the users. Abhishek et al. (2017) call this "disaggregate individual level data".

2.3.3 Types of attribution models

In the advent of attribution modelling, a simple model by the name of "last-touch" attribution was quickly adopted by the industry (Berman, 2018). This model simply attributes 100% of the credit for the conversion to the last "touch", or the last channel that the user interacted with prior to converting (Berman, 2018). However, as pointed out by Shao and Li (2011) and Chandler-Pepelnjak (2009), this model is highly flawed, as it ignores all of the user interaction information prior to the last touch.

Similar to last-touch attribution (LTA), a number of heuristic models have been later developed. This includes the first-touch model (where all the credit gets attributed to the first interaction), linear model (where all touches get attributed an equal share of the conversion), time-decay model (where the last conversion gets the most credit, and each preceding channel gets less credit, depending on a mathematical relation), and others (Y. Zhang, Wei, & Ren, 2014).

Naturally, these "heuristic" models only carry a fraction of the truth about the ability of channels to drive conversions. Over time, research and practice has shifted into multi-touch data-driven attribution models, where more than one touch point can receive a fraction of the credit for a conversion, depending on an underlying (statistical) model (Shao & Li, 2011). This is exactly the work that Shao and Li (2011) build on - but this field is quite broad, as many different multi-touch models have been proposed over the years, ranging a multitude of different mathematical ways of thinking. Just to name a few, researchers have tried to tackle the problem of attribution modelling from a Bayesian point of view, a time-series analysis framework, by using cooperative game theory, and lately even with neural networks and deep learning, as well as ensemble techniques (Bharti, 2020).

2.3.4 Topics within attribution modelling

However, research within the field of attribution modelling isn't simply focused on building different attribution models. Similarly to how this Thesis aims to contribute on the topic of missing data within the field of attribution, many authors have explored a wide range of confined but important problems within the field. A few of these are notable to mention, in order to provide the background of attribution modelling.

Early research in attribution modelling focuses on CLV, customer lifetime value - the mathematics behind spreading a customer's conversions over all their past, present and future orders (Rust, Lemon, & Zeithaml, 2004; Berger & Nasr, 1998).

Many papers focus on the interplay ("effects") between certain variables related to attribution modelling. For instance, the concept of interaction effects introduced by Rosnow and

Rosenthal (1989) has been used extensively to examine interaction effects between channels. Li and Kannan (2014) have built upon this, introducing the carryover effect (how visit involvement impacts future visits through both the same channel and other channels) and the spillover effect (that a visit through one channel may lead to visits from other channels). Xu, Duan, and Whinston (2014) similarly explore "exciting effects", the tendency of an advertising click to impact the probability of clicking on advertisements in the future.

Some researchers treat the problem of attribution in a game-theoretical approach, as a game between multiple publishers. For example, Berman (2018) examines the interplay between advertisers and publishers, investigating ad allocation efficieny, and the concepts of marginality, pay to play, and symmetry within advertising efforts.

2.3.5 Bagging and boosting

Some of the work of Shao and Li (2011) (namely, their novel bagged logistic attribution model) is based on the concept of bagging, which has also seen some use in attribution modelling research. Bagging is applied by generating replicated bootstrap samples of data to improve the predictive power of classifier learning systems (Quinlan et al., 1996). In other words, one iteratively estimates the same model on a different sample of the underlying data in each iteration, in order to improve the model's accuracy. Similar concepts have been introduced, for instance, boosting, which similarly generates samples of the underlying data, but uses all the samples at once by assigning each sample a weight (Quinlan et al., 1996).

2.4 Direct traffic

As part of the customer journey, a user can visit a website multiple times before converting. In fact, a study from 2016 gauged that 96% of the visits to retail websites end up with no conversions (McDowell, Wilson, & Kile Jr, 2016), and this figure is likely increasing by the year. Naturally, this traffic can be generated by companies from many different sources, whether be it organic search, social media, other referring websites, and many others. Among these channels is a particular source that's often misunderstood - called direct traffic.

Direct traffic, as the name might imply, is a category of traffic to websites that usually represents a situation whereby a user enters the URL of the website directly into their browser (Kakalejčík, Bucko, & Danko, 2020). However, the same authors even mention that this traffic can result from offline marketing sources, or improper setup of a company's marketing software and marketing campaigns. In practice, many web analytics frameworks (such as Google Analytics) often use this channel as a bucket for traffic that couldn't be categorized into any other channels.

Due to the novelty of the problem of the ambiguity of direct traffic, there exists no better

recent study that critically examines direct traffic, other than the above-mentioned paper of Kakalejčík et al. (2020). The authors of this study examine direct traffic from a compelling point of view - that of brand awareness. Kakalejčík et al. (2020) use direct traffic as a proxy for brand awareness - finding that a good first impression on the customer will build loyalty, and over time, the customer will start going on the website directly instead of having to be prompted with advertising. "This will brand on the customer's memory so that he will not only repeatedly visit the website from the *Direct Traffic* source but also his customer journey will end with the purchase of the company's products", state (Kakalejčík et al., 2020) in the conclusion.

However, while this is a reasonable conclusion in the case of online retailers, for example; the same effect can be minimal in other cases. For instance, an article on a news publishing website can hardly get any direct traffic, as it's highly unlikely that a user shall enter the entire URL of an article into the search bar. Similarly, visitors are unlikely to enter a website directly through a product page that is dozens of characters long.

The authors mention this as a drawback of their research, calling it the ambiguity of direct traffic. "The *Direct Traffic* source could represent one of the other marketing resources, e.g. a visit from a mobile app (Facebook and Messenger), a browser bookmark or an offline ad such as billboards, leaflets or catalogs" (Kakalejčík et al., 2020). This is precisely the motivation for this Thesis to treat direct traffic as partly "unknown" traffic, and use missing data imputation techniques to attempt to estimate the lost information.

2.5 Missing Data

2.5.1 Definition

Missing data refers to a class of problems that are made difficult by the absence of some part of a familiar data structure (Efron, 1994). Nowadays, it is fairly typical for datasets to contain missing values, especially in the case of big data (Z. Zhang, 2016). This can have large implications on the accuracy of models estimated using such data, and reproducibility of results for studies with missing data. This issue is further exacerbated by many studies failing to report their methodology of dealing with missing data, and popular packages defaulting to different approaches of dealing with the missing data (Z. Zhang, 2016).

2.5.2 Categories of missing data

One of the most famous studies on inference and missing data by Rubin (1976) highlights three categories of missing data. If the data points all have an equal probability of being missing, the data is classified as missing completely at random (MCAR). If this probability can be explained by observed characteristics within the data points, this data is said to be missing at random

(MAR). If neither of these scenarios applies, the data is said to be missing not at random (MNAR) otherwise referred to as not missing at random (NMAR). This is the most complicated case - the probability of data missing varies for reasons unknown to the researcher (Van Buuren, 2018).

2.5.3 Dealing with missing data

Naturally, many methods of dealing with missing data have been developed since the inception of the problem decades ago. For instance, a method familiar to many is complete-case analysis, or listwise deletion: simply said, dropping the observations that contain missing data (Van Buuren, 2018). Another simple method is pairwise deletion - instead of dropping observations with missing values, this method uses as much of the data as possible to compute summary statistics, such as mean and covariance, which are then used for regression, factor analysis or other procedures (Van Buuren, 2018). However, the majority of the methods are based on the concept of imputation.

2.5.4 Missing data imputation

Imputation is a strategy of dealing with missing data by replacing missing values with "imputed" values (Z. Zhang, 2016). The imputed values are are usually computed from the observed data under a set of assumptions, frequently governed by those in Section 2.5.2. For instance, a quick approach would be to use mean imputation: replacing missing values of a specific variable by using their mean value. These are adequate in certain circumstances, such as MCAR and a low threshold of missing data, where the simpler methods such as mean imputation and the aforementioned listwise deletion are an acceptable approach (Azur, Stuart, Frangakis, & Leaf, 2011). However, in practice these often produce biased results (Donders, Van Der Heijden, Stijnen, & Moons, 2006) as the assumptions are too stringent. Naturally, more complex methods exist that operate under a relaxed set of assumptions, such as multiple imputation.

2.5.5 Multiple imputation

As the name suggests, multiple imputation (MI) is a strategy where the data is imputed m times (where $m > 1$ and typically small, usually $m \in [3, 10]$) (Schafer, 1999). MI uses the properties of the distribution of the observed data to estimate a set of plausible values for all the missing values (White, Royston, & Wood, 2011). These values are used to build m separate data sets. The multiple imputed data sets are then used for further analysis, the results of which are pooled together, for instance, by averaging the resulting coefficients and providing a confidence interval for each. Previous studies suggest that multiple imputation is able to successfully deal with datasets that contain from 10% to a staggering 60% of missing data (Barzi & Woodward, 2004).

2.5.6 Multiple Imputation by Chained Equations (MICE)

Multiple Imputation by Chained Equations (MICE) is a technique for MI that procedurally generates imputed datasets by a series of linked equations. The definition is much better described algorithmically, as seen in in Section 4 and Algorithm 4 (Azur et al., 2011). It should be noted that MICE operates under the MAR assumption (see Section 2.5.2), i.e. the fact that missing data can be explained by other observed variables in the dataset (Azur et al., 2011).

For more detail on the theoretical background of MICE, one can consult Azur et al. (2011) or White et al. (2011), where this is covered in extreme detail. Further, this algorithm will be expanded on in Section 4 in order to explore the mathematics behind each of the steps in the process.

3 Data

This Section is devoted to describing the dataset used to demonstrate the results of this Thesis. Since the dataset used by Shao and Li (2011) isn't publicly available, and furthermore hundreds of gigabytes in size, this paper will opt into using a smaller publicly available dataset. This will be used to first build the attribution models developed by Shao and Li (2011), and secondly to demonstrate the results of "direct traffic imputation".

3.1 *ChannelAttribution* dataset

In 2016, researchers at Microsoft have published an R package under the name of *ChannelAttribution* (Altomare, Loris, & Altomare, 2016). This package implements Markov Model solutions for the attribution problem, as well as a number of heuristic models. More importantly for this paper, it also implements a "dummy" dataset that can be used directly to test the aforementioned models. This data will be used in this paper in an attempt to re-create and extend the results of Shao and Li (2011). This was chosen since there are already papers using this very package and data to implement and test their methods. Further, the built-in model can be used as a sanity check for the "accuracy" of the logistic and probabilistic models that will be custom-coded for the purpose of this paper.

3.1.1 Data description

The dataset in the *ChannelAttribution* package (Altomare et al., 2016) consists of 10,000 unique paths and 4 variables: the path name, total active users resulting from that path, total inactive users resulting from that path, and a transformed conversion value. Each path is represented by a character variable: with each channel on the path represented by a Greek letter, and channels separated from each other using the ">" symbol. For example, the first path in the dataset is the path $eta > iota > alpha > eta$, netting 1 active user, 3 inactive users, and a conversion

value of 0.244. Table 1 shows the general descriptive statistics of this dataset.

Table 1: Descriptive statistics of the *ChannelAttribution* dataset.

Variable	**Value**	**Description**
Average path length	5.95	The length of the average user's path
Longest path	89	Longest number of interactions from a single user
Shortest path	1	Smallest number of interactions from a single user
Total number of users	88,387	Number of users in the dataset
Total number of visits	378,209	Number of unique views from all the users
Number of active users	19,785	Number of users with conversion in the dataset
Number of inactive users	68,602	Number of users without conversion
Conversion rate	22.4%	Percentage of converting users in the dataset

As can be seen from the Table, the dataset contains just over 88 thousand users, an average path length of almost 6 interactions, and a conversion rate of 22.4% - quite high for modern standards, as seen in Section 2.4.

Table 2: Descriptive statistics of the channels in *ChannelAttribution* data

Channel	**Users**	**% users**	**Conversions**	**% conversion rate**
alpha	48,285	54.6	10,736	22.2
beta	18,609	21.1	4,378	23.5
gamma	1,421	1.61	328	23.1
delta	42	$4.75 * 10^{-2}$	9	21.4
epsilon	4,768	5.39	1,138	23.9
zeta	2,747	3.11	653	23.8
eta	45,386	51.3	10,378	22.9
theta	14,410	16.3	3,267	22.7
iota	30,378	34.4	6,895	22.7
kappa	1,930	2.18	469	24.3
lambda	8,688	8.83	2,054	23.6
mi	15	$1.70 * 10^{-2}$	4	26.7

Table 2 shows the detailed information about each of the 12 unique channels in the data set. As can be seen, all the channels have a comparable conversion rate. This is interesting, considering there is a high fluctuation in the percentage of users that visited each of the channels, as can be seen from column 3. Notably, *delta* and *mi* have only been visited by a small fraction of the users, which could cause some problems with estimation. *alpha* and *eta* brought in over 10,000 conversions each and have been visited by every one in two users in the entire data set.

4 Methodology

This Section builds the theoretical framework for the results of this Thesis. Firstly, the attribution modelling methodology of Shao and Li (2011) is explained: namely, the probabilistic, logistic and bagged logistic attribution models, as well as the V-A metric. And secondly, the theoretical background of missing data imputation is introduced, the procedure for the simulation of missing values is explained, and the approach to imputation of missing values is provided

with regard to "direct traffic imputation".

4.1 Shao & Li (2011) terminology

In the original paper of Shao and Li (2011), the symbols and terminology are introduced quite sporadically throughout the paper. In order aid the interpretation of this paper, all the terms and variables used in this Section are collected below for reference. Note that some of the definitions vary from the original paper for consistency with further methods introduced in this Thesis.

$y = \{0, 1\}$, whether a specific user has converted or not

x_{ji} = whether a channel i has been encountered along user j's path

x_i = set of all users that encountered channel i, used for the probabilistic approach

$i \in [1, K]$ = a numeric value assigned to each of the K unique channels

$j \in [1, N]$ = a numeric value assigned to each of the N unique users in a sample

K = total number of unique channels in the dataset

N = total number of users in the sample

$S = 100$, number of times the data is re-sampled for estimating the V-A metric

$M = 1000$, number of iterations within each bagged regression

$p_c = \{0.25, 0.5, 0.75\}$, the proportion of covariates sampled within the bagged regression

$p_s = \{0.25, 0.5, 0.75\}$, the proportion of users sampled within the bagged regression

4.2 Probabilistic approach to attribution modelling

The probabilistic model proposed by Shao and Li (2011) is exactly as the model name suggests, a probability-centric approach to attribution modelling. The authors use $x_i, i \in [1, K]$ to indicate whether a specific channel x_i (out of K total channels) is present along a user's conversion path. Further, they define a binary variable y for a specific user's path,

$$y = \begin{cases} 1 & \text{if the user's path lead to a conversion} \\ 0 & \text{if the user's path didn't lead to a conversion} \end{cases} \tag{1}$$

Using these, the authors define a simple probability of conversion, given that a channel x_i was present along the path:

$$P(y \mid x_i) = \frac{N_{positive}(x_i)}{N_{positive}(x_i) + N_{negative}(x_i)} \tag{2}$$

where $N_{positive}(x_i)$ and $N_{negative}(x_i)$ are the number of positive and negative users where channel x_i was present in the conversion path (Shao & Li, 2011). In other words, these are the number of paths where channel x_i was present that respectively either lead to a conversion, or resulted in no conversion.

Similarly, the authors define $N_{positive}(x_i, x_j)$ and $N_{negative}(x_i, x_j)$ as the number of positive and negative users that encountered both channels x_i and x_j in their paths (note that in this scenario, the subscript j refers to a channel, not a user). This allows to define the conditional probability that a user converts, given they encounter both channels x_i and x_j in their path:

$$P(y \mid x_i, x_j) = \frac{N_{positive}(x_i, x_j)}{N_{positive}(x_i, x_j) + N_{negative}(x_i, x_j)} \tag{3}$$

Finally, this allows to define the contribution of channel i, defined as $C(x_i)$, the definition for which is interesting:

$$C(x_i) = p(y \mid x_i) + \frac{1}{2(K-1)} \sum_{j \neq i} \{p(y \mid x_i, x_j) - p(y \mid x_i) - p(y \mid x_j)\} \tag{4}$$

This is interesting, as the authors mention, as the model is "essentially a second-order probability estimation" (Shao & Li, 2011). In other words, this model considers the interaction effects between channels, which ties to earlier research in attribution modelling introduced in Section 2.3.4, namely, the work of Rosnow and Rosenthal (1989).

4.3 Logistic approach to attribution modelling

The second main contribution of Shao and Li (2011) is defining a bagged logistic approach to attribution modelling. The authors themselves don't go into much detail about the underlying logistic approach and focus fully on what makes the bagged approach different. Therefore, for completeness, it is important to define the logistic approach before looking at the bagged approach.

Interestingly, the works that cite Shao and Li (2011) also do not seem to go into much detail of the mathematical definition of a "logistic attribution model". Thus, much of the understanding of this model is left up to interpretation.

To the best of the author's interpretation, the "logistic approach" is nothing more but a logistic regression of whether a user has converted on an intercept and K binary variables of whether each channel has been encountered by the user:

$$\boldsymbol{y} = \Lambda(\boldsymbol{x}\boldsymbol{\beta}) = \Lambda\left(\beta_0 + \sum_{i=1}^{K} x_i \beta_i\right) \tag{5}$$

where $\Lambda(\boldsymbol{x\beta}) = \frac{\exp(\boldsymbol{x\beta})}{1+\exp(\boldsymbol{x\beta})}$ is the inverse logit function. Here, $\boldsymbol{y}_{N\times 1}$ is a binary vector representing whether each user in the sample is active ($y = 1$) or inactive ($y = 0$). $\boldsymbol{x}_{N\times(K+1)}$ is a vector with the first column being the unity vector $\mathbf{1}$ to account for the intercept in the model, and elements of columns 2 to (J+1) representing x_{ji}, whether user j has encountered channel i along their path.

4.4 Bagged logistic approach

As mentioned previously, the second big contribution of the paper of Shao and Li (2011) is the bagged logistic approach to attribution modelling. In short, this approach combines the described above logistic approach together with the concept of bagging, i.e., "a method for generating multiple versions of a predictor and using these to get an aggregated predictor" (Breiman, 1996).

Applied to logistic regression, bagging implies simply fitting the logistic regression on a subset of the data, with a subset of the covariates for a number of iterations, and averaging the resulting coefficients. This is demonstrated in algorithmic form in Algorithm 1, similarly to the work of Shao and Li (2011).

Algorithm 1: Bagged logistic regression

1 Sample a fraction p_s of all the observations in the sample:
  (I) Calculate $N_s = \lfloor N * p_s \rfloor$, the number of data rows to keep for model estimation.
  (II) Randomly pick N_s data points from the initial data without replacement.
2 Sample a fraction p_c of all the covariates in the model:
  (I) Calculate $K_c = \lfloor K * p_c \rfloor$, the number of channels to keep in the regression estimation.
  (II) Randomly pick K_c channels from the full set of channels without replacement.
3 Fit the regular logistic model with the selected users N_s and covariates K_c.
4 Repeat *Steps 1 - 3* for M iterations, saving the regression coefficients after every iteration.
5 Set each coefficient's final estimate in the bagged regression as the average of that coefficient's values from Step 4.

4.5 Channel attributions from the bagged & logistic approaches

It should be noted that a downfall of the paper of Shao and Li (2011) is the fact that they do not provide a way to assign actual numerical values for attribution of each of the channels from the regression output. In other words, there is no way to convert the regression output into an actual "attribution model", that would provide the user with the contributions of each channel, and further allow to compare the (bagged) logistic approaches to the probabilistic approach. This has been criticised in the works that build upon Shao and Li (2011), such as Yadagiri et al. (2015).

However, there exist some ways of converting the output of the (bagged) logistic regression into channel contributions. A natural first reaction would be to estimate the channel contributions via the marginal effects of the model, which can be computed as follows:

$$C(\boldsymbol{x}) = \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \frac{\partial \Lambda(\boldsymbol{x}\boldsymbol{\beta})}{\partial \boldsymbol{x}} = \beta \Lambda(\boldsymbol{x}\boldsymbol{\beta}) \tag{6}$$

where C is just a vector of the contribution contribution of all the channels, similar to the output of the probabilistic model in Formula 4. However, as any β_i can be negative, this could result in negative channel contributions, which are difficult to interpret. Some papers, for instance, that of Kesteren (2015), propose alternative ways of computing attributions from LR and BLR, but none have been set in stone and accepted by others as the "correct" approach.

4.6 Bivariate V-A metric

The third big contribution of Shao and Li (2011) is the so-called "bivariate metric". The motivation for creating this metric is to have a simple form of evaluation of two important features of an attribution model. Firstly, the model's accuracy (or misclassification rate) with regards to predicting a conversion; and secondly, the variability of the model's estimates between different samples of the data. The authors define these as the accuracy measure (A-metric) and variability measure (V-metric), jointly making up the bivariate V-A metric.

4.6.1 A-metric

The A-metric measures the model's accuracy, or more precisely, the misclassification rate. In other words, this metric shows the fraction of times the model wrongly predicts the users' conversion status. This is done by obtaining the fitted values from the computed regression coefficients, and comparing these fitted values to the observed values:

$$\text{A-metric} = \frac{\#\{\hat{y} \neq y_{observed}\}}{\#\{\hat{y} = y_{observed}\} + \#\{\hat{y} \neq y_{observed}\}} = \frac{\#\{\hat{y} \neq y_{observed}\}}{N} \tag{7}$$

where $\#\{\hat{y} = y_{observed}\}$ and $\#\{\hat{y} \neq y_{observed}\}$ are, respectively, the number of users that were correctly and incorrectly predicted as active or inactive.

4.6.2 V-metric

The V-metric, in the words of Shao and Li (2011), measures the stability of the model's estimates. Said otherwise, this metric can be computed by firstly saving all the estimated coefficients from each time the data was re-sampled and a new model was estimated, and secondly, computing

the mean of the standard deviation for each of those coefficients:

$$\text{V-metric} = \frac{1}{K}\sum_{i=1}^{K}\sqrt{\frac{\sum_{s=1}^{S}(\beta_{si}-\bar{\beta}_i)^2}{S}} \tag{8}$$

where β_{si} is the coefficient for the contribution of channel i estimated for the s'th time the data was resampled, and $\bar{\beta}_i$ is the mean value of the coefficient for channel i over all the re-sampling runs.

In order to see how the V-A metric is computed, let us consider how Shao and Li (2011) generated their results in full:

Algorithm 2: Computing the V-A metric from (bagged) logistic regressions

1. Randomly select a sample of N users (without replacement) from the full dataset as training data. The ratio of active to inactive users in this sample should be 1 : 4.
2. Randomly select a sample of N users (without replacement) from the full dataset as testing data. The ratio of active to inactive users is not fixed for this sample.
3. Fit the bagged logistic regression to the training data using Algorithm 1 with specified sample and covariate proportions p_s and p_c, respectively. Record the coefficients.
4. Fit the regular logistic regression to the training data and record the resulting coefficients.
5. Compute the fitted values for the testing data:
   (I) Compute fitted values for the bagged logistic regression on the testing data using coefficients from *Step 3*.
   (II) Compute fitted values for the regular logistic regression on the testing data using coefficients from *Step 4*.
6. Compute the A-metric for each method:
   (I) Compute A-metric for bagged logistic regression
   (II) Compute A-metric for regular logistic regression
7. Repeat *Steps 1 - 6* for S = 100 times
8. Compute the V-metrics for LR and BLR using Equation 8.
9. Compute the resulting A-metrics by averaging the A-metrics from *Step 6* across all the S runs.

4.7 Missing data imputation

The aim of missing data imputation within the context of this Thesis and attribution modelling is to attempt to levy the problem whereby traffic from specific users and specific channels could be in a way "censored" from the marketeer. For instance, in a way it is described in Section 2.4, direct traffic to a website need not always be traffic where a user types in the URL of the website directly - it can also be observed when a user comes from another channel, the tracking of which is either broken or prevented by the user themselves.

Using this example, under some assumptions, it is possible to simulate such a situation within the *ChannelAttribution* dataset by replacing some observations with missing values, and imputing those missing values using MICE described in Section 2.5.6.

4.7.1 Simulating missing values

In order to simulate the described effect of direct traffic, henceforth referred to as the "direct traffic effect", one needs to make some assumptions about where those direct hits could have come from. At the time of writing, there are no studies with explicit descriptions of where direct traffic could come from, or the probabilities of that direct traffic coming from specific channels. Thus, the missing values will be simulated under a set of assumptions described below. The generality of these assumptions shouldn't hinder the interpretability of the results, since the *ChannelAttribution* dataset used to demonstrate the technique is synthetic. In practice, of course, one would encounter "missing data" governed by a much more complex process than two general assumptions.

This Thesis imposes two main assumptions governing where direct traffic could be coming from. Firstly, this traffic could come from specific users that limit the extent to which they can be tracked online, thereby causing some traffic (for example, from social channels, or search) to show up as "direct". Secondly, direct traffic could come from a misconfiguration of a specific campaign, e.g. an email campaign, the tracking for which was broken. Let's make an assumption about how this data would be (un)observed regarding the *ChannelAttribution* dataset:

- Users blocking tracking: assume that 10% of the users in the dataset have installed tracking blocking software. This would obfuscate traffic of channels *beta*, *lambda* and *theta* from the marketer, i.e. that traffic would be observed as "direct". To simulate this, a copy of the dataset was made and the observations of 10% of the users, picked at random, were set as missing (NA) for the affected channels *beta*, *lambda* and *theta*.

- Assume campaigns have a 1% chance to have badly implemented tracking, i.e., each hit in the sample has a 1% chance to be obfuscated from the marketer and be observed as "direct" traffic. To simulate this, the R package *missForest* was used to pick 1% of the values in the dataset, and replace these values with missing values, i.e. NA.

Note that while these assumptions do not imply MAR data, as mentioned in Section 2.5.6, one can assume without loss of generality that this assumption holds for two reasons. Firstly, the missing patterns can be recovered due to the high predictive power of other (non-missing) channel information, and similarity and non-uniqueness of user paths. Secondly, in practice, one would have many other geographic, demographic, and other user-specific variables that can be used for prediction, which are likely to be correlated with the missingness.

The procedure for simulating the above assumptions are presented in Algorithm 3:

Algorithm 3: Simulating direct traffic for data-driven attribution

1. Sample $p_m = 0.1$ proportion of users and set their affected channels $beta, lambda, theta$ to N/A.
2. Sample affected hits with "badly implemented tracking" of $p_n = 0.01$ observations, and set them to N/A.
3. Run Algorithm 4 in order to impute the missing data.

4.7.2 Imputing missing values

The procedure for MICE, introduced in Section 2.5.6, is presented in Algorithm 4, adapted from Azur et al. (2011). The R package *mice* will be used to impute the missing values.

Algorithm 4: Multiple Imputation by Chained Equations

1. All missing points in the data are firstly replaced by "placeholder" data, imputed using a simple algorithm, such as mean imputation.
2. The "placeholder" data for one variable ("var") are set to missing.
3. The observed values for "var" are regressed on the other variables in the imputation model. The choice of regression model and variables depends on the type of MICE model used, and is described in more detail below.
4. The missing values of "var" are replaced by fitted values ("imputations") from the regression model. If "var" is further used for imputing other variables, these fitted values are used.
5. Steps 2-4 are repeated for each variable with missing data. A single iteration imputing the missing values of all the variables with missing data is referred to as a "cycle".
6. Steps 2-5 are repeated for a preset number of cycles, with the imputations being updated at each iteration.

Since the underlying data for estimation is mostly binary, the logical choice for the model in step 3 of the Algorithm would be a simple logistic regression. More specifically, a regression of the following form is run:

$$x_i^* = \alpha + \sum_{h \neq i} \beta_h x_h + \sum_p z_p + \varepsilon_i \tag{9}$$

Here, α is a constant, x_i is the vector of length N that indicates whether channel i was present along the path for each of the N users in the data. z_p is a particular external regressor of length N that contains user-specific data that could be useful for the process of imputation. A simple example used in this Thesis would be the length of the path taken by the user, but

this can get as advanced as the number of previous visits, a factor variable for which country this user is visiting from, and other data.

4.7.3 Procedure for missing data imputation

To sum up, the process of missing data imputation in this Thesis is two-fold. Firstly, the "missing" data for direct traffic needs to be simulated into the data (as shown in Algorithm 3). Secondly, the models of Shao and Li (2011) need to be estimated using the imputed data, and the resulting coefficients pooled together for analysis.

Finally, Algorithm 5 sums up the steps needed to run the models on the imputed data sets and compare these results to the original dataset.

Algorithm 5: Computing model results from imputed data

1. Train a probabilistic model on the full, non-imputed data set using steps in Section 4.2. Save the resulting outputs $C(x_i)$.
2. Train a bagged logistic and simple logistic models on the full data set, for a specified set of coefficients S, M, p_c, p_s, as well as a fixed sample size and ratio of active to inactive users. Note S=1 and sample size is the full sample size, in this case.
3. Repeat Steps 1 & 2 for each of the m imputed data sets, using the same set of parameters.
4. Compute $\overline{C^*(x_i)}$, the mean of the m vectors of $C(x_i)$ resulting from Step 3.
5. Compute $\bar{\beta}^*$, the mean of the m coefficient vectors β from Steps 2 & 3, for both the bagged and logistic regression.
6. Use the coefficients from the Steps above for Tables and box-plots to gauge the efficiency of "direct traffic imputation".

4.7.4 Imputing direct traffic in practice

Finally, there is one thing to consider about the implementation of the "direct traffic imputation" process in practice. The *ChannelAttribution* data doesn't contain a "direct" channel (in fact, the names of the channels are unknown and encrypted to Greek letter names). In practice, however, one would encounter direct traffic that is partly consisting of truly direct traffic (for instance, a click leading a user from one part of a site to another is truly "direct"), and the aforementioned "unknown" direct traffic. One would need to differentiate between the two before applying the methods described in this Section. Explaining how this can be done goes beyond the relevant theory for this Thesis, however, this is possible to do in many ways. One way of achieving this result would be to use referral paths: these contain information about the site referring the user to a specific page. In case of true direct traffic, this would be the same domain of the site the user is currently on. In case of "unknown" direct traffic, this field would be undefined.

5 Results

This Section is devoted to presenting the results of this Thesis. Firstly, the probabilistic and (bagged) logistic models are run on the full dataset without any missing values, in order to gauge their effectiveness without missing data imputation. Secondly, the missing data is simulated according to the procedure in Section 4.7.1. Thirdly, this imputed data is used to once again generate the results from the above attribution models. Finally, the results from the 'full' and 'imputed' data sets are presented and compared.

5.1 Attribution models

The first step to impute the "direct traffic effect" is to implement the bagged, probabilistic and regular logistic models introduced by Shao and Li (2011), in order to get a "feel" for the data. As these attribution models aren't available out-of-the-box from any packages, these were implemented completely from scratch using R. The details on the code for these models can be found in Appendix A.

5.1.1 Probabilistic model

Running the probabilistic model on the *ChannelAttribution* data yields the output presented in Table 3. The channels are sorted descending by their resulting channel contribution $C(x_i)$. Thus, it can be seen that *iota* is the channel with the highest contribution among all the channels in the dataset, while *mi* is the channel with the least credit. In practice, one would likely assign each channel a conversion value of $C(x_i)/\sum_{i=1}^{K} C(x_i)$ percent of the total profits made from the channels. This can then be used in tandem with expenses for, e.g., the ad spend and campaign cost for each channel, to compute the return on investment for each channel.

Table 3: Probabilistic model attribution per channel

Channel	$C(x_i)$
iota	0.117
eta	0.116
epsilon	0.112
zeta	0.112
lambda	0.109
beta	0.106
kappa	0.103
theta	0.102
alpha	0.100
gamma	0.095
delta	0.072
mi	0.044

5.1.2 Logistic & bagged model

As a "sanity check", Table 4 shows the V-A metric comparison of the bagged and logistic approach, aiming to replicate a similar table presented by Shao and Li (2011) in their work. This is done in order to gauge the correctness of the implementation of the aforementioned models, and their ability to deal with the underlying data. These results were produced under S=100 runs, M=1000 bagged logistic iterations, 1:4 active to inactive user ratio, and a sample size of 8,000 users per run.

Table 4: V-A metric comparison of the bagged logistic regression (BLR) and the usual logistic regression (LR), trained on *ChannelAttribution* data

			p_c					
			0.25		0.50		0.75	
			V-metric	A-metric	V-metric	A-metric	V-metric	A-metric
p_s	0.25	LR	0.110	0.224	0.663	0.224	0.109	0.224
		BLR	0.107	0.224	0.649	0.224	0.108	0.224
	0.50	LR	0.112	0.224	0.104	0.224	0.627	0.224
		BLR	0.100	0.224	0.097	0.224	0.535	0.224
	0.75	LR	0.110	0.224	0.633	0.224	0.111	0.224
		BLR	0.099	0.224	0.583	0.224	0.106	0.224

According to Shao and Li (2011), the V-metric for BLR should be significantly lower than for LR, and the accuracy should be lower, but comparable. The results of this Table seem to support the fact that the bagged logistic regression achieves a lower V-metric compared to the regular logistic regression. However, it seems that the A-metric is equal for all the different permutations of p_c and p_s.

Upon further inspection, the cause of this becomes apparent. It appears (bagged) logistic regression isn't an optimal choice for attribution modelling for the *ChannelAttribution* data, as the fitted values $\hat{y}_i$ produced by the (bagged) logistic regressions seem to be highly correlated with the ratio of active to inactive users. In other words, if the ratio of active to inactive users is 1:4, as is the case with Table 4, 20% of the users in the training data will be active. Incidentally, the majority of the fitted values produced by the regression will be around 0.2. This results in the models predicting the user will be inactive in the majority of the cases, resulting in a 22.4% misclassification rate (A-metric), which equal to the sample average conversion rate (i.e., fraction of active users) in Table 1. If the ratio of active to inactive users is changed, the fitted values change in the same direction.

Further, as can be seen in Appendix B, lowering the sample size to 800 users (in order to lessen the probability of each user appearing in multiple of the S=100 different runs of the simulation) yields a V-metric for BLR lower than that of the logistic regression. In other words, bagged regression yields its benefits on larger sample sizes. Sadly, the *ChannelAttribution* data

only has 88,000 users, while Shao and Li (2011) had over 72 million users to work with. Another cause for this can be explained by the same reason as the authors name for the change in the V-metric of the LR regression, namely "random sampling variation" (Shao & Li, 2011). Stated otherwise, some channels are extremely sparse in the *ChannelAttribution* dataset - for instance, *mi* only appears in 2 unique paths and for 15 unique users, as can be seen in Table 2. Similarly, *delta* also appears only every 1 in 2,000 users. It is natural that the standard deviation for these coefficients is higher in the bagged regression, as the concept of bagging, by design, can omit some of these observations. This significantly drives up the V-metric for BLR for channels with less observations.

It should be noted, however, that these results do not impact the conclusions for direct traffic imputation, as an accurate model isn't a requirement for MICE, nor the conclusions that follow.

5.2 Missing data simulation

Before demonstrating the results of the missing data imputation, the "direct traffic", or missing data, was simulated in the *ChannelAttribution* dataset. The results of this are presented in Figure 1.

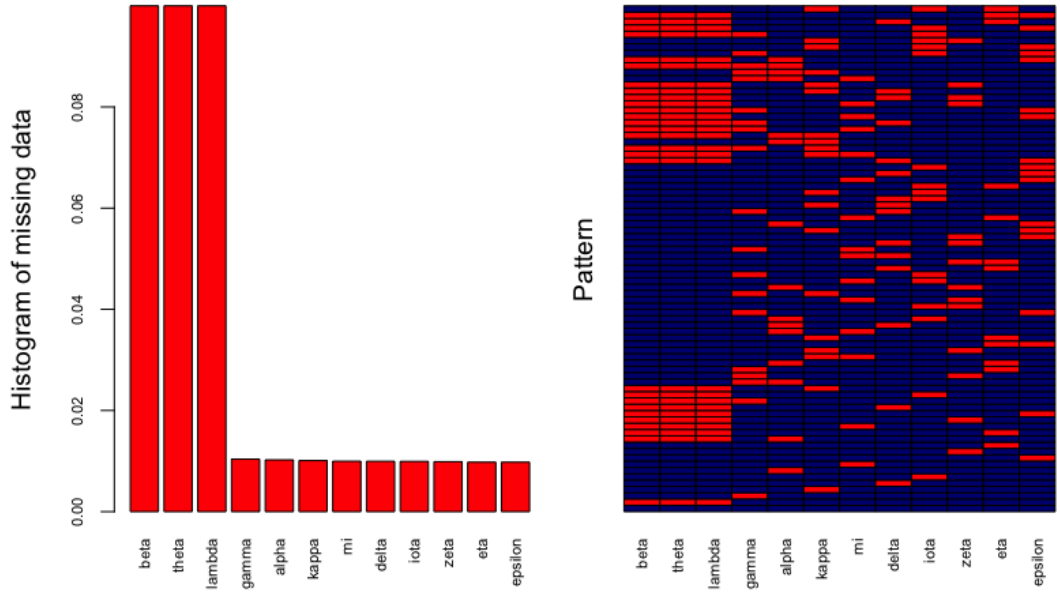

Figure 1: Results of missing data simulation. Left: histogram of fraction of missing data per variable. Right: missingness per variable, visualised for each user.

It can be seen from the left panel that the fraction of missing values is 0.1 for *beta*, *theta* and *lambda*, and approximately 0.01 for all other channels. Further, from the right panel, one can see that the missing observations for *beta*, *theta* and *lambda* are all affecting the same sets of users. In other words, those observations are either all missing for a specific user, or all present

for a specific user. This is the effect of the assumption on users blocking tracking mentioned in Section 4.7.1. Further, it can be seen that the other channels have missing values present rather sporadically, to simulate the effect of broken tracking on specific campaigns. It should be noted that the right panel of the figure is not to scale, as it is trying to represent almost 88,000 users within limited space.

5.3 Missing data imputation

Applying Algorithm 3 with $m = 10$ on the *ChannelAttribution* dataset and running the probabilistic model in Section 4.2 results in channel contributions $\overline{C^*(x_i)}$ in Table 5. The panels of this table depict the channel name, the new contribution $\overline{C^*(x_i)}$, computed by averaging the probabilistic channel attribution over all the imputed data sets, the variance of those estimates (inflated by a factor of 1,000 for conciseness), the initial contribution seen in Table 3, and the percentage difference of $\overline{C^*(x_i)}$ from $C(x_i)$.

Table 5: Probabilistic model attribution on imputed data

Channel	$\overline{C^*(x_i)}$	$V(C^*(x_i)) * 1000$	$C(x_i)$	% change
alpha	0.099	0.001	0.100	-0.80
beta	0.105	0.001	0.106	-0.81
delta	0.069	0.007	0.072	-4.15
epsilon	0.111	0.002	0.112	-1.18
eta	0.117	0.001	0.116	0.88
gamma	0.101	0.180	0.095	6.18
iota	0.118	0.001	0.117	1.11
kappa	0.102	0.003	0.103	-1.18
lambda	0.113	0.189	0.109	3.63
mi	0.068	1.361	0.044	55.4
theta	0.101	0.001	0.102	-1.22
zeta	0.110	0.001	0.112	-1.99

Note: Columns from left to right: channel name, "new" probabilistic attribution under imputed data, variance of probabilistic attribution under imputed data, "original" probabilistic attribution on full data, % change between "original" and "new" probabilistic attribution.

Further, the box-plots of the resulting channel contributions are plotted in Figure 2. From the last panel of Table 5, it can be seen that in the majority of the cases, the 1% missingness of values resulted in a proportional change of around 1% in absolute value of the channel contribution. It is especially reassuring to see the values of $\overline{C^*(x_i)}$ for *beta*, *theta* and *lambda* to be fairly comparable to the respective values of $C(x_i)$ computed using the full data set. It should be noted that the value of *lambda* has changed by 3.63%, but as can be seen from the Figure, this is caused by one outlier value of $C^*(x_{lambda}) = 0.15$.

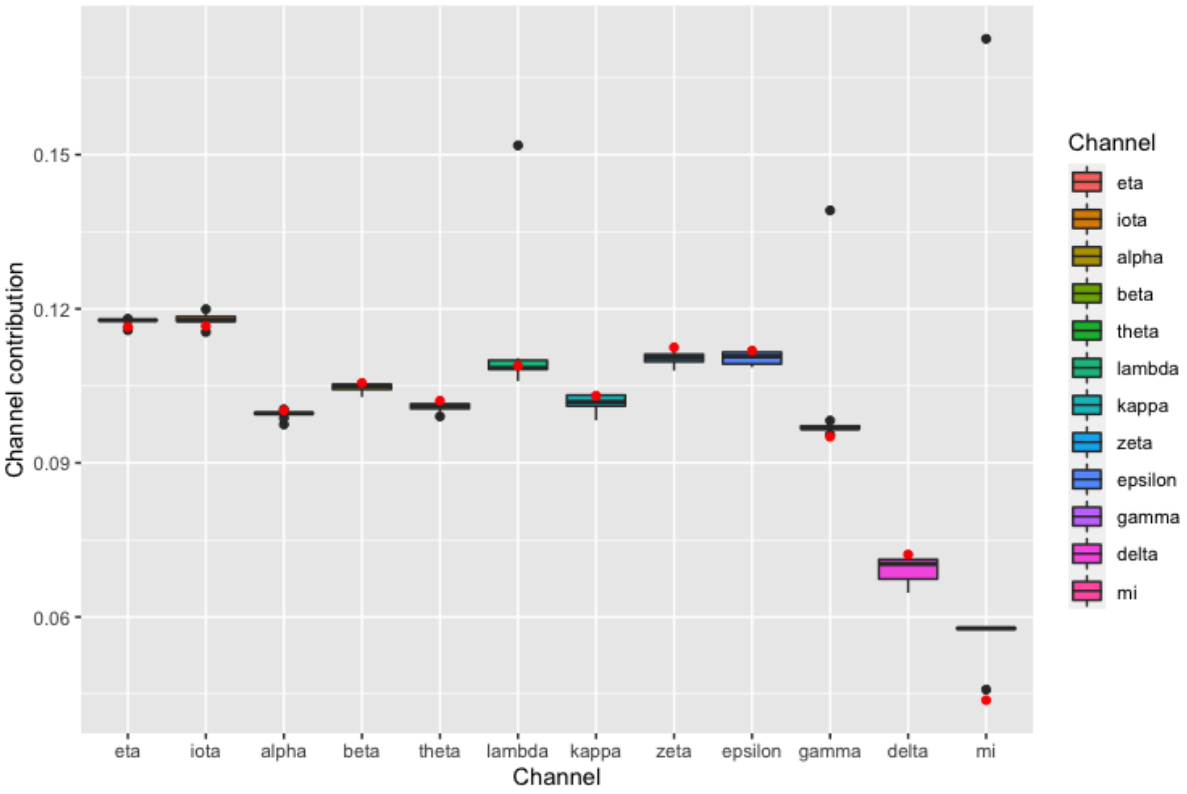

Figure 2: Probabilisitic model attribution on imputed data
Note: box-and-whisker plots represent "new" probabilistic attribution under imputed data; with outliers plotted as black dots. Red dots represent "original" channel attributions from Table 3, computed using full *ChannelAttribution* data.

Next, the coefficient estimates from the (bagged) logistic regression (from the full data) together with the estimates from the imputed data sets are shown together in Table 6. The values of $p_c = 0.5$ and $p_s = 0.5$ were used for brevity, as Shao and Li (2011) suggest these values as those that maximise the effectiveness of the bagged regression.

Table 6: (Bagged) logistic model attribution on imputed data

	Logistic			Bagged		
Coefficient	$\beta_{logistic}$	$\beta^*_{logistic}$	% change	β_{bagged}	β^*_{bagged}	% change
(Intercept)	-1.457	-1.450	0.43	-1.420	-1.419	0.08
alpha	0.028	0.023	-15.24	0.005	0.005	-4.88
beta	0.053	0.056	6.96	0.069	0.069	-0.79
delta	-4.297	-4.786	-11.38	-5.577	-5.945	-6.61
epsilon	0.054	0.052	-3.34	0.062	0.059	-4.65
eta	0.058	0.045	-23.02	0.055	0.048	-13.33
gamma	0.009	0.020	124.6	-0.003	0.008	359.08
iota	0.020	0.018	-13.86	0.018	0.017	-4.88
kappa	0.076	0.083	8.70	0.077	0.083	8.64
lambda	0.070	0.068	-3.72	0.066	0.066	-0.42
mi	-5.304	-4.390	17.24	-5.673	-4.598	18.95
theta	-0.043	-0.025	42.77	-0.017	-0.005	74.14
zeta	0.018	0.020	6.48	0.039	0.037	-5.96

Note: β represent coefficients computed using the full *ChannelAttribution* data. β^* represent coefficients computed using the imputed data. % change shows percentage change between β and β^*.

Similarly to Figure 2, the coefficients of the logistic attribution model and the bagged model on the imputed data are plotted in Figures 3 and 4 respectively. It should be noted before analyzing these values that these are simply the coefficient values from the (bagged) logistic regressions and aren't actual attributions for the individual channels. Therefore, these results aren't directly comparable to those of the probabilistic model.

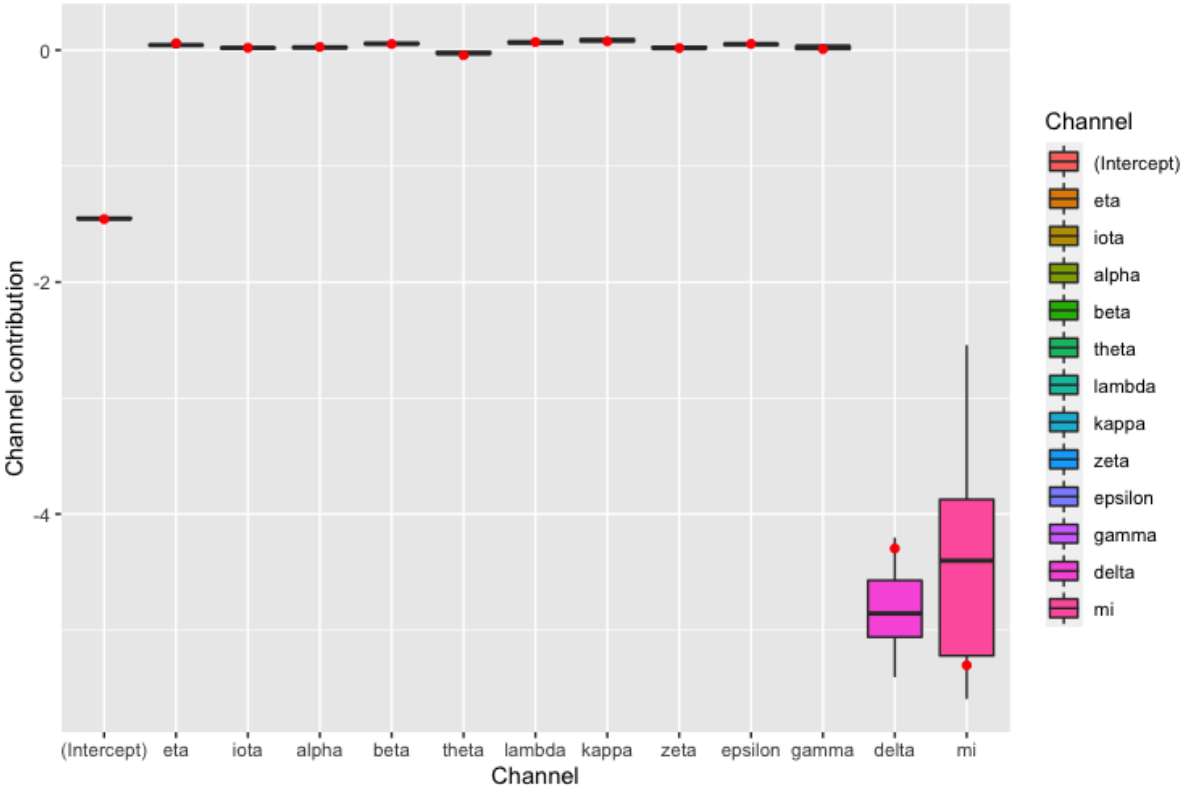

Figure 3: Logistic model attribution on imputed data
Note: box-and-whisker plots represent "new" logistic coefficients under imputed data; with outliers plotted as black dots. Red dots represent "original" logistic coefficients computed using full *ChannelAttribution* data. Detailed Figures can be found in Appendix C

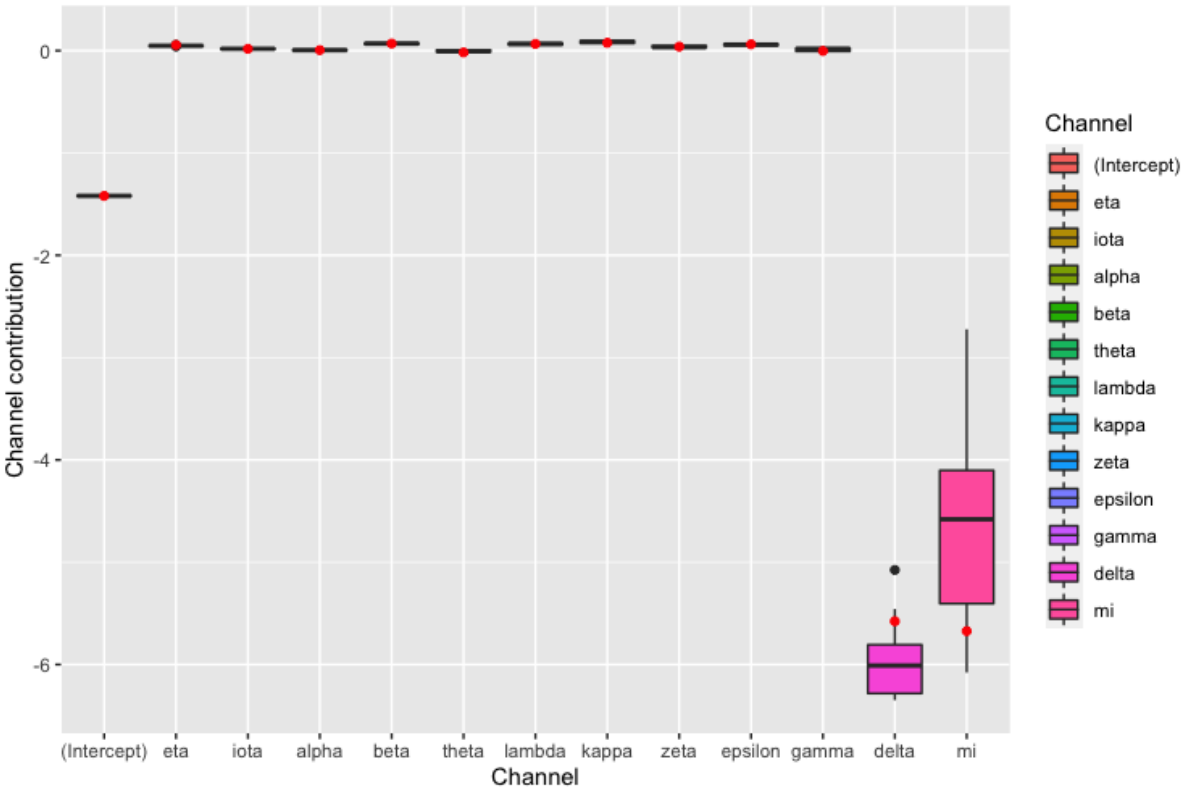

Figure 4: Bagged model attribution on imputed data
Note: box-and-whisker plots represent "new" bagged coefficients under imputed data; with outliers plotted as black dots. Red dots represent "original" bagged coefficients computed using full *ChannelAttribution* data. Detailed Figures can be found in Appendix C.

It can be seen from Table 6 that both the (bagged) logistic regressions are much more affected by imputing missing data. Most notably, the coefficient of *gamma* has more than doubled in either case. This can perhaps be attributed to the fact that this channel is fairly rare in the underlying data, and the 1% missingness could have largely affected it. Most notably, it is a pleasant surprise to see that the values of *beta* and *lambda* have only been affected by a few percentage points in both cases. *theta*, however, seems to be more affected, changing by more than 50% on average in absolute value for both cases.

Inspecting the Figures, however, most of the old coefficients seem to fall into the inter-quartile range of the newly imputed coefficients for all coefficients, but *eta*, *theta*, *mi* and *delta*. This can much better be seen from the individual plots of the variables in Appendix C. As the scaling makes it difficult to fit all the variables on a single plot, the detailed box-and-whisker plots for every variable of the (bagged) logistic regressions can be found in Appendix C.

It is also important to discuss the fact that the coefficients of *delta* and *mi* are largely negative and have much higher variance in both models. The explanation for this is simple, and lies in the sparseness of these variables, as these channels have only been encountered by just over 50 of the 88,000 users in the dataset. This largely affects coefficients of the logistic

regression, explaining the large variance. Further, when those channels are present on the path, there is (on average) four inactive users for every active user. This fact in tandem with the low number of observations leads to a very negative coefficient for both *mi* and *delta*, compared to all the other channels: as essentially, if these channels are present on the path, they can explain why the user has (not) converted much better than the other channels that appear much more frequently. The imputation further contributes to these problems, as a different value for just one user could impact these coefficients significantly, compared to all the other channels.

6 Discussion

Firstly looking at the results of the probabilistic model, the results seem quite promising. Excluding the channel *mi* (the change in $C(x_i)$ for which can be likely linked to an extremely low number of observations for *mi* in the data), the resulting channel attributions from the imputed data differ on average by only 2.1% in absolute value. This average is largely affected by the change in *gamma* and *lambda*'s attribution of 6.18 and 3.63 percentage points respectively. It can be seen from Figure 2 that these percentages are quite heavily affected by two positive outliers in both estimates. Should these be excluded, the absolute percentage change in the final attributions would be even lower. The probabilistic model seems to achieve similar results on the imputed data as on the raw data (compared to LR and BLR), especially on the channels like *mi* and *delta* with lower numbers of observations. This can be partly explained by the fact that the model isn't likelihood-based, compared to (B)LR, meaning it doesn't run into some problems caused by sparse data, such as inverse hessian calculation, for instance.

Considering the (bagged) logistic regressions, the change in coefficients is more drastic than just a few percentage points. Of particular interest are the coefficients of *beta*, *theta* and *lambda*, which have been simulated to have 10% missingness. For *beta* and *lambda*, it is interesting to see that the imputation process has been able to recover enough information such that these coefficients only differ by a few percentage points from those computed using the full data set. This is especially pronounced in the bagged case, where these two differ by less than one percent point in absolute value. The same is not true for *theta*, which seems to have increased by 42 and 74 percent respectively for the simple logistic and bagged logistic model. In total, the coefficients from the imputed data differ in absolute value on average by approximately 12%. Clearly, this value is larger than the above value of approximately two percent for the probabilistic model. The reason for the larger changes has already been mentioned previously: the (bagged) logistic regressions are hardly accurate at predicting the users in the *ChannelAttribution* data, as they predict a user will be inactive the majority of the time. This leads to very low coefficient estimates, and just a small change in these results in a large percentage change in the absolute

value of the coefficient.

As for the differences within the models, the bagged model seems to predict coefficients closer to the ones estimated on the full data, when a part of the traffic is obfuscated and then imputed. This is expected, as the concept of bagging should improve the accuracy of the coefficients. It should be noted once again that these results aren't directly comparable with the probabilistic model, as the coefficient values aren't equal to final channel contributions. As stated previously, Shao and Li (2011) don't provide a way to compute channel contributions from this output.

One should remember, however, that while precision matters, the aim of the "direct traffic imputation" is not to find the most accurate or precise model, but to rather provide the marketeer with another tool to interpret their results. In practice, the marketeer wouldn't know the "real" values of the imputed data, and would therefore be forced to use confidence intervals. And from the Figures in the previous section, as well as Appendix C, it is clear that even using the means and interquartile ranges for the coefficients from just 10 imputations would provide a fair estimate of the coefficients computed using the full data.

6.1 Treating direct traffic as a separate channel

To further demonstrate the importance of taking the extra step to impute the potential effect of direct traffic, one more model will be run on the data. This is to demonstrate what would happen if one were to simply leave the data unprocessed, and treat the direct channel similarly to any other.

Since (bagged) logistic regressions do not provide with a real output of attributions per channel, the probabilistic model will be used for demonstration. This model is exactly the same as in Section 5.1.1, except that all the missing hits were treated as a separate "direct" channel. In other words, the x_{direct} was set to 1 for every user that had a missing data point, and 0 for all others; and all every missing data point was thereby replaced with a 0. The results are presented in Table 7.

It can be seen that while treating direct traffic as its own channel doesn't significantly change the attribution $C(x_i)$ (in fact, the results are comparable to imputing the direct traffic); the problem really arises if one uses direct traffic for computing channel contributions $\frac{100*C(x_i)}{\sum_{i=1}^{K} C(x_i)}$. This figure would, for instance, be used in practice to compute the ROI of each channel; and represents the percentage of total conversion value that each channel is responsible for. It can be seen that the contributions of each channel are negatively biased (on average by -6.57%). This could lead to the conclusion that each channel brought less revenue it really did, reducing their ROI on paper, which could lead to misleading conclusions by the marketer, and ill-driven business decisions.

Table 7: Original probabilistic attribution vs. probabilistic attribution with direct channel

Channel	Attribution $C(x_i)$			Resulting contribution $\frac{100*C(x_i)}{\sum_{i=1}^{K} C(x_i)}$		
	Original	incl. Direct	% change	Original	incl. Direct	% change
direct	-	0.104	-	-	7.90	-
alpha	0.100	0.101	0.306	8.44	7.63	-9.52
beta	0.106	0.106	0.525	8.88	8.05	-9.32
delta	0.072	0.080	10.3	6.07	6.04	-0.505
epsilon	0.112	0.112	0.269	9.41	8.51	-9.56
eta	0.116	0.118	1.30	9.80	8.96	-8.62
gamma	0.095	0.098	2.61	8.00	7.41	-7.44
iota	0.117	0.101	0.61	9.82	8.91	-9.24
kappa	0.103	0.104	1.215	8.67	7.92	-8.70
lambda	0.109	0.109	0.531	9.16	8.31	-9.32
mi	0.044	0.055	26.6	3.69	4.21	14.2
theta	0.102	0.103	0.75	8.59	7.81	-9.12
zeta	0.112	0.110	-2.23	9.47	8.35	-11.8

Note: "original" $C(x_i)$ values taken from Table 3. "incl. Direct" attribution computed using data with simulated missing values, where all missing values are simply assumed to come from a new, "direct" channel.

An attentive reader might point out that simply excluding the direct channel from the calculation of the contributions should bring less biased results. One should remember, however, that direct traffic in practice is often mixed with "true" direct traffic and "unknown" direct traffic; as mentioned previously. In this case, there would still be significant downward bias. Further, separating the "true" and "unknown" direct traffic would also allow for analysis of the "brand awareness direct effect" studied by Kakalejčík et al. (2020).

7 Conclusion

The aim of this Thesis was to develop a technique under which direct traffic to a website could be treated as "unknown" traffic, posing the question: "**To what extent can direct traffic be recovered from marketing attribution data using Multiple Imputation by Chained Equations, and what effect does (not) imputing direct traffic have on various attribution models?**" This effect was simulated into the test data from the *ChannelAttribution* package under a set of arbitrary assumptions on the underlying users. These hits would be treated as "missing values", and the Multiple Imputation by Chained Equations technique was used to generate multiple sets of imputed data. Then, probabilistic, bagged and simple logistic attribution models introduced by Shao and Li (2011) were used to generate attribution output for both the original data, and each of the imputed data sets. It was found that, if the resulting channel contributions from the probabilistic model are averaged, these differ by approximately 2.1% percent from the underlying "full" data. The simple and bagged logistic models were less accurate at predicting the underlying coefficients, with a 12% difference, on average, in the ab-

solute values of the coefficients before and after simulation. Further, avoiding the imputation of the "direct traffic effect" entirely, and simply treating direct traffic as a separate channel would yield the marketer with channel contributions that are downward-biased by -6.57%, according to the probabilistic model. If these values were used for computing channel ROI and future budget allocation decisions, the model would undercut the profit every channel made (compared to the "true" profit if the direct effect were treated), and likely result in the wrong decisions being made.

7.1 Suggestions for future research

It seems that the area of attribution modelling is a culmination of a myriad of statistical techniques. Just to compute a single set of channel contribution coefficients, one might have to use missing data imputation, consider interaction effects, short- and long-term changes in consumer behaviour, dimensionality reduction, deal with extremely large data, run machine learning and clustering models, and much more. In a way, there might always be a way to improve, and this is what is exciting about this field. This Thesis is no exception, thus, here are the suggestions for further research in the area of "direct traffic imputation":

- Attempt to reproduce the results of this Thesis on a real-world dataset. Not only could this provide millions of customer interactions, but also dozens of additional predictors z_p mentioned in Equation 9, such as location data, time data, user demographics, and more; all of which could improve the accuracy of imputing the direct traffic.

- Develop a methodology of deciding which traffic is unknown, and which can actually be classified as direct. This can be done by inspecting the session or hit traffic referrer, and the technical explanation on this goes beyond the scope of this Thesis.

- Attempt to use the "direct traffic imputation" in tandem with more up-to-date attribution models. The probabilistic and (bagged) logistic models introduced by Shao and Li (2011) were all selected due to their simplicity and ease of interpretation. However, in the decade since the inception of that paper, models have been developed higher in both complexity and accuracy. Using such models can provide more stable results.

- Attempt to integrate the process of "direct traffic imputation" into a fully-fledged attribution model, instead of having it be part of the process. As it stands at the moment, the "direct traffic imputation" would be a step a marketeer would need to take before running an attribution model on the dataset. This step could possibly be integrated into an actual attribution model, for instance, a neural network approach. This would lower the complexity of the method, and result in an easier application of the method.

7.2 Research weaknesses

As stated above, the number of factors to consider when building an attribution model are almost limitless. Similarly, the "direct traffic imputation" technique doesn't come without its downfalls, all of which should be considered and investigated further:

- As the approach of Shao and Li (2011) use binary variables (whether or not a specific channel got visited) to compute the attribution models, the imputation could result in a 1 ("true") value imputed, whereas the specific channel wasn't visited in the original data. In essence, this means that the path length of the imputed path could differ from the actual observed path length. This can be fixed by working with a model that uses the number of times a channel was visited in the path, largely reducing this problem. Further, this can be tackled from a different direction, by building an imputation model that would consider each path from a Markovian point of view, for instance, eliminating the problem.

- The "direct traffic imputation" relies on the marketeer being able to identify which traffic should be deemed "missing" in order to be imputed. This is one of the further research areas specified above. Further, an assumption is imposed on which traffic sources direct traffic can come from. In reality, this assumption needs to be validated and further investigated. This traffic is likely to come from many different channels, all with a differing probability. While this is implicitly handled by the regression step in MICE, this can be further investigated and likely identified explicitly. This is better handled by marketeers with many more decades of experience in marketing than the author of this Thesis.

Ultimately, this paper was written with the intent to be a proof of concept for "direct traffic imputation". From the results, it seems that this is a plausible approach. And with the rising privacy concerns in the latest years, only more and more data will be obscured from the marketeer over time. More users will start to be conscious of their data and how it is used, more companies will be forced to give users the freedom to choose how much data they share. It is only logical for marketeers to stop trusting the accuracy of the underlying data, and let statistical techniques take care of the inaccuracies. I truly believe this area of having to use less data to do more will only grow over time, and am exhilarated to see what's to come.

References

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

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

Altomare, D., Loris, D., & Altomare, M. D. (2016). *Package 'channelattribution'.* CRAN.

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, *20*(1), 40–49.

Barzi, F., & Woodward, M. (2004). Imputations of missing values in practice: results from imputations of serum cholesterol in 28 cohort studies. *American journal of epidemiology*, *160*(1), 34–45.

Berger, P. D., & Nasr, N. I. (1998). Customer lifetime value: Marketing models and applications. *Journal of interactive marketing*, *12*(1), 17–30.

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

Bharti, K. (2020). Attribution modelling in marketing: Literature review and research agenda. *Academy of Marketing Studies Journal*, *24*(4).

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140.

Chandler-Pepelnjak, J. (2009). Measuring roi beyond the last ad. *Atlas Institute Digital Marketing Insight*, 1–6.

Donders, A. R. T., Van Der Heijden, G. J., Stijnen, T., & Moons, K. G. (2006). A gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, *59*(10), 1087–1091.

Efron, B. (1994). Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association*, *89*(426), 463–475.

Gruber, K. (2020). *Introduction: Digital communication strategies.* University Lecture.

Kakalejčík, L., Bucko, J., & Danko, J. (2020). Impact of direct traffic effect on online sales. *Journal of Research in Interactive Marketing*.

Kesteren, v. J. (2015). *Multi touch attribution: Searching for the best attribution model* (Master's Thesis). University of Amsterdam.

Lemon, K. N., & Verhoef, P. C. (2016). Understanding customer experience throughout the customer journey. *Journal of marketing*, *80*(6), 69–96.

Li, H., & Kannan, P. (2014). Attributing conversions in a multichannel online marketing

environment: An empirical model and a field experiment. *Journal of Marketing Research*, *51*(1), 40–56.

Lovett, M. J., & Staelin, R. (2016). The role of paid, earned, and owned media in building entertainment brands: Reminding, informing, and enhancing enjoyment. *Marketing Science*, *35*(1), 142–157.

Matthews, J. (2016). *How cooperative game theory can be utilised to enhance marketing analytics attribution* (Unpublished doctoral dissertation). Dublin, National College of Ireland.

McDowell, W. C., Wilson, R. C., & Kile Jr, C. O. (2016). An examination of retail website design and conversion rate. *Journal of Business Research*, *69*(11), 4837–4842.

Mizerski, R. W., Golden, L. L., & Kernan, J. B. (1979). The attribution process in consumer decision making. *Journal of Consumer Research*, *6*(2), 123–140.

MSI. (2020). Research priorities 2020–2022. *Cambridge, Mass.: Marketing Science Institute*.

Naik, P. A., Raman, K., & Winer, R. S. (2005). Planning marketing-mix strategies in the presence of interaction effects. *Marketing Science*, *24*(1), 25–34.

Neslin, S. A., Grewal, D., Leghorn, R., Shankar, V., Teerling, M. L., Thomas, J. S., & Verhoef, P. C. (2006). Challenges and opportunities in multichannel customer management. *Journal of service research*, *9*(2), 95–112.

Nouwens, M., Liccardi, I., Veale, M., Karger, D., & Kagal, L. (2020). Dark patterns after the gdpr: Scraping consent pop-ups and demonstrating their influence. In *Proceedings of the 2020 chi conference on human factors in computing systems* (pp. 1–13).

Quinlan, J. R., et al. (1996). Bagging, boosting, and c4. 5. In *Aaai/iaai, vol. 1* (pp. 725–730).

Reichert, C. (2021, May). *App tracking has only 5% opt-in rate since ios 14.5 update, analyst says.* CNET. Retrieved from `https://cnet.co/2RitSfB`

Rosnow, R. L., & Rosenthal, R. (1989). Definition and interpretation of interaction effects. *Psychological Bulletin*, *105*(1), 143.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Rust, R. T., Lemon, K. N., & Zeithaml, V. A. (2004). Return on marketing: Using customer equity to focus marketing strategy. *Journal of marketing*, *68*(1), 109–127.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical methods in medical research*, *8*(1), 3–15.

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

Sokol, D. D., & Zhu, F. (2021). Harming competition and consumers under the guise of

protecting privacy: An analysis of apple's ios 14 policy updates. *Available at SSRN*.

Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, *30*(4), 377–399.

Wodinsky, S. (2021, Jun). *Apple's biggest privacy flexes of wwdc 2021.* Gizmodo. Retrieved from `https://t.co/0ZZNZJ6Mjm`

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

Yadagiri, M. M., Saini, S. K., & Sinha, R. (2015). A non-parametric approach to the multi-channel attribution problem. In *International conference on web information systems engineering* (pp. 338–352).

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

Zhang, Z. (2016). Missing data imputation: focusing on single imputation. *Annals of translational medicine*, *4*(1).

A Appendix A

This Appendix contains the description of the code used to generate the results and manipulate the data used in this Thesis. Attached to the PDF version of this Thesis is a zip archive of the code used to produce the results, figures and tables in this Thesis. Below is a description of all the files in this archive.

- *setup.R* - loads the required libraries & does basic data manipulations used by other files.
- *expand_data.R* - takes the data from the *ChannelAttribution* package (in which each row represents a unique path), and generates a data frame where the rows represent unique users, and columns represent whether each unique channel was present on the path, and whether the user converted.
- *data_description.R* - generates summary statistics for Section 3.
- *bagged.R* - a function that runs the (bagged) logistic regressions on input data. Inputs: data.frame (generated by *expand_data.R*), S, M, ratio of active to inactive users, p_c and p_s, where all variables are defined in Section 4.1. Output: a list of resulting coefficients from logistic regression, coefficients from bagged regression, A-metric of logistic regression, A-metric of bagged regression, V-metric of logistic regression, and V-metric of bagged regression, exactly in this order.
- *probabilistic.R* - a function that runs the probabilistic attribution model on the input data.
- *imputation.R* - simulates missing values in the data, imputes new data sets to run attribution models on, runs the attribution models, and plots the required results.
- *direct_channel.R* - converts all missing hits to a new "direct" channel for Section 6.1, and generates required output.

B Appendix B

Table 8: Comparison of the bagged logistic regression (BLR) and the usual logistic regression (LR) in terms of the V-A metric (sample size = 800)

			p_c					
			0.25		0.50		0.75	
			V-metric	A-metric	V-metric	A-metric	V-metric	A-metric
p_s	0.25	LR	0.924	0.224	0.791	0.226	0.704	0.224
		BLR	1.29	0.223	1.22	0.226	1.23	0.224
	0.50	LR	0.838	0.224	0.740	0.225	0.892	0.224
		BLR	0.954	0.224	0.940	0.225	1.06	0.224
	0.75	LR	0.775	0.223	0.624	0.224	0.584	0.225
		BLR	0.816	0.224	0.740	0.224	0.699	0.226

C Appendix C

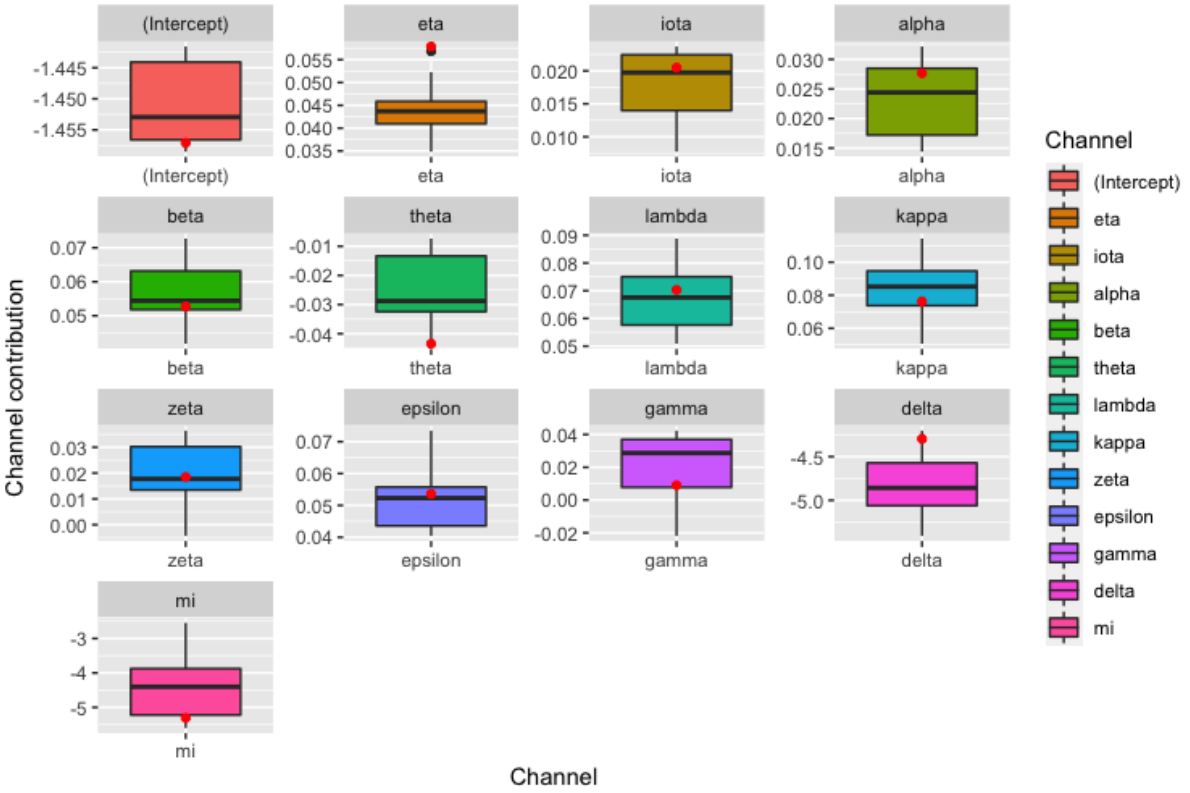

Figure 5: Logistic model attribution on imputed data per variable
Note: box-and-whisker plots represent "new" logistic coefficients under imputed data; with outliers plotted as black dots. Red dots represent "original" logistic coefficients computed using full *ChannelAttribution* data.

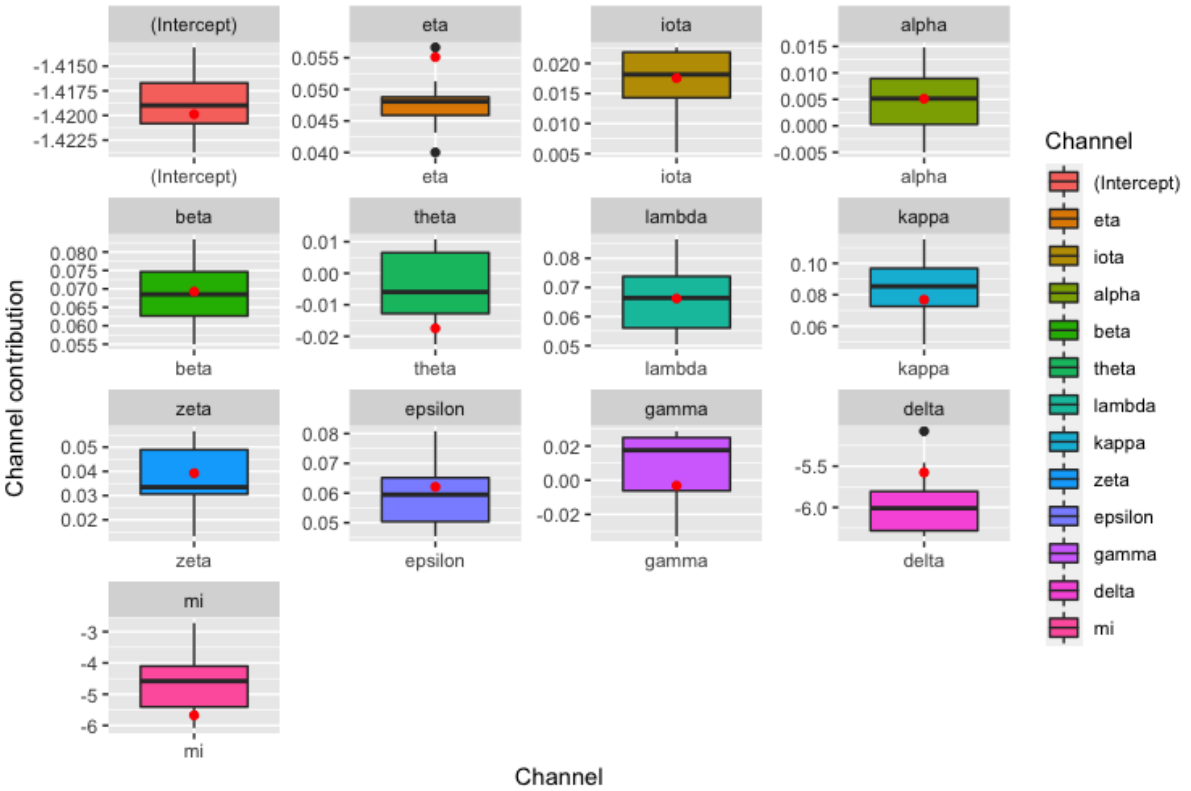

Figure 6: Bagged model attribution on imputed data per variable
Note: box-and-whisker plots represent "new" bagged coefficients under imputed data; with outliers plotted as black dots. Red dots represent "original" bagged coefficients computed using full *ChannelAttribution* data.