Evaluating attribution models on predictive accuracy, interpretability, and robustness

Joep van der Plas

SNR: 1259666

ANR: 284239

Supporting Company: GfK

02-07-2019

Master Thesis Data Science: Business and Governance

Tilburg School of Humanities and Digital Sciences

Tilburg University

Master Thesis Supervisors:	Martijn van Otterlo, Ph.D.	Tilburg University
	Grzegorz Chrupala, Ph.D.	Co-reader Tilburg University
	Peter van Eck, Ph.D.	GfK

Preface

In front of you lies the Master thesis “Evaluating attribution models on predictive accuracy, interpretability, and robustness”. It is written to earn the MSc degree in Data Science: Business and Governance at Tilburg University.

I was engaged in researching and writing this thesis from January 2018 to July 2018. The thesis is executed in collaboration with GfK, with the aim to evaluate attribution models that are often used in practice to capture the true conversion attribution.

Special thanks to Martijn van Otterlo, my supervisor on behalf of Tilburg University, for providing me with great feedback and teaching me how a Hidden Markov model works. The initial proposal was to construct a Hidden Markov model. An explanation about how to build a hidden Markov model based on the notion of a conversion funnel and why this idea is not feasible can be found in the Initial Idea section in the Appendix.

In addition, I would like to thank GfK, for making the required data accessible to address the research question, and, in particular, Peter van Eck for the highly helpful weekly meetings.

Last but not least, I would like to thank my friends, parents, and girlfriend for their support during this process.

Joep van der Plas
July 2018

Summary

To the best of my knowledge, no research has compared and evaluated multiple data-driven attribution models on diverse aspects. This thesis fills the gap by evaluating the heuristic-based attribution models, the Shapley Value solution, the logistic regression, and the Markov chain on interpretability, predictive accuracy, and robustness. These data-driven attribution models are selected since they are relatively easy to interpret and implement, yet are a substantial improvement in the attempt to capture the true conversion attribution in comparison with heuristic-based models. Furthermore, they are often used in practice. The travel agency dataset is provided by GfK and is collected within their Crossmedia Link panel. The dataset contains various touchpoints and whether or not the customer eventually converts. The results show that none of the attribution models outperforms the others on all three aspects. The Shapley Value solution has the highest predictive accuracy and has a good interpretability but is not robust. The logistic regression has a good predictive ability and robustness when the bagging and regularization procedure are applied, yet does not score high on interpretability as the model does not aim to reflect the contribution of a touchpoint. The Markov chain is robust and moderately interpretable, but the model does not score well on predicting conversion. In addition, the data-driven models produce different intermediate results from which different types of information can be obtained. The different types of information are interesting as such, yet combining these results provides additional insights into understanding and influencing the customer journey. However, the generalizability is limited since the analyses are conducted on one single travel agency dataset. The properties, size, and domain of the data may affect the findings. Notwithstanding its limitations, going from heuristic-based models to data-driven models is a considerable improvement in the attempt to capture the genuine attribution. The primary contribution of this thesis is that enterprises can decide which attribution model fits their needs the best. All the evaluation criteria are important, but none of the attribution models is superior. Hence, a direct implication is that enterprises should make a trade-off. Creating transparency by evaluating the models encourage enterprises to abandon heuristic-based models and adopt data-driven models. On the basis of the evaluation of the existing data-driven attribution models, future research could enhance an attribution model on one specific evaluation criterion or develop a multifaceted novel attribution model that is easy to interpret, has a high predictive accuracy, and is robust.

Contents

1. Introduction

The first part of this section provides general information about the attribution modeling landscape. Section 1.2 describes in more depth what this study attempts to address and provides the research question of this thesis. Eventually, section 1.3 concludes with the scientific and practical relevance and an outline for the remainder of this thesis.

1.1 General

Online advertising has grown exponentially in the last decade. The marketing mix of enterprises has shifted from offline to online (Goldfarb, 2014). In contemporary society, enterprises use a wide variety of online media to influence potential customers at different stages of their journey to purchase, including displays, retargeting displays, pre-roll ads, affiliates, and e-mails. Moreover, customers can easily gather information about the product of interest on the web (Yardeni, 1996). For example, through information websites, comparison websites or websites of direct competitors. Any contact that influences the perception of the product is a touchpoint, regardless of whether it was initiated by a company or by the customer (Meyer & Schwager, 2007). All touchpoints belonging to the path followed by the customer before making a purchase decision is referred to as the customer journey.

Enterprises wish to know the contribution of each touchpoint to optimally allocate their advertising budget. This is not a new problem. The problem already existed in offline advertising such as print and television. Traditionally, marketers and scholars tackled the problem by making use of aggregated methods such as marketing mix models. Marketing mix models estimate the effect of various marketing strategies on sales or market share (Constantinides, 2002). However, new opportunities have emerged due to the internet. In addition to the opportunity to reach more customers and to tailor ads to individuals, the internet allows enterprises to track online behavior (Ur, Leon, Cranor, Shay, & Wang, 2012). Technologies such as cookies and tags enable enterprises to gather more granular data. These technologies register the touchpoints within the customer journey. A conversion occurs when the customer journey ultimately ends in a purchase. Models that make use of rich individual-level data to assign the credit of a conversion to the right touchpoints are known as attribution models.

Despite the individual-level data of multiple touchpoints, over-simplistic attribution models based on heuristics are generally employed in practice (Berman, 2017). To illustrate, Google Analytics, a leading platform, uses attribution models based on predefined rules (Clifton, 2012). An often-used heuristic is last touch attribution. The model assigns all credit to the last touchpoint. Nevertheless, there are several data-driven attribution models with varying complexity. However, to the best of my knowledge, no research has compared and evaluated multiple data-driven attribution models on diverse aspects. This thesis evaluates popular attribution models on three aspects – interpretability, predictive accuracy, and robustness.

The first data-driven alternative for the heuristic-based attribution models is proposed by Shao and Li (2011), which they call a simple probabilistic model. This intuitive model estimates the attribution by computing the first and second order conditional probabilities. Dalessandro et al. (2012) extend the model and show that it is equivalent to the Shapley Value solution in cooperative game theory (Shapley, 1988). The simple probabilistic model and its extension are invented to solve the attribution problem and hence the output of the model is directly the credit assigned to each touchpoint.

Furthermore, an often-used binary classifier is logistic regression (Hosmer, Lemeshow, & Sturdivant, 2013). The predicted coefficients are not directly interpretable as contributions and need to be transformed into conversion attribution. One way is to derive the marginal effects. Both the Shapley Value solution and logistic regression do not incorporate temporal dynamics. However, the input of the logistic model can be adjusted by making binary features of the *t* latest touchpoints of the customer journey to capture the dynamics.

A model that inherently includes temporal dynamics is the Markov chain (Anderl, Becker, Von Wangenheim, & Schumann, 2016). The Markov chain captures the sequential nature by calculating the probability of going from one touchpoint to the next. Once the probabilities are calculated, a feature referred to as the removal effect is computed to estimate the contribution of the touchpoints. The removal effect is the decrease in conversion when the touchpoint is removed from the customer journey network. In addition, the Markov assumption is relaxed and higher-order Markov chains are generated.

Another more complex method is a Hidden Markov model.[1] Abhishek, Fader, and Hosanagar, (2015) employed a hidden Markov model anchored by the notion of a conversion funnel. The latent stages of the Markov model reflect the engagement of the customer through the conversion funnel (i.e. disengaged, active, engaged, conversion). Through the use of the conversion funnel, touchpoints can be assessed within the engagement stage. Some advertisements may be more effective in earlier stages and some in later stages.

Alternative complex attribution models exist. Xu, Duan, and Whinston, (2014) propose a mutually exciting point process that includes individual heterogeneity in a hierarchical Bayesian fashion. Zhang, Wei, and Ren (2014) employ an attribution model based on the survival theory, where "death" denotes a customer journey that ends in a non-purchase. Li and Kannan (2014) apply a Bayesian model to compute the carryover and spillover effects at various purchase stages. More complex models are however not better by definition, especially in the attribution modeling context. In the next section, a motivation is given which attribution models are chosen and why.

[1]The initial idea was to construct a Hidden Markov model. An explanation about how to build a hidden Markov model based on the notion of a conversion funnel and why this idea is unfeasible can be found in the Initial Idea section in the Appendix.

1.2 Towards the research question

The Shapley Value solution, the logistic regression, and the Markov chain and their above-mentioned extensions are estimated in this thesis. These models are chosen because of three reasons. Firstly, leading attribution platforms generally use models based on predefined rules. These heuristic techniques for assigning contribution to touchpoints have ingrained biases that make them inherently flawed (Clifton, 2012). A rationale for the use of these heuristics is their ease of implementation and interpretation. Going from heuristic-based models to data-driven models is a substantial improvement in the attempt to capture the true conversion attribution. Hence, the Shapley Value solution, the logistic regression, and the Markov chain are selected because they are relatively easy to implement and interpret.

Secondly, this thesis follows the Occam's Razor principle (Gamberger & Lavrač, 1997). This means that when there exist multiple models to measure an event, the simpler models are preferred over the complex models because more complex models make generally more assumptions. Evaluating the data-driven models that are relatively easy to implement and interpret will help enterprises to build more trust in these models.

Thirdly, the Shapley Value solution, the logistic regression, and the Markov chain are attribution models that are often used in practice. To give an indication, the marketing attribution software of Abakus and Google Attribution 360 use a version of the Shapley Value solution (Hülsdau & Teuteberg, 2018). The logistic regression is one of the most popular binary classifiers and in an attribution context used by enterprises such as Neustar, Convertro, and Nielsen (Sequent Partners, 2018). Windsor.AI, a marketing analytics platform, assigns contribution to touchpoints based on the Markov chain (Windsor.AI, n.d.). Hence, the estimated data-driven attribution models are actually implemented by large enterprises and platforms.

To the best of my knowledge, no one examined and compared different attribution models like this study does. When novel models are proposed, they are often compared against another simple model on the predictive accuracy. For example, Shao and Li (2011) compared the attribution assigned by the Shapley Value solution to the last touch attribution model. Nevertheless, no study evaluated multiple attribution models on diverse aspects. Furthermore, little information can be found on how models perform and on how to evaluate attribution models. This thesis fills this gap by evaluating the Shapley Value solution, the logistic regression, and the Markov chain and its extensions on interpretability, predictive accuracy, and robustness.

According to Dalessandro et al. (2012), an attribution model has a good interpretability when it is "generally accepted by all parties with material interest in the system, on the basis of its statistical merit, as well as on the basis of intuitive understanding of the components of the system" (p. 32). This definition is adopted in this thesis. Yet, the center of attention goes to the intuitive understanding of the components as this is key for creating transparency in the data-driven attribution models.

Although there is a difference between prediction and attribution, assessing the predictive performance of attribution models helps to evaluate the model. In addition, it assists to persuade marketers of the model's trustworthiness (Lodish, 2001). The predictive accuracy is assessed with two measures. Firstly, the predictive accuracy is measured by the area under the receiver operating characteristic (ROC) curve in the training and validation set. The ROC curve is insensitive to whether the class label (i.e. conversion) is unequally distributed. As a second measure for the predictive performance, the top-decile lift is computed. The top-decile lift shows how much more likely the predicted top ten percent is going to convert in comparison with the average customer (Neslin, Gupta, Kamakura, & Mason, 2006).

Attribution models are robust when they are not prone to deviations in the data, as it conveys the ability to render reproducible and stable outcomes (Box, 1979). The robustness of the attribution models is also assessed with two measures. Firstly, the predictive performance described in the previous paragraph should be consistent when resampling from the underlying data-generating process. Secondly and more important for our present purpose, the contribution assigned to each touchpoint should be stable across resamples. This is pivotal because the marketing budgets of enterprises are based on these results. Fragile results indicate a weak attribution measure that is questionable (Wooff & Anderson, 2013).

This leads to the following research question:

> *To what extent are the heuristic-based attribution model, Shapley Value solution, the logistic regression, and the Markov chain easy to interpret, robust, and accurate?*

The analyses are conducted on a travel agency dataset. Online advertisement plays an increasingly pivotal role in the travel market (Park & Oh, 2012). In addition, consumers spend generally lots of time researching vacations online, sometimes spread over a long time span, and are consequently exposed to various touchpoints (Pabel & Prideaux, 2016). As a disclaimer, the research question is solely addressed on one dataset which limits the generalizability.

1.3 Motivation

As a marketing pioneer, John Wanamaker stated - "I know half the money I spend on advertising is wasted; but I can never find out which half" (as quoted in Mayer, 1991, p. 138). By implementing data-driven models, more accurate and additional information will be gained and, therefore, the marketing expenses can be optimized. More importantly, assessing the models assists enterprises to choose the most suitable model to address their problem. One may be concerned with predictive accuracy and another may be more risk averse and want stable results. Creating transparency by evaluating the models will encourage enterprises to abandon heuristic-based models and adopt data-driven models.

Studying attribution modeling is crucial from a practical perspective as enterprises spend a tremendous amount of money on online advertising. Applying more precise models helps enterprises allocate their advertising budget more efficiently. This has not only a positive effect on the enterprise but also improves the efficiency in the marketplace since, eventually, the consumer pays for the ad. Hence, more efficient advertising leads to fewer costs and a higher consumer welfare (Mitra & Lynch, 1996).

Emphasizing the significance, attribution modeling is highlighted as most important research priority (2016-2018) by the Marketing Science Institute (MSI Research Priorities, n.d.). Before creating novel models, existing models need to be evaluated. Based on the results of the evaluations, aspects and directions for novel models or extensions of models can be proposed.

In addition, most papers that model attribution are based on cookie-level data, such that they identify individuals, in the best scenario, on an individual device level (e.g. Abhishek et al., 2015; Anderl et al., 2016). This study contains a more comprehensive dataset on a fine-grained level as it uses panel data where individuals are traced between devices.

The outline of the remainder of this thesis is as follows. In section two of this study, the Shapley Value solution, the logistic regression, and the Markov chain are described in-depth and the evaluation criteria are defined and operationalized. Section three provides the experimental set-up regarding the models and evaluation metrics used in this thesis. In section four, the heuristic-based models, the Shapley Value solution, the logistic regression, and the Markov chain are estimated. The contribution assigned to the touchpoints of these models are compared. Furthermore, these models are evaluated on the ease of interpretation, predictive accuracy, and robustness. In section five, the findings of section four are discussed. Moreover, the limitations of this study are acknowledged, the contribution and implications are provided, and suggestions for future research are given.

2. Background

This section begins with specifying the area of research and provides context for the research focus. In section 2.2 the Shapley Value solution, the logistic regression, and the Markov chain are exhaustively discussed. In the last section, the evaluation criteria are defined and operationalized.

2.1 Digital landscape

The internet has radically changed the advertising landscape as of three reasons. Firstly, the potential reach of the internet is enormous. The number of internet users increased from 2017 to 2018 with seven percent to more than four billion around the globe (We are social, 2018). Secondly, the World Wide Web provides the opportunity to distinguish potential customers at a more granular level, enabling enterprises to tailor ads to individuals which enlarges its effectiveness. Tailoring ads have been studied extensively in the literature (e.g. Bleier, De Keyser, & Verleye, 2018). Thirdly, the internet enables enterprises to monitor and record online behavior (Ur et al., 2012). An enterprise can decide to shift the advertising spending based on historical data. This can be accomplished by making use of various statistical models.

Traditionally enterprises have implemented marketing mix models to measure the effectiveness of advertisements on an aggregated level (Aras, Syam, Jasruddin, Akib, & Haris, 2017). Marketing mix models generally use sales, pricing, and advertising and promotion information of a particular time period as input to measure for example the effectiveness of an advertising campaign. However, these models disregard the variation created by individuals. More recently, a new class of models has become popular and are referred to as attribution model in the popular press (Nisar & Yeung, 2017). These models make use of rich online individual-level data to assign the credit of a conversion to the right touchpoint. More specifically, the data used for attribution models exist of the browsing history of potential customers. Attribution models allow enterprises to understand the touchpoints a customer encounters during their journey to purchase. Gaining knowledge about the touchpoints can be leveraged to make managerial decisions.

The registered browsing history of potential customers consists of multiple touchpoints. In this thesis, all touchpoints belonging to the path followed by the customer are referred to as the customer journey. A customer journey ends with a conversion or non-conversion. However, in reality, the proceedings that occur between touchpoints are part of the customer journey as well. In addition, not every contact is recorded. For example, word of mouth communication is not recorded yet might have a positive or negative effect on the propensity to convert (Hudson, Roth, Madden, & Hudson, 2015). The registered set of touchpoints is nonetheless a comprehensive representation of the truth in comparison with marketing mix models.

A precondition to construct an attribution model is the capability to detect individual customers on the web. Once these individuals can be identified, touchpoints can be devoted to a specific customer and customer journeys can be derived based on a set of rules. The identification

process is commonly accomplished by using cookies (Anderl et al, 2016). A cookie is a unique identifier that a web server place on an individual's browser when visiting a webpage. The web server retrieves this unique identifier again when the web page is loaded at a later moment in time (Englehardt et al., 2015). The cookies are stored locally on a device. A drawback of this approach is that customers can use multiple browsers or devices in their path to purchase, which will be classified as separate customer journeys. Furthermore, tracking cookies can be turned off by individuals, especially with the entrance of the general data protection regulation (Young, 2017). Another less commonly used approach is tagging, which is a JavaScript snippet of code loaded when a webpage is retrieved (Silverbauer, 2017). The information is stored as a log file in a database and must be linked with an individual user. This thesis makes use of a panel where every individual is known and hence can be linked to the appropriate log file. A large benefit from this approach is that customers are tracked across devices, leading to comprehensive customer journeys.

These customer journeys are thoroughly described in the literature (Herhausen, Kleinlercher, Emrich, Verhoef, & Rudolph, 2017). Within the customer journey, distinctive stages of the purchase decision process can be identified. The process of walking through the customer journey and eventually purchasing is the conversion funnel (Kotler & Armstrong, 2010). Several models exist to describe the stages that occur. However, the most well-known model is AIDA (i.e. attention, interest, desire, and action; Strong, 1925). The customer starts at an unaware stage and goes on to the attention stage when a touchpoint is encountered. Interest in the product may be generated and the customer becomes more engaged and as a consequence goes on to the subsequent stage. The interest may become irrevocable and in the final stage, the customer can decide to convert.

This thesis exclusively focuses on the online interactions of the customer journey. The customer journey is established by means of tagging and a GfK custom-designed plugin for passive measurement. Individual panel members are known and tracked across devices. The customer journeys serve as input for the attribution models. Several attribution models are described in the next section.

2.2 Attribution models

This section provides a thorough review of the Shapley Value solution, the logistic regression, and the Markov chain. A motivation why these data-driven attribution models are selected is discussed in the introduction. In addition, before diving into data-driven attribution models, two heuristic-based attribution models are outlined to provide a baseline.

2.2.1 Heuristic-based models

Despite rich online behavioral data, over-simplistic attribution models based on heuristics are generally employed in practice (Berman, 2017). To illustrate, Google Analytics, a leading platform, uses attribution models based on predefined rules (Clifton, 2012). Heuristic-based models are not

determined by the data, they rather have a pre-specified distribution of parameters. Several heuristic-based attribution models exist. A predominantly used heuristic is last touch attribution. This model assigns all credit to the last touchpoint. Another heuristic-based attribution model is first touch attribution and devotes the full contribution to the first touchpoints. The popularity of these models is probably due to the relative ease of implementation and interpretation. Nevertheless, heuristic methods for devoting credit to touchpoints have ingrained biases that make them inherently flawed (Clifton, 2012). Going from heuristic-based models to data-driven models is a substantial improvement in the attempt to capture the true conversion attribution. In addition, heuristic-based models only consider customer journeys that eventually end in conversion, disregarding a lot of information of customer journeys that do not convert. On the contrary, data-driven attribution models use both customer journeys leading up to conversion and non-conversion. The data-driven models are discussed below.

2.2.2 Shapley Value solution

The first data-driven alternative for the heuristic-based attribution models is proposed by Shao and Li (2011), which they call a simple probabilistic model. The model takes the individual touchpoints and the interaction between touchpoints into account. The computation of the model consists of two steps. The first step is to calculate the conditional probability of conversion given the individual touchpoints and the conditional probability of conversion given the interaction between two touchpoints. The two formulas are provided below:

$$P(y|x_i) = \frac{N_{conversion}(x_i)}{N_{conversion}(x_i) + N_{non\ conversion}(x_i)}$$

where y is a binary label denoting whether the user converts or not. $x_i, i, \ldots, k$, denote k the different touchpoints in the customer journey. $P(y|x_i)$ is the probability that a label occurs given touchpoint i is encountered. $N_{conversion}(x_i)$ and $N_{non\ conversion}(x_i)$ denote the number of converted and non-converted customer journeys exposed to touchpoint i, respectively.

$$P(y|x_i, x_j) = \frac{N_{conversion}(x_i, x_j)}{N_{conversion}(x_i, x_j) + N_{non\ conversion}(x_i, x_j)}$$

where $i \neq j$. The same notation of the first formula is applied, however, now pair-wise conditional probabilities are computed and therefore the notation is generalized. Hence, (x_i, x_j) denotes the interaction between two touchpoints. In other words, the customer has to be exposed to these two touchpoints.

In step two of the simple probabilistic model, the attribution for each touchpoint is computed at an individual level. The credit assigned to touchpoint i is calculated for each customer journey leading up to a conversion. The formula is provided below:

$$A(x_i) = P(y|x_i) + \frac{1}{2(k-1)} \sum_{j \neq i} [\, P(y|x_i, x_j) \; - \; P(y|x_i) - \; P(y|x_j) \,]$$

where $A(x_i)$ is the attribution for touchpoint i and k is the number of touchpoints encountered during a particular customer journey. Hence, for a particular customer journey leading up to conversion, the attribution for touchpoint i is the contribution of the individual touchpoint as well as the synergy effects of combinations between touchpoint i and the other touchpoints. The synergy effect is the impact of both touchpoints without their individual influence. Thus, attribution is the contribution of the individual touchpoint and the synergy effects. After estimating the attribution for each customer journey leading up to a conversion, the total attribution for touchpoint i can be determined by summing up the attributions assigned to touchpoint i. As the last step, the attribution of all touchpoints is normalized to facilitate interpretation and comparison.

This estimation is, in essence, a second order probability model. It is critical to include the second order term because some touchpoints might not have a strong effect on its own, however, they may have a substantial effect by including them in combination with another touchpoint. Shao and Li (2011) argue that one could, in theory, go beyond the pair-wise conditional probabilities and include triple-wise conditional probabilities. Nonetheless, the number of customer journeys with identical third-order interactions decline sharply. Sparse conditional probabilities of particular combinations of touchpoints yield volatile results which are undesirable. In addition, Shao and Li (2011) comment that they choose the most basic feature construction schema for the model (i.e. they encode the presence of the features with binary descriptors irrespective of the number of times a touchpoint occur within a customer journey). Despite that it is theoretically possible to include the frequency that a touchpoint occurs, this is not preferred for the same reasons as mentioned above (i.e. sparse conditional probabilities).

Dalessandro et al. (2012) extend the model and show that it is equivalent to the Shapley Value solution in cooperative game theory. The Shapley Value solution is the concept of fairly distributing gains among the players working in coalition. Conceptually, the Shapley Value solution assigns the average marginal contribution to each player after considering all possible combinations (Shapley, 1988). Dalessandro et al. (2012) incorporate, despite the advice of Shao and Li (2011) to keep only first and second order probabilities, *(k-1)* interactions (i.e. the number of unique touchpoints present in the journey minus one). The model becomes quickly unwieldy when having more than ten touchpoints since the number of Shapley Value solutions to compute increases exponentially. More specifically, the number of Shapley Value solutions to compute is two to the power of unique touchpoints (i.e. $2^{\text{touchpoints}}$). Nevertheless, the work of Dalessandro et al. (2012) is paramount because it establishes trust in the model since the Shapley Value solution has multiple fair properties. For example, if touchpoints are completely equivalent they should get an identical attribution and if touchpoints do not contribute anything their attribution should be zero. To highlight the fairness of distributing attribution based on

the Shapley Value solution, the economist Lloyd Stowell Shapley won the Nobel Prize in 2012 for inventing the Shapley Value solution. Nowadays several large enterprises such as Abakus and Google Attribution 360 have developed an attribution model founded on a version of the Shapley Value solution (Hülsdau & Teuteberg, 2018).

2.2.3 Logistic regression

Another prevailing model to classify binary labels is logistic regression. A logistic regression estimates the log-odds of the probability of an event from a linear combination of features. The logit transformation of the probability imposes an s-shape and ensures that the predicted outcome is between zero and one. The parameters of the logistic regression do not have a closed form solution as in a linear regression but are generally estimated with the maximum likelihood method. The maximum likelihood method is an iterative procedure, converging when the parameters have found the minima. The equation of the logistic regression is provided below:

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \sum_{i=1}^{k} \beta_i \, x_i$$

where P is the probability to convert. x_i, i, …, k, is the number of times a touchpoint occurs within a customer journey. β_0 is the bias and β_i, i, …, k, is the corresponding parameter for the touchpoint. A difference between the input data for the Shapley Value solution and the logistic regression is that the logistic regression takes frequency into account, where the Shapley Value solution only checks whether or not a touchpoint occurs. By taking the frequency into account, more information is exploited which generally leads to more accurate estimates. However, neither the Shapley Value solution nor the logistic regression takes the temporal ordering into account.

The logistic regression is a predictive model used for various kinds of problems. Chatterjee, Hoffman, and Novak (2003) were the first to predict online customer behavior on a webpage with a logistic regression. The labels used in the study by Chatterjee et al. (2003) are whether the user clicked on a banner and features are the number of pages visited, banners seen so far, the time between browsing events and so forth. In this thesis, the logistic regression predicts whether or not a customer will convert based on the frequency of encountering touchpoints. Hence, the output of the model is the probability that a customer journey converts.

A major difference between the logistic regression and the Shapley Value solution is the objective of the model. The Shapley Value solution naturally provides an attribution value for each touchpoint, yet the logistic regression predicts whether a customer eventually converts as accurate as possible. The parameters of the logistic regression need to be translated to attribution. From a theoretical perspective, the most obvious approach is to estimate the marginal effects. More specifically, one can compute the average marginal effect implying the average expected increase in the probability of conversion when the corresponding touchpoint increases by one (Greene, 2012). A drawback of the average marginal effect is that when the estimated coefficient of a predictor is

negative, the average marginal effect will be negative which in turn leads to a negative attribution. An alternative method to estimate attribution is proposed by Rentola (2014), the formula to compute attribution for touchpoint *i* is given below:

$$A(x_i) = \left(\frac{1}{1 + e^{-(\beta_0 + \beta_i)}}\right)$$

The same notation as for the equation above is applied. The formula to compute attribution for touchpoint *i* is both the bias (β_0) as well as the corresponding coefficient (β_i) for touchpoint *i* inserted into the logit function. In other words, the attribution of a particular touchpoint is the effect of the bias and touchpoint in a non-linear function. As a final step, the attribution of touchpoints is normalized. The approach by Rentola (2014) is a rather practical perspective. A major advantage of this approach is that it does not suffer from the problem of negative coefficients. Large negative coefficients will be driven down to zero.

An extension that can be applied to logistic regression is bagging. Shao and Li (2011) adopt this approach and estimate a bagged logistic regression to classify conversion. Bagging, also known as bootstrap aggregation, is a meta-algorithm which takes *M* subsamples with replacement from the dataset (Błaszczyński & Stefanowski, 2015). The model is trained on subsamples of the dataset and generally an average of the predictions is taken. It is usually applied as of two main reasons. Firstly, it reduces variance since it takes random samples from the underlying data generating process. Secondly, it avoids overfitting because the random fitted noise will be averaged. Overfitting might cause a problem when there are a lot of features and relatively few observations. Bagging does in general not affect the bias significantly (Błaszczyński & Stefanowski, 2015).

Another method to prevent overfitting is regularization. Different types of regularization exist. To prevent generalized linear regressions from overfitting, either Lasso or Ridge regularization are usually applied (MartíNez-MartíNez, 2011). In both Lasso and Ridge regularization, the loss function is penalized by adding an additional term to the regression. The additional term consists of a constant factor, generally denoted as lambda, and a norm vector of the coefficients. The additional term is depicted below. The larger the value of the coefficients the higher the cost function. This prevents the equation from getting large coefficients and hence overfitting. In addition, lambda is a hyperparameter that can be set, the higher the lambda the more the coefficients are suppressed.

$$+ \sum_{i=1}^{k} \lambda \, \beta_i$$

Lasso regularization uses the L1-norm also known as the least absolute deviations. Ridge regularization apply the L2-norm of squared errors. The main difference between the two methods is that Lasso tends to pick one of the features and discards the others (MartíNez-MartíNez, 2011). Hence,

Lasso regularization is often used for feature selection. Ridge regularization does also prevent the regression from overfitting by shrinking the coefficients to zero.

Both the Shapley Value solution and logistic regression do not incorporate temporal dynamics. However, the input of the logistic model can be adjusted by making binary features of the *t* latest touchpoints of the customer journey to capture the dynamics (Anderl et al., 2016). Instead of encoding a feature for every touchpoint one can encode a feature for every possible touchpoint at that time instance. It is unfeasible to include features for all time instances as the number of features will become unwieldy large. One could also incorporate the first time instances. However, Wooff and Anderson (2013) demonstrate that the last touchpoints have more explanatory power than first touchpoints. The logistic regression with the one hot encoded *t* latest touchpoints is given below:

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \sum_{t=1}^{z}\sum_{i=1}^{k} \beta_{it}\, x_{it}$$

where the same notation is applied as in the previous logistic regression equation, but now an additional summation sign is incorporated which represents the number of *t* latest time instances to include. In addition, one of the touchpoints for each time instance need to be removed to prevent the multicollinearity issue. In order to calculate attribution from the parameters, both the average marginal effect and the approach proposed by Rentola (2014) can be employed. But the only difference is that all coefficients corresponding to the touchpoints need to be taken into consideration.

2.2.4 Markov chain

A model that inherently includes temporal dynamics is the Markov chain (Anderl et al., 2016; Norris, 1998). The Markov chain is a mathematical system that computes the probability of hopping from one state to the next (Keilson, 2012). A state in the context of attribution modeling can be a touchpoint or an end state of the customer journey. The customer journey either ends in a conversion or no behavior for a specified period of time. The end state is the absorbing state meaning that it is impossible to leave when reached. A collection of the possible states is the state set:

$$S = \{s_1, \dots, s_n\}$$

The first-order Markov assumption states that the information captured at time *t* is fully explained by the feature at time *t-1* implying that observations before *t-1* do not matter (Keilson, 2012). Based on this assumption, the Markov chain estimates the probability of transitioning from one state to another. The transition probabilities are calculated with the following formula:

$$w_{ij} = \mathrm{P}\left(X_t = s_j | X_{t-1} = s_i\right), 0 \le w_{ij} \le 1, \sum_{j=1}^{N} w_{ij} = 1 \forall i$$

where w_{ij} is the transition probability of hopping from state *i* to state *j*. This is computed as the empirical probability of going to state *j* given state *i*. In addition, the computation of the transition

probabilities has two properties. Firstly, the transition probability must be between zero and one. Secondly, the summation of the probabilities to all possible states given a particular state must be one. A simplified example of four customer journeys is depicted in Figure 1 below.

Customer journey 1	Touchpoint 1 -> Touchpoint 3 -> Conversion
Customer journey 2	Touchpoint 2 -> Touchpoint 3 -> Non-conversion
Customer journey 3	Touchpoint 2 -> Touchpoint 1 -> Touchpoint 3 -> Conversion
Customer journey 4	Touchpoint 1 -> Non-conversion

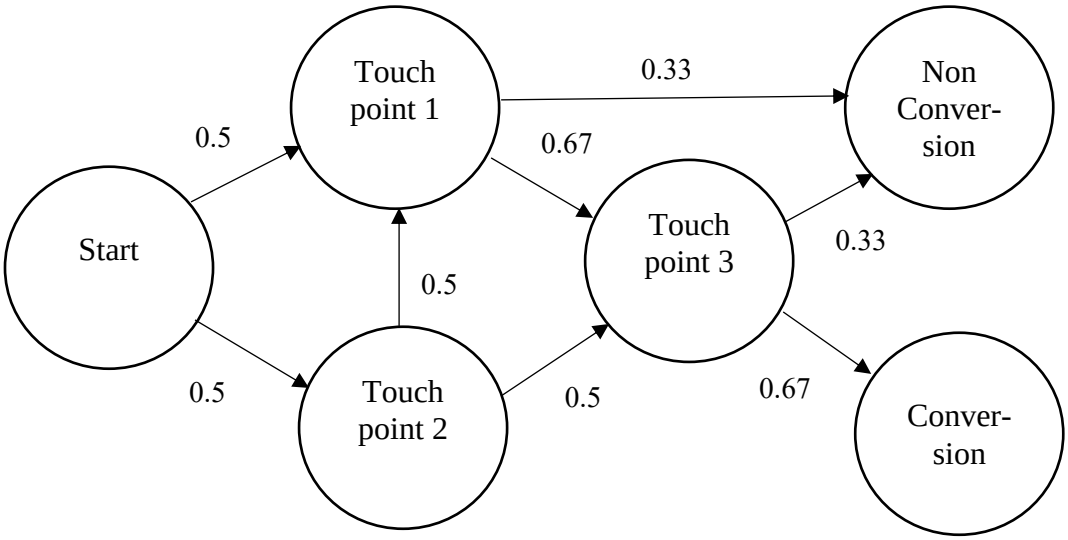


Figure 1. A list of four customer journeys and their graphical representation is provided. The nodes represent the states, the arrows indicate the direction, and the probability of hopping from one state to the next is given.

The transition probabilities are commonly presented in a transition matrix, representing a map of followed paths by customers. These transition probabilities express the sequential nature of the customer journey rather than an aggregated collection of touchpoints. The transition matrix is useful for discovering rarely or frequently walked paths that drive to conversions. The transition matrix allows for identifying structural correlations between touchpoints to construct an attribution model. More specifically, attribution is estimated as the change in probability to reach the conversion state from $t = 0$ when removing s_i from the matrix. Anderl et al. (2016) refer to this as the removal effect. The formula of the removal effect is given below:

$$A(x_i) = 1 - \left(\frac{Converion\ rate\ without\ touchpoint\ i}{Conversion\ rate}\right)$$

In other words, the removal effect provides the change in conversion if state s_i is completely removed, enabling to perform a counterfactual analysis for computing attribution. After estimating the removal effect, the contribution assigned to the touchpoints will be normalized to assist interpretation and comparison.

First-order Markov chains imply that the current state solely depends on the previous touchpoint and not on earlier touchpoints. An extension of the Markov chain is to relax the first-order assumption to higher-order assumptions. Anderl et al. (2016) adopt this approach and take the *t* latest touchpoints of the customer journey into consideration. They estimate the first-, second-, third-, and fourth-order Markov chains. By relaxing this assumption, the state becomes a sequence of touchpoints. A generalization of the provided formulas for the Markov chain is applied. Higher-order Markov chains are generated to incorporate longer temporal dynamics which may lead to better performance.

2.3 Evaluation criteria

The first evaluation criterion is that attribution models should be interpretable. Interpretability is a hot topic in machine learning, however, not all articles define exactly what it encompasses (e.g. Doshi-Velez & Kim, 2017; Vellido, Martín-Guerrero, & Lisboa, 2012). Lipton (2016) suggests that interpretability consists of more than one concept, reflecting multiple distinct notions. One view is to think of interpretability as intelligibility or understandability (Caruana et al., 2015). This view is about grasping how models work. Comprehensible models are referred to as transparent and opaque models as black-boxes. Thus, this perspective regards attribution models as easy to interpret if they are intrinsically interpretable. On the other hand, some argue that the model itself does not have to be highly interpretable, but the predictions need to be elucidated (Molnar, 2018). Explaining how the model predicts without explaining the components of the model might suffice for some areas of expertise. This view is often referred to as post-hoc interpretability since the model is made interpretable after the model is computed.

Yet, this thesis adopts the former perspective that attribution models should be intrinsically interpretable as of three reasons. Firstly, attribution models do not have a label with the true attribution and prediction is only one evaluation criterion. Interpreting the predictions with post-hoc explanations only partially solves the problem. Secondly, the goal of attribution models is to assign credit to touchpoints and therefore attribution needs to be evaluated on interpretability. Hence, assessing all steps of computing attribution provides a comprehensive picture of the interpretability of the attribution model. Thirdly, post-hoc interpretability is generally used if models have a deeply nested structure as neural networks and intrinsic interpretability is unfeasible (Lipton, 2016). In addition, creating transparency in the data-driven attribution models is key for building trust. Marketers and managers may feel more confident with an intrinsically interpretable model. Hence, they may rely on the results of the attribution model more easily. Therefore, in this thesis, a model that is easy to interpret should have a clear and intuitive understanding of the components of the attribution model, while the statistical procedure should be valid.

The second evaluation criterion is that the model should predict conversion accurately. Although attribution is distinct from prediction, it is highly likely that a model that can classify

individuals that convert from individuals that do not convert is also able to assign the appropriate contribution to the touchpoints. It is hard to argue that models that do not predict well are good in attributing. In addition, it assists to persuade managers of the model's trustworthiness (Lodish, 2001).

The binary classification task to predict conversion is usually skewed as there are more individuals who do not convert. Hence due to the imbalanced labels, standard evaluation metrics as the classification accuracy or log-likelihood are inappropriate (Kuncheva, Arnaiz-González, Díez-Pastor, & Gunn, 2018). The predictive accuracy measured by the area under the receiver operating characteristic (ROC) is insensitive to the skewness of the data. The ROC can be decomposed in the true positive rate and the false positive rate, which are represented on the vertical and horizontal axis, respectively. All correct predicted conversions divided by all true conversion is the true positive rate. All incorrectly predicted conversions divided by all true non-conversions is the false positive rate. The true positive rate and the false positive rate are irrespective of the actual conversion in the dataset. The area under the curve represents the predictive accuracy and ranges from 0.5 to 1. A larger area under curve is better. An additional benefit is that no hard threshold of the classifier has to be specified.

As a second measure for predictive accuracy, the top-decile lift is calculated. The top-decile lift is computed in three steps. First, the top ten percent of the customers with the highest predicted probability is taken. From this ten percent, the average probability to convert is calculated. As the last step, the average probability to convert of the top percent is divided by the average probability to convert of the entire dataset. Hence, the top-decile lift shows how much more likely the predicted top ten percent converts in comparison with the average customer (Neslin et al., 2006). A high top-decile lift indicates a good ability to predict, it demonstrates the power of an attribution model to beat a random model. The top-decile lift is comparable between models. The top-decile lift is a prevailing metric in the customer churn and targeting industry (Zhu, Baesens, Backiel, & van den Broucke, 2018).

The last evaluation criterion is whether the model is robust. This thesis embraces the definition by Shao and Li (2011) that attribution models should generate stable and consistent results. The robustness of the attribution models is assessed with two measures. Firstly, the predictive performance described in the previous paragraph should be consistent when resampling from the underlying data-generating process. This is measured by the standard deviation of both the area under the ROC curve and the top-decile lift.

Secondly and more important for our present purpose, the contribution assigned to each touchpoint should be stable across resamples. This is crucial because the marketing spending of enterprises depends on the results of the model. Fragile results suggest a weak attribution measure that is questionable (Wooff & Anderson, 2013). The robustness of attribution is measured with the coefficient of variation (McAuliffe, 2015). The coefficient of variation of a model is determined in three steps. Firstly, the mean and standard deviation of attribution of each touchpoint are estimated. Secondly, the coefficient of variation of each touchpoint is calculated by dividing the standard

deviation of a touchpoint by its mean. Lastly, the average coefficient of variation of all touchpoints is computed to make the coefficient of variation readily comparable between models. Hence, the coefficient of variation is, unlike the standard deviation, not proportional to the mean of the credit assigned to a particular touchpoint nor to the attribution model.

3. Method

The first part of this section gives a description of the dataset in detail. Section two starts with cleaning the dataset and provides the descriptive statistics. The last section presents the experimental procedure by explaining how the models are implemented and evaluated.

3.1 Dataset

The travel agency dataset is provided by GfK and is collected within their Crossmedia Link panel. Hence, there are repeated observations for different individuals i = 1, ..., N observed over t = 1, ..., T periods. The temporal aspect of the dataset can be leveraged since the individuals and timestamps are known. The dataset ranges from January 2015 to October 2016. Each observation in the dataset consists of one single touchpoint. More specifically, the dataset contains various touchpoints, which can be broadly classified in enterprise-initiated contacts and customer-initiated contacts. Enterprise-initiated contacts are divided into banners, retargeting, pre-roll ads, affiliates, and e-mails. The enterprise-initiated contacts are impressions meaning that the customer did not necessarily clicked on it. All enterprise-initiated touchpoints are measured passively by tagging and the GfK custom-designed plugin. Relevant customer-initiated contacts are websites, apps, and search engine terms, and can, in turn, be divided into information/comparison, accommodation, airline, competitive travel agencies and the focal travel agency. Which websites, apps, and search engine terms are relevant and to what class they belong to is determined in collaboration with the focal company. The enterprise-initiated touchpoints are measured passively by the GfK custom-designed plugin. Additionally, bookings at the focal travel agency are registered through the confirmation page of the websites, tagging, and surveys.

After obtaining the data, the definition of a customer journey is based on a set of rules determined by both the focal company and GfK. A customer journey ends when a customer makes a purchase or does not have any contact initiated by the customer for four weeks. Moreover, the customer journey can be extended when an enterprise-initiated touchpoint is encountered within two weeks before the start of the customer journey to incorporate potentially triggers of enterprise-initiated touchpoints. However, if no active behavior occurs within the two weeks after the enterprise-initiated touchpoint, the enterprise-initiated touchpoint is not part of the customer journey and therefore removed from the data. All touchpoints up to 6 months before the end of the customer journey are considered part of the customer journey.

After figuring out how the data is obtained and how the customer journeys are defined, a description of the touchpoints as well as whether the touchpoints are initiated by the enterprise or customer is provided as list 1 below.

Touchpoint	Description	Type
Affiliate	Visiting a website of an affiliate. Affiliate marketing is a commission-based method that rewards the sender for referring the customer to the focal website.	Enterprise-initiated
Banner	Encountering a banner. Banners are graphical web-advertisements with the purpose to increase traffic or sell more.	Enterprise-initiated
Email	Opening an email. E-mails in the travel industry are generally discount offers.	Enterprise-initiated
Pre-roll	Viewing a short promotional video before watching the intended video.	Enterprise-initiated
Retargeting	Encountering a banner after the first interaction has occurred.	Enterprise-initiated
Accommodation website	Visiting a relevant website where one can book a place to stay.	Customer-initiated
Accommodation app	Visiting a relevant application on either a mobile or tablet where one can book a place to stay.	Customer-initiated
Accommodation search	Searching for one of the relevant accommodation websites on a search engine.	Customer-initiated
Information/comparison website	Visiting a relevant website where one can find information about a journey or where one can compare offers.	Customer-initiated
Information/comparison app	Visiting a relevant application on either a mobile or tablet where one can find information about a journey or where one can compare offers.	Customer-initiated
Information/comparison search	Searching for one of the relevant information or comparison websites on a search engine.	Customer-initiated
Travel agent website	Visiting a relevant website where one can book a trip or vacation.	Customer-initiated
Travel agent app	Visiting an application on either a mobile or tablet where one can book a trip or vacation.	Customer-initiated
Travel agent search	Searching for one of the relevant travel agent websites on a search engine.	Customer-initiated
Focal website	Visiting the focal website where one can book a trip or vacation.	Customer-initiated
Focal search	Searching for the focal website of the travel agent on a search engine.	Customer-initiated
Airline company website	Visiting a relevant website where one can book a flight.	Customer-initiated
Airline company app	Visiting a relevant application on either a mobile or tablet where one can book a flight.	Customer-initiated
Airline company search	Searching for one of the relevant airline company websites on a search engine.	Customer-initiated
Generic search	Searching for a relevant term related to a trip or vacation on a search engine.	Enterprise-initiated

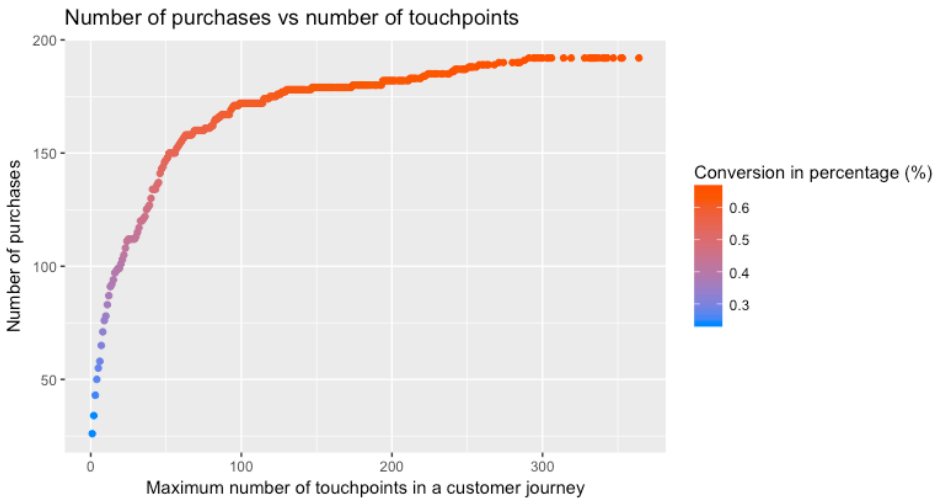
List 1. A description of the touchpoints as well as whether the touchpoints are initiated by the enterprise or customer. Note 1: all enterprise-initiated touchpoints are impressions, meaning that the customer did not necessarily clicked on it. Note 2: relevant means that it is classified as important by the focal company. Note 3: accommodation enterprises, information/comparison websites, travel agents, and airline companies are pre-defined by both the focal company and GfK.

3.2 Data cleaning and descriptive statistics

Before presenting the descriptive statistics, two notable observations need to be highlighted. Firstly, the data shows that a lot of consecutive touchpoints are the same and occur within a short time span.

For example, a particular customer has visited an accommodation app approximately 100 times in three minutes. This is not an extraordinary customer journey. Since the frequency of registration of touchpoints differs between touchpoints and it is highly unlikely that customer visits the same touchpoint over and over in a remarkably short time period, it is chosen to exclude consecutive touchpoints that are the same. This is performed in consultation with GfK. With the objectives in mind, to measure and evaluate attribution, it seems to be a plausible assumption. For clarity, this does not mean that the same touchpoint cannot occur multiple times within a customer journey. After excluding the same consecutive touchpoints, the average number of touchpoints within a customer journey drops from 84.67 to 11.58.

To stress the second notable observation, graph 1 is provided below. The horizontal axis represents the maximum number of touchpoints in a customer journey and the vertical axis the corresponding number of purchases. Furthermore, the color indicates the conversion in percentage as shown in the legend. The graph demonstrates that the conversion rate increases as the number of touchpoints within a customer journey increases. Customer journeys with solely one touchpoint have on average a conversion rate of approximately 0.25%. Customer journeys with 10 touchpoints have on average a 0.1 percentage point higher chance to convert and customer journeys with 25 touchpoints are already twice as likely to convert as customers with one customer interaction. When the customer journeys consist of 100 touchpoints the conversion rate is 0.6%, after which the conversion rate increases slowly to a maximum of 0.66% at 192 touchpoints. Hence, the graph shows a diminishing increase in conversion rate as the number of touchpoints increases. A possible rationale for the large increase is that short customer journeys are not genuine journeys. For example, it could be that a panel member searches for "Amsterdam" because a friend works there. This may be an erroneously classified generic search term. Nevertheless, another explanation is that customers who interact more frequently are more engaged and therefore their propensity to purchase is higher (Pansari & Kumar, 2017). Since there are several possible explanations, no customer journeys are excluded from the analysis.



Graph 1. Showing that the conversion rate increases as the number of touchpoints within a customer journey increases.

The descriptive statistics are presented after excluding the same consecutive touchpoints. These statistics are summarized in Table 1 and will be discussed now. This research is conducted in the travel agency industry where online advertising plays an increasingly important role (Park & Oh, 2012). Customers generally invest lots of time in figuring out vacations and are thereby exposed to various customer-firm interactions (Pabel & Prideaux, 2016). Nevertheless, the results of this research should be interpreted with care, since the analyses are only conducted on a single travel agency dataset to address the research question.

There are twenty unique touchpoints in this research. A distinction can be made between customer-initiated and enterprise-initiated touchpoints as explained in the previous section. Furthermore, the total number of journeys in this dataset is 29,011. To give an indication about the distribution of the number of touchpoints within a customer journey, the number of customer journeys with > 2 touchpoints is 22,974 and > 5 touchpoints is 19,154, implying that there are a lot of short journeys with one or two touchpoints and that the number of customer journeys with more touchpoints diminishes. Moreover, the total number of touchpoints in the dataset is 336,001. The average number of touchpoints within a customer journey is 11.58 and has a standard deviation of 32.65. The standard deviation is relatively large and indicates that there are still a lot of deviations from the mean even after removing the same consecutive touchpoints. More specifically, the deviations are negatively skewed since customer journeys must consist of at least one touchpoint. A customer journey ultimately ends in a purchase or after a blackout period of four weeks. The total number of customer journeys ending in a conversion is 192 in this dataset. Viewing this from a customer journey perspective, the conversion rate in percentage is 0.66%.

Description	
Industry	Travel agency
Number of different touchpoints	20
Customer-initiated touchpoints	15
Enterprise-initiated touchpoints	5
Total number of journeys	29,011
Thereof with length > 2	22,974
Thereof with length > 5	19,154
Total number of touchpoints	336,001
Number of touchpoints within a customer journey	11.58 (SD = 32.65)
Number of conversions	192
Conversion rate in percentage per customer journey	0.66%

Table 1. Presenting some descriptive statistics.

To assess the distribution of the number of touchpoints, Table 2 is provided below. The table presents information about the rate of occurrence of touchpoints in this dataset. The first four columns are self-explanatory and the last column is calculated by dividing the number of times a touchpoint occurs by the total number of customer journeys. Hence, the values in the last column indicate the number of times an average customer journey encounters a specific touchpoint. There are several noteworthy

observations in the table. Firstly, 2.39% of the dataset are enterprise-initiated touchpoints and the remaining 97.61% are customer-initiated, suggesting the distribution is highly skewed. Expressing this in the average number of enterprise-initiated touchpoints encounter per customer journey is 0.28, meaning that on average a quarter of the customer journeys is affected by enterprise-initiated touchpoints. Secondly, 78.48% of the touchpoints are coming from visiting a website. The major stakeholders of this are accommodations websites and travel agent websites. These contribute 28.77% and 28.25% respectively. Thirdly, the average number of visits on the focal website per customer journey is 0.39, thereby receiving rank 6. Combining the frequency of visits of the focal website and the total number of conversion (displayed in the previous table), the percentage of customer journeys leading up to a conversion is 1.71% (=192/11,217) given they have visited the focal website. In other words, about 1 in 60 customer journeys convert when landing on the focal website.

Touchpoint	**Frequency**	**Share of touchpoints**	**Rank**	***Frequency*** / ***Total customer journeys***
Affiliate	529	0.16%	19	0.02
Banner	595	0.18%	18	0.02
Email	1,643	0.49%	13	0.06
Pre-roll	811	0.24%	17	0.03
Retargeting	4,449	1.32%	8	0.15
Accommodation website	96,658	28.77%	1	3.33
Accommodation app	2,949	0.88%	10	0.10
Accommodation search	5,319	1.58%	7	0.18
Information/comparison website	38,264	11.39%	3	1.32
Information/comparison app	2,631	0.78%	11	0.09
Information/comparison search	1,462	0.44%	15	0.05
Travel agent website	94,921	28.25%	2	3.27
Travel agent app	913	0.27%	16	0.03
Travel agent search	2,573	0.77%	12	0.09
Focal website	11,217	3.34%	6	0.39
Focal search	451	0.13%	20	0.02
Airline company website	33,853	10.08%	4	1.17
Airline company app	1,549	0.46%	14	0.05
Airline company search	4,335	1.29%	9	0.15
Generic search	30,879	9.19%	5	1.06

Table 2. Presenting information about the rate of occurrence of touchpoints.

3.3 Experimental procedure

Conversion is a binary label hinting that it is a classification task. However, unlike many classification tasks, the main objective is not to predict. The task is to assign credit to the touchpoints and compare the models on the evaluation criteria. Hence, the goal is not to optimize the parameters on this dataset but rather to evaluate the attribution models as precisely as possible. The models are evaluated on the ease of interpretation, predictive accuracy, and robustness. Information about the models that are compared and the selected parameters are provided below. In addition, the analyses are conducted in the statistical program R version 3.4.4 (R Core Team, 2017).

3.3.1 Heuristic-based models

First, two heuristic-based attribution models are implemented to provide a baseline to compare attribution and to compare the evaluation criteria. To be more specific, last touch attribution and first touch attribution are implemented. Last touch attribution assigns all credit to the last touchpoint. On the contrary, first touch attribution assigns all credit to the first touchpoint. These heuristic-based models provide as output an attribution score. To compare the models for predictive accuracy, they are transformed into a predictive model. The predicted probability to convert for the last touch attribution model is the normalized attribution to the last touchpoint. Following the same logic, the predicted probability to convert for the first touch attribution model is the normalized attribution to the first touchpoint. These two models are easy to program in R.

3.3.2 Shapley Value solution

The Shapley Value solution proposed by Shao and Li (2011) takes the individual touchpoints and the interaction between touchpoints into account. The two-step computation provided in section 2.2.2 is followed. First, the conditional probability of conversion given the individual touchpoints and the conditional probability of conversion given the interaction between two touchpoints are computed on the entire dataset. Secondly, the credit assigned to the touchpoints is calculated by taking the contribution of the individual touchpoints as well as the synergy effects into consideration. Lastly, the attribution is normalized. In addition, the model can be made predictive since the normalized attribution sums up to one and the features are one hot encoded. The predictive probability for a customer journey is the sum of the devoted attribution to the unique touchpoints present in the customer journey. Since the model does not include interactions higher than pair-wise, no Shapley Value solution package is used to implement the model.

Dalessandro et al. (2012) extend the model by incorporating more than two interactions and show that it is equivalent to the Shapley Value solution in cooperative game theory (Shapley, 1988). The model becomes quickly unwieldy when having more than ten touchpoints since the number of Shapley Value solutions to compute increases exponentially. The number of touchpoints in this study is twenty, meaning that 1,048,576 (2^{20}) Shapley Value solutions need to be estimated. The code to run the model is implemented however it is unfeasible due to computation time. A rough calculation shows that it would take extremely long.[2] Consequently, only the model proposed by Shao and Li (2011) is estimated.

3.3.3 Logistic regression

Another popular model that is implemented to classify binary labels is logistic regression. Logistic regression predicts whether or not a customer will convert based on the frequency of encountering

[2] One Shapley Value solution takes approximately 15 seconds on an iMac Late 2013 with a 2.7GHz quad-core Intel Core i5 processor and an 8GB of 1600MHz DDR3 memory. This means that 1,048,576 Shapley Value solutions will take extremely long.

touchpoints. The estimated coefficients of the logistic regression can be interpreted as the increase or decrease in log-odds of the probability (Greene, 2012).

The literature describes two approaches to translate the coefficients of the logistic regression to attribution. From a theoretical perspective, the most natural approach is to estimate the average marginal effect implying the average expected increase in the probability of conversion when the corresponding touchpoint increases by one (Greene, 2012). A disadvantage of this approach is when the coefficient is negative the attribution will be negative. An alternative approach to estimate attribution is to insert the bias and corresponding coefficient for a touchpoint in the logit formula (Rentola, 2014). For example, when the bias is -2 and the coefficient of the focal website is 1.5, the unnormalized attribution is 0.378. The approach is the solution to the issue of negative coefficient as it sets large negative coefficients virtually to zero. The logistic regression is implemented with the stats package in R, the average marginal effect with the margins package, and the approach from Rentola (2014) is coded without a package (R Core Team, 2017; Leeper, 2018).

An extension of the logistic regression to reduce variance and prevent the model from overfitting is bagging. Bagging takes M subsamples with replacement from the dataset. The estimated coefficients of the logistic regressions are averaged. After averaging, attribution is determined and the evaluation measures are calculated. To clarify, the predictive accuracy is measured with the average estimated coefficients. However, in some studies, predictions are made first and then the average predicted probability is taken. Yet in this thesis, the coefficients are first averaged since attribution is also determined on the average coefficients. To perform the bagging procedure, two hyperparameters need to be selected; the size of the subsample and the number of times the subsample is taken (Błaszczyński & Stefanowski, 2015). The two parameters are not optimized with a grid search or an alternative technique as the goal is not to optimize the parameters but rather to evaluate the attribution models. Optimizing the model will improve the fit of this dataset. Yet, the objective is to generalize the conclusions of the evaluation criteria to a broader perspective. As in the study by Louppe and Geurts (2012), the number of replications is set to 50. Moreover, the size of the subsample is two-thirds of the entire dataset. One may argue that the size of the subsample is relatively high. However, a small subsample may lead to zero or little conversions per repetition since the dataset is skewed. Hence due to the imbalanced dataset, the subsample is chosen to be two-thirds of the entire dataset. No package in R is used to implement the bagging procedure.

Regularization is a machine learning technique applied to models that fit the training data to well. Generally, Lasso or Ridge regularization is used for generalized linear regressions. The main difference between the two methods is that Lasso tends to pick one of the touchpoints and discard the others, which is an undesirable property as the goal is to evaluate the attribution models as precise as possible (MartíNez-MartíNez, 2011). Consequently, when the logistic regression overfits, Ridge regression will be used. The package in R by Friedman, Hastie, and Tibshirani (2010) simplifies the implementation of Ridge regularization for generalized linear regressions.

A drawback of both the Shapley Value solution and the logistic regression is that they do not incorporate temporal dynamics. However, the input of the logistic model can be adjusted by making binary features of the *t* latest touchpoints of the customer journey to capture the dynamics (Anderl et al., 2016). Instead of encoding a feature for every touchpoint, a feature for every possible touchpoint at the last *t* time instances is encoded. As it is unfeasible to include features for all time instances, three models are computed. A model with only the last touchpoints, the two last touchpoints, and the three last touchpoints. The model with the three last touchpoints already consists of 60 parameters and therefore models with more time dynamics are not estimated. In addition, a reference level needs to be chosen to estimate the model. The reference level of the model with only the last time instance is the touchpoint "generic search" since the other enterprise-initiated touchpoints consist of websites, apps, and search engine terms and there is for generic solely search. Models that include earlier time instances have as reference level "none" as it compares the touchpoints to no touchpoint. Ideally, "none" is also selected for the last time instance, however, customer journeys must consist of at least one touchpoint and therefore the category "none" does not exist as a last touchpoint.

3.3.4 Markov chain

A model that inherently include temporal dynamics is the Markov chain (Anderl et al., 2016). The first-order Markov assumption states that the state at the current time only depends on the state at the previous time point and that the past is irrelevant (Keilson, 2012). Based on this assumption, the Markov chain estimates the probability of transitioning from one state to another. The transition probabilities are presented in a transition matrix, representing a map of followed paths by customers. Moreover, the transition matrix facilitates computing the removal effect to measure attribution (Anderl et al., 2016). The removal effect computes the decrease in conversion rate if a particular touchpoint is disregarded from the network. A simplified transition matrix is provided in Table 3 below to explain how the removal effect is computed.

	Conversions	Non conversion	Touchpoint 1	Touchpoint 2	Touchpoint 3
Touchpoint 1	0	0.3333	0	0	0.6667
Touchpoint 2	0	0	0.5	0	0.5
Touchpoint 3	0.6667	0.3333	0	0	0
Start	0	0	0.5	0.5	0

Table 3. Example of a transition matrix. The vertical axis represents the previous state and the horizontal axis the current state.

The formula provided in section 2.2.4 is applied to compute the removal effect. The conversion rate without removing any touchpoint is 0.5 (=0.5*0.67*0.67 + 0.5*0.5*0.67 + 0.5*0.5*0.67*0.67). Hence, the unnormalized removal effect for touchpoint 1 is 0.67 (=1- ((.5*0.5*0.67) / 0.5)), touchpoint 2 is 0.55 (=1- ((0.5*0.67*0.67) / 0.5)), and touchpoint 3 is 1 (=1- ((0) / 0.5)). Normalizing the removal

effect leads to an attribution of 0.30, 0.25, and 0.45 for touchpoint 1, touchpoint 2, and touchpoint 3, respectively.

Altomare and Loris (2018) developed the ChannelAttribution package to compute attribution with the Markov chain and the removal effect. This package is customized to do these estimations. Nevertheless, the package does not provide any indication about the predictive accuracy or robustness of the model. The model is made predictive by using the probability to convert of the last touchpoint of the customer journey. Hence, the probability only depends on the last touchpoint in the sequence, yet this is in line with the definition of the Markov assumption.

In addition, the first-order Markov chain is extended to higher-order Markov chains by relaxing the Markov assumption. More specifically, the second-, third-, and fourth-order Markov chains are estimated to incorporate more information in the model. By relaxing this assumption, the state becomes a sequence of touchpoints. The logic of making predictive models from higher-order Markov chains is the same, however, the probability depends on the last sequence of touchpoints. Higher-order Markov chains are generated to incorporate longer temporal dynamics which may lead to better performance.

3.3.5 Evaluation criteria

The heuristic-based model, the Shapley Value solution, the logistic regression, and the Markov chain are evaluated on the ease of interpretation, predictive accuracy, and robustness. The procedure to evaluate the models is universal. Generally, for classification tasks, the data is partitioned into three datasets: the training set, validation set, and test set. However, the purpose of this thesis is not to predict conversion, yet the task is to devote credit to touchpoints and evaluate the models. Consequently, the dataset is partitioned into a training set and validation set. The training set to learn the model and the validation set to measure attribution and evaluate the model. A test set where one can assess the best performing model with optimized parameters is not required. The training and validation set will be split by means of stratified 10-fold cross-validation. To train and validate the model, conversion is needed in both sets. Hence, stratified cross-validation is chosen to ensure that both sets contain some conversions (Refaeilzadeh, Tang, & Liu, 2016). As the task is binary, each fold contains roughly 19 conversions. When regular cross-validation was applied, some folds might not have any conversion or little conversions.

The first evaluation criterion, interpretability is a more subjective criterion and therefore no statistical test can be exploited. However, the definition is indispensable to achieve more transparency in the data-driven attribution models. Interpretability is in this thesis defined as having a clear and intuitive comprehension of how attribution is determined, while the statistical procedure should be valid. Most data-driven attribution models consist of several steps. The first steps to compute intermediate results and the last step to go from these results to attribution. All these steps are scrutinized to evaluate the models on interpretability.

The second evaluation criterion, predictive accuracy, is measured with the area under the receiver operating characteristic curve and the top-decile lift. Firstly, the area under the ROC curve is computed for each iteration on both the training and validation set and then averaged. The ROC can be decomposed in the true positive rate and the false positive rate, which are represented on the vertical and horizontal axis, respectively. The ROC graph is provided to visually assess whether models outperform one another at combinations between the true positive rate and the false positive rate. Intuitively, the larger the area under the curve the more accurate an attribution model can predict conversion and non-conversion. The ROCR package by Sing, Sander, Beerenwinkel, and Lengauer, (2005) is used to compute the area under the curve and construct the ROC graphs.

As a second measure for predictive accuracy, the top-decile lift is calculated on each repetition on both the training and validation set and then averaged. Conceptually, a top-decile lift of three means that the customer journeys with the highest predicted top ten percent are three times as likely to convert than the average customer journey in the dataset. Hence, when a model has a higher top-decile lift than another model, it can better predict whether the top ten percent is going to convert. (Neslin et al., 2006). To compute the top-decile lift, the "lift" package is applied (Hoornaert, Ballings, & Poel, 2015).

The last evaluation criterion, robustness means that the attribution model should produce stable and consistent results when resampling from the underlying data-generating process. Robustness is assessed with two metrics. Firstly, the predictive accuracy measured with both the area under the curve and top-decile lift should be stable across cross-validations. Hence, the standard deviation of the area under the curve and top-decile lift is calculated across cross-validations on both the training and validation set. Intuitively, a relatively small standard deviation implies that the model produces stable results in distinguishing conversion from non-conversion.

Secondly and even more important is to evaluate the robustness of the contribution assigned to the touchpoints. The robustness of attribution is measured with the coefficient of variation (McAuliffe, 2015). First, the mean and standard deviation of attribution of each touchpoint across cross-validations is computed. To compute the coefficient of variation, the standard deviation is divided by the mean which makes the metric scale-free. After computing the coefficient of variation for every touchpoint, the average is taken to get a scaler. The coefficient of variation enables to compare attribution models on robustness.

4. Results

In the first section, attribution models are estimated and noticeable intermediate results are provided. In the second section, the results of the evaluation criteria are compared between models.

4.1 Model estimation

The models are computed with stratified 10-fold cross-validation. However, intermediate results as conditional probabilities, coefficients or transition probabilities are estimated on the entire dataset to give a holistic overview of the data. To compare the data-driven models to a baseline on the evaluation criteria both first touch attribution and last touch attribution are implemented.

4.1.1 Shapley Value solution

The computation of the Shapley Value solution consists of two steps (Shao & Li, 2011). The attribution results are provided in the next section, however, there are interesting intermediate results in the computation of step one and two.

Firstly, the conditional probabilities of conversion given the individual touchpoints are shown in Table 1 in the Appendix. The touchpoints: focal website, focal search, email, and retargeting have a conditional probability higher than 0.05. Secondly, the conditional probabilities of conversion given the interaction between two touchpoints are provided in Table 2 in the Appendix. Thirty of the 190 possible combinations between touchpoints have a value of zero, meaning that they either do not occur in the data or have a genuine conditional probability of zero. The sparse conditional probabilities indicate that the chosen feature construction schema for the model is appropriate. Moreover, ten interactions between touchpoints have a conditional probability higher than 0.1. An interesting finding is that a lot of high conditional probabilities occur at an interaction between the focal website, focal search, or retargeting and another touchpoint. The highest conditional probability is between the focal website and retargeting with a value of 0.186. Thirdly, the synergy effects between possible combinations of touchpoints are calculated and are shown in Table 3 in the Appendix. A synergy effect is the effect of both touchpoints without their individual effect. The computations are as follow:

$$P(y|x_i, x_j) - P(y|x_i) - P(y|x_j)$$

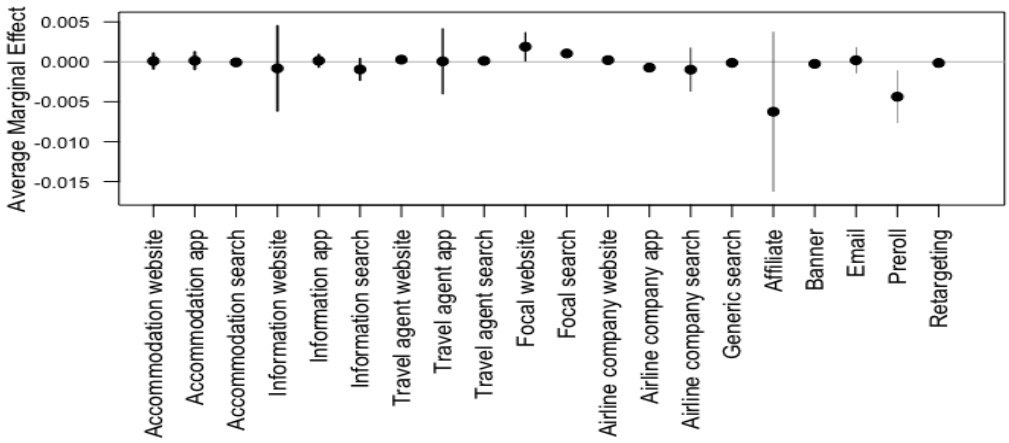
Sixteen of the 190 combinations between touchpoints have an absolute value higher than 0.05, meaning that they have a strong positive or negative synergy effect. More specifically, five have a positive synergy effect and eleven have a negative synergy effect (i.e. antagonism effect). Three of the five positive synergy effects occur in combination with the focal website. A lot of negative synergy effects occur at a combination between either a banner or pre-roll ad and another touchpoint.

4.1.2 Logistic regression

A vanilla logistic regression is fitted. Estimated coefficients are provided in Table 4 in the Appendix. Signs of the coefficients are directly interpretable (Greene, 2012). The airline company website, generic search, focal website, and focal search have a significant positive effect on the probability to

convert.[3] The information app and travel agent search have a significant negative effect on the probability to convert. Moreover, the bias has a large significant negative effect due to the skewness in the data. However, the magnitude of the touchpoints is difficult to interpret as the logistic regression estimates the log-odds of the probability. The coefficients of the logistic regression need to be transformed into attribution.

There are two approaches to transform the coefficients as discussed earlier. First, the average marginal effect is estimated since this is from a theoretical perspective the most natural approach (Leeper, 2018). Graph 2 below demonstrates the average marginal effect of each touchpoint as well as the accompanying standard error. A lot of the point estimates of the marginal effect are negative which in turn results in a negative attribution. In order to solve this problem, an alternative approach to estimate attribution is applied. The approach computes the attribution of a specific touchpoint by inserting the bias and the corresponding coefficient for that specific touchpoint into the logit function (Rentola, 2014). In other words, the contribution assigned to a touchpoint is a non-linear function of the bias and that touchpoint. An advantage of this practical solution is that it does not suffer from the problem of negative coefficients as large negative coefficients will be driven down to zero since the exponential of a number cannot be negative.



Graph 2. The average marginal effect of each touchpoint as well as the accompanying standard error. The values of the marginal effects used to create the graph are provided in Table 5 in the Appendix.

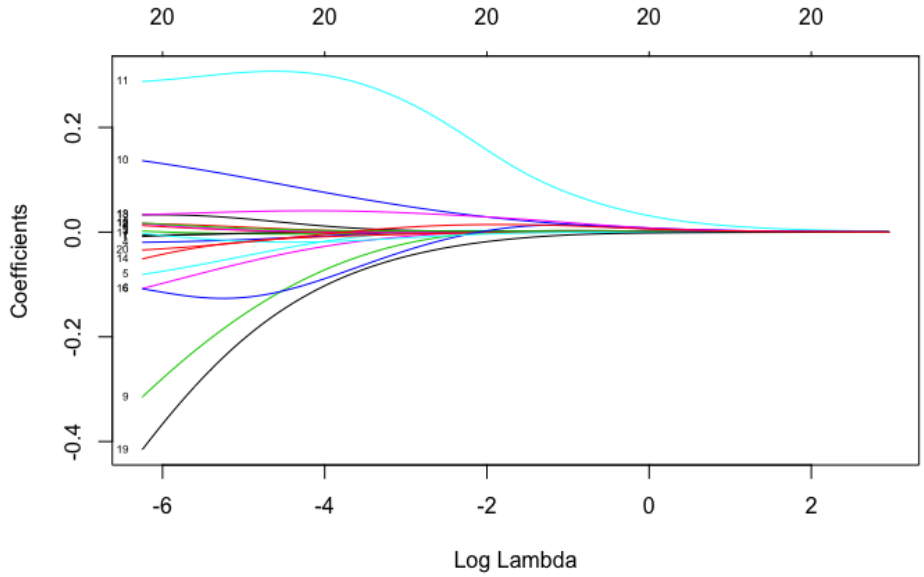
4.1.3 Bagged logistic regression

A bagged logistic regression is fitted to reduce the variance of attribution (Błaszczyński & Stefanowski, 2015). With the objectives in mind, the hyperparameters are fixed. The number of iterations per cross-validation is set to 50 and the size of the subsample is two-thirds of the entire dataset (Louppe & Geurts, 2012). However, at some iterations, the logistic regression separate the

[3] Significant at $\alpha = 0.05$.

labels perfectly due to the reduced sample size. Consequently, the estimated coefficients become extremely large or small. These coefficients will affect attribution drastically.

To prevent the coefficients from getting extreme, regularization is applied. More specifically, Ridge regularization is implemented as it does, unlike Lasso regularization, not tends to pick one of the touchpoints and discard the others (MartíNez-MartíNez, 2011). As explained in the background section 2.2.3, the regularization term consists of a constant factor, denoted as lambda, and a norm vector of the coefficients. Lambda cannot be learned from the model and need to be chosen in advance (i.e. is a hyperparameter). As decided earlier, the dataset is only partitioned in a training and validation set and the parameters are not optimized. Hence, lambda is chosen arbitrarily, yet it is important that there still is variation in the coefficients of the logistic regression as they serve to compute attribution. Graph 3 below is created to assess whether the coefficients are not driven down to zero. More specifically, the graph illustrates the magnitude of the coefficients at different lambdas. The horizontal axis displays the lambda values on a logarithmic scale. The vertical axis exhibits the magnitude of the coefficients of the logistic regression. Above the graph, the numbers of coefficients are presented. The number of coefficient remains 20 as Ridge regularization does not select features. Lambda 1.0e-06 is selected as there is still a lot of variation between touchpoints. A small lambda prevents the coefficients from getting extreme, yet do not shrinks the coefficients to zero. The coefficients of the bagged regularized logistic regression are shown in Table 6 in the Appendix.



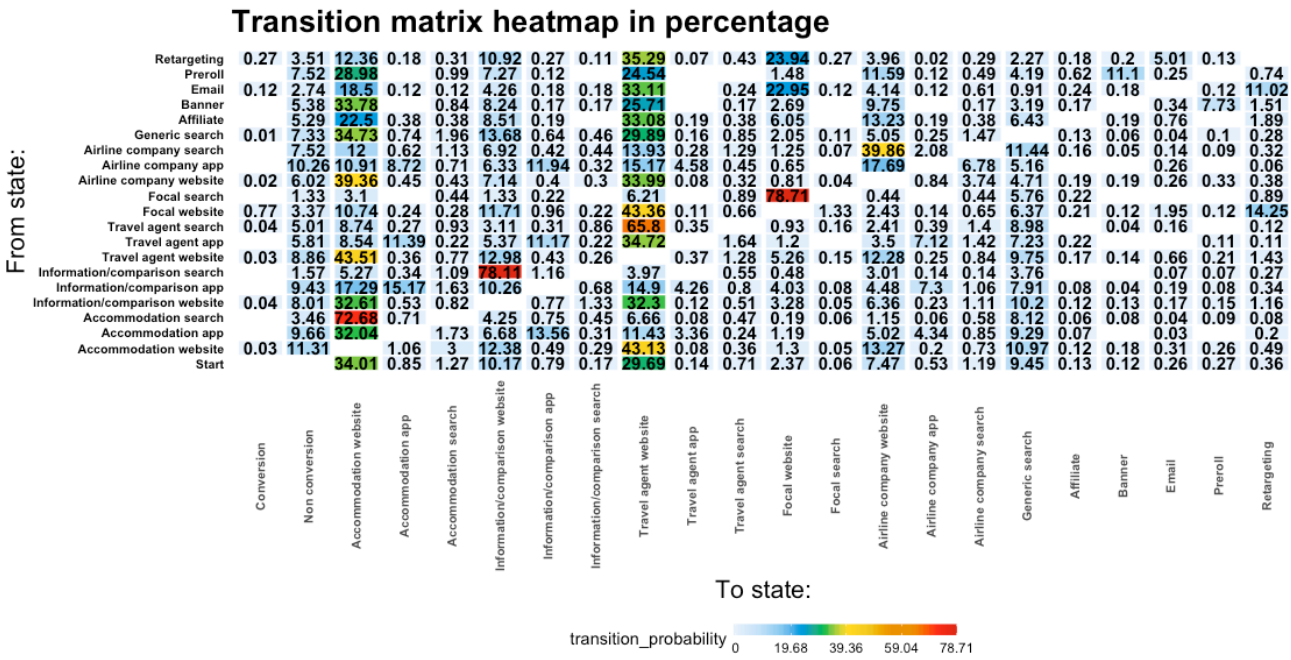
Graph 3. Observing the magnitude of the coefficients at different lambdas.

4.1.4 Dynamic logistic regression

The logistic regression can be adapted to incorporate temporal dynamics by making binary features of the latest touchpoints of the customer journey. Including features for all time instance would lead to a lot of parameters which is unfeasible to estimate. For example, estimating a dynamic logistic regression with the last ten time instances results in a model with 200 parameters. Problems with estimating parameters at early time instances arise since the number of long customer journeys is small. Therefore, three dynamic logistic regressions are estimated. A model with only the last touchpoints, the two last touchpoints, and the three last touchpoints. The dynamic logistic regressions are estimated without bagging and regularization. The coefficients of the three dynamic logistic regressions are provided in Table 7, 8, and 9 in the Appendix. Two interesting intermediate findings are obtained. Firstly, the coefficients of the touchpoints at the last time instance do not differ substantially in both sign and magnitude between the three dynamic models. This indicates that adding features at earlier time instances do not alter features at later time instances. Secondly, the coefficients for several same touchpoints at different time instances differ. For example, in the dynamic model with the three last touchpoints, email has a positive coefficient at *t-2* and a negative coefficient at *t-1*. Hence, the dynamic logistic regression suggests that the effect of a touchpoint is contingent on when it occurs in the customer journey.

4.1.5 Markov chain

A Markov chain inherently includes temporal dynamics (Norris, 1998). The Markov chain estimates the probability of hopping from one state to another. The transition probabilities of the first-order Markov chain are presented in a transition matrix heatmap, see Graph 4 below. It is a map of followed paths by the customer. The vertical axis represents the previous state and the horizontal axis the current state. The colors correspond to the regions of different probability. A row in the matrix has to sum up to one. Three interesting findings are reported. Firstly, there are four transition probabilities > 0.5. They represent the transition from a search term to a corresponding website in the categories: accommodation, information/comparison, travel agent, and the focal company. Secondly, there are several empty cells meaning that the transition did not occur in the data. Thirdly, the columns with the highest accumulated probability are accommodation websites and travel agent websites. These touchpoints occur with high frequency.

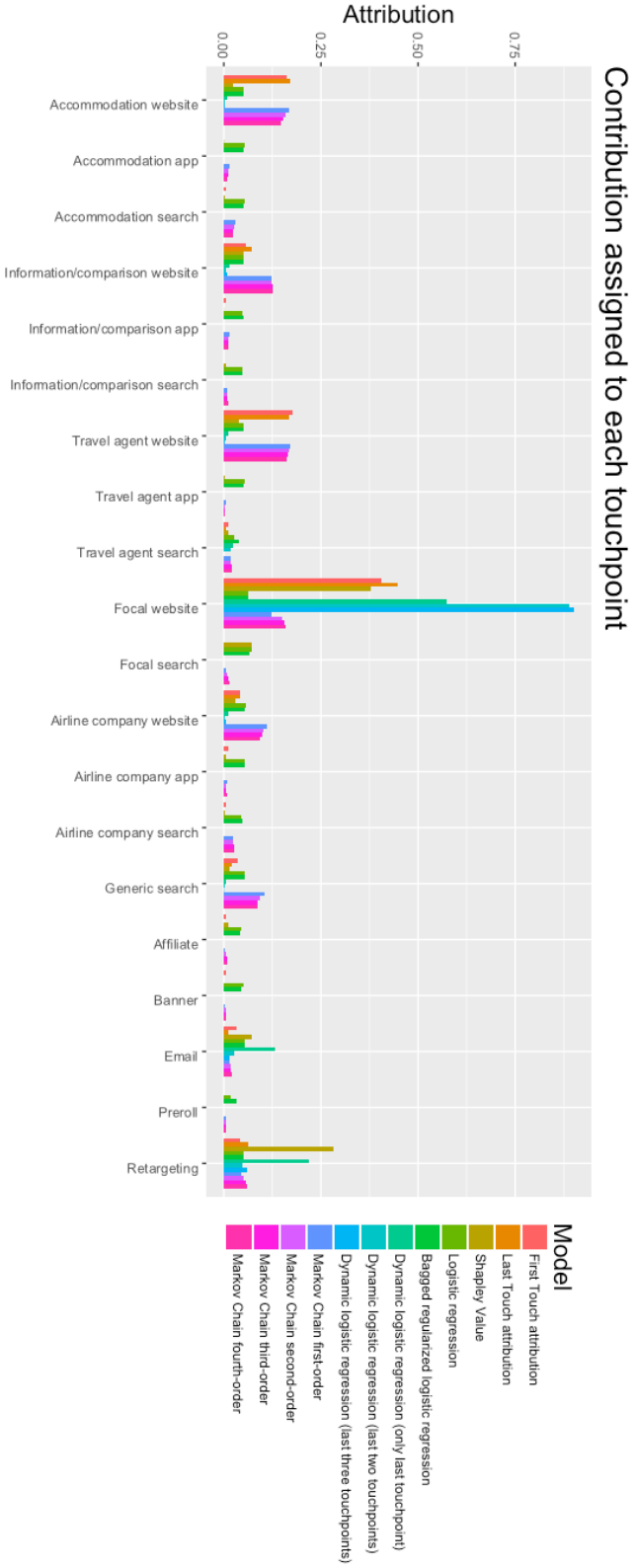


Graph 4. Transition matrix representing a followed paths by the customer. The probabilities are presented in percentage to take up less space.

Based on these transition probabilities, the removal effect is applied to estimate attribution (Anderl et al., 2016). Furthermore, the principle for higher-order Markov chains is the same. However, the state becomes a sequence of touchpoints.

4.2 Model evaluation

The heuristic-based models, the Shapley Value solution, the logistic regression, and the Markov chain are evaluated on the ease of interpretation, predictive accuracy, and robustness. In addition, the contribution assigned to each touchpoint by the estimated models is presented in Graph 5 below.



Graph 5. The contribution assigned to each touchpoint by the estimated models. The data to produce the graph is provided in Table 10 and Table 11 in the Appendix.

4.2.1 Interpretability

To evaluate the interpretability of the models, all steps to compute attribution are examined and the (intermediate) findings are assessed. Attribution presented in Graph 5 above is taken into consideration when evaluating interpretability.

Estimating heuristic-based models only consists of one step and therefore no intermediate output is generated. A pre-defined rule determines attribution. Heuristic-based attribution models disregard customer journeys leading up to non-conversion. A drawback is that when the number of conversions is low, some touchpoints may get zero attribution. The question whether this is fair is debatable. Inspecting the normalized attribution of the last touch attribution model, one can see that eleven of the twenty touchpoints get an attribution score of zero. One of the touchpoints that gets zero attribution is the touchpoint focal search. Moreover, the touchpoints accommodation website and competitive travel agency receive an attribution of 0.1719 and 0.1667, respectively. These two touchpoints occur with a high frequency in the dataset. Hence, it seems to be that touchpoints occurring frequently get assigned a lot of contribution.

The Shapley Value solution consists of two reasonable steps. In the first step, the individual and pair-wise conditional probabilities are computed. The individual conditional probabilities do not show odd results. It assigns a high conditional probability to the focal website, focal search, email, and retargeting. Furthermore, no individual conditional probability is zero. Assessing the pair-wise conditional probabilities, approximately fifteen percent has a value of zero, meaning that they either do not occur in the data or have a true conditional probability of zero. Moreover, the highest conditional probabilities occur at interactions between the focal website and other touchpoints, which seems to be reasonable. In the second step, attribution is determined by adding the contribution of the individual touchpoint and the synergy effects. The touchpoint with the highest normalized attribution is the focal website with a value of 0.4479. The second, third, and fourth highest attribution are assigned to retargeting, email, and focal search, respectively. The other touchpoints receive a normalized attribution below 0.05. Hence, the assigned attribution seems to be plausible.

The logistic regression estimates the log-odds of the probability of conversion from a combination of touchpoints. The optimization procedure used is maximum likelihood. Maximum likelihood attempts to find the coefficients that maximize the likelihood function given the touchpoints. The computation is somewhat more complex, but the intuition to find the best fit is easy. The signs of the coefficients are instantaneously interpretable, yet the magnitude is rather hard to interpret as they are expressed in log-odds of the probability. Interpreting the results of the vanilla logistic regression, one can see that the bias has a large negative value suggesting that the average probability to convert is low. Moreover, a lot of touchpoints have a negative effect, including retargeting with an effect of -0.0421 which is regarded as important by the Shapley Value solution. Hence, the interpretation is difficult due to two reasons. The effects of the touchpoints are expressed in

log-odds of the probability and are relative to the reference level. An approach to compute attribution is through the average marginal effect (Greene, 2012), however, a lot of touchpoints will have a negative attribution when this approach is applied. For example, when the marginal effect for the bagged regularized logistic regression is applied, fourteen of the twenty touchpoints have a negative attribution. Hence in this thesis, all logistic regressions use the logit function to determine attribution (Rentola, 2014). Assessing the normalized attribution of the vanilla logistic regression, the highest attribution assigned to a touchpoint is 0.0717 and the lowest is 0.0187. The distance between the highest and lowest becomes even smaller with the bagged regularized logistic regression. The highest normalized attribution is 0.0648 and the lowest is 0.0331. Hence, the attribution values are close together since the bias has a large negative value. Assessing the normalized attribution of the dynamic logistic regression, twelve touchpoints have an attribution value of approximately zero as they have a large negative coefficient. Moreover, when more temporal dynamics are included, large positive attribution values tend to become larger and small attribution values tend to become smaller. For example, the focal website has an attribution of 0.5731 when only the last time instance is included and 0.902 when the last three time instances are included. Thus, the dynamic logistic regression with the last three time instances assigns a major share to the focal website since all three coefficients of the focal website are large. Hence, the logistic regression has difficulties in interpretation.

In order to compute attribution from the Markov chain, two steps can be distinguished. First, the probabilities of hopping from one state to another are calculated. The intuition to compute the transition probabilities is easy to understand as it is essentially calculating the probability of going to a particular state given a state. The output of the first step can be visually presented in a transition matrix. Assessing the transition matrix heatmap, one can quickly identify frequently and rarely walked paths by the customer. For example, the transition probability of going from a search term to a corresponding website is high. More interesting, the probability to convert given another state can be assessed. For example, the transition probability of going from an email state to a conversion state is the third highest with a probability of 0.0012. In the second step, the removal effect is computed to determine attribution. The removal effect basically computes the decrease in conversion rate if a particular touchpoint is disregarded from the network. This approach seems to be intuitive. Assessing the normalized attribution, attribution of first- and higher-order Markov chains are almost identical. Nevertheless, some normalized attribution results seem to be erroneous. For example, for the first-order Markov chain, attribution of the focal website is 0.1219, attribution for the competitive travel agency website is 0.1721, and attribution for the accommodation website is 0.1671. The latter two attributions are relatively high in comparison with attribution for the focal website. Hence, as the competitive travel agency website and accommodation website have both a substantial share in occurrence, it seems to be that attribution is not only affected by the contribution of the touchpoints but also the frequency.

4.2.2 Predictive accuracy

In order to assess the models on predictive accuracy, two metrics are calculated. Firstly, a ROC graph is produced and the areas under the curve are computed. Secondly, the top-decile lifts are computed. Table 4 shows the area under the curve and the top-decile lift for both the training and validation set. Models that have a higher predictive performance for the training set than the validation set, tend to overfit. In addition, Graph 6 illustrates the ROC curves of the models with the true positive rate on the vertical axis and the false positive rate on the horizontal axis.

The best performing heuristic-based model on both the area under the curve and the top-decile lift is the last touch attribution model. It has an area under the curve for the training set of 0.7085 and for the validation set of 0.6858. Moreover, the top-decile lift of the last touch attribution model is approximately 4.8 for both the training and validation set. Hence, the model does not overfit and is considered as the baseline to compare other models.

The Shapley Value solution has the highest area under the curve on both the training set and validation set with 0.8848 and 0.8839, respectively. This is a considerable improvement of the baseline. Assessing the ROC graph, the Shapley Value solution reaches a high true positive rate, implying that the model is good in predicting the customer journeys ending in conversion. Nevertheless, the false positive rate increases quickly in the beginning relative to the other attribution models, suggesting that the model is less accurate in distinguishing customer journeys ending in non-conversion. Furthermore, the Shapley Value solution has the highest top-decile lift for the validation set.

The vanilla logistic regression has an area under the curve for the training set of 0.8197 and for the validation set of 0.8004, which is a satisfactory improvement over the baseline. Yet, the bagging and regularization procedure improve the predictive ability considerably. The area under the curve for the validation set increases with approximately 0.05 to 0.8501 and the top-decile lift for the validation set increases from 7.0815 to 7.7581. Moreover, the bagged regularized logistic regression outperforms the Shapley Value solution at some combinations of the true positive rate and false positive rate.

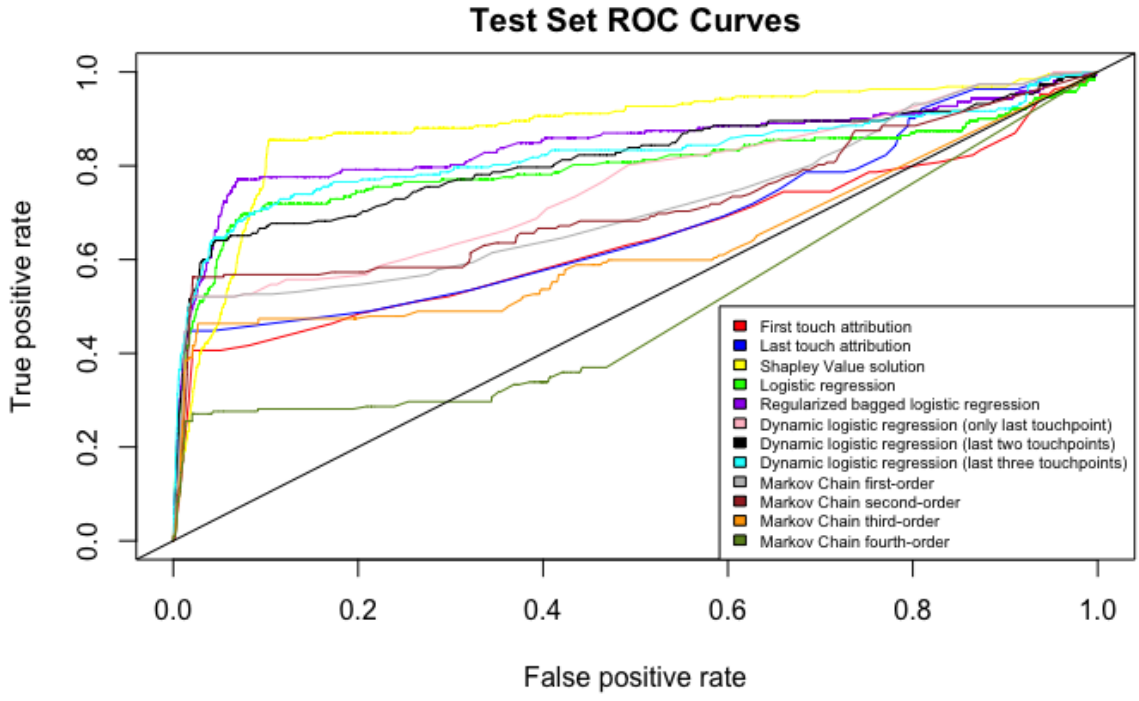
The best performing dynamic logistic regression both on the area under the curve and the top-decile lift is the model with the last three time instances. At validation time, it has an area under the curve of 0.8277 and a top-decile lift of 7.0262. However, comparing the dynamic logistic regression between the training and validation set on both predictive metrics, one can see that the model tends to overfit. This becomes more severe when the model becomes more complex.

The first-order Markov chain has a better predictive ability than higher-order Markov chains on the area under the curve for the validation set. However, the difference with the second-order Markov chain is small. On the contrary, the second-order Markov chain outperforms the first-order Markov chain on the top-decile lift for the validation set with the values 5.7203 and 5.3097, respectively. Furthermore, Higher-order Markov chains tend to overfit immensely. For example, the

fourth-order Markov chain has an area under the curve for the training set of 0.9357 and for the validation set of 0.5046, which is virtually the performance of the random model. Taking a closer look at the ROC curve, the random model outperforms the fourth-order and third-order Markov chain at some combinations of the true positive rate and false positive rate.

	Area under the curve on the training set	Area under the curve on the validation set	Top-decile lift on the training set	Top-decile lift on the validation set
First touch attribution	0.6719	0.6556	4.4038	4.4832
Last touch attribution	0.7085	0.6858	4.8264	4.8359
Shapley Value	0.8848	0.8839	7.5983	7.8659
Vanilla logistic regression	0.8197	0.8004	7.2802	7.0815
Bagged regularized logistic regression	0.8773	0.8501	8.1135	7.7581
Dynamic logistic regression (only last touchpoint)	0.7894	0.7784	5.7465	5.6204
Dynamic logistic regression (last two touchpoints)	0.871	0.8184	7.1989	6.7181
Dynamic logistic regression (last three touchpoints)	0.8984	0.8277	7.8181	7.0262
Markov chain first-order	0.7781	0.7304	5.6596	5.3097
Markov chain second-order	0.8339	0.7241	6.7304	5.7203
Markov chain third-order	0.8971	0.634	7.4649	4.6305
Markov chain fourth-order	0.9357	0.5046	8.2696	2.7588

Table 4. The models evaluated on the area under the curve and top-decile lift on both the training and validation set.



Graph 6. The ROC curves of the models on the validation set.

4.2.3 Robustness

To measure the robustness of the models, two metrics are employed. Firstly, the standard deviations of the metrics used to measure predictive performance are computed. The standard deviations are computed across folds of the cross-validation. Secondly and more important for our present purpose, the average coefficient of variation of attribution is estimated. Table 5 illustrates the standard deviation and the average coefficient of variation of the models. To give a more in-depth view into the coefficient of variation of attribution, Graph 7 is provided below. The graph shows the boxplots of the coefficients of variation of touchpoints per model. Hence, the coefficients of variation in the graph are not averaged over touchpoints to get a single value for each model.

The last touch attribution model has a standard deviation of the area under the curve for the validation set of 0.0855 and a standard deviation of the top-decile lift for the validation set of 0.197. Moreover, it has an average coefficient of variation of attribution of 13.1514. Graph 7 illustrates that the first quartile and median are both zero which decreases the average coefficient of variation. In addition, the graph demonstrates that there are two outliers. The values of the last touch attribution model are taken as the baseline.

The Shapley Value solution has a relatively small standard deviation on both the area under the curve and the top-decile lift for the validation set with values of 0.0506 and 0.5225, respectively. This suggests that the predictive performance is stable. Nevertheless, the average coefficient of variation of attribution is 31.2133, which is much higher than the baseline and other data-driven attribution models. Assessing the graph, one can see that the Shapley Value solution has the largest coefficient of variation of a single touchpoint. More specifically, banner has a coefficient of variation of 206.28. Altogether, the contribution assigned to touchpoints by the Shapley Value solution is sensitive to small departures in the underlying data-generating process.

The vanilla logistic regression has a standard deviation below the baseline and has an average coefficient of variation of attribution of 9.6206. This is an improvement in robustness in comparison with the baseline. Yet the bagged regularized logistic regression has even a smaller standard deviation and has the second smallest average coefficient of variation of all estimated models with a value of 4.8702. Moreover, the graph clearly shows that the dispersion of the coefficients of variation of touchpoints decreases. Hence, the bagging and regularization procedure has a positive effect on the robustness.

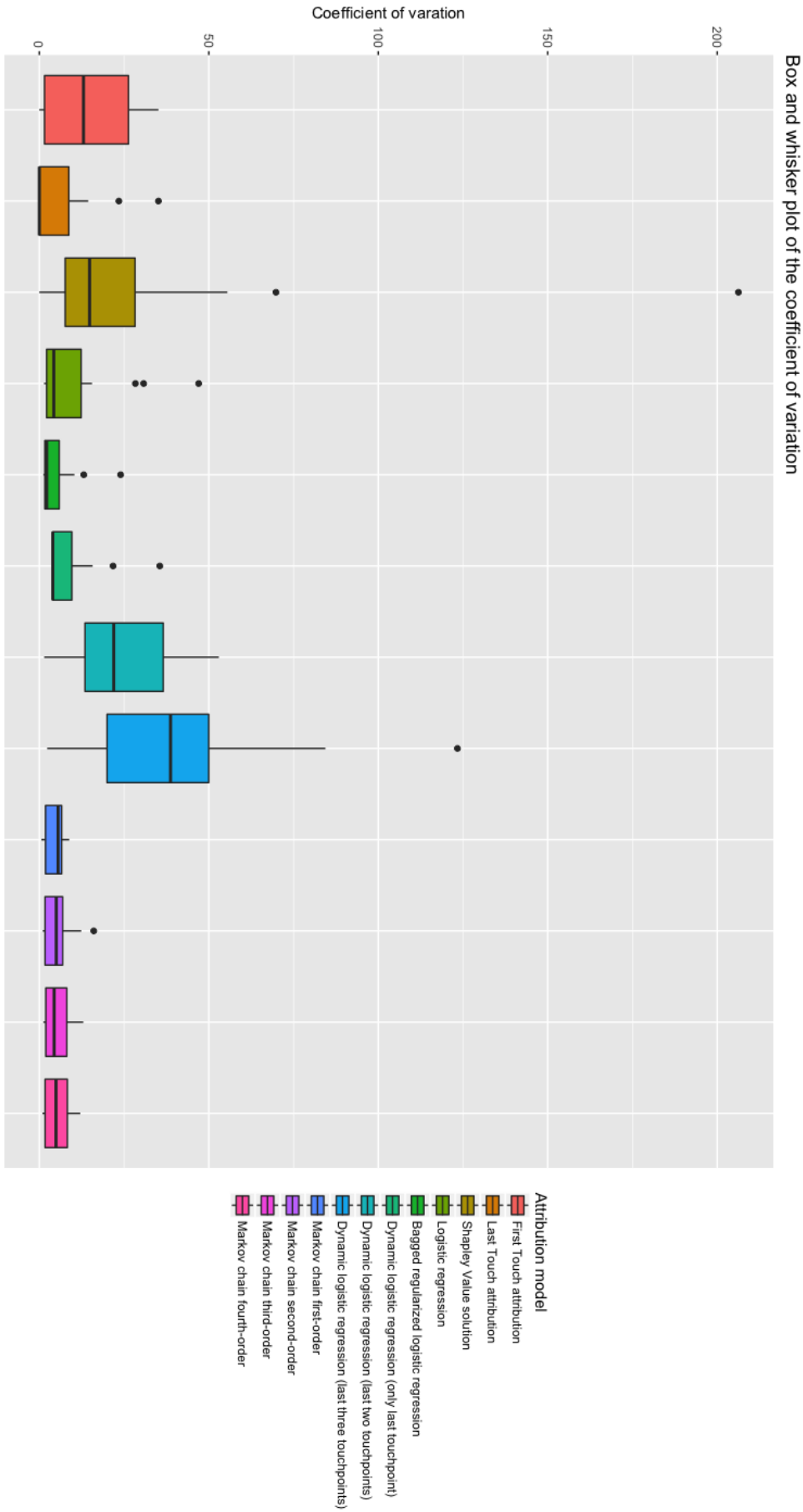
The dynamic logistic regressions have diverse results with regards to the robustness. More specifically, the standard deviation of the top-decile lift for the validation set tends to decrease as more time instances are included in the dynamic logistic regression. On the contrary, the average coefficient of variation of attribution increases sharply as the dynamic logistic regression becomes more complex. For example, the average coefficient of variation of the dynamic logistic regression with the last three time instances has a value of 40.2148, which is the highest average coefficient of variation of all

models. The graph supports this finding. The dispersion of the coefficients of variation of touchpoints increases considerably as the box and the whiskers become larger.

All Markov chains score well on the robustness aspect. The standard deviation of both the area under the curve and the top-decile lift for the validation set are somewhat lower than the baseline. Furthermore, the average coefficient of variation of attribution of all estimated Markov chains is small. The first-order Markov chain has a value of 4.588 which is the smallest average coefficient of variation of all attribution models. This result is supported by the boxplot. All boxes of the coefficient of variation of the Markov chains are small, yet the boxes increase a little when the Markov assumption is relaxed.

	Standard deviation of the area under the curve on the training set	Standard deviation of the area under the curve on the validation set	Standard deviation of the top-decile lift on the training set	Standard deviation of the top-decile lift on the validation set	Coefficient of variation of attribution
First touch attribution	0.0071	0.0625	0.0985	0.8006	20.2376
Last touch attribution	0.0128	0.0855	0.197	1.3042	13.1514
Shapley Value	0.0055	0.0506	0.075	0.5225	31.2133
Vanilla logistic regression	0.0186	0.0619	0.2155	1.0098	9.6206
Bagged regularized logistic regression	0.0123	0.0403	0.1501	0.7165	4.8702
Dynamic logistic regression (only last touchpoint)	0.0071	0.0665	0.1599	1.3992	8.2216
Dynamic logistic regression (last two touchpoints)	0.0049	0.1008	0.0545	1.2483	25.4451
Dynamic logistic regression (last three touchpoints)	0.0053	0.0902	0.0771	1.154	40.2148
Markov chain first-order	0.0107	0.0805	0.1828	1.2402	4.588
Markov chain second-order	0.0067	0.0796	0.165	1.081	5.1614
Markov chain third-order	0.006	0.084	0.0923	1.1891	5.3305
Markov chain fourth-order	0.0044	0.0602	0.1073	0.871	5.3717

Table 5. The standard deviation on both the area under the curve and top-decile lift on both the training and validation set and the coefficient of variation of attribution. Note: the coefficient of variation is multiplied by hundred to express the value in percentages.



Graph 7. Boxplot of the coefficients of variation of the touchpoints per attribution model.

5. Discussion

In the results section, the heuristic-based models, the Shapley Value solution, the logistic regressions, and the Markov chains are estimated and intermediate results are presented. Moreover, the models are evaluated on the ease of interpretation, predictive accuracy, and robustness. The findings are discussed in section 5.1 and 5.2. Discussing the findings directly addresses the research question by comparing the models on the evaluation criteria. Furthermore, section 5.3 provides the limitations, emphasize the importance of this thesis and gives suggestions for future research.

5.1 Discussing intermediate results

The Shapley Value solution, the logistic regression, and the Markov chain produce different intermediate outputs from which different types of information can be obtained.

The Shapley Value solution generates three relevant intermediate outcomes (Shao & Li, 2011). Firstly, the conditional probabilities of conversion given the individual touchpoints indicate which touchpoints are important to increase the likelihood that a customer journey ends in conversion. Secondly, the conditional probabilities of conversion given the interaction between two touchpoints demonstrate which combinations of touchpoints are effective to increase the probability to convert. Lastly, the synergy effects between possible combinations of touchpoints show the added value due to the interactions between touchpoints. For example, in this dataset, there is a high synergy effect when a customer searches for the focal website and receives an email. A possible rationale is that customers are triggered by receiving an email, go on the web and search for the focal company and eventually convert. Another explanation is that the customer is already moderately interested and has searched for the focal website, yet get enticed by the email and convert. As this example demonstrates, the temporal order in the proceedings cannot be acquired from the Shapley Value solution, nevertheless, prominent touchpoints or combinations between touchpoints can get obtained. Hence, marketers and managers can decide which combination of touchpoints should be part of the marketing media mix.

The logistic regression produces coefficients instead of probabilities, meaning that it explores relations between touchpoints and the likelihood to convert (Greene, 2012). The signs of the coefficients are directly interpretable. For example, it can be seen that when a customer enters a generic search term, it has a significant positive effect on the probability to convert. However, the magnitude of the touchpoints is difficult to interpret as the values are expressed as the log-odds of the probability (Leeper, 2018). Manipulating the input data of the logistic regression by making binary features of the latest touchpoints of the customer journey results in a dynamic logistic regression. It is unfeasible to include features for all time instances. As the results indicate, the coefficients at the last time instance do not differ substantially in both sign and magnitude between the three estimated dynamic logistic regressions. This suggests that the effects of the coefficients are consistent, regardless of the number of time instances incorporated in the model. However, the coefficients for several same touchpoints at different time instances vary. In other words, the effect of a touchpoint on the

propensity to purchase depends on when it occurs in the customer journey, which is in line with findings by Lemon and Verhoef (2016). In addition, the results show that the bagging and regularization procedure enhances the performance of the model. More specifically, it improves the predictive ability and robustness on all measured facets and it prevents the model from overfitting.

In order to estimate attribution from the Markov chain, the probabilities of going from one touchpoint to the next are computed (Norris, 1998). Based on the probabilities, the removal effect is applied to calculate attribution (Anderl et al., 2016). The transition probabilities are also interesting as such. Generally, a transition matrix is generated to inspect the probability of hopping from one state to another. Based on the transition matrix, one can map often and rarely followed paths by the customer. By including domain and firm knowledge, the transition matrix can assist marketers and managers to make decisions. In addition, by combining the findings of the Shapley Value solution with the Markov chain, one can find which combinations of touchpoints are effective and in what sequence the customer most likely encounters touchpoints. This enables marketers and managers to affect customer behavior in real-time by showing the customer effective ads based on previously encountered touchpoints. Hence, combining attribution models can provide additional insights into understanding and influencing the customer journey. In other words, it is beneficial to compute multiple attribution models as they provide supplementary intermediate results.

5.2 Discussing results

The findings of the interpretability, predictive accuracy, and robustness section are discussed below.

5.2.1 Interpretability

The heuristic-based attribution models are based on pre-defined rules which makes them easy to understand. Yet after inspecting the results, the findings seem to be erroneous. Despite interpreting is rather subjective, an attribution of zero for searching the focal company on the web, a high attribution to the accommodation website, and a high attribution to the competitive travel agency are doubtful. The pre-defined rules have ingrained biases that make the model inherently flawed. For example, last- and first touch attribution models tend to give no attribution to touchpoints in the middle of a customer journey such as searching for the focal website on the web. Moreover, frequently encountered touchpoints seem to drive attribution rather than true contribution. The method to compute attribution is easy to grasp however the statistical procedure may not be valid. Hence, by following the definition of interpretability, the models do not score high on the interpretability aspect since the model is not grounded on the basis of its statistical merit.

The Shapley Value solution is simple to understand as the first step to compute conditional probabilities and the second step to compute attribution make intuitive sense. The results of both the individual conditional probabilities and pair-wise conditional probabilities do not show peculiar results. Comparing the Shapley Value solution with the last touch attribution model, attribution of the

Shapley Value solution seems to be more plausible. For example, focal search gets, unlike the heuristic based models, attribution assigned. Furthermore, the Shapley Value solution is not affected by touchpoints that do occur frequently. Thus, the model is easy to interpret as the components of the Shapley Value solution are easy to grasp and the statistical procedure is valid.

The logistic regression is relatively difficult to interpret as of five reasons. Firstly, the optimization procedure is more difficult to understand mathematically however the intuition is quite easy. Secondly, the effects of the touchpoints are expressed in log-odds of the probability. Hence, the sign is instantaneously interpretable, yet the magnitude not. Thirdly, the coefficients of the logistic regression are relative to reference level which make them more difficult to interpret. Fourthly, it is ambiguous how to determine attribution from the logistic regression as the model does not aim to reflect the contribution of a touchpoint. Lastly, when the logit function is used to determine attribution, the results seem to be erroneous. For example, with the vanilla logistic regression and bagged regularized logistic regression, attribution values are all close together. Including useless touchpoints seems to distort attribution. Another example with dynamic logistic regressions, when more time instances are incorporated, attribution values becomes more extreme. Extrapolating this finding, when including the last ten time instances, virtually all attribution is assigned to the focal website. Hence, it is not easy to interpret attribution from the logistic regression as the model does not have a clear and intuitive comprehension of how attribution is determined.

The first step of the Markov chain to determine the transition probabilities is clear and understandable as the computation is easy to grasp. The findings of the transition probabilities are plausible. Furthermore, the transition matrix heatmap provides a visual overview of the paths followed by the customers which makes it easier to interpret. In the second step, attribution is determined by the decrease in conversion rate if a particular touchpoint is disregarded from the network. This seems to be reasonable. However, after inspecting the results, it appears to be that attribution is not only driven by the touchpoints' contribution to convert, yet also by the frequency of occurrence of touchpoints. Yet, this is in line with the definition of the removal effect as the entire customer journey does not lead to conversion if one touchpoint is removed. For example, when almost every customer that convert is exposed to a particular touchpoint, the attribution of that touchpoint is high, even if the touchpoint does not truly contribute anything. Hence, the model scores moderate on the ease of interpretation as the steps are easy to comprehend, however, attribution seems to be affected by the frequency of occurrence of touchpoints.

In summary, the Shapley Value solution scores highest on the interpretability aspect. The Markov chain second highest as it seems to be affected by the frequency of occurrence of touchpoints. The Heuristic-based models third highest as the rules are inherently flawed and attribution is driven by frequently encountered touchpoints. The logistic regression ends last as it has multiple difficulties in interpretation.

5.2.2 Predictive accuracy

Consistent with past research, the heuristic-based model with the highest predictive performance is the last touch attribution model (Wooff & Anderson, 2013). The last touch attribution model has an area under the curve of 0.6858 and a top-decile lift of 4.8264 for the validation set. These performance scores amply suppress the random model.

Despite the Shapley Value solution uses the most basic feature construction scheme, it has the highest predictive ability on all metrics. It is a considerable improvement of the last touch attribution model with a value of 0.8839 on the area under the curve and a value of 7.8659 on the top-decile lift for the validation set. Hence, this implies that simpler constructed models with less information can outperform overly complex models that make use of all information.

The dynamic logistic regression has a better predictive performance than the regular logistic regression. From the dynamic logistic regressions, the best performing model is the most complex one with the three latest time instances. However, as obtained from the data, more complex models tend to overfit. In addition, both a bagging and regularization procedure is applied to the regular logistic regression, which substantially improved the predictive ability. More specifically, the bagged regularized logistic regression outperforms the Shapley Value solution at some regions of the ROC graph. Meaning that depending on the desired threshold cutoff between the true positive rate and false positive rate, the model may be superior.

The Markov chain is less well in predicting conversion in comparison with other data-driven models. However, the first-order and second-order Markov chain still improve the heuristic-based models. The first-order Markov chain performs better on the area under the curve for the validation set and the second-order Markov chain performs better on the top-decile lift for the validation set, which makes the model comparable with regard to predictive accuracy. Yet, higher-order Markov chains tend to overfit drastically. The fourth-order Markov chain has a predictive ability just above the random model. A rationale for overfitting is that the sequence of touchpoints (i.e. state) occurs rarely and is therefore not generalizable to new unseen examples.

In summary, the order of models with the best predictive accuracy from high to low is as follows: the Shapley Value solution, the logistic regression, the Markov chain, and the heuristic-based models.

5.2.3 Robustness

The robustness of a model is measured on two aspects. The robustness of the predictive accuracy and the robustness of assigning contribution to the touchpoints. The latter is more important as the goal is to evaluate attribution models. Heuristic-based models have an average robustness in comparison with the data-driven models. To give an indication, the last touch attribution model has an average coefficient of variation of attribution of 13.1514.

The Shapley Value solution has a relatively small robustness score on the predictive ability measures. Nonetheless, the average coefficient of variation of attribution is 31.2133, which is the second highest robustness score of all estimated models. Hence, the predictive performance is stable, however, the attribution is unstable to small deviations in the underlying data-generating process. An explanation for the high robustness score of attribution is that in the second step of the model only the customer journeys leading up to conversion are taken into consideration. Hence, when the data is highly skewed, it contains a few customer journeys leading up to conversion which, in turn, leads to volatile results.

The regular logistic regression has a better robustness score than the heuristic-based models both on the predictive ability and attribution. The dynamic logistic regression with the last two or last three time instances has a substantially worse robustness score of attribution. More specifically, the dynamic logistic regression with the last three time instances has the highest average coefficient of variation of attribution. Increasing the complexity of the dynamic logistic regression leads to unstable attribution. In addition, the bagging and regularization procedure has a positive effect on the robustness. The bagged regularized logistic regression has the second smallest average coefficient of variation of all estimated models.

The Markov chain is the best performing model on the robustness aspect. The robustness score of the predictive ability is lower than the heuristic-based models. More importantly, the average coefficient of variation of attribution of all estimated Markov chains is low and the differences between lower- and higher-order models are small. Yet, the Markov chain with the lowest average coefficient of variation of attribution is the first-order model. An explanation for the high robustness score of the Markov chain is that the model computes the transition probabilities on the entire dataset by exploiting the temporal sequence.

In summary, the order of models regarding their robustness from high to low is as follows: the Markov chain, the logistic regression, the heuristic-based models, and the Shapley Value Solution.

5.2.4 Overall results

In conclusion, the results indicate that none of the attribution models are superior on all three aspects. The Shapley Value solution has the highest predictive accuracy and has a good interpretability but is not robust. The logistic regression has a good predictive ability and robustness when the bagging and regularization procedure are applied, yet does not score high on interpretability as the model does not aim to reflect the contribution of a touchpoint. The Markov chain is robust and moderately interpretable, but the model does not score well on predicting conversion. Nevertheless, all data-driven models are better than the heuristic-based models since they outperform heuristic-based models on at least two of the three aspects. See Figure 2 below for a graphical summary of the data-driven attribution models. In addition, the data-driven models produce different intermediate results from

which different types of information can be obtained. The different types of information are interesting as such, yet combining these results provides additional insights into understanding and influencing the customer journey.

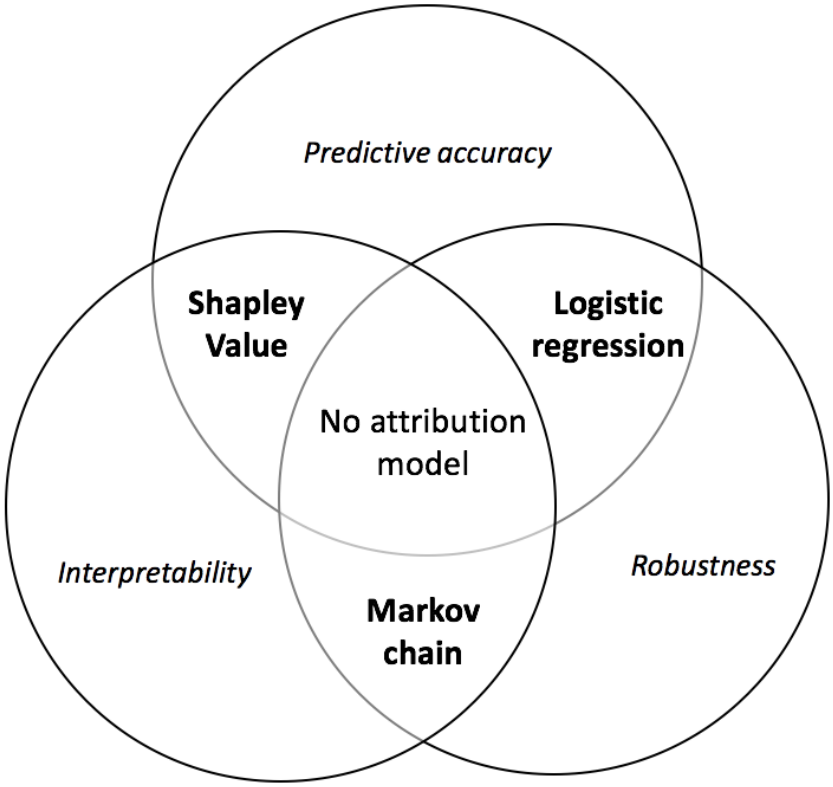


Figure 2. Venn diagram of the data-driven attribution models.

5.3 Limitations, importance, and future research

This thesis has several limitations. First of all, the results are solely based on one dataset. The properties, size, and domain of the data may affect the findings. A lot of customer journeys in this dataset are short as a high number of customer journeys only consists of one or two touchpoints. More sophisticated data-driven models are not required since in the case of short customer journeys, heuristic-based models deliver quite satisfactory performances. As the results demonstrate, the predictive accuracy of the last touch attribution model appears to be reasonable. Nevertheless, when longer and more complex customer journeys are present, more sophisticated data-driven models are indispensable. Moreover, the enterprise-initiated touchpoints in the dataset are impressions meaning that the customer did not necessarily clicked on it. Data of clicks may be more informative and lead to different outcomes. Furthermore, the analyses are conducted on a travel agency dataset. Despite online advertisement plays a pivotal role and customers spend generally lots of time researching vacations or trips (Pabel & Prideaux, 2016; Park & Oh, 2012), the results may not be generalizable to other industries.

Even so, the attribution models themselves have their limitations. To make more fine-grained decisions of the contribution of touchpoints, one could consider incorporating more information as the

revenues from the conversion or the cost of the touchpoints. Berman (2017) proposed a model that incorporates the revenue generated by the customer. Another issue of devoting credit to touchpoints is that attribution models are endogenic. Put differently, alternative explanations exist for the relationship between the touchpoints and conversion. For example, retargeting customers who already interacted with the enterprise have inherently a higher propensity to purchase as they have already shown interest. To establish a causal inference, elimination of extraneous variables is required. Dalessandro et al. (2012) proposed a causal attribution model, however, they also acknowledged the impracticality of estimating a fully causal model.

Notwithstanding its limitations, this study fills a gap in the literature by comparing and evaluating the heuristic-based attribution models, the Shapley Value solution, the logistic regression, and the Markov chain. Changes should be made gradually and going from heuristic-based models to data-driven models is a considerable improvement in the attempt to capture the genuine attribution. It is often unattainable and impractical to make radical changes (Burke, 2017). The primary contribution of this thesis is that enterprises can decide which attribution model fits their needs the best. In other words, it assists enterprises to choose the adequate attribution model. All the evaluation criteria are important, but none of the attribution models is superior. Hence, a direct implication is that enterprises should make a trade-off. Some enterprises may be more concerned with the predictive ability and other may be more risk averse and want stable results. Yet, creating transparency by evaluating the models encourage enterprises to abandon heuristic-based models and adopt data-driven models. Another implication is that it is valuable to estimate multiple data-driven attribution models since the intermediate outcomes generate different types of information. These different types of information are interesting as such, however, combining the outcomes provides supplementary insights into understanding and influencing the customer journey. On the basis of the evaluation of the existing data-driven attribution models, future research could take several directions. More specifically, one could propose a new attribution model or a modification to an existing attribution model that outperforms the evaluated attribution models on one specific evaluation criterion. Alternatively, one could attempt to develop a multifaceted novel attribution model that is easy to interpret, has a high predictive accuracy, and is robust.

6. Conclusion

The aim of this thesis is to evaluate attribution models that are often used in practice to capture the true conversion attribution. Better attribution models lead to fewer costs and higher profits which eventually results in a higher economic welfare. This leads to the following research question: to what extent are the heuristic-based attribution model, Shapley Value solution, the logistic regression, and the Markov chain easy to interpret, robust, and accurate? The results show that none of the attribution models outperforms the others on all three aspects. The Shapley Value solution has the highest predictive accuracy and has a good interpretability but is not robust. The logistic regression has a good predictive ability and robustness when the bagging and regularization procedure are applied, yet does not score high on interpretability as the model does not aim to reflect the contribution of a touchpoint. The Markov chain is robust and moderately interpretable, but the model does not score well on predicting conversion. In addition, the data-driven models produce different intermediate results from which different types of information can be obtained. The different types of information are interesting as such, yet combining these results provides additional insights into understanding and influencing the customer journey. However, the generalizability is limited since the analyses are conducted on one single travel agency dataset. The properties, size, and domain of the data may affect the findings. Notwithstanding its limitations, enterprises can decide which attribution model fits their needs the best. All the evaluation criteria are important, but none of the attribution models is superior. Enterprises should make a trade-off between which aspects they regard as most important. Yet, going from heuristic-based models to data-driven models is a considerable improvement in the attempt to capture the genuine attribution.

References

Abhishek, V., Fader, P., & Hosanagar, K. (2015), "Media Exposure through the Funnel: A Model of Multi-stage Attribution," Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2158421

Altomare, D., & Loris., D, (2018). ChannelAttribution: Markov Model for the Online Multi-Channel Attribution Problem. *R package version 1.12.*

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

Aras, M., Syam, H., Jasruddin, J., Akib, H., & Haris, H. (2017). The Effect of Service Marketing Mix on Consumer Decision Making. In *International Conference on Education, Science, Art and Technology* (pp. 108-112).

Berman, R. (2017), "Beyond the Last Touch: Attribution in Online Advertising". Available at SSRN: http://ssrn.com/abstract=2384211

Błaszczyński, J., & Stefanowski, J. (2015). Neighbourhood sampling in bagging for imbalanced data. *Neurocomputing*, *150*, 529-542.

Bleier, A., De Keyser, A., & Verleye, K. (2018). Customer engagement through personalization and customization. In *Customer Engagement Marketing* (pp. 75-94). Palgrave Macmillan, Cham.

Box, G. E. (1979). Robustness in the strategy of scientific model building. *Robustness in statistics* (pp. 201-236).

Burke, W. W. (2017). *Organization change: Theory and practice*. Sage Publications.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1721-1730). ACM.

Chatterjee, P., Hoffman, D. L., & Novak, T. P. (2003). Modeling the clickstream: Implications for web-based advertising efforts. *Marketing Science*, *22*(4), 520-541.

Clifton, B. (2012). *Advanced web metrics with Google Analytics*. Indianapolis, IN: John Wiley & Sons.

Constantinides, E. (2002). The 4S web-marketing mix model. *Electronic commerce research and applications*, *1*(1), 57-76.

Dalessandro, B., Perlich, C., Stitelman, O., & Provost, F. (2012, August). Causally motivated attribution for online advertising. In *Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy* (p. 7). ACM.

Disatnik, D., & Sivan, L. (2016). The multicollinearity illusion in moderated regression analysis. *Marketing Letters*, *27*(2), 403-408.

Doshi-Velez, F., & Kim. B. (2017) Towards A Rigorous Science of Interpretable Machine Learning. *arXiv preprint arXiv:1702.08608.*

Englehardt, S., Reisman, D., Eubank, C., Zimmerman, P., Mayer, J., Narayanan, A., & Felten, E. W. (2015). Cookies that give you away: The surveillance implications of web tracking. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 289-299). International World Wide Web Conferences Steering Committee.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1), 1-22.

Gamberger, D., & Lavrač, N. (1997). Conditions for Occam's razor applicability and noise elimination. In *European Conference on Machine Learning* (pp. 108-123). Springer, Berlin, Heidelberg.

Goldfarb, A. (2014). What is different about online advertising?. *Review of Industrial Organization*, *44*(2), 115-129.

Greene, W. H. (2012). Econometric analysis, 71e. *Stern School of Business, New York University*.

Herhausen, D., Kleinlercher, K., Emrich, O., Verhoef, P., & Rudolph, T. (2017). Customer Journeys and their Effects on Customer Satisfaction and Loyalty. In *Proceedings of the European Marketing Association 46th Annual Conference*.

Hoornaert, S., Ballings, M., & Poel, D. (2015). Lift: Compute the Top Decile Lift and Plot the Lift Curve. *R package version 0.0.2.*

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Hudson, S., Roth, M. S., Madden, T. J., & Hudson, R. (2015). The effects of social media on emotions, brand relationship quality, and word of mouth: An empirical study of music festival attendees. *Tourism Management*, *47*, 68-76.

Hülsdau, M., & Teuteberg, F. (2018, March). Towards a taxonomy of algorithmic attribution models– Which is the right model to measure, manage and optimize multiple campaigns? Paper presented at *MKWI 2018*. Retrieved from http://mkwi2018.leuphana.de/programm/sessions/

Keilson, J. (2012). *Markov chain models—rarity and exponentiality* (Vol. 28). Springer Science & Business Media.

Kotler, P., & Armstrong, G. (2010). *Principles of marketing.* Pearson education.

Kuncheva, L. I., Arnaiz-González, Á., Díez-Pastor, J. F., & Gunn, I. A. (2018). Instance Selection Improves Geometric Mean Accuracy: A Study on Imbalanced Data Classification. *arXiv preprint arXiv:1804.07155*.

Leeper, J. (2018). margins: Marginal Effects for Model Objects. *R package version 0.3.20.*

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. K. (2014). Attributing conversions in a multichannel online marketing environment: An empirical model and a field experiment. *Journal of Marketing Research*, *51*(1), 40-56.

Lipton, Z. C. (2016). The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*.

Louppe, G., & Geurts, P. (2012). Ensembles on random patches. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 346-361). Springer, Berlin, Heidelberg.

Lodish, L. (2001). Building marketing models that make money. *Interfaces: Special Issues on Marketing Engineering.* 31(3), Part 2 S45-S55.

MartíNez-MartíNez, J. M., Escandell-Montero, P., Soria-Olivas, E., MartíN-Guerrero, J. D., Magdalena-Benedito, R., & GóMez-Sanchis, J. (2011). Regularized extreme learning machine for regression problems. *Neurocomputing*, *74*(17), 3716-3721.

Mayer, M. (1991). *Whatever happened to Madison Avenue?: Advertising in the'90s*. Little Brown.

McAuliffe, R. E. (2015). Coefficient of variation. *Wiley Encyclopedia of Management*.

Meyer, C., & Schwager, A. (2007). Customer experience. *Harvard business review*, *85*(2), 116-126.

Mitra, A., & Lynch, J. G. (1996). Advertising effects on consumer welfare: prices paid and liking for brands selected. *Marketing Letters*, *7*(1), 19-29.

Molnar, C. (2018). Interpretable machine learning. A guide for making black box models explainable. Retrieved from https://christophm.github.io/interpretable-ml-book/

MSI Research Priorities 2016-2018. (n.d.). *Marketing Science Institute*. Retrieved from https://www.msi.org/uploads/articles/MSI_RP16-18.pdf

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of marketing research*, *43*(2), 204-211.

Netzer, O., Ebbes, P., & Bijmolt, T. H. (2017). Hidden Markov Models in Marketing. In *Advanced Methods for Modeling Markets* (pp. 405-449). Springer, Cham.

Nisar, T., & Yeung, M. (2017). Attribution modelling in digital advertising: An empirical investigation of the impact of digital sales channels. *Journal of Advertising Research*, *57*(4).

Norris, J. R. (1998). *Markov chains* (No. 2). Cambridge university press.

Pabel, A., & Prideaux, B. (2016). Social media use in pre-trip planning by tourists visiting a small regional leisure destination. *Journal of Vacation Marketing*, *22*(4), 335-348.

Pansari, A., & Kumar, V. (2017). Customer engagement: the construct, antecedents, and consequences. *Journal of the Academy of Marketing Science*, *45*(3), 294-311

Park, J., & Oh, I. K. (2012). A case study of social media marketing by travel agency: The salience of social media marketing in the tourism industry. *International Journal of Tourism Sciences*, *12*(1), 93-106.

R Core Team (2017). R: A language and environment for statistical computing. *R Foundation for*

Statistical Computing, Vienna, Austria.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77(2)*, 257-286.

Rentola, O. (2014), Analyses of Online Advertising Performance Using Attribution Modeling, *MSc. Thesis, Aalto University, Helsinki.*

Refaeilzadeh, P., Tang, L., & Liu, H. (2016). *Cross-validation. Encyclopedia of database systems*, 1-7.

Sequent Partners (2018). A Comparison of Leading Providers of Media Performance Analyses. Retrieved from http://cimm-us.org/wp-content/uploads/2012/07/CIMM-ROI-Attribution-Providers-Guide_February-2018.pdf

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

Shapley, L. S. (1988). A value for n-person games. *The Shapley value*, 31-40.

Silverbauer, J. (2017). Tracking website data with Google Tag Manager. *Journal of Brand Strategy*, 6(3), 242-249.

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). "ROCR: visualizing classifier performance in R." *Bioinformatics*, 21(20), pp. 7881.

Speekenbrink, M., & Visser, I. (2013). Types and states: Mixture and hidden Markov modelling for the cognitive sciences. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 35, No. 35).

Strong, E. K. (1925). *The psychology of selling and advertising*. McGraw-Hill book Company, Incorporated.

Ur, B., Leon, P. G., Cranor, L. F., Shay, R., & Wang, Y. (2012). Smart, useful, scary, creepy: perceptions of online behavioral advertising. In *proceedings of the eighth symposium on usable privacy and security* (p. 4). ACM.

Vellido, A., Martín-Guerrero, J. D., & Lisboa, P. J. (2012). Making machine learning models interpretable. In *ESANN* (Vol. 12, pp. 163-172).

Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological methods*, 17(2), 228.

We are social (2018). Digital in 2018: World's internet users pass the 4 billion mark. Retrieved from https://wearesocial.com/blog/2018/01/global-digital-report-2018

Windsor.AI (n.d.), *Optimise marketing ROI with multi-touchpoint attribution modelling.* Retrieved from https://www.windsor.ai/attribution-modelling/

Wooff, D., & Anderson, J. (2013) Time-weighted attribution of revenue to multiple e-commerce marketing channels in the customer journey., Working Paper. *Durham Research Online*.

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.

Yardeni, E. (1996). Economic consequences of the internet. *Deutsche Morgan Grenfell, Topical.*

Young, T. (2017). PRIMER: General Data Protection Regulation. *International Financial Law Review*.

Zhang, Y., Wei, Y., & Ren, J. (2014, December). Multi-touch attribution in online advertising with survival theory. In *Data Mining (ICDM), 2014 IEEE International Conference on* (pp. 687-696). IEEE.

Zhu, B., Baesens, B., Backiel, A. E., & van den Broucke, S. K. (2018). Benchmarking sampling techniques for imbalance learning in churn prediction. *Journal of the Operational Research Society*, *69*(1), 49-65.

Appendix

Touchpoints	Probability in percentage
Accommodation website	0.69
Accommodation app	0.70
Accommodation search	0.67
Information/comparison website	1.13
Information/comparison app	0.58
Information/comparison search	1.36
Travel agent website	0.86
Travel agent app	1.2
Travel agent search	1.6
Focal website	5.28
Focal search	7.45
Airline company website	1.07
Airline company app	1.45
Airline company search	0.94
Generic search	0.81
Affiliates	2.41
Banner	0.82
Email	6.91
Pre-rolls	0.35
Retargeting	8.43

Table 1. A table representing the conditional probability of conversion given an individual touchpoint. Note 1: the probabilities are expressed in percentage. Note 2: the highlighted values have a probability higher than 0.05.

0.67																			
0.98	0.72																		
5.26	2.74	2.49																	
2.75	1.52	2.07	12.5																
1.61	1.13	1.23	2.47	1.67															
1.48	1.1	1.22	1.61	2.02	1.03														
0	0	0.6	0	0	2.5	0													
0	0	6.7	7.69	0	0	6.21	0												
4	8.33	8.29	11.76	5.88	4.44	9.77	0	22.22											
3.88	3.17	4.82	6	5.13	3.78	5.18	2.25	13.48	7.92										
0.93	0.88	0.97	4.83	2.3	1.21	1.49	0.68	8.57	7.94	4.45									
1.09	0.68	0.9	6.25	2.4	1.38	0.76	0	0	4.35	2.74	0.89								
1.59	2.14	1.55	6.67	2.33	1.98	1.96	0	14.29	8.57	4.02	1.67	1.27							
1.08	1.08	1.28	2.39	2.96	1.41	1.52	0.87	8.72	9.55	5.37	1.33	1.07	1.41						
0	0	0.39	0	0	0	0.63	0	0	0	0.92	0	0	0	0.6					
7.69	5.93	7.57	6.58	13.33	5.51	8.33	5.56	18.52	13.95	8.98	7.4	8.57	8.97	8.8	2.63				
0	2.94	1.69	0	2	2.17	1.15	0	0	14.29	3.92	2.13	1.32	3.85	2.04	0	9.09			
1.15	1.38	1.73	2.99	1.75	2.15	2.25	0	15.79	7.56	4.38	1.62	1.06	3.11	2	0	7.69	2.17		
1.15	0.91	1	2.2	1.77	1.2	1.32	0.65	7.32	8.26	4.93	1.12	0.95	1.59	1.48	0.42	8.46	1.61	1.67	

Table 2. A matrix representing conditional probability of conversion given the interaction between two touchpoints. The matrix will become too large when the names are presented on the axis. The touchpoints are sorted in alphabetical order. The order is: accommodation app, accommodation search, accommodation website, affiliates, airline company app, airline company search, airline company website, banner, email, focal search, focal website, generic search, information/comparison app, information/comparison search, information/comparison website, pre-rolls, retargeting, travel agent app, travel agent search, and travel agent website. Note 1: the probabilities are expressed in percentage to take up less space. Note 2: only the left-hand side of the diagonal of the matrix is filled since the right-hand side of the diagonal is the mirror image. Note 3: the highlighted values have a probability higher than 10 in percentage.

-0.7																			
-0.41	-0.64																		
2.15	-0.34	-0.61																	
0.6	-0.6	-0.07	8.64																
-0.03	-0.48	-0.4	-0.88	-0.72															
-0.29	-0.64	-0.54	-1.87	-0.5	-0.98														
-1.52	-1.49	-0.91	-3.23	-2.27	0.74	-1.89													
-7.61	-7.58	-0.9	-1.63	-8.36	-7.85	-1.77	-7.73												
-4.15	0.21	0.15	1.9	-3.02	-3.95	1.25	-8.27	7.86											
-2.1	-2.78	-1.15	-1.69	-1.6	-2.44	-1.17	-3.85	1.29	-4.81										
-0.58	-0.6	-0.53	1.61	0.04	-0.54	-0.39	-0.95	0.85	-0.32	-1.64									
-0.19	-0.57	-0.37	3.26	0.37	-0.14	-0.89	-1.4	-7.49	-3.68	-3.12	-0.5								
-0.47	0.11	-0.5	2.9	-0.48	-0.32	-0.47	-2.18	6.02	-0.24	-2.62	-0.5	-0.67							
-0.75	-0.72	-0.54	-1.15	0.38	-0.66	-0.68	-1.08	0.68	0.97	-1.04	-0.61	-0.64	-1.08						
-1.05	-1.02	-0.65	-2.76	-1.8	-1.29	-0.79	-1.17	-7.26	-7.8	-4.71	-1.16	-0.93	-1.71	-0.88					
-1.44	-3.17	-1.55	-4.26	3.45	-3.86	-1.17	-3.69	3.18	-1.93	-4.73	-1.84	-0.44	-0.82	-0.76	-6.15				
-1.9	1.07	-0.2	-3.61	-0.65	0.03	-1.12	-2.02	-8.11	5.64	-2.56	0.12	-0.46	1.29	-0.29	-1.55	-0.54			
-1.15	-0.89	-0.56	-1.02	-1.3	-0.39	-0.42	-2.42	7.28	-1.49	-2.5	-0.79	-1.12	0.15	-0.73	-1.95	-2.34	-0.63		
-0.41	-0.62	-0.55	-1.07	-0.54	-0.6	-0.61	-1.03	-0.45	-0.05	-1.21	-0.55	-0.49	-0.63	-0.51	-0.79	-0.83	-0.45	-0.79	

Table 3. A matrix representing the synergy effects between touchpoints. The matrix will become too large when the names are presented on the axis. The touchpoints are sorted in alphabetical order. The order is: accommodation app, accommodation search, accommodation website, affiliates, airline company app, airline company search, airline company website, banner, email, focal search, focal website, generic search, information/comparison app, information/comparison search, information/comparison website, pre-rolls, retargeting, travel agent app, travel agent search, and travel agent website. Note 1: the values are multiplied by 100 percent to take up less space. Note 2: only the left-hand side of the diagonal of the matrix is filled since the right-hand side of the diagonal is the mirror image. Note 3: the highlighted values have an absolute synergy effect higher than 5.

	Estimate	Std. Error	z value	Pr(>|z|)
Bias	-5.1263	0.0825	-62.1410	<0.0001***
Accommodation website	-0.0115	0.0171	-0.6730	0.5010
Accommodation app	0.0166	0.0815	0.2040	0.8384
Accommodation search	0.0251	0.0917	0.2740	0.7842
Information/comparison website	-0.0215	0.0160	-1.3420	0.1797
Information/comparison app	-0.1217	0.0310	-3.9270	0.0001***
Information/comparison search	-0.1638	0.2313	-0.7080	0.4789
Travel agent website	-0.0231	0.0154	-1.5040	0.1327
Travel agent app	0.0333	0.1369	0.2430	0.8081
Travel agent search	-0.7340	0.2760	-2.6600	0.0078**
Focal website	0.1759	0.0202	8.7150	<0.0001***
Focal search	0.3176	0.1505	2.1100	0.0349*
Airline company website	0.0462	0.0222	2.0770	0.0378*
Airline company app	0.0236	0.0668	0.3540	0.7236
Airline company search	-0.1595	0.1156	-1.3800	0.1677
Generic search	0.0360	0.0173	2.0730	0.0382*
Affiliates	-0.1374	0.4556	-0.3020	0.7629
Banner	0.0109	0.3478	0.0310	0.9750
Email	0.0209	0.0317	0.6580	0.5108
Pre-rolls	-1.0498	0.8519	-1.2320	0.2178
Retargeting	-0.0421	0.0239	-1.7600	0.0784

Table 4. Estimated coefficients of the vanilla logistic regression. Significance codes: *** = P<0.001, ** = P<0.01, * = P<0.05.

	Average marginal effect	Std. Error	z value	Pr(>|z|)
Accommodation app	0.0001	0.0005	0.2040	0.8384
Accommodation search	0.0001	0.0005	0.2738	0.7843
Accommodation website	-0.0001	0.0001	-0.6722	0.5014
Affiliate	-0.0008	0.0027	-0.3016	0.7630
Airline company app	0.0001	0.0004	0.3536	0.7236
Airline company search	-0.0009	0.0007	-1.3747	0.1692
Airline company website	0.0003	0.0001	2.0600	0.0394*
Banner	0.0001	0.0021	0.0314	0.9750
Email	0.0001	0.0002	0.6574	0.5109
Focal search	0.0019	0.0009	2.0927	0.0364*
Focal website	0.0010	0.0001	7.7563	<0.0001***
Generic search	0.0002	0.0001	2.0512	0.0402*
Information/comparison app	-0.0007	0.0002	-3.8240	0.0001***
Information/comparison search	-0.0010	0.0014	-0.7074	0.4793
Information/comparison website	-0.0001	0.0001	-1.3360	0.1816
Pre-roll	-0.0062	0.0051	-1.2273	0.2197
Retargeting	-0.0003	0.0001	-1.7531	0.0796
Travel agent app	0.0002	0.0008	0.2429	0.8081
Travel agent search	-0.0044	0.0017	-2.6145	0.0089
Travel agent website	-0.0001	0.0001	-1.4985	0.1340

Table 5. Average marginal effects of the vanilla logistic regression. Significance codes: *** = P<0.001, ** = P<0.01, * = P<0.05.

	Estimate
Bias	-5.1533
Accommodation website	-0.0061
Accommodation app	-0.0051
Accommodation search	-0.0384
Information/comparison website	-0.0128
Information/comparison app	-0.0293
Information/comparison search	-0.0475
Travel agent website	-0.0156
Travel agent app	-0.0302

Travel agent search	-0.3501
Focal website	0.1556
Focal search	0.2475
Airline company website	0.0146
Airline company app	0.0238
Airline company search	-0.0768
Generic search	0.0064
Affiliates	-0.2781
Banner	-0.1842
Email	0.0181
Pre-rolls	-0.4382
Retargeting	-0.0083

Table 6. Estimated coefficients of the bagged regularized logistic regression.

	Estimate
Bias	-6.3386
Accommodation website	0.5356
Accommodation app	-13.2275
Accommodation search	-13.2275
Information/comparison website	0.9495
Information/comparison app	-13.2275
Information/comparison search	-13.2275
Travel agent website	0.7676
Travel agent app	-13.2275
Travel agent search	1.4788
Focal website	4.8580
Focal search	-13.2275
Airline company website	0.7988
Airline company app	-13.2275
Airline company search	-13.2275
Affiliates	-13.2275
Banner	-13.2275
Email	3.2251
Pre-rolls	-13.2275
Retargeting	3.7736

Table 7. Estimated coefficients of the dynamic logistic regression with only the last touchpoint. The reference level is the touchpoint “Generic search”.

	Estimate
Bias	-7.1306
Coefficients of features last time instance:	
Accommodation website	0.5084
Accommodation app	-13.8874
Accommodation search	-14.1950
Information/comparison website	0.9285
Information/comparison app	-14.2880
Information/comparison search	-14.2793
Travel agent website	0.5797
Travel agent app	-14.1352
Travel agent search	1.5048
Focal website	4.8258
Focal search	-16.4847
Airline company website	0.7546
Airline company app	-14.3268
Airline company search	-14.1744
Affiliates	-14.5394
Banner	-14.3782
Email	2.1885
Pre-rolls	-14.6787
Retargeting	1.8213
Coefficients of features one before last time instance:	
Accommodation website	0.6817
Accommodation app	2.4079
Accommodation search	0.3602
Information/comparison website	1.0870
Information/comparison app	-14.8270
Information/comparison search	-14.7501
Travel agent website	1.0727
Travel agent app	-13.6584
Travel agent search	1.4193
Focal website	3.5030
Focal search	0.4279
Airline company website	0.4904

Airline company app	-14.2594
Airline company search	0.9544
Generic search	0.6796
Affiliates	2.5838
Banner	2.6030
Email	1.2005
Pre-rolls	-13.8815
Retargeting	2.1168

Table 8. Estimated coefficients of the dynamic logistic regression with the two last touchpoints. The reference level at the last time instance is "Generic search" and one before last time instance is "None".

	Estimate
Bias	-6.8705
Coefficients of features last time instance:	
Accommodation website	0.5345
Accommodation app	-13.6694
Accommodation search	-13.9636
Information/comparison website	0.7585
Information/comparison app	-14.9646
Information/comparison search	-13.8453
Travel agent website	0.3739
Travel agent app	-14.6754
Travel agent search	1.6445
Focal website	4.2904
Focal search	-16.4172
Airline company website	0.3758
Airline company app	-15.8760
Airline company search	-14.3111
Affiliates	-14.5110
Banner	-13.6259
Email	1.7401
Pre-rolls	-14.2220
Retargeting	0.7237
Coefficients of features one before last time instance:	
Accommodation website	-0.5265
Accommodation app	1.1791

Accommodation search	-0.6847
Information/comparison website	-0.0677
Information/comparison app	-16.3165
Information/comparison search	-16.0607
Travel agent website	-0.2685
Travel agent app	-15.2946
Travel agent search	0.1426
Focal website	2.1807
Focal search	-1.0127
Airline company website	-0.5669
Airline company app	-15.2187
Airline company search	-0.5657
Generic search	-0.5715
Affiliates	1.3386
Banner	1.3366
Email	-0.3333
Pre-rolls	-15.1145
Retargeting	0.2187
Coefficients of features two before last time instance:	
Accommodation website	0.5870
Accommodation app	-12.9599
Accommodation search	-13.6082
Information/comparison website	1.4991
Information/comparison app	2.0338
Information/comparison search	-13.6993
Travel agent website	1.1955
Travel agent app	-11.2772
Travel agent search	-14.4097
Focal website	2.4429
Focal search	1.3208
Airline company website	1.9274
Airline company app	3.8495
Airline company search	1.6158
Generic search	0.6277
Affiliates	3.8859

Banner	-14.5672
Email	1.2212
Pre-rolls	-14.3466
Retargeting	3.1397

Table 9. Estimated coefficients of the dynamic logistic regression with the three last touchpoints. The reference level at the last time instance is "Generic search", one before last time instance is "None", and two before last time instance is also "None".

	First Touch attribution	**Last Touch attribution**	**Shapley Value**	**Vanilla logistic regression**	**Bagged regularized logistic regression**	**Dynamic logistic regression (only last touchpoint)**
Accommodations website	0.1615	0.1719	0.0233	0.0522	0.0517	0.0093
Accommodation app	0	0	0.001	0.0529	0.0517	0
Accommodation search	0.0052	0	0.0016	0.0537	0.0505	0
Information website	0.0573	0.0729	0.0495	0.0518	0.0511	0.0141
Information app	0.0052	0	0.0005	0.0471	0.0496	0
Information search	0	0	0.0053	0.0469	0.0488	0
Travel agent website	0.1771	0.1667	0.0384	0.0514	0.0513	0.0117
Travel agent app	0	0	0.0013	0.0539	0.0509	0
Travel agent search	0.0104	0.0052	0.0122	0.0262	0.0386	0.0235
Focal website	0.4063	0.4479	0.3777	0.0637	0.0615	0.5731
Focal search	0	0	0.0707	0.0717	0.0648	0
Airline company website	0.0417	0.0417	0.0288	0.0557	0.0529	0.0121
Airline company app	0.0104	0	0.0047	0.0544	0.0533	0
Airline company search	0.0052	0	0.004	0.0444	0.0483	0
Generic search	0.0365	0.0208	0.0154	0.0548	0.0524	0.0055
Affiliate	0.0052	0	0.0113	0.045	0.0411	0
Banner	0.0052	0	0	0.0509	0.0437	0
Email	0.0312	0.0104	0.0726	0.0539	0.0537	0.1308
Pre-roll	0	0	0	0.0187	0.0331	0
Retargeting	0.0417	0.0625	0.2815	0.0507	0.0509	0.2199

Table 10. The contribution assigned to each touchpoint by the estimated models. Graph 5 in section 4.2 of the main report provides a graphical depiction of the attribution scores. Note: the table is split into two parts, table 11 is the accompanying table.

	Dynamic logistic regression (last two touchpoints)	**Dynamic logistic regression (last three touchpoints)**	**Markov chain first-order**	**Markov chain second-order**	**Markov chain third-order**	**Markov chain fourth-order**
Accommodations website	0.003	0.0019	0.1671	0.1587	0.1527	0.1477
Accommodation app	0	0	0.0159	0.0125	0.0108	0.0094
Accommodation search	0	0	0.0292	0.0255	0.0241	0.0229
Information website	0.007	0.0094	0.1224	0.1225	0.1247	0.1262
Information app	0	0	0.0148	0.0116	0.0111	0.011
Information search	0	0	0.0093	0.0094	0.0101	0.011
Travel agent website	0.0048	0.0038	0.1721	0.1682	0.1643	0.1616
Travel agent app	0	0	0.0054	0.0039	0.0034	0.0032
Travel agent search	0.0164	0	0.0166	0.019	0.0199	0.0205
Focal website	0.8892	0.9026	0.1219	0.1496	0.1566	0.1593
Focal search	0	0	0.0059	0.0082	0.0103	0.0137
Airline company website	0.0032	0.006	0.1094	0.1007	0.0974	0.094
Airline company app	0	0	0.0086	0.0064	0.0067	0.0073
Airline company search	0	0	0.0239	0.0242	0.0262	0.0258
Generic search	0.0018	0.0011	0.1056	0.0942	0.0877	0.088
Affiliate	0	0	0.0036	0.0063	0.0098	0.01
Banner	0	0	0.0039	0.0047	0.0046	0.0044
Email	0.0276	0.014	0.0161	0.0182	0.0188	0.0201
Pre-roll	0	0	0.0049	0.0052	0.0054	0.0048
Retargeting	0.0469	0.0611	0.0435	0.0509	0.0555	0.0591

Table 11. The contribution assigned to each touchpoint by the estimated models. Graph 5 in section 4.2 of the main report provides a graphical depiction of the attribution scores. Note: the table is split into two parts, table 10 is the accompanying table.

Initial idea - a Hidden Markov model

The initial goal of this thesis was to assess to what extent there is a difference in the value of ads between stages of the purchase decision process. The idea was derived from the paper by Abhishek et al. (2012) who employed a hidden Markov model (hereafter abbreviated as HMM) anchored by the notion of a conversion funnel. Within the customer journey, distinctive stages of the purchase decision process can be identified. The process of walking through the customer journey and eventually purchasing is the conversion funnel (Kotler & Armstrong, 2010). The latent stages of the Markov model reflect the engagement of the customer through the conversion funnel (i.e. disengaged, active, engaged, conversion). Through the use of the conversion funnel, touchpoints can be assessed within the engagement stage. Some advertisements may be more effective in earlier stages and some in later stages.

As earlier indicated in the dataset section in the main report, the dataset can be broadly classified in enterprise-initiated touchpoints and customer-initiated touchpoints. Enterprise-initiated touchpoints are divided into displays, retargeted displays, pre-roll ads, affiliates, and e-mails. Relevant customer-initiated touchpoints are websites, apps, and search engine terms, and can, in turn, be divided into information/comparison, accommodation, airline, competitive travel agencies and the focal company. In collaboration with the focal company is determined which websites, apps, and search engine terms are relevant and belongs to which of above-mentioned categories. The distinction between enterprise-initiated touchpoints and customer-initiated touchpoints is not implemented in the main report, but when a HMM was to be applied, it would be performed.

The emission of the latent stages of the HMM is multivariate, as it would output both a customer-initiated touchpoint and whether the consumer has converted at a particular point in time. This latent stage of the customer journey is not present in the data but can be inferred from the multivariate output (i.e. customer-initiated touchpoint and conversion). The transition between the latent stages is a function of enterprise-initiated touchpoints, which can be interpreted as ads (Netzer, Ebbes, & Bijmolt, 2017). Based on the estimated parameters in the model, conversion attribution can be determined. The number of latent stages is not fixed, but could be determined by an information criterion (e.g. BIC), however, the last stage needs to be the conversion stage (Vrieze, 2012). An example of how the model would look like with four stages is depicted in Figure 3 (Abhishek et al., 2012, p. 9).

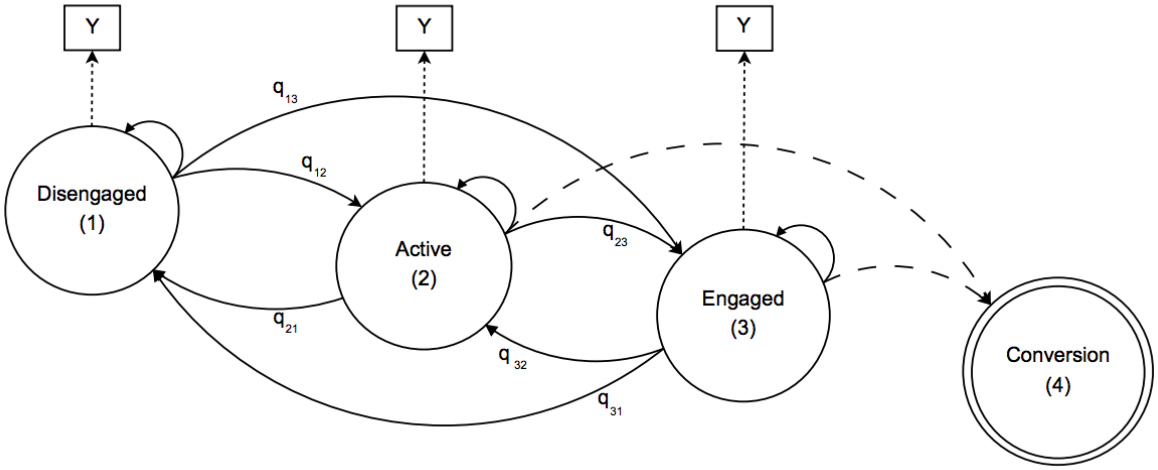


Figure 3. Diagram representing the latent stages and outcomes of a HMM. Reprinted from Media exposure through the funnel: A model of multistage attribution. (p. 9), by Abhishek, V., Fader, P., & Hosanagar, K.

The objective was to assess to what extent there is a difference in the value of ads between stages of the purchase decision process. Since the idea was grounded on the conversion funnel, the propensity to purchase must increase as the customer moves down the funnel. Hence, it is vital to impose (at the least) the constraint that the propensity to purchase must increase as the customer walks through the funnel because otherwise the model fits something arbitrary and without this constraint, it is impossible to address the question whether there is a difference in the value of ads between stages of the purchase decision process (Abhishek et al., 2012). To the best of my knowledge, there is solely one package in R (and none in Python) that can impose constraints and has all other required properties (i.e. have discrete distributions, manage panel data, possibility to add covariates/factors to model the transition probabilities, and has a multivariate output). This package is the DepmixS4 package in R (Speekenbrink & Visser, 2013). Unfortunately, the HMM employed with the DepmixS4 package could not find a solution with the constraint that the propensity to purchase must increase as the customer moves down the funnel. When these constraints are imposed, the model does not fit and a warning message is provided.

Conducting the model without constraints results in a HMM that seems to fit interests in different types of products rather than a conversion funnel, although the results are difficult to interpret. Before diving into the output of the DepmixS4 package, I will elaborate on the HMM. The model consists of three components: the initial stage probabilities, the transition probabilities, and the emission probabilities (Rabiner, 1989). The initial stage distribution is the probability distribution that a customer journey starts at a particular stage. A reasonable assumption is that the consumer starts in a disengaged stage and becomes more engaged when encountering touchpoints. Ideally, a constraint has to be imposed that the consumer begins in the first stage (Abhishek et al., 2012). For instance, when

there are four stages the initial stage distribution would be {1, 0, 0, 0}. The transition probabilities between the hidden stages is a function of time-varying ads. More specifically, the ads are assumed to follow a multinomial logistic regression, where each stage can loop to itself or another stage. In addition, a constraint that the customer can only move down the funnel would improve the concept of the conversion funnel as this is natural to the funnel (Abhishek et al., 2012). Lastly, the emission probabilities are twofold as the output of the stages are both customer-initiated touchpoints as well as whether or not the consumer convert. The probabilities of the customer-initiated touchpoints have to sum up to one and the probability of conversion and non-conversion has to sum up to one. The last stage in the conversion funnel is the conversion stage and hence the probability to convert has to be 1 at this stage. Moreover, the propensity to purchase must increase as the customer moves down the funnel. The latter constraint is required for the identification of the stages since otherwise there exists no funnel (Abhishek et al., 2012).

The output of the HMM without constraints can be found in the Appendix of the initial idea. More specifically, the output consists of three components: the initial stage probabilities, the transition coefficients of the multinomial logit model for each stage and the corresponding probabilities at zero values of the enterprise-initiated touchpoints, and the twofold emission probabilities of the customer-initiated touchpoints as well as whether or not the consumer convert. It is considerably difficult to identify and interpret the hidden stages of the HMM as the model attempt to fit the data perfectly. Nonetheless what the model fits is unknown and is extremely hard to comprehend as the concept of the conversion funnel has vanished. One could argue that the HMM fit interests in different types of products. For example, the probability of visiting an accommodation website given it is in stage one is 0.852, which indicate that customers in stage one are only interested in accommodations. Another argument to support this claim is that the probability to convert in stage one, stage two, and stage three is <0.001, which suggest that the customers are not interested in booking a trip at a travel agency. Another notable fact is that the probability to purchase is extremely low (0.004), which makes it even harder for the model to find a good fit. Nevertheless, the interpretation of the stages without a theoretical background is rather subjective and one might come up with a different interpretation of the stages.

In conclusion, imposing constraints will resolve the identification and interpretation problem but there is no HMM package that could fit the data with these constraints. As indicated in the previous paragraph, the model does seem to fit interests in different types of products rather than a conversion funnel. Hence, the whole concept of the conversion funnel has vanished. It is impossible to measure to what extent there is a difference in the value of ads between stages of the purchase decision process, without imposing constraints on the HMM. Therefore, a new feasible research question is chosen which I address in the main sections of this thesis.

Appendix - Initial idea

The Output of a HMM without constraints performed with the package DepmixS4.

Initial stage probabilities			
$P(s_1)$	$P(s_2)$	$P(s_3)$	$P(s_4)$
0.392	0.181	0.343	0.084

P*(Si)* is the probability of starting in stage *i*.

Transition coefficients of the multinomial logit model for stage 1				
To:	S_1	S_2	S_3	S_4
Bias	0	13.553	14.064	9.323
Affiliate	0	5.455	5.754	-1.290
Banner	0	-13.026	-12.729	-11.397
Retargeting	0	-13.443	-13.168	-9.469
Pre-roll	0	14.955	14.746	17.671
Email	0	-13.392	-13.387	-9.677

The values in column S_i are the coefficient for the multinomial logit function of going from stage 1 to stage *i.* Moreover, the first column is parametrized to zero for the base category.

Probabilities at zero values of the enterprise-initiated touchpoints for stage 1			
$P(s_1)$	$P(s_2)$	$P(s_3)$	$P(s_4)$
<0.001	0.373	0.622	0.005

P*(Si)* is the probability of going from stage 1 to stage *i* at zero values of the enterprise-initiated touchpoints.

Transition coefficients of the multinomial logit model for stage 2				
To:	S_1	S_2	S_3	S_4
Bias	0	-9.005	-0.225	-5.202
Affiliate	0	8.960	0.907	2.438
Banner	0	7.664	-0.151	-3.920
Retargeting	0	8.339	0.422	4.348
Pre-roll	0	7.974	0.173	-4.434
Email	0	-8.026	0.050	-1.096

The values in column S_i are the coefficient for the multinomial logit function of going from stage 2 to stage *i.* Moreover, the first column is parametrized to zero for the base category.

Probabilities at zero values of the enterprise-initiated touchpoints for stage 2			
$P(s_1)$	$P(s_2)$	$P(s_3)$	$P(s_4)$
0.554	<0.001	0.443	0.003

P*(Si)* is the probability of going from stage 2 to stage *i* at zero values of the enterprise-initiated touchpoints.

Transition coefficients of the multinomial logit model for stage 3				
To:	S_1	S_2	S_3	S_4
Bias	0	-0.673	-6.846	-4.378
Affiliate	0	-0.047	1.487	2.632
Banner	0	-0.535	2.923	3.539
Retargeting	0	-0.025	5.364	3.308
Pre-roll	0	-0.028	4.450	-4.102
Email	0	-0.783	5.052	1.946

The values in column S_i are the coefficient for the multinomial logit function of going from stage 3 to stage *i.* Moreover, the first column is parametrized to zero for the base category.

Probabilities at zero values of the enterprise-initiated touchpoints for stage 3			
$P(s_1)$	$P(s_2)$	$P(s_3)$	$P(s_4)$
0.656	0.335	<0.001	0.008

P*(Si)* is the probability of going from stage 3 to stage *i* at zero values of the enterprise-initiated touchpoints.

Transition coefficients of the multinomial logit model for stage 4				
To:	S_1	S_2	S_3	S_4
Bias	0	-0.058	0.558	4.783
Affiliate	0	-0.217	-6.071	-4.119
Banner	0	-0.108	-0.949	-4.208
Retargeting	0	-0.647	-0.222	-0.819
Pre-roll	0	-0.441	-2.818	-4.930
Email	0	-1.047	-2.854	-1.301

The values in column S_i are the coefficient for the multinomial logit function of going from stage 4 to stage *i*. Moreover, the first column is parametrized to zero for the base category.

Probabilities at zero values of the enterprise-initiated touchpoints for stage 4			
$P(s_1)$	$P(s_2)$	$P(s_3)$	$P(s_4)$
0.008	0.007	0.014	0.970

P*(Si)* is the probability of going from stage 4 to stage *i* at zero values of the enterprise-initiated touchpoints.

Emission probabilities of the customer-initiated touchpoints				
	$P(s_1)$	$P(s_2)$	$P(s_3)$	$P(s_4)$
Accommodations website	0.852	0.005	<0.001	0.068
Accommodation app	0.003	0.011	0.018	<0.001
Accommodation search	<0.001	0.029	0.031	0.001
Information/comparison website	0.044	0.208	0.083	0.220
Information/comparison app	0.009	0.011	0.006	0.004
Information/comparison search	0.003	<0.001	0.004	0.016
Travel agent website	<0.001	<0.001	0.787	0.343
Travel agent app	0.006	0.004	<0.001	<0.001
Travel agent search	0.009	0.010	<0.001	0.018
Focal website	0.008	0.020	0.002	0.190
Focal search	<0.001	<0.001	<0.001	0.010
Airline company website	0.004	0.451	0.001	0.014
Airline company app	0.008	0.003	0.005	<0.001
Airline company search	0.020	0.004	0.016	0.005
Generic search	0.033	0.244	0.045	0.112

P*(Si)* is the probability of encountering a specific customer-initiated touchpoint when being in stage *i*.

Emission probabilities of the whether the potential customer convert				
	$P(s_1)$	$P(s_2)$	$P(s_3)$	$P(s_4)$
Non-conversion	>0.999	>0.999	>0.999	0.996
Conversion	<0.001	<0.001	<0.001	0.004

P*(Si)* is the probability of converting or non-converting when being in stage *i*.