DEGREE PROJECT IN THE FIELD OF TECHNOLOGY
INDUSTRIAL ENGINEERING AND MANAGEMENT
AND THE MAIN FIELD OF STUDY
MATHEMATICS,
SECOND CYCLE, 30 CREDITS
STOCKHOLM, SWEDEN 2021

Insights on Creating a Growth Machine Using Attribution Modelling

HANNES KINDBOM

VIKTOR REINECK

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF ENGINEERING SCIENCES

Insights on Creating a Growth Machine Using Attribution Modelling

HANNES KINDBOM

VIKTOR REINECK

Degree Project in Optimization and Systems Theory (30 credits)
Master's Programme, Industrial Engineering and Management (120 credits)
Date: June 24, 2021
Host company: Hedvig AB
School of Engineering Sciences
Supervisor and examiner at KTH: Per Enqvist
Supervisor at Hedvig: John Ardelius
Swedish title: Insikter kring skapandet av en tillväxtmaskin med attributionsmodellering

Insights on Creating a Growth Machine Using Attribution Modelling / Insikter kring skapandet av en tillväxtmaskin med attributionsmodellering

Abstract

Given access to detailed tracking data, the problem of attribution modelling has recently gained attention in both academia and the industry. Being able to determine the influence of each marketing channel in driving conversions can help advertisers to allocate their marketing budgets accordingly and ultimately increase their customer base and achieve a higher Return On Investment (ROI). However, Last-Touch Attribution (LTA), the current industry standard to approach the problem, has been criticized for oversimplification.

In this degree project, two data-driven attribution models are therefore compared to the LTA model on real data from an insurance company, with the objective to optimize for customer base growth and ROI. Raw attributions for each channel are obtained after training the models to predict conversion or non-conversion. By using a linear function to obtain a Customer Lifetime Value (CLV) estimate, the attributions are then adjusted to the ROI of each channel and finally validated through an attribution based budget allocation and historical marketing data replay.

The experimental results demonstrate that all models reach approximately 82% accuracy on balanced data, just below the calculated theoretical maximum. While current research consistently argues for more complex data-driven Multi-Touch Attribution (MTA) models, this project provides a nuance to this field of research in showing that the LTA model may, in fact, be suitable in some cases. A new approach to develop specialized models based on correlations between conversion and contextual variables, then shows that attribution models for mobile users specifically yield higher accuracy. The sum of such unnormalized attributions function as indicators for the conversion strength of contextual variables and can further assist decision making.

Keywords

Digital marketing, Attribution modelling, Last-touch attribution, Multi-touch attribution, Simple probabilistic, Logistic regression

Sammanfattning

Givet tillgång till detaljerad spårningsinformation har attributionsmodellering nyligen fått uppmärksamhet i både akademin och näringslivet. Att kunna förstå påverkan varje marknadsföringskanal har på att driva konverteringar, kan underlätta för annonsörer att fördela marknadsföringsbudgetar och i slutändan öka antalet kunder samt uppnå en högre avkastning på investeringen. Last-Touch-Attribution (LTA), den nuvarande branschstandarden för att angripa problemet, har emellertid kritiserats för att vara överförenklande.

I det här examensarbetet jämförs därför två datadrivna attributionsmodeller med LTA på verklig data från ett försäkringsbolag med målet att optimera för kundbastillväxt och avkastning. Råa attributioner för varje kanal erhålls efter att modellerna tränats på att prediktera konvertering eller icke-konvertering. Genom att estimera kundens livstidsvärde med en linjär funktion, justeras attributionerna sedan med avkastningen på investering för varje kanal och valideras slutligen genom en attributionsbaserad budgetallokering och uppspelning av historisk marknadsföringsdata.

De experimentella resultaten visar att alla modeller når ungefär 82% träffsäkerhet på balanserad data, strax under det beräknade teoretiska maximivärdet. Medan aktuell forskning konsekvent argumenterar för mer komplexa datadrivna multi-touch-modeller, ger det här projektet en nyans till forskningsfältet genom att visa att LTA i vissa fall kan vara lämplig. Ett nytt tillvägagångssätt för att utveckla specialiserade modeller baserade på korrelationer mellan konvertering och kontextuella variabler, visar sedan att attributionsmodeller för enbart mobilanvändare ger högre träffsäkerhet. Summan av sådana onormaliserade attributioner fungerar som indikatorer på konverteringsstyrkan för kontextuella variabler och kan ytterligare underlätta beslutsfattandet.

Nyckelord

Digital marknadsföring, Attributionsmodellering, Last-touch attribution, Multi-touch attribution, Simpel probabilistisk, Logistisk regression

Acknowledgments

We would first like to express our sincere appreciation to our supervisor at KTH Royal Institute of Technology, Per Enqvist, for support and feedback throughout this project. We would also like to acknowledge the endless support and guidance from our friends at Hedvig. We are indebted to our excellent supervisor John Ardelius for his skills, ideas and advice. He has been truly instrumental as part of this project. Growth geniuses Carl Lager, Sonny Andersson and Emil Wallerstedt all deserve maximum credit as well.

Special thanks also to our peer reviewers and class mates for both valuable and inspiring comments, advice and suggestions on how to shape and improve this project.

Finally, we are incredibly grateful for unfailing support and continuous counsel from our friends, families and loved ones throughout the entire process of conducting this project. Our most heartfelt thank you!

Stockholm, June 2021
Hannes Kindbom and Viktor Reineck

Contents

1 Introduction **1**

1.1 Problem Background 2

1.1.1 Problem and Definition 3

1.2 Purpose and Goals 3

1.3 Delimitations and Assumptions 4

1.4 Outline 4

2 Background **6**

2.1 Google Analytics 6

2.2 Attribution Models 7

2.2.1 Simple Probabilistic 7

2.2.2 Logistic Regression 8

2.2.3 Last-Touch Attribution 8

2.3 Evaluation 9

2.3.1 Cross-Validation 9

2.3.2 Performance Metrics 10

2.4 Related Work 11

2.4.1 Modelling 11

2.4.2 Data Format 12

2.4.3 Models 12

3 Method **14**

3.1 Data 14

3.1.1 Data Collection 14

3.1.2 Data Processing 15

3.1.3 Exploratory Data Analysis 16

3.2 Model Implementation 17

3.3 Customer Lifetime Value Estimation 19

3.4 Control Variables 20

3.4.1 Conversion Strength of Control Variables 21
3.5 ROI-Based Attributions 21
3.6 Evaluation Protocol 22

4 Results 24
4.1 Exploratory Data Analysis 24
4.2 Control Variable Correlations 26
4.3 Predictive Performance 26
4.4 Attributions 28
4.5 Evaluation Protocol 30

5 Discussion 32
5.1 Interpretation of Results 32
5.1.1 Data 32
5.1.2 Predictive Performance 32
5.1.3 Attributions 33
5.1.4 Control Variables 33
5.1.5 Evaluation Protocol and ROI 34
5.2 Limitations 34
5.2.1 Data 34
5.2.2 Feature Engineering 36
5.2.3 Evaluation 37
5.3 Conclusions 37
5.4 Future Work 38

References 39

List of Acronyms

CLV	Customer Lifetime Value
CPC	Cost Per Click
CV	Cross-Validation
LR	Logistic Regression
LTA	Last-Touch Attribution
MTA	Multi-Touch Attribution
p.p.	percentage points
ROI	Return On Investment
SP	Simple Probabilistic

Chapter 1

Introduction

Following the rise in the number of internet users around the world, the interest for marketing via the internet has grown all the same [1]. Such digital marketing methods generally offer marketeers greater abilities to both target and track marketing efforts and individual campaigns than has ever been possible with traditional offline marketing. Prominent *channel* types in digital marketing include for instance social media, search engines and e-mail advertising. For these channels, data are generally much more accessible than for marketing with traditional channels, such as TV and billboards.

The goal of marketeers in digital marketing is typically to direct people, belonging to a target group or cohort, to a company's website. This applies particularly in e-commerce settings, where the ultimate goal often is to sell the visitor a product or service, commonly referred to as a *conversion*. However, it has been found that upwards of 95-97% of website visits do not result in a conversion [2]. Since paid marketing channels typically charge a *Cost Per Click (CPC)* on an ad, this raises the question on how to optimize marketing budgets to maximize the number of customers or *Return On Investment (ROI)*, for instance.

Clearly, it is undesirable to spend money on a marketing channel that yields *clicks*, but no immediate nor future conversion. If however a person clicks on multiple, possibly different, channels before converting, all those channels should perhaps be given some recognition in supporting the conversion. Thus, it is desirable to be able to track what is referred to as the *customer journey* of a person, i.e., the sequence of ads a user has been interacting with before either converting or not. A customer journey leading to a conversion is referred to as a *positive* customer journey, and vice versa as a *negative* customer journey. An example of a positive customer journey is provided in Figure 1.1. In digital marketing, thanks to the access to relevant data, it is often possible to map these journeys to varying degrees of detail.

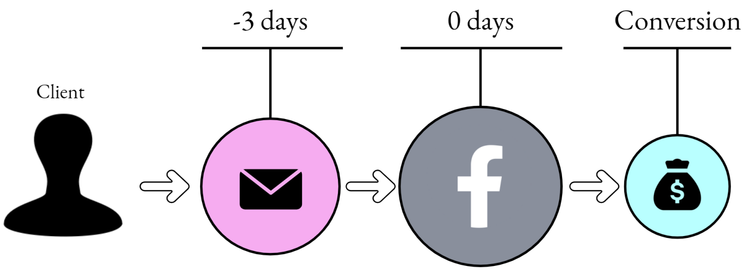


Figure 1.1 – An example of a positive customer journey where the user clicked on a link in an email three days before converting via a Facebook ad

Given access to detailed data, the problem of *attribution modelling* has increasingly gained attention in both academia and industry [1, 3, 4]. In this marketing context, attribution modelling refers to the problem of determining *attributions*, i.e., the influence of different marketing channels in driving conversions. Being able to estimate such a weight or importance of each channel can help marketeers in the allocation of marketing budgets and aid the development of informed strategies.

This problem has traditionally been approached by rule-based heuristic methods, where *Last-Touch Attribution (LTA)* is the current industry standard [4, 5]. It attributes the entire value of a conversion to the last interaction channel prior to a conversion. It is however frequently criticized for oversimplifying the problem and thus yielding skewed results [5]. Particularly, it may overestimate channels that naturally tend to be closer to a conversion, such as search engines, and underestimate others. This has motivated the birth of more sophisticated data-driven *Multi-Touch Attribution (MTA)* algorithms in accounting for all clicks in a customer journey, and valuing their positive and negative conversion outcomes. This type of method to allow for optimizing marketing spend is the subject to be explored in this degree project.

1.1 Problem Background

The Swedish insurance company Hedvig is currently undergoing a stage of bold growth objectives where increasing the number of customers is one of their main goals. On behalf of the project partner Hedvig, this degree project is therefore focused on how to find the optimal budget allocation among digital marketing channels through attribution modelling. Given the growth phase that Hedvig is in, the importance of assigning marketing spend optimally is eminent. However, the problem of how to assign such a budget is multifaceted. Adding complexity to the problem is the trust business-aspect of insurance. Customer journeys likely differ between one-time purchase e-commerce and subscription-based services. The subscription aspect of insurance also motivates the need for estimating *Customer Lifetime Value (CLV)* in an attribution model. Conversions are not necessarily valued equally, as CLVs likely vary greatly between individual customers in terms of, for instance, premium and loyalty metrics, such as subscription length. It is moreover probable that customers to a greater extent have to trust an insurance company before becoming a customer, when compared to a product company. A hypothesis termed by Hedvig is thus that these customer journeys span a longer time period and contain more interactions than for other types of products or services.

1.1.1 Problem and Definition

The original problem formulation as stated by Hedvig is to investigate how an attribution model may be created for a trust reliant business. Moreover, sub-problems include exploring how such a model can be optimized with regards to ROI and customer base growth, respectively.

This leads to the following research question:

> *How to allocate a budget over a number of marketing channels to optimize for ROI and customer base growth, respectively?*

Answering this research question is one part of building, what could be referred to as, a *growth machine*. This abstract concept refers to a marketing strategy, whose benefits exceed its costs, that can be used for growth until the market is saturated.

1.2 Purpose and Goals

The overall purpose of this thesis is to explore the subject of attribution modelling from an optimization perspective. Particularly, with the inclusion of aspects such as CLV and *control variables*, the aim is to provide an edge and contribute to current attribution modelling research. The goal is further to construct a model that calculates an optimal proportion of marketing spend to be allocated among available channels, given a set of specified arguments and conditions. Moreover, such a model is to fulfil a number of requirements as set in association with Hedvig, such as interpretability and accuracy. Ultimately, the goal is to provide tools and insights for Hedvig to allow for increasing future growth. To achieve this in an economically sustainable way, the maximization of return on marketing investments needs to be considered as well.

Since the majority of data used in this degree project derive from free and widely used software, specifically Google Analytics, our aim is further to allow for companies and organizations in similar positions to benefit from these results in the maximization of returns on marketing spend.

1.3 Delimitations and Assumptions

Firstly, this degree project is delimited to the Swedish market, for which the majority of the tracking data exist. Since individual web traffic data collection was initiated in February 2021, data collected during a time window of approximately two months were used throughout the project. This time window is further limited by the life span of cookies. This means that seasonal variations are to some extent disregarded.

Furthermore, only limited cross device tracking is possible with Hedvig's current implementation of Google Analytics, in practice limiting this project to only single device tracking. Additionally, this setup only allows for data on clicks, while data on *impressions*, i.e., views of an ad, are unavailable. The models in this project are also developed without differentiating between individual advertising campaigns.

1.4 Outline

The remainder of this report starts with a background in Chapter 2 and an overview of attribution modelling and relevant models, ending with a section about related work. Subsequently, Chapter 3 provides a description of the data and models used in answering the research question. Chapter 4 follows with a review of the various results obtained. Lastly, Chapter 5 concludes the report with a discussion, conclusion and suggestions for future work. Overall, the workload has been divided equally between the authors, with only minor deviations.

Hannes Kindbom is also writing another degree project report in parallel, which is building upon the results of this project. It investigates a different attribution modelling approach, using a deep learning model on two datasets and focuses on explicitly validating its attributions. It is further mentioned in the Future Work section 5.4 and may preferably be read as a continuation of this report.

Chapter 2

Background

This chapter is meant to give the reader the prerequisites necessary for understanding the report. It begins with a brief introduction to Google Analytics, specifically in the context of attribution modelling. Thereafter, the theoretical background of relevant attribution models is provided together with evaluation methods. A section of related work concludes the chapter.

2.1 Google Analytics

The most widely used tool in the industry for tracking and analyzing website traffic is Google Analytics [6]. The penultimate generation of Google Analytics, called Universal Analytics, is used in this project. It is *session* based, meaning that website visits are registered based on how much time the visitor spends on the website's different pages. More precisely, a session consists of a user's website interactions appearing within a certain time interval. An interaction may for instance be a page view or a transaction. A session is generally terminated after 30 minutes of inactivity, at midnight or if a user arrives via a campaign, leaves and arrives via a different campaign.

In addition to sessions, the concepts *source* and *medium* are relevant within Google Analytics. The source refers to the origin of the visitor while medium refers to how the visitor arrived at the website. Examples of sources include Google and Facebook, while Organic and cpc are examples of mediums. In this context, cpc does not denote Cost Per Click, but paid traffic from search engines. Oppositely, Organic refers to free traffic coming from search engines.

On a technical level, traffic tracking is enabled by adding a page tag, effectively a snippet of JavaScript code, to each page of the website. Together with Urchin Tracking Module (UTM) tags, which allow for logging which website or ad a visitor originates from, it is then possible to use Google Analytics to collect detailed data on each visitor. The tags can also assign HTTP cookies to the visitors' browsers. Different cookies have different life lengths, ranging from seconds to two years. Using cookies and a client id, Google Analytics can thus be set up to track individual users over time. Specifically, a pseudonymous client id is assigned to every browser instance [7].

2.2 Attribution Models

Data-driven attribution modelling is commonly approached as a classification problem with binary response y_i for click sequences, equivalent to customer journeys, $i = 1, 2, ..., n$. A positive label $y_i = 1$ shows that the sequence resulted in conversion and a negative label $y_i = 0$ represents a *non-conversion*. Furthermore, a sequence of length m_i corresponds to a series of channel clicks made by a unique user. The channel attributions can then be obtained either from trained model parameters or some additional algorithm. The theoretical background to models used in this degree project is presented in this section.

2.2.1 Simple Probabilistic

In an attempt to achieve improved interpretability and low estimation variability, a first version of the *Simple Probabilistic (SP)* model was introduced in 2011. It was claimed to be the industry's first commercially available data-driven MTA model [8]. SP is based on conditional conversion probabilities $P(Y = 1 \mid k)$, which are empirically estimated as in Equation 2.1 for each channel k. The probability $P(Y = 1 \mid k)$ forms the attribution Attr'_k for channel k. $n_+(k)$ and $n_-(k)$ refer to the number of occurrences of channel k in sequences resulting in conversion and non-conversion respectively.

$$P(Y = 1 \mid k) = \frac{n_+(k)}{n_+(k) + n_-(k)} \tag{2.1}$$

The estimated conversion probability for a sequence i may then be calculated as in Equation 2.2 [9]. Intuitively, it defines the complement to not converting based on any of the clicks in the sequence, when assuming independence.

$$P(Y_i = 1 \mid \boldsymbol{c_i}) = 1 - \prod_{j=1}^{m_i} (1 - P(Y = 1 \mid c_{ij})) \tag{2.2}$$

$\boldsymbol{c_i}$ refers to a sequence, which is a list of chronologically ordered channels, corresponding to clicks in the customer journey. Moreover, c_{ij} is the channel k of the j:th click and m_i is the length of sequence i.

2.2.2 Logistic Regression

Logistic Regression (LR) may be used for binary classification problems, with the response variable $y_i \in \{0, 1\}$ for each sample i. It is also assumed that Y_i is a Bernoulli random variable that takes value 1 with probability p and 0 with probability $1 - p$. To ensure that the model output $\hat{y}_i$ is on the interval $[0, 1]$, the nonlinear logistic function is applied as in Equation 2.3.

$$P(Y_i = 1 \mid \boldsymbol{x_i}) = \hat{y}_i = \sigma(\boldsymbol{x_i}) = \frac{1}{1 + \exp(-\boldsymbol{\beta}^\top \boldsymbol{x_i})} \tag{2.3}$$

where $\boldsymbol{\beta}$ is a vector containing trainable coefficients and $\boldsymbol{x_i}$ is the feature vector, padded with a 1 to account for the intercept, if included [10]. An advantage of the LR model is the interpretability of its parameter $\boldsymbol{\beta}$. By rearranging Equation 2.3, an expression for the odds (Equation 2.4) and log-odds (Equation 2.5) respectively can be stated.

$$\frac{P(Y_i = 1 \mid \boldsymbol{x_i})}{1 - P(Y_i = 1 \mid \boldsymbol{x_i})} = \exp(\boldsymbol{\beta}^\top \boldsymbol{x_i}) \tag{2.4}$$

$$\log\left(\frac{P(Y_i = 1 \mid \boldsymbol{x_i})}{1 - P(Y_i = 1 \mid \boldsymbol{x_i})}\right) = \boldsymbol{\beta}^\top \boldsymbol{x_i} \tag{2.5}$$

To train the model, some numeric optimization technique is used to maximize the log-likelihood function (defined in Equation 2.6), under the assumption of n independent observations. The exact optimization solver depends on the implementation.

$$\log L(\boldsymbol{y}, \boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \boldsymbol{\beta}^\top \boldsymbol{x_i} - \sum_{i=1}^{n} \log[1 + \exp(\boldsymbol{\beta}^\top \boldsymbol{x_i})] \tag{2.6}$$

2.2.3 Last-Touch Attribution

The industry standard attribution model LTA is essentially a simpler version of the SP model as described in Section 2.2.1. The LTA model takes into consideration only the last click when estimating the conversion probability for a sequence i, as in Equation 2.7.

$$P(Y_i = 1 \mid \boldsymbol{c_i}) = P(Y = 1 \mid c_{im_i}) \tag{2.7}$$

where m_i is the length of click sequence $\boldsymbol{c_i}$. Moreover, the conversion probability given click on some channel k is defined as in Equation 2.8. $n_{LT+}(k)$ and $n_{LT-}(k)$ refer to the number of occurrences of channel k as the last click in sequences resulting in conversion and non-conversion, respectively. Similarly to SP, $P(Y = 1 \mid k)$ becomes the attribution Attr'_k for channel k.

$$P(Y = 1 \mid k) = \frac{n_{LT+}(k)}{n_{LT+}(k) + n_{LT-}(k)} \tag{2.8}$$

2.3 Evaluation

2.3.1 Cross-Validation

By partitioning the data into two disjoint sets, models can be evaluated by being trained on one set and tested on the other. In *l-fold Cross-Validation (CV)*, the dataset is split into l equally or nearly equally sized sets, for which one is left for evaluation, whereas the remaining $l - 1$ sets combined form the training set. These sets cross-over in successive iterations to make sure every data point is used in the test set once [11]. This way, l different results are obtained, allowing for e.g. central tendency and dispersion metrics in obtaining aggregate results. As such, CV is used as a model validation tool, while also assessing model generalizability. $l = 5$ or 10 are common choices in machine learning to balance the bias-variance trade-off, where a smaller value of l results in a lower variance but higher biased estimate and vice versa [12, 13].

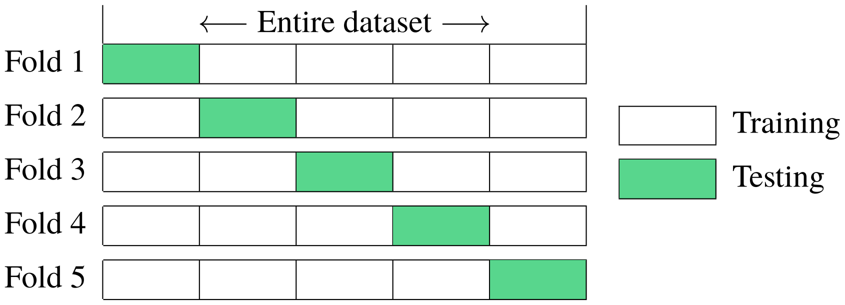


Figure 2.1 – An overview of the steps in 5-fold cross-validation

2.3.2 Performance Metrics

In binary classification settings, a number of metrics are available to evaluate the performance of a model. Many of these stem from the elements of a *confusion matrix*, in which test data are separated into four categories, as in Figure 2.2.

		Actual	
		Positive	Negative
Predicted	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Figure 2.2 – Confusion matrix

If the total number of samples in the dataset is $n = TP + TN + FP + FN$, then some of the more common metrics to evaluate model performance may be listed as in Equation 2.9a - 2.9d.

$$\text{Accuracy} = \frac{TP + TN}{n} \tag{2.9a}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.9b}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.9c}$$

$$\text{F}_1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} = 2\,\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.9d}$$

In all above cases, scores are between 0 and 1 (equivalent to 0% and 100%, respectively), where a higher value suggests higher performance. Some of the metrics alone may be misleading, for instance in situations with unbalanced data. Altogether, they do however provide a detailed view of the model performance. F_1 in particular does provide a balance between both precision and recall, while also being better suited to handle class imbalance [14].

2.4 Related Work

2.4.1 Modelling

It is common in academia to approach attribution modelling similarly to how a *feature importance* problem typically would be approached. In the context of machine learning, this refers to the analysis of input features in their usefulness of predicting some output [15]. In many cases of attribution modelling studies, researchers have created a model to predict the outcome of a customer journey, before subsequently analyzing the influence of each channel in making such predictions. By the nature of prediction models, the approach in these studies has been to train and evaluate a model on sequences of clicks and/or impressions. For simpler models, independent variables are based on those sequences, with a binary response. The objective has thus been to create a model that predicts whether a sequence of clicks results in a conversion or not. The output value $\hat{y}_i$ of such a model would then be transformed to a binary conversion prediction by rounding, $\lfloor \cdot \rceil$, such that

$$\lfloor \hat{y}_i \rceil : [0, 1] \rightarrow \{0, 1\}$$

With a suitable method that yields desirable predictive performance, some method to extract the features' influence on prediction has subsequently been utilized to yield attributions, further discussed in Section 2.4.3. Often, these attributions are then normalized to allow for comparisons between models or model configurations.

Some of the information in click sequences is however lost when using such binary classification methods, particularly the cost and value associated with each customer journey. The inclusion of this information may potentially translate to different channel attributions. To account for this problem, an evaluation protocol has been proposed [9]. By first estimating the ROI of a channel based on money spent and total *conversion value*, and then combining with raw attributions, ROI-adjusted attributions were obtained. This way, both cost, value and attributions were taken into account. This allocation could then be used in an evaluation protocol algorithm on historical data, to allow for evaluating how a model would perform in a real scenario given a certain budget. The results from such a replay of data gives a lower bound on the true performance of attributions [9]. This proposed method in attempting to validate the effectiveness of obtained attributions is one of few found in academia. *A/B-testing* stands as a costly alternative, in which hypothesis testing is conducted based on results from real-life trials.

To improve usefulness of attribution results, researchers have proposed a method that also includes contextual features in estimating a set of new attributions for certain categories [16]. With automatic segment detection, the authors explored variables for which attributions were statistically significantly different from attributions gained when evaluating all data. This way, attributions on a more refined level is available to the marketeer. In their study on real data, they found that attributions within mobile operating systems corresponded to the biggest statistical differences in attributions when evaluated using Akaike Information Criterion (AIC).

2.4.2 Data Format

In applied attribution modelling, studies using data from web analytic suites such as Google Analytics [17] and Adobe Analytics [16] are somewhat common. A few studies have used aggregated data, in which the number of conversions, non-conversions and total conversion value per unique sequence of channels are available [17]. Seldom have synthetic data been used [18]. Most prominent in research are data on individual customer journeys with varying degrees of detail about the user, sometimes including contextual features. These may include information about device type, e.g. mobile or desktop, web browser, e.g. Google Chrome or Apple Safari and region, e.g. Europe or North America. Many such features are available alongside click data obtained in web analytics suites. Since clicks are more easily logged and obtainable in a large scale, only a few identified studies include impression data [8, 9, 19].

2.4.3 Models

One of the earliest studies in the subject suggested using a bagged LR model [8]. In such an ensamble model, a LR model is constructed a number of times for independent samplings of the original dataset, to allow for more accurate and stable results when combined. With each channel modelled as a feature, averaged coefficients could then be interpreted as attribution weights if the model was used to predict conversions. When compared to a combined first- and second order SP model, the two yielded similar results in terms of attributions. In such a simpler model, the conditional probabilities themselves allow for direct interpretation as attribution weights, in turn allowing for great interpretability. Both models were evaluated on real-world data and the SP model was found to offer the lowest variability of the two.

After criticizing the LR method for non-trivial interpretation of coefficients in attribution modelling, some researchers conducted a study using ideas from cooperative game theory [18]. Based on the positioning of attribution modelling as a causal estimation problem, two alternative models were suggested, in which Shapley values were integral. As Shapley values normally are used to evaluate the importance of different players in a cooperation game, it was found to be an applicable method in attribution modelling. In the study, it was also found that advertisers would tend to assign credit to themselves when the user was, in fact, willing to convert no matter the exposure of a certain ad. By estimating causal relationships, the authors suggested their proposed models would be able to bypass such issues.

By considering channels as states in a system, a Markovian approach has been taken on attribution modelling in a few studies [17, 20]. The higher-order Markov chains were augmented with states for start, conversion and non-conversion. Aggregate data were then used to construct transition probabilities in higher-order Markov chains. Such models have the benefit of taking into account potential spill- and carryover effects where users' previous clicks affect future clicks. Attribution weights were estimated using removal effects as the change in conversion probability following the removal of a channel state. It was found that a majority of customer journeys in an e-commerce dataset resulted in a conversion after more than five clicks [17]. When compared to rule-based alternatives, both studies found that results varied greatly between models, suggesting the Markov model better captured patterns. Moreover, the graph-based nature of such models allow for great interpretability compared to some other models.

Different deep learning models have been used to a great extent in recent attribution modelling research novel examples including DARNN [9] and DNAMTA [19], both based on Long Short-Term Memory (LSTM) neural networks. As a benchmark, more primitive models were evaluated on the same data in both studies. When compared to models such as LTA, LR and SP, the two neural network models scored approximately 4-8 percentage points (p.p.) higher in Area Under the receiver operating characteristic Curve (AUC). Additionally, LTA achieved the lowest accuracy. This suggests that the more advanced models outperformed other models. As for extracting attribution weights from the models, the two studies employed different methods. However, the validity of such attributions may be questioned and are difficult to confirm. Particularly when taking into account properties such as complexity, interpretability and maintainability, these results may suggest that simpler models are, in fact, adequate for the task of attribution modelling.

Chapter 3

Method

In this chapter, the general method used in this degree project is described. Starting by outlining data collection, processing, exploratory data analysis and model implementation steps follow. Subsequently, methods to estimate CLV and to find suitable control variables are proposed. Finally, a description of the method used for evaluating the models concludes the chapter.

Note that while the data used cannot be made publicly available, the code used is available on Github *.

3.1 Data

3.1.1 Data Collection

As for clicks and customer journeys, suitable data were collected from February 2021 onwards. Particularly, the Google Analytics feature for collecting `client_id` data was only then enabled, allowing for differentiating individual users and their respective journeys. Together with the collection of click data on website visitors and their click history, were also some contextual data, such as device type and location. Conversions via platforms other than web, e.g. apps, were not included in this data.

To analyze costs, marketing spend data were obtained from Funnel, a marketing data collection software. The most detailed data format available was total spend per day per channel. To then estimate CPCs, this cost was divided by the total number of clicks from a certain channel during each day.

For converted users, additional user data were available in Mixpanel, a business analytics software. Using detailed features such as premium and

* `https://github.com/hkindbom/attribution-modelling`

number of co-insured people, it was then possible to form estimates of CLVs.

In total, data were collected during 64 days and subsequently used in this project, resulting in a raw data sample size of 87 882 clicks from 73 326 unique users.

3.1.2 Data Processing

The Python library Pandas 1.2.1 was used for data manipulation and processing. In a first step, duplicate data samples were removed. Additionally, any clicks made after a conversion were removed to make sure that the models only captured behaviour leading up to a conversion in positive customer journeys.

To further make sure that the models captured full user behaviour and to minimize the number of false negatives, the dataset was filtered for a time-specific cohort. Given click data for some time period (t_0, T), the objective was to make sure that only customer journeys in their entirety were contained in such data. Thus, only users with a first click within a subinterval (t_A, t_B) were kept in the cohort, see Figure 3.1. In this degree project, data collection was initiated on $t_0 =$ 2021-02-03 and continued until $T =$ 2021-04-07. The length of the time interval between $t_A =$ 2021-02-24 and $t_B =$ 2021-03-17 was considered a reasonable trade-off between amount of data and minimization of false negatives and noise. Since Hedvig's budget allocation changes over time but is considered static in this project, a too long time interval (t_A, t_B) would decrease the accuracy of the results with respect to model assumptions. However, when using these attribution models in applied settings, this time interval should ideally be a sliding window such that models are retrained regularly.

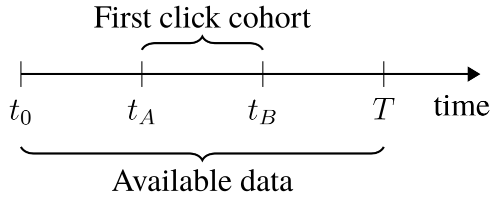


Figure 3.1 – Description of the data availability period and cohort time interval

As Hedvig had only recently started collecting suitable data, there were a limited number of data samples available for some channels. These channels have either received less funding or were simply less effective in gaining clicks. To make sure that the models would be trained on a dataset large enough for each channel, the most uncommon channels, in terms of number of clicks, were removed. In practice, to not keep incomplete customer journeys, journeys who contained one or more clicks from these uncommon channels were removed in their entirety. Specifically, sequences containing clicks on channels other than the top 10 most common channels were dropped.

Lastly, classes were balanced for model training purposes. Before balancing, there were approximately 2.85% positives (811), i.e., conversions, and 97.15% negatives (27 663), i.e., non-conversions, in the cohort. Classification models that do not take into account such class imbalances tend to be overwhelmed by the majority class and subsequently disregard the minority class [21, 22]. While many methods of varying complexity on countering class imbalance exist, down- or undersampling of the majority class was considered adequate in this context. It has moreover been shown to be superior to up- or oversampling [23]. Thus, classes were balanced such that the data contained equal amounts of positive and negative customer journeys.

These data processing steps altogether, including balancing, resulted in a total sample size of $n = 1622$, whereof 811 positive customer journeys contained 1116 clicks, and 811 negative customer journeys contained 886 clicks. Some fictional data on the data format used can be seen in Table 3.1.

Table 3.1 – Example (dummy) rows from processed dataset

Client	Session	Timestamp	Channel	Device	Conversion	Conversion value
10435	16142	2021-02-25 10:10:38	TikTok	Mobile	0	0
28588	29441	2021-03-03 17:02:15	Facebook	Desktop	0	0
	32817	2021-03-11 08:33:54	Google Paid	Desktop	0	0
	33902	2021-03-12 10:40:32	Google Organic	Desktop	1	139

These data were then transformed to form click sequences with a corresponding binary outcome, see Equation 3.1a - 3.1b.

$$\boldsymbol{c}_1 = \left[\text{TikTok}\right], \qquad y_1 = 0 \tag{3.1a}$$

$$\boldsymbol{c}_2 = \left[\text{Facebook}, \quad \text{Google Paid}, \quad \text{Google Organic}\right], \quad y_2 = 1 \tag{3.1b}$$

3.1.3 Exploratory Data Analysis

To improve quality of validation of the results, a brief exploratory data analysis was conducted. The objective behind this method is to gather the main characteristics of the data at hand primarily by using various charts and tables

[24]. In this project, this analysis was also used to aid the decision making in selecting adequate models to apply on the data. Doing so further allowed for familiarizing with the data at hand. Specifically, to evaluate the complexity of customer journeys, their respective lengths were charted in a histogram, as well as the customer journey duration as time between first to last click. A table containing the number of clicks, conversions and spend per channel was also created as part of this analysis.

To be able to objectively assess the models in relation to the highest possible predictive performance, a theoretical maximum accuracy on the dataset was calculated. This was possible since the models were only taking into account click sequences, meaning that the highest accuracy would be obtained when predicting the outcomes as the most common label of each unique sequence. The theoretical maximum was thus computed by first counting the number of positive and negative labels for each unique sequence in the dataset. Subsequently, the predicted label for each sequence was set to its most common label, i.e., the label with the highest count, which then resulted in the maximum possible accuracy. Using the sequences from Equation 3.1a - 3.1b as an example, the corresponding calculation for sample values would follow as below:

- Seq. 1 [TikTok] occurs 15 times in the data, 2 of which result in a conversion while 13 do not. The optimal prediction follows as $\lfloor \frac{2}{15} \rceil = 0$.
- Seq. 2 [Facebook, Google Paid, Google Organic] occurs 10 times in the data, 6 of which result in a conversion while 4 do not. The optimal prediction follows as $\lfloor \frac{6}{10} \rceil = 1$.
- Theoretical maximum accuracy for this dummy dataset is then $\frac{13+6}{15+10} = 76\%$

3.2 Model Implementation

The MTA models LR and SP, as described in Section 2.2, were found to be most suitable for this project. The motivation for this choice is multifaceted. Firstly, these models had shown promising results in previous studies, particularly when compared to other, more complex models. Secondly, they were deemed adequate due to the lack of complexity in the data as per the evaluation in the exploratory data analysis. Lastly, the models demonstrated a suitable trade-off between interpretability and complexity in relation to the requirements and objectives of this degree project. LTA was further included as a benchmark for the industry standard. Additionally, the purpose of this degree project was not to improve the predictive performance of models, but rather to find and motivate suitable channel attributions.

In the construction of the models, Python 3.7 with packages NumPy 1.19.5 and Scikit-learn 0.24.1 were used. To initialize and train the LR model, the Scikit-learn function `LogisticRegression()` was used. Specifically, the quasi-Newton limited-memory Broyden–Fletcher–Goldfarb–Shanno (lm-BFGS) algorithm was used in estimating LR coefficients. The model was constructed including an intercept β_0 and otherwise with default Scikit-learn parameters. Each dimension was modelled as an indicator variable corresponding to the presence of each channel k in a customer journey. Moreover, a time dimension was added to the data, in an attempt to differentiate clicks at different timesteps. This could have allowed for the model to find possible relationships related to sequence ordering. Thus, a continuous time value, counting from the first click as 0, was added to each indicator variable with a corresponding click. This way, later clicks in the sequence were weighted higher. The reverse method of weighting earlier clicks higher was attempted as well. However, due to no significant performance improvements, the inclusion of a time weight was omitted.

The SP model was constructed as per Equation 2.1. In calculating the conditional probabilities $P(Y_i = 1 \mid k)$, the entire dataset was iterated over such that all clicks were added toward the respective channel counts. In LTA, the probabilities in Equation 2.8 were calculated similarly, albeit only taking into account the last click in each sequence.

As an initial evaluation, the models were all tested on their predictive performance. Given a new sample sequence, the models all yielded conversion probabilies as in Equation 2.2 for SP, Equation 2.3 for LR and Equation 2.7 for LTA. By the nature of probabilities and the logistic function, all such values $\hat{y}_i$ were on the interval $[0, 1]$. This value was subsequently transformed to a binary conversion prediction by rounding, $\lfloor \hat{y}_i \rceil$. This prediction was further evaluated against the true outcome of the customer journey y_i, i.e., whether the user actually converted or not. In this fashion, models were evaluated on all performance metrics described in Section 2.3.2.

Channel attributions from the SP model were obtained directly using the respective empirical channel conditional probabilities $P(Y = 1 \mid k)$ as described in Equation 2.1. Similarly, attributions for the LTA model were obtained from empirical probabilities as in Equation 2.8. Attributions from the LR model were obtained using the channel coefficients β_k directly. However, since negative coefficients correspond to a decrease in conversion probability with the presence of those variables, such coefficients were given an attribution of 0 as per

$$\text{Attr}'_k = \max(0, \beta_k)$$

As an improvement to the basic SP model, pairwise conditioning was considered as has been in previous studies [8]. However, due to limited access to data in all combinations of channels as well as the non-trivial extraction of individual attribution weights from pairwise probabilities, this idea was abandoned.

Further, l-fold CV was used on all models to estimate dispersion and to evaluate generalizability with a 80/20 training-test split, i.e., for $l = 5$. Attributions were obtained in each such fold, before averaging and normalizing the attribution weights obtained in each model as per

$$\text{Attr}_k = \frac{\text{Attr}'_k}{\sum_{\forall q} \text{Attr}'_q}$$

for each channel k. Attr'_k refers to an attribution value, which was obtained as $\max(0, \beta_k)$ from the LR model and as $P(Y_i = 1 \mid k)$ from the SP and LTA models. The normalization procedure allowed for budget allocation and visualization with the ability to compare model results. This process further yielded the final channel attribution values Attr_k for each channel k. In all iterations of the CV, models were evaluated on all performance metrics as described in Section 2.3.2.

3.3 Customer Lifetime Value Estimation

A CLV was estimated for each converted user by using detailed information on such users. While there are ways to properly calculate CLV values using, for instance, present value methods, a less advanced model was used to provide a heuristic estimate. This method allowed for the marketeer to customize weights on various user properties, e.g. depending on what the marketeers, company or perhaps investors value. Examples of properties to include are age, premium and number of claims made, but also categorical properties such as insured object type. In this project, a reasonable example of the weights on a few selected user properties were used since efforts on finding ideal weights and properties were considered outside of the scope.

$$\text{CLV} = w_1 \cdot \text{premium} + w_2 \cdot \text{\#co-insured} + w_3 \cdot \text{is\_student} \tag{3.2}$$

where is_student is 1 if the client is a student and 0 else. Weights used were $w_1 = 24$, $w_2 = -300$ and $w_3 = -100$. w_1 was set to 24 as an example corresponding to two years of mean customer retention with the premium charged monthly. The signs of the example values on w_2 and w_3 indicate a decrease in CLV given the corresponding properties of a user. Note that user properties could only be obtained in Mixpanel for 92.49% of the converted users in the dataset.

3.4 Control Variables

To explore possible relationships among conversions and control variables, the contextual data already obtained from Google Analytics were used. Contextual variables such as desktop, mobile or tablet (device type category) and e.g. Stockholm or Gothenburg (location category) were added as control variables to the processed dataset containing customer journeys. Correlations were subsequently calculated between clicks with each control variable on each channel and conversion. Correlation values were estimated using the Pearson product-moment correlation coefficient in Equation 3.3 as a function of the covariance and the standard deviations σ. Each binary component of the vector $\boldsymbol{X}_k$ indicates if a click from channel k resulted in a conversion, and each binary component of the vector $\boldsymbol{Y}_k$ is an indicator of whether or not the same click was from a control variable. This means that $\boldsymbol{X}_k$ and $\boldsymbol{Y}_k$ are of the same length, equivalent to the total number of clicks via channel k.

$$\rho_{\boldsymbol{X}_k,\boldsymbol{Y}_k} = \frac{\mathrm{Cov}(\boldsymbol{X}_k, \boldsymbol{Y}_k)}{\sigma_{\boldsymbol{X}_k}\sigma_{\boldsymbol{Y}_k}} \tag{3.3}$$

The benefit of exploring these control variables was that the users then could be divided into target groups behaving similarly, as determined by these control variables. This way, a new set of attributions could be formed for each such target group data containing only journeys with clicks on a specific control variable. Thus, the models could, perhaps easier, find different attributions compared to when trained on the entire dataset. Additionally, it could allow for improved predictive performance.

To find the most relevant control variables, absolute correlation values were sorted and values above a threshold of 0.15 were extracted. Control variables present in less than 20% of all clicks were filtered out to have enough target group data for models to be specifically trained on. For reference, the proportion of clicks within a channel of each control variable was also recorded. For example, if all clicks from Facebook come from mobile devices, no conclusions can be drawn about the influence of the device type for Facebook users.

3.4.1 Conversion Strength of Control Variables

As it is possible to analyze unnormalized model-derived attributions directly, they may also be suitable for assessing the strength of control variables in driving conversions. A higher unnormalized attribution for a certain channel indicates that the channel was more active in influencing users to convert on the dataset it was calculated on. The sum of unnormalized attributions was therefore recorded as a summarizing metric for each model and control variable. In each case, 80% of the data was used for training.

3.5 ROI-Based Attributions

Since no type of conversion value had been included in the models thus far, the attributions were subsequently modified based on ROI values, i.e., taking into account both cost and value. The ROIA_k of channel k was estimated as in Equation 3.4 by including both v_i as the value of a conversion and Attr_k as the model-derived attribution of channel k obtained in Section 3.2. Note that v_i corresponds to conversion value, rather than CLV, since the parameters in Equation 3.2 were selected only as a demonstrative example. However, v_i can easily be replaced by CLV, once it is more accurately estimated.

$$\text{ROIA}_k = \frac{\sum_{\forall i: y_i=1} v_i \sum_{j=1}^{m_i} \text{Attr}_k \mathbb{1}(c_{ij} = k)}{\text{Spend}_k} \tag{3.4}$$

$\mathbb{1}(\cdot)$ is the indicator function and Spend_k denotes the total spend for all data on channel k in the cohort. The new ROI-based attributions were then normalized and budget attributions b_k's calculated as in Equation 3.5, where B is the total budget.

$$b_k = \frac{\text{ROIA}_k}{\sum_q \text{ROIA}_q} B \tag{3.5}$$

Note that this approach was only applied to channels with some associated *direct cost*, corresponding to channels charging a CPC or a commission for each successful conversion. This means that channels such as organic search and direct traffic were not included and naturally needed no budget allocation.

3.6 Evaluation Protocol

To allow for an evaluation of all models, and particularly their attributions, an evaluation protocol was formed based on previous research [9]. The protocol simulated a real life setting by using available click data with a simulated budget B. In particular, the protocol was used to estimate the total number of conversions (Y), total conversion value (Z), total CLV (V), and total cost (O) obtained if a certain attribution was applied to a marketing budget. A new variable y_{ij} was introduced, taking the value 1 if click j was the last click in a positive customer journey i, and else 0.

To allow for completely emptying all channel budgets when replaying the entire dataset, it was necessary to set the total budget B less than real life total spend. The ROI-adjusted budget allocations obtained as described in Section 3.5 were then used as channel budgets b_k, with one set of budget allocations per model.

Limited by the available data, the total budget was set to $B = 20\ 000$. The motivation behind this budget was to have it large enough to be able to distinguish between the result from each model, but small enough not to buy all clicks from any one channel. Even though the total spend for channels such as TikTok was below this budget, no ROIA was assigned to such channels. Instead, Adtraction constituted the lowest total real life spend among channels with direct costs, which subsequently limited the total budget B.

The algorithm was run for each set of budget allocations obtained from each model. All clicks in the unbalanced cohort were then iterated, such as to mimic a real situation. A click was bought if the specific channel budget allowed. If not, the corresponding user was blocked such that no future clicks from the user could be bought. This method of blocking users is one of the main assumptions of the evaluation protocol [9]. The details of the protocol algorithm are further outlined in Algorithm 1.

Algorithm 1: Evaluation protocol on historical data

```
Input: Budget allocations $b_1, \dots, b_K$ and a dataset containing clicks ordered chronologically
Output: Total cost, number of conversions, conversion value and CLV
Initialize total cost as $O = 0$, total conversions as $Y = 0$, total conversion value as $Z = 0$, total CLV as $V = 0$ and user block list as $\mathcal{B} = \{\}$
for each click $i, j$ in data do
    if user $i$ not in block list $\mathcal{B}$ then
        if channel-specific budget $b_k \geq o_{ij}$ then
            subtract CPC $o_{ij}$ from $b_k$
            add CPC $o_{ij}$ to $O$
            add conversion $y_{ij}$ to $Y$
            add conversion value $z_i \cdot y_{ij}$ to $Z$
            add CLV $v_i \cdot y_{ij}$ to $V$
        else
            add user to block list $\mathcal{B}$
```

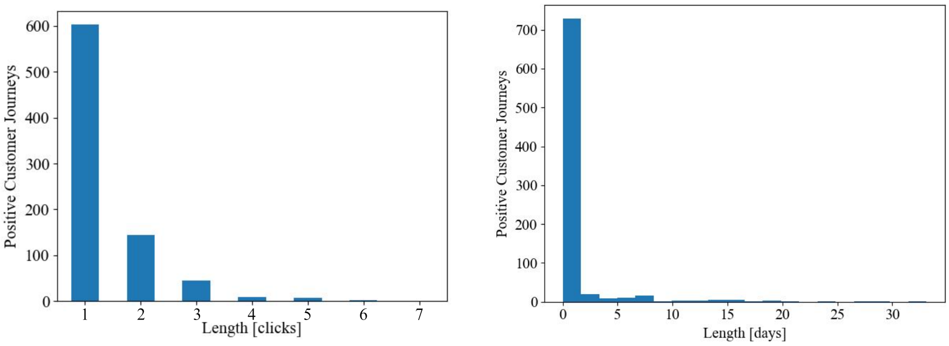
Chapter 4

Results

This chapter presents the results from the various steps of the method chapter. After displaying results from the exploratory data analysis, the top correlations with respect to control variables are shown. Furthermore, results on the predictive performance of all three models are presented, followed by the attributions obtained from model training. Results from the evaluation protocol conclude the chapter.

4.1 Exploratory Data Analysis

Using the method of exploratory data analysis on the processed dataset, class balancing excluded, yielded various descriptive results. These include graphs over customer journey length distributions, as the number of clicks in a positive customer journey, in Figure 4.1a, and duration as time between first to last click in the customer journey, displayed in Figure 4.1b. It can be observed that a majority of the users convert after merely one click. However, approximately 11.1% of users clicked at least two times on at least two different channels before converting, meaning that 11.1% of the conversion click information is lost when using LTA.



(a) Customer journey length distribution as measured by number of clicks

(b) Customer journey duration measured in days from first to last click

Figure 4.1 – Two plots showing the distribution of positive customer journey lengths and duration, respectively

On a channel specific level, Table 4.1 shows how common different channels are in terms of total clicks together with the percentage of those clicks resulting in conversion. This corresponds to the proportion of all clicks from the channel that were in a positive customer journey. Additionally, spend is listed together with a mean CPC value, i.e., channel spend divided by the number of clicks. As can be seen, the marketing budget is concentrated to a few channels and only a few channels form the majority of clicks. Facebook yielded the most clicks while the proportion of conversion clicks is the highest for Newsletter.

Table 4.1 – Occurrences of clicks and percentage leading to conversion for the 10 most common channels in the entire dataset. Spend and mean CPC, both in SEK, are also included.

Channel	Clicks	Proportion conversion clicks	Spend	Mean CPC
Facebook	10950	0.32%	308643	28.19
TikTok	6471	0.00%	1162	0.18
Google Organic	5113	6.66%	0	0
Google Paid	3663	9.01%	264231	72.14
Direct	2487	5.87%	0	0
Studentkortet	783	7.15%	22000	28.10*
Snapchat	639	0.00%	871	1.36
Adtraction	638	11.08%	18631	29.20*
LinkedIn	576	0.00%	0	0
Newsletter	574	27.07%	0	0
Total	31894	3.50%	615538	26.60

Mean total CPC is calculated as the total spend divided by total number of clicks from paid channels. Total proportion of conversion clicks is calculated as the number of clicks in positive customer journeys divided by total clicks.

Note that channels use different pricing strategies. While most channels charge a CPC, channels marked with * use a commission strategy, in which each conversion costs a certain amount. For this table specifically, the mean CPC was however calculated using the same method for all channels. Moreover, additional channels appeared in certain experiments later on, particularly for control variable subsets. These channels were not included in the exploratory data analysis.

4.2 Control Variable Correlations

Correlations above the threshold 0.15 between control variables and conversions, as described in Section 3.4 are illustrated in Table 4.2. The control variable category with the highest correlations is Device, which was analyzed separately to yield control variable-specific attributions. As can be seen from the table, conversion via the channel Newsletter in combination with Desktop as device had a positive correlation with value 0.23, implying that users clicking via Newsletter using Desktop tend to convert to a larger extent than via Mobile, which instead had a negative correlation. Possible reasons behind such correlations remain unexplored.

Table 4.2 – Most prominent correlations between conversions and clicks on a channel with the control variable

Channel	Correlation Coefficient	Control Variable	Proportion in Data	Proportion in Channel
Newsletter	0.23	Desktop (Device)	24.33%	27.62%
Newsletter	-0.21	Mobile (Device)	73.65%	70.90%
Studentkortet	0.19	Desktop (Device)	24.33%	1.30%
Studentkortet	-0.19	Mobile (Device)	73.65%	98.70%
Adtraction	0.17	Desktop (Device)	24.33%	51.11%
Adtraction	-0.16	Mobile (Device)	73.65%	46.68%

4.3 Predictive Performance

The models' predictive performance on the entire dataset as measured using various metrics can be observed in Table 4.3. It can be noted that both SP and LTA achieve the highest values on all metrics. However the differences in performance are small and all models achieve an accuracy close to the calculated theoretical maximum of 82.31%, when evaluated on data with balanced classes. This theoretical maximum is described thoroughly in Section 3.1.3.

Table 4.3 – Predictive performance of models when trained and evaluated with 5-fold CV on the entire dataset with balanced classes

Model	Accuracy	Precision	Recall	F_1
SP	82.16%	75.10%	96.16%	84.33%
LR	82.04%	75.09%	95.91%	84.23%
LTA	82.16%	75.10%	96.16%	84.33%

The predictive performance of the same models on subsets of the data are shown in Table 4.4. These new datasets are divided according to the identified control variables from Section 4.2. As can be seen, all models yielded improved performance metrics on the mobile specific data in comparison to when trained on the entire dataset, with the exception of recall. However, the predictive performance decreased on the desktop specific dataset. Note that experiments on the tablet specific dataset are omitted due to limited amounts of data, which can be seen in Table 4.5. As a reference, maximum theoretical accuracy on all mobile data with balanced classes was 86.17%, and 68.41% on corresponding desktop data.

Table 4.4 – Predictive performance of models when trained and evaluated with 5-fold CV on subsets of the data with control variables and balanced classes

Model	Control variable	Accuracy	Precision	Recall	F_1
SP	Mobile	85.80%	80.40%	94.56%	86.91%
	Desktop	66.70%	61.07%	92.16%	73.46%
LR	Mobile	85.80%	80.74%	94.40%	87.04%
	Desktop	66.14%	61.06%	89.77%	72.68%
LTA	Mobile	85.80%	80.40%	94.56%	86.91%
	Desktop	66.25%	61.06%	89.67%	72.65%

Note that the union of the subsets in Table 4.5 does not equal the full dataset as described above. This follows since the 10 most common channels differ between the entire dataset and the control variable subsets.

Table 4.5 – Number of customer journeys and clicks in each control variable dataset

Control variable	Positive journeys	Clicks in positive journeys	Negative journeys	Clicks in negative journeys
Mobile	423	543	423	472
Desktop	440	672	440	539
Tablet	8	10	8	11

4.4 Attributions

A bar chart showing normalized attributions from the three models when trained on the entire dataset with balanced classes is illustrated in Figure 4.2. Along with the mean attribution, the standard deviation for each channel is shown as an error bar, obtained with 5-fold CV. All three models assign the highest attribution to Newsletter. Furthermore, the LR model has the highest standard deviation throughout its positive attribution channels.

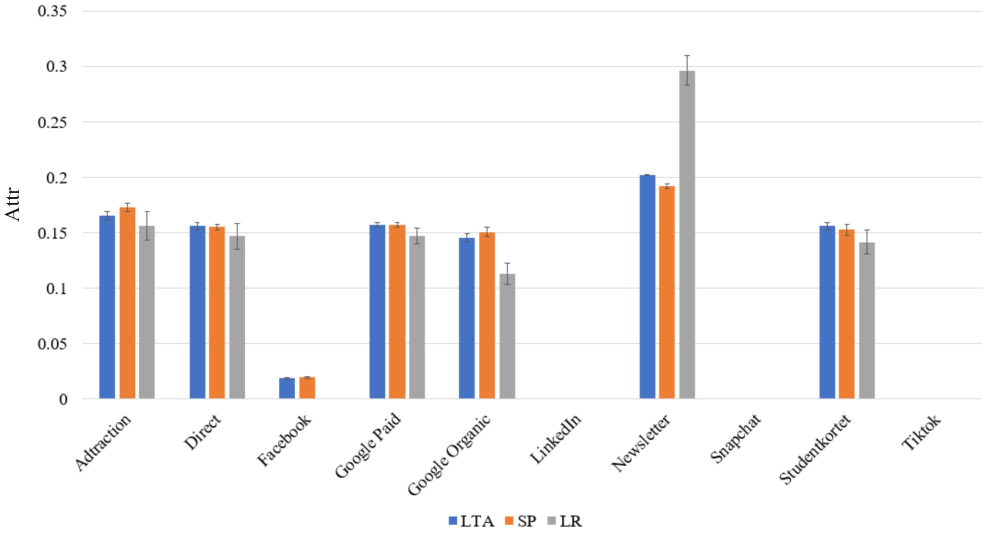


Figure 4.2 – Attributions when trained on the entire dataset, with balanced classes

Attributions obtained when trained on control variable subsets are shown in Figure 4.3 and 4.4. Customer journeys containing other clicks than with the specific control variable were removed from this dataset. As may be observed, the top 10 channels have changed since other channels are more common on these subsets. For instance, Mecenat is new on Mobile, while Benify, Match2One and Bing are new for Desktop.

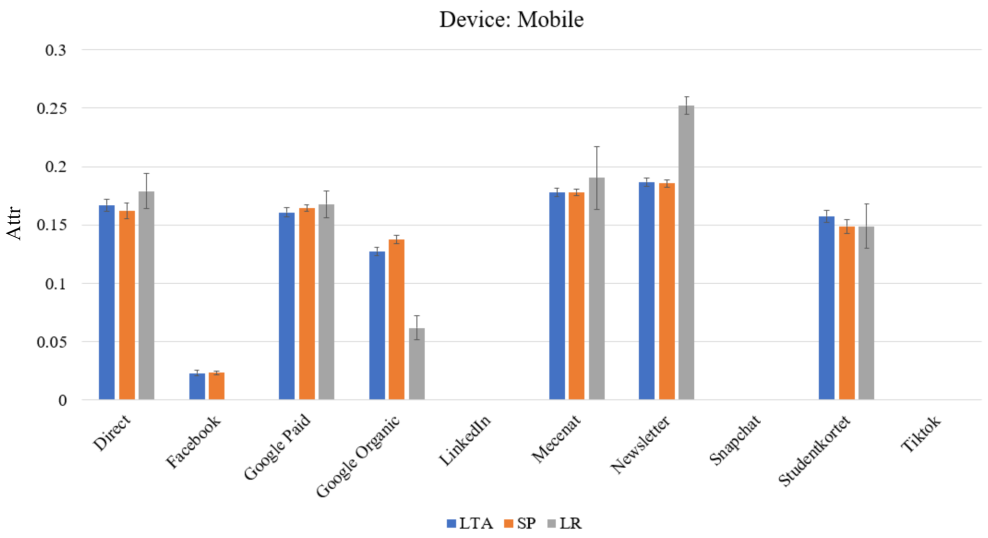


Figure 4.3 – Attributions for clicks from a mobile device only, when evaluated on data with balanced classes

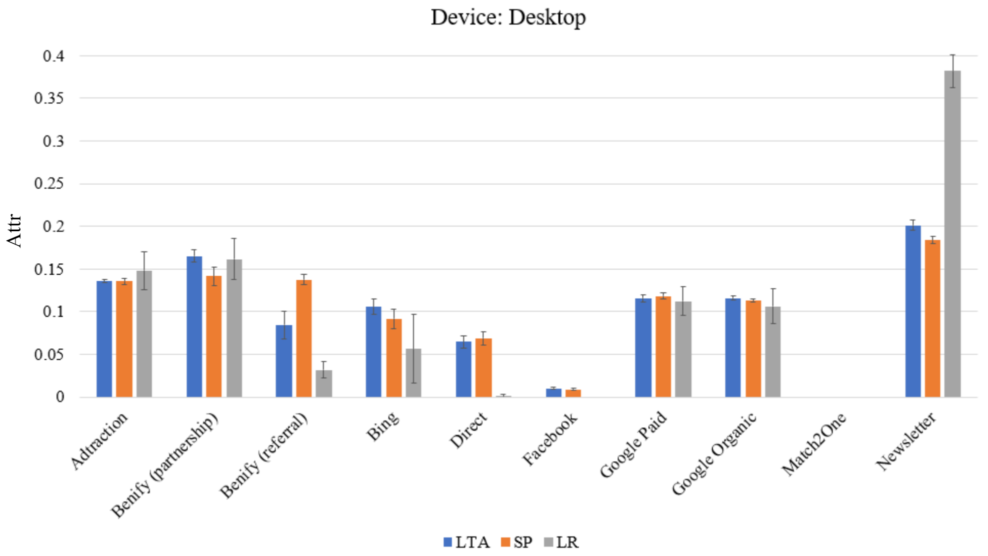


Figure 4.4 – Attributions for clicks from a desktop device only, when evaluated on data with balanced classes

Unnormalized attributions were subsequently summed to assess the conversion strength of each control variable, as shown in Table 4.6. Note that these values were calculated without CV, but with training on 80% of the respective control variable subsets.

Table 4.6 – Sum of unnormalized attributions for control variables showing conversion strength of each control variable. The largest sum of attributions for each model is marked in bold.

Model	Control variable	Sum of attributions
SP	*None*	4.84
	Mobile	4.95
	Desktop	**5.64**
LR	*None*	8.16
	Mobile	**10.14**
	Desktop	6.51
LTA	*None*	4.65
	Mobile	4.88
	Desktop	**5.06**

4.5 Evaluation Protocol

The ROI adjusted attributions after training the models on the entire dataset with balanced classes are shown in Figure 4.5. Although models trained on Mobile data only achieved higher predictive performance compared to when trained on all data, that data subset was too small to run the evaluation protocol on. Hence, the computations involving ROI were only performed with the entire dataset.

Clearly, Adtraction received the highest ROI adjusted attribution of all paid channels, while Snapchat and Tiktok received 0, whereas Facebook received close to 0. Note that free channels received 0 ROIA by default and that these ROI adjusted attributions do not necessarily sum to 1 in accordance with Equation 3.4.

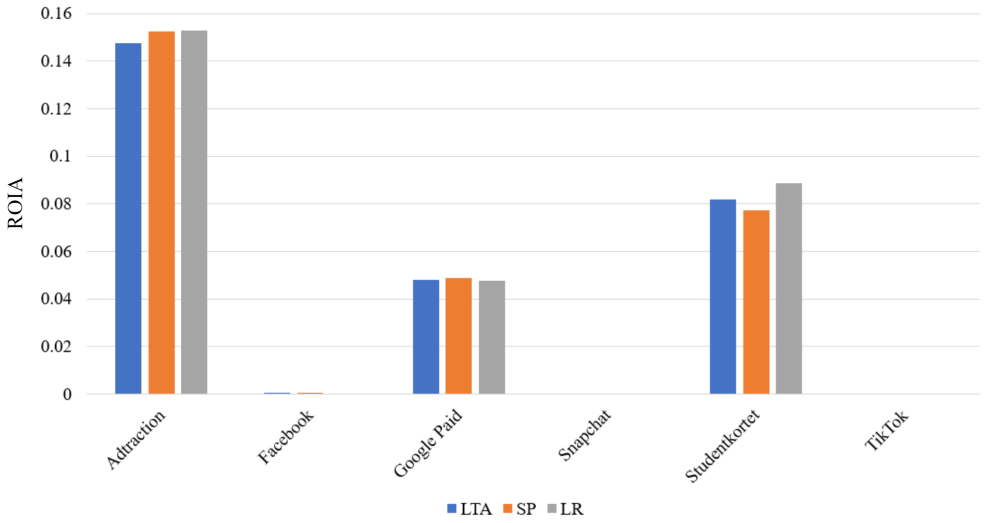


Figure 4.5 – ROI adjusted attributions for entire dataset, with balanced classes and 5-fold CV. Only paid channels are shown.

The results from running the evaluation protocol are shown in Table 4.7, with an additional naive baseline attribution of 100% to Facebook, due to the maximum number of clicks in the exploratory data analysis in Table 4.1. While the three models performed similarly, they all performed better than the naive baseline.

Table 4.7 – Results from the evaluation protocol with model-specific budget allocations

Model	Conversion value	CLV	Cost	Number of Conversions	CLV - Cost	CLV ROI
SP	65319	1440020	19690.8	478	1420329.2	73.13
LR	65239	1438300	19495.4	478	1418804.6	73.78
LTA	65160	1436504	19208.8	477	1417295.2	74.78
Only FB	59408	1308660	19993.0	439	1288667.0	65.46

Chapter 5

Discussion

This chapter begins with a discussion of how the results are interpreted. This is followed by a discussion of the limitations of the study, a brief conclusion and an overview of areas for future research.

5.1 Interpretation of Results

5.1.1 Data

The results from the exploratory data analysis in Section 4.1 suggest that most customers converted after just one click, which is conflicting evidence against Hedvig's trust business hypothesis. In short, the hypothesis was that customer journeys would span a longer time period and over more clicks than for other types of products or services, since trust for an insurance company has to be built prior to conversion. The short sequences also serve as a possible explanation for why the predictive performance of LTA was as high as for the other models in Section 4.3. The fact that LTA performed at least as high with respect to all prediction metrics indicates that the extra 11% of information it neglected was not helpful for predictions.

5.1.2 Predictive Performance

As can be seen in Section 4.3, all three models SP, LR and LTA perform similarly in terms of predictive performance on the entire balanced dataset with approximately 82% accuracy, all close to the theoretical maximum. These results differ from the related works in Section 2.4.3. The fact that LTA performs at least as high as the MTA models is remarkable, possibly due to the above mentioned short sequence data.

5.1.3 Attributions

In contrast to predictive performance, the calculated attributions differ significantly between models. The Newsletter channel attribution is for instance just below 0.3 for LR, while it is just below 0.2 for SP when trained on the entire dataset. Newsletter is also the channel consistently receiving the highest attribution, likely since newsletters were only sent to prospective customers. It is moreover notable that attributions between folds vary more for LR than for other models. This is consistent with previous research, as described in Section 2.4.3. Since the predictive performance of the models are similar, there is no unambiguous way to decide which is more accurate in this scenario. However, a potential solution is to compromise by basing marketing decisions on some type of mean between the attributions of the three models.

Regarding general reliability of the calculated attributions, they may be compared to the proportion of conversion clicks in Table 4.1, although these numbers are not normalized and are based on unbalanced data. The calculated attributions are coherent with the proportion of clicks, indicating that the attribution computations are free of errors. In fact, the unnormalized attributions from SP are calculated in the same way as the proportion of conversion clicks column in Table 4.1.

5.1.4 Control Variables

All models' accuracy increased approximately 3.6-3.8 p.p. when trained on customer journeys with device type Mobile, even though less data were available. Since digital marketing campaigns often can be targeted specifically to mobile users, it allows for a more accurate budget allocation for mobile users. On the contrary, the predictive accuracy decreased with roughly 15-16 p.p. on Desktop only. A possible reason to this accuracy difference is that users may share desktops, creating more noise in that data. Additionally, the difference in attributions between channels who performed well and bad, respectively, were bigger on the Mobile only subset, which create more distinct patterns in the data. These results align with some of the related work in Section 2.4.1, in which significant attribution differences were found when separating on device type.

The sums of unnormalized attributions in Table 4.6 can be used to further aid budget allocation in between these target groups, as they may reflect the purchasing power of each target group. Nonetheless, the ordering of these sums differed between models, which complicates the allocation of such budget. SP and LTA both suggest desktop users as the more attractive target group, while LR suggests mobile users. The relatively low predictive performance of all models on Desktop makes the results additionally unsure. Since clicks on Desktop showed positive correlations with conversion and vice versa for Mobile in Table 4.2, it indicates that SP and LTA may be more accurate in Table 4.6.

5.1.5 Evaluation Protocol and ROI

The ROI-adjusted attributions in Figure 4.5 are naturally only applicable for paid channels, making it necessary to consider the attributions in Figure 4.2 to develop a complete marketing strategy. How these two results are combined is not trivial and may be left for the individual marketeer to decide. It is however clear that Adtraction and Studentkortet gain proportions of the budget in the ROI-adjusted attributions, seemingly by being more cost efficient.

Regarding the results from the evaluation protocol algorithm, all models show similar results across metrics. As expected, their performances are all better than the naive baseline of only investing in Facebook marketing. The similar performance among models is most likely a result of the similar ROI-adjusted attributions distributed across few channels. It should be noted that the budget size B affects the absolute difference in all these metrics, including CLV ROI. Since the used budget size of 20 000 SEK only corresponds to approximately 3% of Hedvig's real marketing spend for this period, the absolute differences would have been larger in reality. Interestingly, the CLV and conversion values are correlated, indicating that the shortsighted conversion value is sufficient to base decisions on with the definition of CLV used in this degree project.

5.2 Limitations

5.2.1 Data

A general concern is the trade-off between quality of data and privacy or integrity of users. From this perspective, the GDPR (General Data Protection Regulation) regulates what data can be collected and stored [25]. Such regulations and technical constraints combined have led to data being a major limitation in this degree project. Firstly, the data contained only clicks and no impressions. The inclusion of such impression data would perhaps have allowed for different results. Some channels are likely more successful in driving conversions from impressions rather than from clicks solely. Secondly, there were no data on conversions via the Hedvig app nor from cross device interactions, i.e., when the same individual uses different devices to access the Hedvig webpage. These limitations, combined with a limited cookie life span, likely decreased the length of customer journeys, as explored in Section 4.1. The trust business hypothesis thus remained unanswered. It is possible that true customer journeys in this setting both span a longer time interval and are more complex in terms of containing clicks from various channels deriving from different devices. Therefore, this dataset may not capture long term marketing and branding effects.

Additionally, all data on interactions were collected using Google Analytics and the data included clicks from two Google channels. This third party reliance may potentially induce some bias towards these Google channels.

As for the cost mapping of channels, only direct costs related to paid channels were considered. In reality, there are numerous indirect costs related to both paid and free channels. For instance, there are likely costs associated with forming a newsletter and optimizing organic search engine performance. While these costs are not necessarily trivial to account for, they could have been included to allow for ROI-adjusted attributions with respect to more channels.

An assumption made in this project was that all campaigns were equally effective in driving conversions, and that only the channel itself had a corresponding positive or negative effect. Ideally, campaigns should have been taken into consideration, perhaps as a control variable, as various campaigns undoubtedly had an associated effect. Due to the combination of missing and low-quality campaign data, this was however omitted.

Variations in marketing strategy and spend were also omitted to some extent. By evaluating clicks from a cohort of three weeks, no variations during this interval were considered. Moreover, there is a trade-off between dataset size and time interval to be evaluated. By shortening the interval, less, albeit more coherent, data is collected. With a shorter interval, less noise is induced due to limited variations in campaigns and marketing spend. In this degree project, a cohort time span of three weeks was however deemed a suitable interval with respect to this trade-off. With a bigger company or a corresponding website with more traffic, this naturally becomes less of a problem.

Also related to the choice of cohort time interval is a limitation related to false negatives. The number of false negatives in the data remains unknown, since any non-converted user in the cohort may, theoretically, convert at any time after the last date of data collected, T. Since the exploratory data analysis suggests the data consists solely of short customer journeys, this is, however, perhaps not as big of an issue.

In digital marketing in general, it is further possible to make a distinction between three types of users, describing users with various intent and outcomes after reacting to an ad. Firstly, there are users who reacted to an ad but did not convert. Then there are users who click on an ad and

a. convert because they saw the ad, or
b. convert, but would have converted anyway

If a user converts after having clicked on an ad, the assumption in the model setup used in this project is that the specific channel had a positive impact on the conversion. Ideally, clicks from users of type b should thus not be part of the data. This type of conversion is referred to as a consequence of the *selection effect*, contrary to *advertising effect* type a conversions, in which the ad actually influenced the user to convert in a causal manner [26]. Type a conversions are clearly desirable to analyze, whereas type b conversions simply increase noise in the data. However, separating the two types is likely a daunting, if not impossible, task, since knowledge about the underlying causalities is required. Yet, this constitutes a major limitation in current attribution modelling in general.

5.2.2 Feature Engineering

In the data processing for all models, input data were formed as a sequence of channels corresponding to the ordered clicks of a user. However, the inclusion of continuous time or other contextual features may have allowed for an increased theoretical maximum accuracy and improved predictive performance among models. Only limited efforts were spent on exploring these aspects in this degree project. In practice, this would correspond to an increased dimensionality of the input space. This could ultimately have translated to more accurate attributions.

5.2.3 Evaluation

As for evaluation of the attributions themselves, the evaluation protocol was used as described in Section 3.6. While this gives some idea on the performance of the models, it is by no means a complete evaluation protocol. Ideally, validation should relate the model performance to some known attribution or similar.

5.3 Conclusions

Throughout this degree project, it has been shown that the simple LTA model is in fact sufficient as an attribution model, given the short sequences and limited data quality without cross-device tracking and impression level data. While current research consistently argues for more complex data-driven MTA models, this project provides a nuance to this field of research. The choice of attribution models seems to be rather data dependent. Additionally, the cohort time interval should be chosen based on amount of data and dynamics in the marketing strategy. A dynamic marketing strategy, which is updated regularly, would require a shorter interval, if enough data is available. Thus, it is recommended to begin with an exploratory data analysis, as has been done in Section 4.1, to decide if a simple rule-based model should be sufficient, or if it is worthwhile to invest time in complex model development.

Altogether, the results, disregarding the limitations, suggest that Hedvig's current short term marketing efforts should be distributed fairly evenly across Newsletter, Adtraction, Direct, Google Paid, Google Organic and Studentkortet, when optimizing for customer base growth. Facebook, LinkedIn, Snapchat and TikTok should not be used further, from a short term perspective. Among paid channels, Hedvig may want to invest additionally in Adtraction when optimizing for ROI, as it receives the largest ROI adjusted attribution. As for unpaid channels, increased efforts could involve spending more resources on crafting newsletters and developing campaigns, for instance. Clearly, these attributions should be updated on a rolling basis to account for time dependencies.

Models trained on subsets of the data, as based on the analysis of correlations with respect to control variables, yielded improved predictive performance on device type Mobile. Hedvig may therefore attempt to assign separate marketing budgets for each of the separate device types. Since these obtained attributions could not be evaluated using the evaluation protocol, and the predictive performance for Desktop was relatively low, this may be done cautiously. In general, however, it is recommended to evaluate correlations to potentially find suitable control variables, allowing for separate attribution models and subsequently separate targeted marketing strategies.

5.4 Future Work

As most of the limitations in this degree project follow due to limitations in data, the authors suggest future researchers focus on obtaining high quality data, while still adhering to GDPR regulations. Particularly, the need to capture as much of true customer journeys as possible is imminent. Perhaps Google Analytics is not adequate for the task of collecting data to be used for attribution modelling in some circumstances. For websites requiring logins from all users, the Google Analytics feature `user_id` may however instead be used to collect cross-device data, and thus improve data quality. Forced logins may however not be desirable for other reasons. If, however, longer and more complex sequences are obtained, more complex MTA models are more likely to outperform simple models such as LTA. In such cases, the data quality may further be evaluated to estimate the number of false negatives, or a suitable cohort time interval, perhaps using a regression model to extrapolate to the unknown future.

To allow for more thorough analysis in the comparison of paid and unpaid channels, further research may also strive for a complete mapping of direct and indirect costs derived from each marketing channel. Additionally, this is ideally to be combined with more experimental variation in campaigns beforehand to provide more complete and varied data on all channels [5]. This method also allows for a more thorough control variable analysis. Perhaps additional features, such as time, can be added to improve predictive performance of models. All these improvements would likely lead to more substantiated insights as support for revisions of marketing strategies.

An additional suggestion is to not only focus on digital marketing channels, but also on channels in traditional media to take into account long-term branding effects. This could include radio, TV, newspaper and billboards, for instance. Due to limited access to data in such cases, a different type of model architecture is likely required, perhaps on an aggregated rather than on an individual level. Going further, it would clearly be desirable to allow for comparing attributions between traditional and digital media.

In estimating the value of customers by using CLV values, only example parameters were set in this degree project. A suggestion for future research is thus to estimate CLVs statistically, and evaluate how attributions then differ when compared to models optimized with respect to basic conversion values. Thus, it can be argued if estimating CLV is even worthwhile in attribution modelling.

Additional room for improvement concerns validation. Only limited focus has been spent on the validation of attributions in this degree project, as has also been the case in much of previous research. Future research may thus attempt more explicit and direct validation, perhaps using simulated data. As previously mentioned in Section 1.4, this is explored in a parallel degree project conducted by one of the authors of this report [27].

> *Finally, the one who builds upon this project and implements these suggestions will be one step closer to creating, what we would call, a growth machine.*

References

[1] J. Gaur and K. Bharti, "Attribution Modelling in Marketing: Literature Review and Research Agenda," *Academy of Marketing Studies Journal*, vol. 24, no. 4, 2020.

[2] C. Jones, *The Multichannel Retail Handbook: A Guide to Planning, Implementation, Operation and Enhancement, 2016 Edition*. Redsock Management Ltd, 2015. ISBN 978-1-326-47257-3

[3] P. Kannan, W. Reinartz, and P. C. Verhoef, "The Path to Purchase and Attribution Modeling: Introduction to Special Section," *International Journal of Research in Marketing*, vol. 33, no. 3, pp. 449–456, Sep. 2016. doi: 10.1016/j.ijresmar.2016.07.001

[4] D. Yang, K. Dyer, and S. Wang, "Interpretable Deep Learning Model for Online Multi-touch Attribution," 2020. [Online]. Available: https://arxiv.org/abs/2004.00384

[5] P. J. Danaher and H. J. van Heerde, "Delusion in Attribution: Caveats in Using Attribution for Multimedia Budget Allocation," *Journal of Marketing Research*, vol. 55, no. 5, pp. 667–685, Oct. 2018. doi: 10.1177/0022243718802845

[6] W3Techs. Usage Statistics and Market Share of Traffic Analysis tools for Websites. Accessed: 2021-04-02. [Online]. Available: https://w3techs.com/technologies/overview/traffic_analysis

[7] Google. Google Developers: Google Analytics. Accessed: 2021-03-22. [Online]. Available: https://developers.google.com/analytics

[8] 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*, ser. KDD '11. New York,

New York, USA: Association for Computing Machinery, Aug. 2011. doi: 10.1145/2020408.2020453. ISBN 978-1-4503-0813-7 p. 258–264.

[9] K. Ren, Y. Fang, W. Zhang, S. Liu, J. Li, Y. Zhang, Y. Yu, and J. Wang, "Learning Multi-Touch Conversion Attribution with Dual-Attention Mechanisms for Online Advertising," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, ser. CIKM '18. New York, New York, USA: Association for Computing Machinery, Oct. 2018. doi: 10.1145/3269206.3271677. ISBN 978-1-4503-6014-2 p. 1433–1442.

[10] D. Montgomery, E. Peck, and G. Vining, *Introduction to Linear Regression Analysis*, 5th ed. Hoboken, New Jersey, USA: John Wiley & Sons, 2012. ISBN 978-0-470-54281-1

[11] P. Refaeilzadeh, L. Tang, and H. Liu, *Cross-Validation*. Boston, Massachusetts, USA: Springer, 2009, pp. 532–538. ISBN 978-0-387-35544-3

[12] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, New York, USA: Springer, 2009. ISBN 978-0-387-84857-0

[13] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. New York, New York, USA: Springer, 2013. ISBN 978-1-4614-7137-0

[14] R. Cruz, K. Fernandes, J. S. Cardoso, and J. F. Pinto Costa, "Tackling Class Imbalance with Ranking," in *2016 International Joint Conference on Neural Networks (IJCNN)*. New York, New York, USA: Institute of Electrical and Electronics Engineers, Jul. 2016. doi: 10.1109/IJCNN.2016.7727469. ISBN 978-1-5090-0621-2 pp. 2182–2187.

[15] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, "A Benchmark for Interpretability Methods in Deep Neural Networks," 2019. [Online]. Available: https://arxiv.org/abs/1806.10758

[16] R. Sinha, S. Saini, and N. Anadhavelu, "Estimating the Incremental Effects of Interactions for Marketing Attribution," in *2014 International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC 2014)*. New York, New York, USA: Institute of Electrical and

Electronics Engineers, Oct. 2014. doi: 10.1109/besc.2014.7059518 pp. 1–6.

[17] L. Kakalejčík, J. Bucko, P. A. Resende, and M. Ferencova, "Multichannel Marketing Attribution Using Markov Chains," *Journal of Applied Management and Investments*, vol. 7, no. 1, pp. 49–60, Feb. 2018.

[18] 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*, ser. ADKDD '12. New York, New York, USA: Association for Computing Machinery, 2012. doi: 10.1145/2351356.2351363. ISBN 978-1-4503-1545-6

[19] N. Li, S. K. Arava, C. Dong, Z. Yan, and A. Pani, "Deep Neural Net with Attention for Multi-channel Multi-touch Attribution," 2018. [Online]. Available: https://arxiv.org/abs/1809.02230

[20] E. Anderl, I. Becker, F. von Wangenheim, and J. H. Schumann, "Mapping the customer journey: Lessons learned from graph-based online attribution modeling," *International Journal of Research in Marketing*, vol. 33, no. 3, pp. 457–474, Sep. 2016. doi: 10.1016/j.ijresmar.2016.03.001

[21] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: Special Issue on Learning from Imbalanced Data Sets," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 1–6, Jun. 2004. doi: 10.1145/1007730.1007733

[22] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou, "On the Class Imbalance Problem," in *2008 Fourth International Conference on Natural Computation*, vol. 4. New York, New York, USA: Institute of Electrical and Electronics Engineers, 2008. doi: 10.1109/ICNC.2008.871. ISBN 978-0-7695-3304-9 pp. 192–201.

[23] C. Drummond and R. C. Holte, "C4.5, Class Imbalance, and Cost Sensitivity: Why Under-sampling beats Over-sampling," in *Workshop on Learning from Imbalanced Datasets II*, 2003. doi: 10.1.1.68.6858 pp. 1–8.

[24] M. N. Saunders, P. Lewis, and A. Thornhill, *Research Methods for Business Students*, 8th ed. Harlow, United Kingdom: Pearson Education, 2019. ISBN 978-1-292-20878-7

[25] Publications Office of the European Union. (2016, Apr.) Regulation (EU) 2016/679 of the European Parliament and of the Council. Accessed: 2021-05-04. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj

[26] Forbes. (2020, Mar.) The Potential Dangers Of Online Advertising. Accessed: 2021-04-27. [Online]. Available: https://www.forbes.com/sites/forbesbusinesscouncil/2020/03/16/the-potential-dangers-of-online-advertising/

[27] H. Kindbom, "Investigating the Attribution Quality of LSTM with Attention and SHAP: Going Beyond Predictive Performance," Master's thesis, KTH Royal Institute of Technology, Stockholm, Sweden, Jun. 2021.

www.kth.se