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

Exploring the Capabilities of Random Forests and Shapley Value: Attribution Modeling in Online Advertising

Sophie Gehrels (476366sg)

Supervisor:	Kathrin Gruber
Second assessor:	Markus Mueller
Date final version:	1st July 2023

Abstract

As companies increasingly invest in online advertising, understanding the impact of different online media channels on conversion probability has become crucial. This paper addresses the attribution problem in online advertising by introducing new modeling techniques for prediction and post-hoc interpretation while preserving the order of consumer paths. Predictive models, including the benchmark Markov Chain (MC), traditional Random Forest (RF), and the Moving Random Forest (MRF) that accounts for time dependency, are discussed. Results demonstrate that RF outperforms MC and MRF in terms of prediction accuracy, achieving an AUC value of 0.6. Furthermore, various attribution metrics are explored. Besides the common heuristic methods, Incremental Value Attribution (IVA) and Shapley Value (SV) are computed using simulated Markov graphs. Subsequently, the Symmetric Shapley Value (SSV) and the Asymmetric Shapley Value (ASV) are employed to enhance the post-hoc interpretability of the RF. A comparative analysis of SSV and ASV reveals the preference for ASV, as it not only takes into account the causal structure of the data, but also produces only positive outcomes, making it more aligned with the desirable marketing perspective. The study highlights computational complexities associated with the Shapley Value in attribution modeling. However, combining RF with ASV offers a feasible approach that considers the sequential nature of the data and delivers interpretable results. These findings contribute to understanding attribution in online advertising, providing marketers with reliable measures for decision-making.

Contents

1 Introduction

1.1 Research Problem

HotWire sold the first banner ads to various advertisers in 1994, which marked the beginning of online advertising (Evans, 2008). Since then, the online advertising industry is continuously growing and evolving as new platforms and technologies are developed. Examples of new technological advances are innovative algorithms that companies can use to present ads to users, based on browsing history and cookies. Other developments are search engine snippets that companies apply to ensure the alignment of keywords and phrases with users' search queries(Thomaidou & Vazirgiannis, 2011). The amount of money circulating in this emerging online advertising world is huge. Statista stated that in 2021 the spending in online advertising worldwide was 522.5 billion US dollars (Department, 2023). Hence, it's important for advertisers to understand how to allocate their advertising budget, such that these decisions can maximize their return on investment (ROI). The ROI strongly depends on the number of ad-interactions that lead to a conversion. A *conversion* can be defined as the event in which a customer purchases the product that has been advertised (Singal, Besbes, Desir, Goyal & Iyengar, 2019). The decision on conversion can be influenced by various advertisements on different media channels. Quantifying the impact of ads on channels is referred to as *advertisement attribution*. More formally, advertising attribution can be identified as a method to assign conversion value to a media channel as a result of an interaction with an online user (Dalessandro, Perlich, Stitelman & Provost, 2012).

The topic of attribution is widely debated within the advertising industry, with currently numerous proposed attribution models in the literature, including rule-based models and algorithmic models. However, determining the most effective approach to measure the contribution of each media channel remains a topic of ongoing discussion. The Shapley Value (SV), a metric from cooperative game theory that has been used in the online advertising industry, is one promising metric for attribution. The SV ensures fair distribution of value across channels. This is especially important in scenarios involving the collaborative impact on achieving conversion outcomes, by acknowledging the interconnected nature of channels (Dalessandro et al., 2012). Singal et al. (2019) propose an innovative approach to attribution in online advertising by introducing an axiomatic framework called the Counterfactual Adjusted Shapley Value (CASV). This metric accounts for counterfactual actions and aligns with the unique-uniform attribution scheme. Singal et al. (2019) claim it is important to incorporate the counterfactual value in attribution modeling, as it provides a more accurate understanding of their individual contributions to conversion outcomes. The authors state that attribution involves measuring the extra benefit of showing an advertisement compared to what would have happened if no advertisement was shown. This comparison, known as the *counterfactual* scenario, helps to understand the impact of ads on conversion. Theoretical comparative analyses demonstrate the superiority of this novel attribution scheme over commonly used methods such as uniform weights, or last touch attribution. While the CASV seems desirable, its practical implementation is challenging for most datasets. Computing the SV requires exploring all possible combinations of variables in the dataset, and to additionally observe the counterfactual action for each data point is unrealistic. It becomes even more difficult when the specific order of the consumer path to conversion needs to be considered, which is the requirement for an unbiased estimator. A *consumer path* can be illustrated as the stages of a consumer's search and purchase journey. The different stages represent interactions with advertisements across various media channels. Hence, as the number of channels increases, this approach becomes unfeasible. Singal et al. (2019) attempted to address these challenges by employing a Markov model to imitate the consumer path, and using the SV to compute channel attribution within this framework. However, this solution does not fully resolve the problem as the Markov model itself becomes the limiting factor, introducing its own performance limitations in the number of channels which are computationally feasible. As the SV is not dependent on the Markovian assumption, it is interesting to investigate how to characterize the SV under non-Markovian models.

In contrast to Markov models, which face limitations due to computational feasibility when employing a high dimensional covariate space, machine learning (ML) techniques excel in their ability to process an extensive number of variables and data. A notable model that can be utilized in combination with the SV as another predictive model is the Random Forest (RF) algorithm developed by Breiman (2001). It has achieved remarkable success as a versatile method for classification and regression tasks (Biau & Scornet, 2016). As such, RF seems a desirable approach in the context of assigning attribution value to media channels. However, it is crucial to consider a model's ability to incorporate 'secret information,' specifically, the order of channels within the consumer path. The interaction between channels holds valuable insights that should not be disregarded, as it significantly contributes to the predictive power of the model. In other words, the sequence of advertising interactions can influence a user's decision to convert.

In their research Goehry, Yan, Goude, Massart and Poggi (2021) propose an innovative machine learning technique used for the construction of RF, called *block bootstrapping*, to maintain the time-dependent structure of the data. An approach that has not been considered yet in the context for attribution modelling. Another approach to address the sequence issue is to leverage an interpretability model, such as the desirable SV, to tackle this task. Further research can be conducted by combining the predictive RF model with the SV for interpretability while accommodating the sequential structure of the advertising data. This leads to the following research question:

What is the optimal approach for incorporating the Shapley Value as an interpretability measure in Random Forests for attribution in online avdertising, considering the inclusion of the sequential order of channels in the consumer path, and how does this approach compare to applying the Shapley Value to a Markov Chain in terms of prediction performance, and attribution value?

1.2 Motivation

The increasing interest in advertising attribution has led to the rise of various different attribution methods. However, it remains uncertain what is the most effective approach to measure the attribution of each channel. Common approaches rely on heuristic models. Examples are Last Touch Attribution (LTA) and Uniform Weight Attribution (UWA). LTA is an unfair approach since it does not distribute value to channels building product awareness. UWA might solve this problem, however, remains very general and is not data-driven (Abhishek, Fader & Hosanagar, 2012). Also, both methods ignore the potential value of users who have not interacted with the media channel, known as the counterfactual value (Singal et al., 2019).

Other more advanced data-driven models are Incremental Value Heuristics (IVH), and SV. IVH outperforms LTA and UWA in the sense that it will somewhat account for the counterfactual, however, it assigns more value than the actual advertising network generates, as proven by Singal et al. (2019). Shapley Value is a model based on cooperative game theory that Nisar and Yeung (2015) employ in their non-parametric approach to attribution. SV is a desirable approach as it possesses several beneficial properties: *linearity, null-player, symmetry, and efficiency*, which will be discussed in detail in Section 4.3 (Singal et al., 2019). However, as briefly mentioned earlier, there is a notable pitfall associated with this metric, as the computation involves dealing with computational complexity due to the need to explore all possible combinations of players (channels). This leads to infeasible computation times. In their study Singal et al. (2019) seek to circumvent the computational intractability by using the Markov model, which models the consumer path, and subsequently combine this with the SV. However, by doing this they shift the intractability problem of the SV to the Markov model, and now the predictive model becomes the bottleneck. As the number of states or variables in Markov models increases, the complexity of the models also increases, even though they do not require searching through all possible combinations in the data like the SV. Hence, also with this set up, the number of channels has to be reduced, or aggregated, in order to manage computation. In attribution modelling, not much is known about how to characterize the SV with another predictive model, like the RF. This limitation causes a knowledge gap that needs to be filled, to give marketers more reliable and precise attribution measures, that can be applied in various contexts. The research is structured as follows. First, the conventional Markov model will be employed, incorporating various attribution metrics, including the Shapley Value (SV). This setup serves as the baseline for comparison in this study. Second, two predictive models will be implemented and compared: the traditional Random Forest and a Random Forest utilizing a block bootstrap technique to account for data ordering. Finally, an Asymmetric Shapley Value (ASV) will be constructed to interpret the results of a standard Random Forest model, providing a causal framework that preserves the sequential nature of the data. The ASV will then be compared to the usual SV.

1.3 Relevance

The topic of conversion attribution is widely debated within the advertising industry. Being able to correctly attribute value to specific channels is of great importance for various stakeholders, assisting to improve the effectiveness and efficiency of advertising campaigns. The advertising eco-system contains several players, including marketers, vendors, search engines, publishers, and consumers (Dalessandro et al., 2012).

First of all, marketers need accurate information about advertising effectiveness, in order to make correct and accountable decisions on their budget allocation. They only have a limited amount of money to spent, hence it's important to understand what marketing campaigns will maximize their ROI. Geyik, Saxena and Dasdan (2014) analyse budget allocation in online advertising and ROI performance. However, they utilize last touch and multi touch models, and do not consider the analysis of other attribution metrics.

Second, to prove the worth and efficiency of their advertising solutions, online advertising platforms and technology providers can use attribution modeling. With accurate attribution, they can demonstrate how their communication channels have a positive impact on conversions and illustrate how their advertising tools have a positive impact on the success of campaigns. For example, in the study of Singal et al. (2019) they show that e-mail activity increases the probability for a user to move from an 'aware' state to an 'interest' state by 5 times. Such information is useful for advertisement agencies to show their clients how effective their campaigns are, attracting more clients and defending their charges.

Thirdly, publishers and media channels also have an interest in attribution allocation. They need to understand the contribution of their channels in the conversion process to demonstrate their value to advertisers, negotiate fair pricing for ad inventory, and optimize their content and ad placements to maximize conversions. Attribution methods help advertisers understand which publishers contribute the most to their success. By using these methods, advertisers can make better decisions and boost their profits (Berman, 2018).

Lastly, consumers themselves can benefit from accurate attribution. Brajnik and Gabrielli (2010) claim that advertisements can serve as additional channels to fulfill informational or emotional needs, but they can also create challenges and complexities for users. By analyzing user behavior and attribution data, advertisers can personalize and improve the user experience, presenting more relevant and targeted advertisements. This enhances the overall online experience for users and reduces the likelihood of irrelevant or intrusive advertising.

In summary, accurate attribution in online advertising is important for advertisers, advertising platforms, publishers, and consumers alike. It enables informed decision-making, increasing revenues, optimization of advertising strategies, fair pricing, and enhanced user experiences.

The literature on the methods used to model consumer paths, and on the different attribution metrics is reviewed in the next section. Section 3 describes the characteristics, and the necessary pre-processing steps of the dataset. Subsequently, Section 4 covers all methods used in this study. The results are presented in 5, followed by a discussion and conclusion in Section 6.

2 Literature

The literature review provides an overview of the research done on advertising attribution modelling. We will first focus on the Markov model, followed by the analysis of a traditional Random Forest, and a RF in the context of sequence-dependent data. Thereafter, several several attribution metrics are analysed, and in specific the desirable Shapley Value.

2.1 Markov Model

Markov models offer a valuable framework for presenting the sequential nature of the consumer path in advertising attribution, allowing for the analysis of how different channels interact and influence conversion outcomes. A lot of research has been done one this topic, utilizing different approaches. This brings us to the following research question:

How do several different Markov models perform in modeling the consumer path to conversion in online advertising attribution?

A Markov model is a mathematical model that helps us understand how a sequence of events or observations is influenced by previous events. In the context of advertising, it can be used to simulate and analyze the progression of a consumer path, taking into account the dependencies and transitions between different states or channels. Abhishek et al. (2012) were the first to introduce a Hidden Markov Model (HMM) to create the consumer path. The states were defined as the unobserved levels of interest during the decision process of a user. The authors demonstrated that the decision on the moment to expose a user an advertisement, makes a lot of difference in conversion rate. The idea of defining the states in a Markov model as unobserved levels of interests, contrasts to the approach of this research, where states are defined as the interaction with different media channels. The "media-channel" approach is easier to implement, as it's hard to collect 'level-of-interest' data. Hence, a more useful approach for all parties who want to optimize their online advertising performance. Anderl, Becker, Von Wangenheim and Schumann (2016) choose a higher order Markov model, combined with the "Removal Effect" metric, which will be described later in this Section, to tackle the attribution problem. Compared to standard Markov models, higher-order models, not solely consider dependency on the current state but also on a specified number of past states. The authors conclude higher order Markov models perform better than standard Markov Models in modelling consumer paths. However, the authors fail to address the complexity a higher order brings to the table. The higher the order of the Markov model, the more parameters need to be estimated. This increases the complexity, and makes it computationally intractable. As such, their employed data set consists of short user journeys, a lot of which are only a one time interaction. This restricts their attribution model's capacity to be used with customer journeys that are more complicated.

The study by Kakalejčík, Bucko, Resende and Ferencova (2018) builds upon the research on Markov models. They compare the results of their Markov model, in terms of credit assigned to various channels, to that of models based on heuristics. They further enhance prior research by investigating the optimal order for their Markov model, determining that an order of 4 yields the most effective analysis of consumer paths. However, it is worth noting that the results of Kakalejčík et al. (2018) may have certain limitations. One limitation is the exclusion of consumer paths that do not end up in a conversion. In contrast, this present research encompasses non-conversion paths, recognizing their significance in increasing the predictive accuracy of the model. This is particularly relevant considering the substantial number of consumer paths that do not result in immediate purchases but may still contribute to future conversions as potential buyers. By incorporating these non-conversion paths into the analysis, a more comprehensive understanding of the consumer path is achieved.

Finally, Singal et al. (2019) elaborate on the existing Markovian models in literature to illustrate the consumer path. Where Abhishek et al. (2012), Anderl et al. (2016) and Kakalejčík et al. (2018) could be seen as unique examples of the model, they outline their Markov approach in more general terms. Rather than solely using past interactions to map the consumer path , they compose a combined state action space. They define their state space as the user moving from being unaware about the product, to possibly buying the product, including some intermediate states. Their action space includes doing nothing, sending an email, showing ad display, etc. It's notable that the definitions of state and action spaces is different for all aforementioned papers and can be altered to suit the needs of the advertisers, the related context, or objective of the research. However, for the abstract definition of Singal et al. (2019), it's hard to collect appropriate data for their state-action framework, which indicates a limitation in the implementation of their method. Furthermore, (Singal et al., 2019) distinguish themselves by combining the Shapley Value with the Markovian network, as a way to add interpretability to the predictive Markov model. As discussed in Section 1.3, this set-up sounds desirable, however, doesn't solve the computation issue, especially when too many channels are considered.

2.2 Machine Learning Models

Given the substantial amount of literature on Markov models for attribution in online advertising, it is evident that their ability to capture the consumer paths, which contain valuable information, has sparked significant interest. However, due to their limitation on covariate space, it is equally important to explore the potential of machine learning models in incorporating sequential dependence within the data. Hence, the next sub-question is as follows:

How do Machine Learning models perform in the context of online advertising attribution?

Machine learning models have emerged as transformative non-parametric statistical models, enabling powerful data driven analysis and decision making across various sectors. The application of machine learning methods in the context of online advertising has been introduced by Arava, Dong, Yan, Pani et al. (2018). In their study, the authors propose the utilization of a Deep Neural Network (DNN) to assess the influence of advertising channels in online marketing. This approach aims to tackle challenges related to channel interaction, user characteristics, and time dependency. What sets their model apart from other attribution models is the incorporation of user data, including demographics, which helps reduce estimation bias. Additionally, to enhance the interpretability of the effects from various media channels, they introduce an attention mechanism. The mechanism relies on giving different levels of importance to different interactions with media channels at different times. While this gives insights to the importance of different ad-interactions, it doesn't offer an accurate and quantitative measure of assigning channel attribution. It's more suitable for enhancing the prediction performance than providing quantitative contributions per channel. The Shapley Value is, therefore, superior as an interpretation method. Du, Zhong, Nair, Cui and Shou (2019), on the other hand, do use the SV as an interpretation method to assign credit to channels. Their approach contains two steps, very similar to the framework of this research. First, fit a DNN model to construct the consumer path to conversion. Second, use this model to assign the incremental value of the conversion to a specific channel, by means of the SV. Compared to Markov models that require lower-dimensional states, this approach is particularly desirable as it utilizes the capabilities of a DNN to effectively handle the high dimensionality associated with over 200 different ad types. Furthermore, they are able to preserve the sequential dependence in the data. However, the their novel approach also has some limitations when applied in a practical setting. DNN can be computationally demanding and need a lot of data for training, which could be problematic in situations where there are few observations, like real-world marketing scenarios. Furthermore, deep learning models use specific techniques such as Back Propagation, which are computationally very demanding (Cao, Wang, Ming & Gao, 2018).

Hence, it's interesting to investigate other machine learning methods. Random Forests (RF) are a type of machine learning algorithm used for making predictions, which is in this case conversion or non-conversion. They were initially developed by Breiman (2001) and nowadays used in many different settings. RF combine decision trees to improve forecast accuracy and address overfitting. It outperforms general decision trees by reducing variance and is a prominent method in the ML spectrum. However, in the advertising attribution industry not much research regarding the RF has been published as it's not sophisticated enough to incorporate the sequential nature of the data. Lechner and Okasa (2019) propose a novel approach for this problem and develop an Ordered Random Forest Model (ORFM). The model is based on the conventional RF and thus inherits the advantage of the ability to deal with a high-dimensional covariate space. However, this study considers the outcome variables Y to be ordered. In the scope of this research we don't deal with ordered outcome variables, but with sequential covariates. In other words, the interactions with the media channels that impact the outcome, are sequentially dependent. Hence, the set up of Lechner and Okasa (2019) is not desirable in the advertising industry. Davis and Nielsen (2020) theoretically analyse RF in the context of time series. They built regression trees for nonlinear autoregressive data and support their findings by several simulation studies. However, the article specifically focuses on nonlinear autoregressive processes of order p, which are not suitable for modeling consumer paths.

In their work, Goehry et al. (2021) propose a novel technique called block-bootstrapping, which builds upon the standard bootstrap method to preserve the sequential nature of user data. The standard bootstrap involves randomly selecting observations with replacement from a single sample, but it fails to account for data dependencies. Block-bootstrapping addresses this limitation by sub-sampling time series during the tree construction phase. This approach mitigates the loss of prediction accuracy caused by breaking the data structure. 3 different block-bootstrapping methods are discussed, non-overlapping, which involves constructing non-overlapping blocks, the moving block bootstrap which chooses blocks of successive observations at random, and the circular block bootstrap which corrects the moving block bootstrap for potential bias. Based on empirical results, Goehry et al. (2021) conclude that both the moving and circular perform better compared to the standard random forest. They found that both methods improve the RMSE up to 11% compared to the traditional RF. The non-overlapping version, however, requires a large number of observations to yield reliable outcomes. A clear limitation of their work is that they have only analysed their results in one specific field of application, namely the load forecasting of a building, i.e. predicting the energy consumption of a building. It's therefore of interest to apply their method in a different context with a dependent data structure, for example the prediction of conversion in attribution modelling.

2.3 Attribution methods

The prediction models discussed above can be used to construct the consumer path. Consequently, attribution metrics are used to interpret these models and thus, to analyse the impact one specific media channel has on the conversion outcome. This information is of great importance in the marketing industry. Therefore the next sub-question is:

How are the different attribution metrics defined and how do they compare to each other?

2.3.1 Heuristic Attribution Metrics

To put value to the attribution of online advertisement channels to a conversion, different models have been developed over the years. The different models can be divided into two groups: rule-based metrics and data-driven algorithmic metrics. In this section the several models that fall under these groups will be discussed.

The first rule-based attribution metric is the Last Touch Attribution (LTA) metric. For this method the advertiser gives all the value of conversion to the channel that has last been 'touched' (or interacted with) on the consumer path to conversion. LTA seems as an unfair metric since it doesn't give any credit to channels that might have contributed to the awareness of the user (Singal et al., 2019). The uniform weight metric might be a solution for this. The uniform weight metric gives equal weights to all channels used in the consumer path. This model is easy to implement, however, there is no obvious justification for linear attribution. Singal et al. (2019) describe that a clear limitation of both methods is that they don't account for the counterfactual value. This means that you only observe the data with a positive conversion value attached to them, hence you throw away a lot of valuable data. Furthermore, Archak, Mirrokni and Muthukrishnan (2010) stated that heuristic methods do not adequately consider interdependence among channels. Hence, algorithmic based attribution metrics were explored.

2.3.2 Algorithmic Attribution Models

Algorithmic models are data-driven rather than being based on fixed and predetermined rules. Abhishek et al. (2012) develop Hidden Markov graphs to construct the consumer path. Subsequently, they employ these graphs to compute a metric which is based on the incremental increase in the probability of a customer making a purchase because they saw the ad. This algorithmic approach is called Incremental Value Attribution (IVA), or *Removal Effect*. The incremental value is estimated by removing a specific ad from the consumer path, followed by comparing the eventual conversion levels. The IVA was initially put forth as an attribution metric in Markov graphs by (Archak et al., 2010) and was modified by Anderl et al. (2016) in their higher-order Markov graphs. Instead of using Markov graphs to describe the consumer path, Arava et al. (2018) employ neural networks to model this behavior, followed by computing the Removal effects. As discussed earlier, their model is called the Deep Neural Net With Attention multi-touch attribution model. Both models are used to predict the incremental value of an add action, however, there is no certainty for IVA being a good attribution scheme (Singal et al., 2019). They might be more granular than rule-based methods, however, require more data and expertise to implement. In their article Singal et al. (2019) demonstrate that IVA can lead to incorrect attribution. The authors show that even though IVA accounts for counterfactual scenarios, the method is not budget feasible. This is because the metric distributes more value to a specific channel than the entire system actually creates.

Singal et al. (2019) state the the widely recognized Shapley Value provides a solution to the aforementioned limitation of IVA. Dalessandro et al. (2012) suggested that using SV as an approximation for the problem of estimating causality in attribution. SV is a post-hoc interpretability method based on game theory. It explains the prediction of the conversion of a consumer, by giving each player their credit in a cooperative game. In an advertisement attribution setting the attribution value refers to the outcome of the game, where the interaction with different media channels are viewed as the players of the game. According to the analysis of Berman (2018), Shapley Values can be regarded as a more fair attribution metric compared to LTA. Furthermore, SV has several beneficial properties, such as *efficiency, linearity, symmetry and null player*, which will be described in Section 4.3 (Singal et al., 2019). However, this interpretation method has two major drawbacks. First of all, estimating SV is computationally exhaustive as you have to search for all the possible combinations in the set of variables. In their paper, Singal et al. (2019) aim to overcome this problem by first modelling the consumer path with the use of a Markov chain. However, this simply shifts the problem to the Markov model, which now becomes the bottleneck. Second, it doesn't account for the counterfactual value. Dalessandro et al. (2012) define the counterfactual value in an advertising framework as the effect of an ad on the consumer conversion rate. In other words, what is the quantification of the impact of showing an ad on a user's conversion behavior by comparing their conversion outcomes with and without exposure to the ad. Singal et al. (2019) tackle these problems and contribute to the existing literature with a new measure for advertisement attribution: the Counterfactual Adjusted Shapley Value (CASV). Theoretically, this metric seems to overcome all the above mentioned limitations. However, as discussed in Section 1.3, it's very hard to detect the counterfactual value in a dataset, and hence not a tractable method to implement.

3 Data

3.1 Data Analysis

The dataset utilized for this research is the "Attribution Data" available on Github, which was uploaded by Mahima Kaushiva (2020). The dataset used in this study comprises 586,737 observations, each containing six variables: *cookieID*, *time stamp*, *interaction*, *conversion*, *conversion value*, and *channel*. This data is typically obtained through cookie tracking, where websites store user information in text files within the browser for analysis purposes. The *cookieID* variable captures this information, while the *time stamp* variable records the exact time of a touch-point by a specific *cookieID*. The *interaction* variable indicates whether the observation corresponds to an *impression* or an actual *conversion*. The *conversion* variable is binary, with a value of 1 indicating a conversion and 0 indicating only an impression of the media channel. The *conversion value* is a numerical representation of the impact of a specific conversion on a business. However, for this dataset, the specific impact referred to by the "conversion value" is unclear, and therefore it is not used for analysis. Lastly, the *channel* variable encompasses five distinct media channels: Facebook, Paid Search, Online Video, Instagram, and Online Display.

Figure 1 illustrates a histogram depicting the distribution of interactions across different channels. Notably, Facebook exhibits the highest number of interactions, accounting for approx-

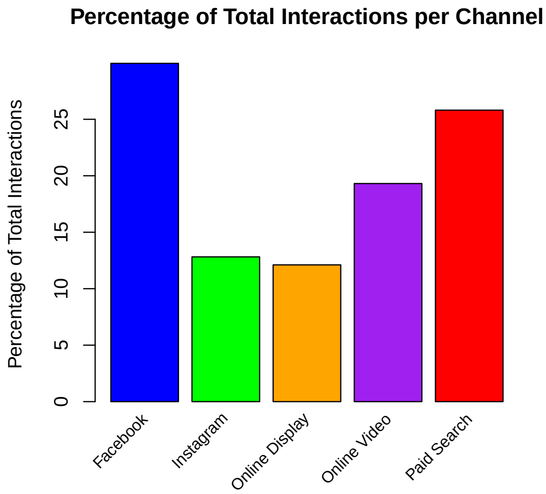

Figure 1: Histogram of the interactions per channel in percentage

imately 30.0% of the entire dataset. Paid Search follows closely with 25.8% of the interactions. These two channels demonstrate a wider range of interaction values compared to the less active channels. Online Display has the lowest interaction percentage at 12.1%, while Instagram is slightly higher at 12.8%. Online Video performs moderately, representing 19.3% of the total number of observations in terms of interactions.

In Table 1, the statistics for each channel are presented to provide an overview of their performance. These statistics include the total number of interactions, interaction rate (percentage of total interactions per channel), total number of conversions, and conversion rate per channel. The total number of conversions represents the interactions that resulted in a conversion for each channel, while the conversion rate indicates the percentage of interactions of a specific channel, that led to a conversion.Remarkably, the percentage of interactions that lead to a conversion is consistently around 3% across all five channels. This observation suggests that it's of interest to focus on channels with a high number of total interactions, such as Facebook and Paid Search, in terms of maximizing conversion outcomes.

Table 1: Descriptive statistics of the *Attribution Data* per channel

Media Channel	Total Interactions	Interaction Rate	Total Conversions	Conversion Rate
Facebook	175741	30.0	5301	3.02
Instagram	75201	12.8	2244	2.98
Online Display	71053	12.1	2139	3.01
Online Video	113302	19.3	3408	3.01
Paid Search	151440	25.8	4547	3.00

3.2 Data Transformation

In order to use the data set for this research, we modify the data using the coding language R. First, the data transformation of the Markov Model will be discussed, followed by the data transformation of the Random Forest.

To implement a Markov Model, consumer paths for all users in the dataset need to be constructed. The different users are recognized by the *cookieID*, which serves as a unique user ID, and can subsequently be ordered by *time stamp*. An example of a consumer path could be $(Facebook > PaidSearch > Instagram > Conversion)$, however, paths could also lead to non-conversion. The modified dataset contains 11,374 unique path instances, which lead to 17,639 conversions in total. One should keep in mind that each specific path instance could possibly lead several users to conversion. Each instance consist of a sequence of channels, a "path", and is accompanied by three variables: *total conversions*, *total null*, and *total conversion value*. The first two attributes represent the number of users that either made a conversion or did not, respectively. The last attribute denotes the total value of conversions made by consumers following that particular path. However, as mentioned earlier, *total conversion value* was not used for the analysis of this research.

To facilitate the implementation of the RF model, several other data preprocessing steps are undertaken. Firstly, five new categorical variables are created to serve as explanatory variables in the model, representing the five media channels. Following this, the number of interactions of each media channel is recorded for each unique *cookieID*. Finally, a column is created to aggregate and track the total interactions per channel. The instances which only have 1 total interaction, or more than 15 interactions, are deleted. The reason behind this is that the time required to calculate Shapley Values when a lot of variables are involved (long consumer paths), can lead to a significant practical drawback of the approach. This problem can be mitigated by decreasing the path length. Filtering the data like this does not disregard large amounts of data, as presented in Figure 2. Moreover, paths of length of 1 are removed. Findings by Anderl et al. (2016), Abhishek et al. (2012), and Berman (2018) suggest that one-touch-paths are not suitable to determine the attribution of channels, or derive carry- and spillover effects between channels.

4 Methods

In this section, the used methods for this research are discussed. Starting with the Markov Chain (MC), the Random Forest (RF) and the moving RF, which are used to predict the outcome of a conversion path. Followed by an outline of the different attribution metrics, Last Touch Attribution (LTA), Uniform Weight Attribution (UWA), Incremental Value Attribution (IVA), and two different approaches to compute the Shapley Value (SV). Lastly, an evaluation method to compare the predictive models is described.

Multiple interactions are involved in the consumer path to conversion. In the setting of this paper, interactions are seen as an engagement (impression or conversion) that occurs when a user comes across an advertisement displayed on a specific channel, including Facebook, Instagram, Paid Search, Online Video and Online Display. Three predictive models are employed to utilize these channel interactions as features for predicting conversion. The heuristic metrics discussed in Section 2.3.2 are non-probabilistic measures. Non-probabilistic metrics do not rely on historical data used for predictive models to estimate the probability of conversion in order

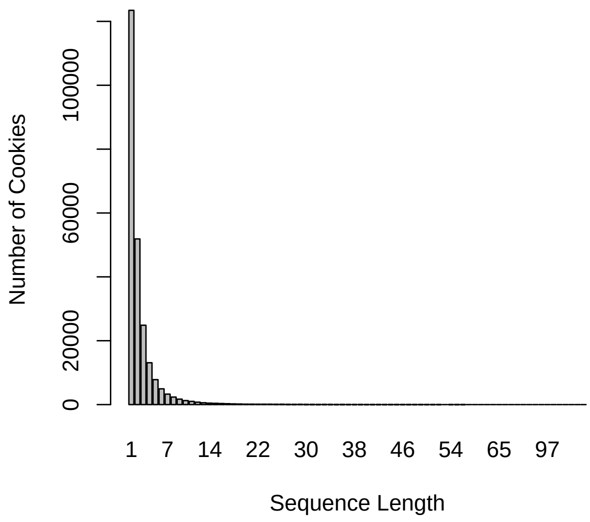

Figure 2: Histogram of consumer path sequence lengths

to attribute value to media channels. Hence, they are not dependent on any predictive model. On the other hand, the algorithmic metrics, namely IVA and SV, are data-driven metrics that utilize predictive models. Consistent with the findings of Singal et al. (2019), both SV and IVA are computed based on a MC. This research contributes to the existing literature in two main ways. Firstly, it involves the implementation of a traditional RF model and a modified version called the moving RF, which will be compared to the predictive MC model. This comparison allows for a comprehensive evaluation of the different modeling approaches. Secondly, the research explores the application of SV attribution using the traditional RF model. In addition to the standard SV, an asymmetric SV is also utilized to account for the time sequence present in the consumer paths. Preserving the order of user paths is essential, as it is a fundamental characteristic of the MC model. To address this, both the adjusted asymmetric SV and the moving RF, which consider the sequential nature of the data, are tested.

4.1 Predictive Models

4.1.1 Markov Model

Described as the probability to transition from one state to another, the Markov Chain can be characterized by three components. First, the set of states $S = s_1, ..., s_n, s_c, s_{nc}$, which can be characterized as the interaction with all 5 media channels C, and also includes the conversion state c and the non-conversion state nc. Second, the transition matrix T, which indicates the probability of moving from one state to another. $t_{i,j}$ presents the probability of moving from state i to state j, where $0 < t_{i,j} < 1$, and $\sum_{j=1}^{n} t_{i,j} = 1$ for all i. The final necessary component is the arrival state, X_0. Consumer paths can thus be constructed by multiplying the arrival state X_0 with the transition matrix T, resulting in the consumer path X_0, X_1, X_2, ..., X_{t-1}, X_t over time. In line with Singal et al. (2019) it's assumed that each row of matrix T is absorbing, which means that each path either leads to conversion or non-conversion. Furthermore, because a path does not begin with either conversion or non-conversion, s_c and s_{nc} in X_0 are equal to zero. Different orders of Markov models indicate the number of prior observations that have had an impact on the current state, and that impact the transition to the following state. For the purpose of this research the objective is to replicate some of the results by Singal et al. (2019). They propose a Markov model in which the transitions only depend on the current state, in other words, a model of order 1. This leads to the following 1st order Markov model, where X_t is a random variable at time t, s_j, $s_i \in$ S, and the probability of moving from state i to state j can be denoted by:

$$t_{i,j} = P(X_t = s_j | X_{t-1} = s_i) \tag{1}$$

4.1.2 Random Forest

Random Forests (RF) are desirable machine learning models, popular due to their high prediction accuracy, robust to outliers and noise, short computation time, and helpful internal estimates of error, correlation, variable importance and strength (Breiman, 2001). The objective is to create a number of rather small decision trees, and subsequently take the average over all trees to use for the final prediction of instance **x**, where **x** represents the input values of the used features. The advantage of RF is that the path dependent structure of usual decision trees tends to overfit the data and leads to a very high variance. By taking the average of multiple trees, the danger of overfitting the data is overcome. This paper follows the methods outlined in Lechner and Okasa (2019). In addition to averaging over multiple trees, the Random Forest algorithm utilizes the bootstrapping technique. This technique involves randomly sampling the training data with replacement to create B bootstrap samples $B_i(X_i, Y_i)$ of size N_i, replicating the sample distribution. For each of these samples, a decision tree $\hat{T}_b$ is constructed by iteratively splitting the leaves. However, instead of considering all variables for each split, only a subset of the variables is randomly chosen. This process continues until a minimum leaf size or minimum leaf variance is reached. Finally, the predictions of all the trees are averaged to obtain the final result. Mathematically, this can be expressed as follows, where $\hat{T}_b(x)$ represents a single decision tree:

$$\hat{RF}^B(x) = \frac{1}{B} \sum_{b=1}^{B} T_b\hat{(}x) \tag{2}$$

With,

$$\hat{T}_b(x) = \frac{1}{|\{i : X_i \in L_b(x)\}|} \sum_{i:X_i \in L_b(x)} Y_i \tag{3}$$

Where $L_b(x)$ presents a leave for instance x, and Y_i is a response variable.

4.1.3 Block Bootstrapping

A limitation of the application of Random Forests in the prediction of consumer paths, is their inability to incorporate the sequence dependency of channels in the consumer paths. This leads to a significant loss of information loss, as the timing of channel exposure impacts the probability of conversion. Goehry et al. (2021) propose a solution to this problem with an innovative machine learning technique used for the construction of random forests, called *Block Bootstrapping.* The concept involves replacing the conventional bootstrap with a dependent bootstrap, known as a block bootstrap, to sample subsets of time series data during the construction of decision trees. With the objective of creating user paths, the moving block bootstrap (MBB) technique will be utilized, which chooses blocks of consecutive observations at random. Each block represents an individual consumer path. In line with the methods discussed in Goehry et al. (2021), let $B_{i,l_n} = ((X_i, Y_i), \ldots, (X_{i+l_n-1}, Y_{i+l_n-1}))$ denote a block of size l_n starting with the observation (X_i, Y_i), where $i \in (1, \ldots, n - l_n + 1)$. The procedure involves randomly selecting K indices $(I_j)1 \leq k \leq K$ uniformly from the set $1, \ldots, n - l_n + 1$ and linking one block with each index. This results in $(B_{I_1}, \ldots, B_{I_K})$ as the bootstrap set, denoted as $D_n = (B_1, \ldots, B_k)$. Consequently, the bootstrapped set is used for the construction of the trees of the RF in the conventional way. The RF with the moving block bootstrap will be called the moving RF.

4.2 Attribution Metrics

4.2.1 Heuristic Attribution Metrics

In this section the methods used to compute the Last Touch Attribution (LTA) and the Uniform Weights Attribution (UWA) will be discussed. Equation 4 and 5 describe the value π_r generated by channel r according to attribution metrics UWA and LTA, respectively.

$$\pi_r^{UWA} = \frac{1}{n} \sum_{i \in M} \frac{n_{i,r}}{|R|} \tag{4}$$

Here, n denotes the total number of paths. $n_{i,r}$ represents the number of times channel r occurs in path $i \in M$, where M is the set of paths that lead to conversion. Lastly, $|R|$ denotes the number of channels in path i, which are not necessarily unique.

$$\pi_r^{LTA} = \frac{\phi_r}{n} \tag{5}$$

In equation 5, ϕ_r is equal to 1 if channel r is the last channel in the converted path $i \in M$. Equation 4 and 5 are in line with the notations of Singal et al. (2019).

4.2.2 Incremental Value Attribution

The Incremental Value Attribution metric or *Removal Effect*, is an attribution metric commonly employed in conjunction with Markov Chains, as demonstrated by Anderl et al. (2016). According to their methodology, the IVA is determined by the the shift in conversion probability when a specific channel r_i is eliminated from the i-th consumer path. This reflective analysis distinguishes the IVA compared to other methods as it takes into account counterfactual information. Following the approach outlined by Anderl et al. (2016) the computation of the IVA a given channel r_i can be expressed as follows:

$$\pi_{r_i}^{IVA} = Visit(r_i) * Conversion(r_i) \tag{6}$$

Where $Visit(r_i)$ is the probability of passing through channel r_i on a random walk, and $Conversion(r_i)$ is the probability of conversion from a give channel r_i.

4.2.3 Shapley Value

As described in Section 2.3.2, according to Dalessandro et al. (2012), Shapley Values, a concept rooted in cooperative game theory, are frequently employed to distribute the value created by a coalition to its members. Specifically, in the context of this study, SV refers to the change in attribution value related to a certain channel. The Markov model will be interpreted using the conventional SV, which is mathematically notated in this subsection. The set of players (channels), denoted as C in this research framework, reflects interactions with numerous channels that work together to achieve a single goal: user conversion. The coalition is represented by H and refers to a group of participants in C. The characteristic function $v(k)$ maps a coalition $HsubseteqC$ to a numeric value, where the value reflects the conversion attained by customers interacting with channels in H. The value assigned to an empty coalition, denoted as $v(0)$, is normalized to 0. For a given channel r, the Shapley Value can be mathematically expressed as follows:

$$\pi_r^{SV} = \sum_{H \subseteq C \setminus \{r\}} w_{|H|} \times \{v(H \cup \{r\}) - v(H)\} \tag{7}$$

where

$$w_{|H|} = \frac{|H|!(|C| - |H| - 1)!}{|C|!} \tag{8}$$

SV have several advantageous characteristics, which make it a well considered attribution method. This research follows the mathematical framework of Singal et al. (2019) to outline the properties. The properties of the SV are listed below:

1. *Symmetry*: If $v(H \cup r) = v(H \cup r')$ then $\pi_r = \pi_{r'}$.
2. *Efficiency*: $\sum_{r \in C} \pi_r = v(C)$.
3. *Null Player*: If $H \subseteq C$, $v(H \cup \{r\}) = v(H)$ then $\pi_r = 0$.
4. *Linearity*: Consider two characteristic functions s(.) and v(.). For all channels $r \in C$, it holds that $\pi_r(s + v) = \pi_r(s) + \pi_r(v)$ and $\pi_r(bs) = b\pi_r(s)$ for all $b \in R$.

The SV is the only metric that combines these four properties. Due to these inherent properties, SV are commonly regarded as a *fair* attribution method. The detailed proofs of these properties can be found in the work of Singal et al. (2019)

4.2.4 Shapley Additive Explanations for Average Attributions (SHAPAA)

Addressing the challenge of interpretability in complex Machine Learning models, Lundberg and Lee (2017) developed the SHapley Additive exPlanations (SHAP) model. This framework, rooted in cooperative game theory and inspired by Shapley Values, provides a comprehensive approach for understanding predictions in various predictive models, including those based on decision trees. In this study, the SHAP model is employed to quantify the contribution of each media channel to conversion events predicted by the traditional RF model. Expanding on this work, Biecek and Burzykowski (2021) extended the methodology to handle the sequential nature of datasets. Their objective was to address the challenge posed by ordered data by computing the average attribution value of a feature across all possible orderings or a substantial number of them, resulting in the SHAP for Average Attributions (SHAPAA) method. This technique overcomes the limitations of traditional RF models (with standard bootstrap), which are unable to capture the sequential information that can be captured by Markov Models. Therefore, given the characteristics of our dataset, SHAPAA is a suitable approach for disregarding the impact of the ordering of dierent channels within the user path, to interpret the traditional Random Forest.

Biecek and Burzykowski (2021) propose to mathematically express the SHAPAA value for a particular instance x_* of predictive model $f()$ as follows:

$$\Theta(x_*, j) = \frac{1}{p!} \sum_J \Delta^{j|\pi(J,j)}(x_*) \tag{9}$$

In simple terms $\Theta(x_*, j)$ represents the average of the variable-importance measures over all potential orderings of explanatory variables. The covariate ordering is represented by J, which is a permutation of the set of indices $(1, 2, ...p)$. $\pi(J, j)$ represents all the indices which are positioned before the variable with index j in the ordering. $\Delta^{j|\pi(J,j)}$ denotes the deviation in expected model prediction between the expected prediction when setting the covariates of the indices of set $j \cup \pi(J, j)$ to (x_*) and the expected prediction when only the covariates corresponding to the index j are set to (x_*). Fortunately, this SHAPAA approach inherits all the desirable properties of the usual Shapley Value *(Symmetry, Linearity, Efficiency, and Null Player)*.

4.2.5 Asymmetric Shapley Values (ASV)

While Biecek and Burzykowski (2021) address the ordering within the data by averaging over all possible orderings, Frye, Rowat and Feige (2020) propose a different approach to handle the ordering issue. Instead of ignoring, as in traditional Shapley Values, or removing the sequential structure of the data, as in the SHAPAA method, they introduce a more flexible framework called the Asymmetric Shapley Value (ASV). This framework allows for incremental interpretations in time-series models by considering the sequential nature of the data. In the context of this research, the ASV is another interesting method to interpret the traditional RF model while accounting for the order of the consumer paths. The consumer paths are constructed based on time notations in the variable *time*. This causal information can be used to implement the ASV framework.

The general idea of the ASV is to relax one of the four properties of the SV: *Symmetry*. Frye et al. (2020) argue that imposing this condition of *Symmetry* in model interpretability, can obscure established sequential relationships present in the data. For this study it is important to preserve the structural information in the data, where the interaction with one media channel may be influenced by another. For instance, a user's exposure to a Facebook ad may trigger a search for the product online, resulting in a Paid Search interaction. Frye et al. (2020) demonstrate that by disregarding the symmetry condition, the uniform distribution $w(\pi)$, which is placed over the orderings π in the traditional SV, is no longer valid. To develop a clearer understanding of the distribution selection $w(\pi)$ in the ASV, it is important to note that $w(\pi)$ assigns non-zero weights exclusively to permutations where time t_i precedes t_j. This creates a causal understanding of the model. Specifically, the ASV estimates how variable x_i, where x_{ti} is an observation in the data at time t for channel i, influences the model's output when variable x_{tj} is unknown. Conversely, the j-th ASV assumes that x_i is already known and estimates the impact of x_j.

This formulation of the ASV, proposed by Frye et al. (2020), can be expressed as follows:

$$w(\pi) = \begin{cases} 1 & \text{if } \pi(t_i) < \pi(t_j) \text{ for all } t_i < t_j \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

In this research, the assumption is that if an interaction with channel i occurs before an interaction with channel j, this observation will be assigned a weight of 1. By considering the sequential order of interactions, the ASV benefits from additional information that enhances its attribution accuracy. Through their study, (Frye et al., 2020) discovered that Shapley Values distribute importance across the entire user path, while ASVs tend to concentrate feature importance at the beginning of the user path. The package *shapFLEX* in R, can be used to implement the ASV model and incorporate sequential constraints in the model algorithm.

4.3 Performance Evaluation

After implementing different models to construct the consumer path to conversion or non-conversion, it is of interest to compare the model fit of the predictive models. Specifically, it is evaluated which model is best in predicting whether a user converts to buying a product. The AUC value under the ROC curve is a common metric for binary prediction problems .(Gonçalves, Subtil, Oliveira & de Zea Bermudez, 2014). It reveals the degree to which the model can distinguish between the two groups, conversion and non-conversion. Based on the research conducted by Huang and Ling (2005), it has been established that the AUC value is a more reliable measure than accuracy for comparing models. The superiority of AUC has been demonstrated both theoretically and empirically. Therefore, AUC is considered a suitable evaluation metric for this research. For the purpose of this research, the methodology proposed by Gonçalves et al. (2014) is adopted to plot the the Receiver Operating Characteristic (ROC) curve. The ROC curve requires two values: the True Positive Rate (TPR) on the y-axis and the False Positive Rate (FPR) on the x-axis. TPR represents the model's sensitivity or its ability to correctly predict instances of the conversion class. A higher TPR indicates better model performance. The calculation of TPR is as follows, where TP denotes True Positive and FN denotes False Negative:

$$TPR = \frac{TP}{TP + FN} \tag{11}$$

The FPR represents the specificity of the model and describes how well the model predicts the observations falling within non-conversion category. The lower the FPR the better the model performs. The FPR can be computed the following, where FP presents False Positive and TN presents True Negative:

$$FPR = \frac{FP}{TN + FP} \tag{12}$$

The ROC curve is constructed by plotting FPR on the x-axis the TPR on the y-axis. A diagonal line at a 45-degree angle represents the performance of a random classifier. The higher the ROC curve is positioned above this diagonal line, the better the model performs in distinguishing between positive and negative cases. However, in this research, the ROC curve itself is not presented as a result. Instead, the focus is on the area under the curve (AUC), which provides valuable information for model evaluation. The AUC value represents the probability that a randomly chosen positive observation will be correctly predicted as a positive case by the model (Fawcett, 2006). AUC values range from 0 to 1, where a value of 1 indicates a model that correctly predicts all outcomes with a probability of 1, and a value of 0 represents a model that incorrectly predicts outcomes with a probability of 1. A random prediction has an AUC value of 0.5, indicating no predictive power.

5 Results

In this section the results will be presented. First, the model fit of the predictive models will be discussed, followed by the results of several attribution metrics.

5.1 Predictive Models

The results of the prediction models that were used to simulate the consumer path are covered in this section. We discuss the model fit of the commonly used Markov model, which serves as the benchmark model, and the fit of 2 different Random Forest models, the traditional RF and the moving RF. The evaluation metric that will be used is the AUC value, as discussed in Section 4.3.

The *ChannelAttribution* package in R was employed to estimate the transition probabilities for the Markov Chain order 1, based on 11,374 path instances from the dataset. Subsequently, the AUC value is computed with the application of this package. The RF model is implemented using the *Ranger* package, and the moving RF is implemented by means of the *Rangerts* package. The *Rangerts* package can be applied in the exact same way as the *Ranger* package, however, this version is modified to implement the RF algorithm using the block bootstrapping technique. Besides that, all used techniques and chosen parameters are similar.

Due to the highly unbalanced nature of the dataset, where only 3% of the observations lead to conversion, training a decision tree or any other machine-learning model becomes challenging. A tree will simply predict all observations as a non-conversion event, since this leads to a desirable prediction accuracy of 97%. To mitigate the adverse effects of this class imbalance, an aggregated random undersampling technique is employed. The decision to utilize random undersampling is supported by the research conducted by Mishra (2017), who compared various techniques for addressing class imbalance in classification problems. Their study revealed that random undersampling consistently outperformed other methods, including the oversampling technique known as SMOTE. In this study, aggregated propensity score matching was also examined, but it exhibited poorer performance than aggregated random undersampling, as illustrated by an AUC value of 0.48. The objective of random undersampling is to create a representative sample that reduces the dominance of the majority class (conversion = 0). A 1:1 ratio of the target variable *conversion* is chosen for each sample. Instead of discarding data, the process is iterated randomly ten times, and the resulting ten samples are aggregated to construct a new dataset. This process ensures that the resulting dataset is balanced, with an equal ratio of instances representing conversion and non-conversion classes. Subsequently, 80% of the data is used for training purposes and 20% is used to test the predictive performance of the model on unseen data. To find the number of trees that corresponds to a stable classifier, RF with different *ntree* values (100, 500 and 1000) are built. As the out of bag (OOB) error rate is similar for a RF with 500 and 1000 trees, a less complex model of *ntree*=500 is chosen. To find the optimal value of the *mtry* parameter, which corresponds to the random subset of variables used to base the splits on in a tree, 5 fold cross validation is performed. This resulted in a value of *mtry*=2.

In the traditional RF approach, each row in the dataset represents a unique *cookieID* and the columns represent the interactions associated with that user. This means that each consumer path is represented as a single row. However, when implementing the moving RF, the data needs to be transformed in a slightly different way compared to what was discussed in Section 3. Instead of structuring the data per *cookieID*, the different time stamps for each *cookieID* are preserved as separate rows. The dataset is arranged by *cookieID* and by time, resulting in a consumer path being represented as a block of rows. After processing the dataset accordingly, and yet again taking care of the class imbalance, the distributions of total interactions per cookie can be found in Figure 3.

In contrsast to the tradition RF, the moving RF requires the block length to be specified. Optimally, this block length should be determined based on the length of each consumer path. However, the employed R package doesn't allow to adjust the block length for each block, according to the length of each path. Constructing the bootstrap and tree algorithm manually has been attempted, however, different block lengths for each cookie led to intractable computation times. Based on the result, illustrated in Figure 3, cookies related to a consumer path larger than 1, and smaller than 10, are discarded. The maximum path-length is smaller compared to the restriction discussed in Section 3.2. This results in less varied path lengths, which contributes to the accuracy of the moving RF model. Furthermore, visual inspection of Figure 3 shows that by filtering accordingly, most data observations are preserved. Goehry et al. (2021) claim that the best rule for selecting a block length parameter, would be to take the least periodicity up to a multiplication factor of two or three. As the least periodicity of a consumer path is 2, a block length of 6 is chosen. Finally, table 5.1 gives the AUC values for all three models.

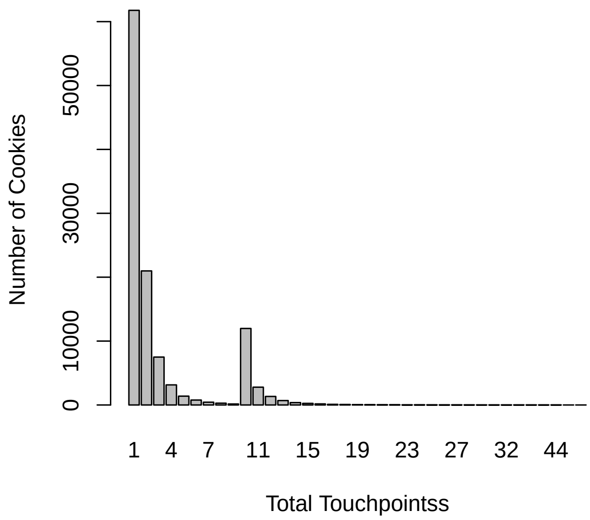

Figure 3: Histogram of consumer path sequence lengths after data preprocessing

Table 2: AUC values for the Markov Chain model, the Random Forest model, and the moving Random Forest model

Model	MC	RF	M-RF
AUC	0.52	0.60	0.52

As already discussed in Section 4.3, the higher the AUC value, the better the model is able to predict the right outcome, where he AUC value of a random prediction is 0.5. Hence, as presented in Table 5.1 the performance of the MC model (0.52) and the moving RF (0.52) is only slightly better than a random classifier. An explanation for the low AUC value of he MC model, which is 0.08 lower than the traditional RF, could be that the order (1) used for the model, in line with the study of Singal et al. (2019), is too low. The higher the order of a MC, the more information is used, which increases the predictive performance. However, the higher the order of the Markov model, the more parameters the model needs to estimate. This increases the complexity, and leads to overfitting of the model. The low AUC value of the moving RF was to be expected concerning the length of the blocks chosen for the moving block bootstrap. The moving RF classifier would more appropriate when clear seasonality is depicted in the dataset, leading to a constant block length. As well as when the block length would alter based on the the number of the total interactions per cookie. However, this leads to a very complicated model which results in high computations times. Hence, due to time constraints this could not be implemented.

It is observed that the traditional RF has an AUC value of 0.6. This is only somewhat improved compared to a random prediction, and would be labeled as poor discrimination according to the standards of Hosmer Jr, Lemeshow and Sturdivant (2013). The inadequate performance of the RF model might be the result of the inability to consider the sequential nature of the data set. The order in which the interactions occur, have an effect on the event of conversion. To illustrate, consider a scenario where a user interacts with a Paid Search advertisement after already being acquainted with a product. In such cases, the impact of the advertisement is likely to be more significant compared to when the ad is presented at the initial stages of the consumer path. Additionally, it is worth noting that the relatively lower AUC value obtained may be attributed to the generation of synthetic data during the random undersampling procedure. Nevertheless, considering that the outcomes of these models do not have life-altering implications and as long as the performance exceeds that of a random classifier, a value of 0.6 can still provide utility. Moreover, the traditional RF model outperforms the MC model, which serves as the benchmark, and thus appears to be the most suitable approach for modeling the consumer path towards conversion.

5.2 Attribution Measures

This section presents the results of various attribution metrics used in the analysis. In the Section *Replication*, the findings of four different attribution metrics are discussed: Uniform Weight Attribution (UWA), Last Touch Attribution (LTA), Incremental Value Attribution (IVA), and Shapley Value (SV). The IVA and SV metrics are computed using Markov Chains. The *ChannelAttribution* package is utilized to calculate the transition probabilities. These transition probabilities are then used to conduct a simulation study of 100,000 iterations to create path instances. These path instances can be used for the SV and IVA computation. Subsequently, the Section *Extension* presents the outcomes of applying the conventional Symmetric Shapley Values (SSV) and Asymmetric Shapley Values (ASV) to a traditional RF model. A comparison is made between these results and the Shapley Values obtained from the benchmark MC model.

5.2.1 Replication

Table 3: Attribution per channel by different attribution metrics in percentages

Media Channel	UWA	LTA	SV-MC	IVA-MC
Facebook	29.6	30.0	26.4	27.9
Instagram	12.8	12.7	16.7	20.6
Online Display	12.0	12.0	12.9	12.7
Online Video	19.0	19.7	17.5	15.9
Paid Search	26.5	25.6	26.6	22.9

Table 5.2.1 provides an overview of the attribution results obtained from the four attribution metrics for each channel. To facilitate comparison, all results have been normalized by dividing them by the total attribution value, and percentages have been computed. Notably, Facebook is considered as the most prominent channel across all attribution metrics. This observation is unsurprising considering the high prevalence of Facebook interactions in the dataset, as illustrated in Figure 1, and the highest number of total conversions per channel, as presented in Table 1. It is interesting to note that SV allocates the least (26.4%) attribution to Facebook, followed by

IVA (27.9%). This is expected since SV takes into account the fact that certain paths contain channels that appear more than once, by only considering unique channels (Singal et al., 2019). Furthermore, Online Display receives the least attribution across all four metrics. This finding aligns with the relatively low number of interactions associated with the Online Display channel. Remarkably, SV attributes the highest value (12.9%) to Online Display, closely followed by IVA (12.7%). These results demonstrate the ability of SV and IVA to allocate value to a channel, even in cases where an interaction does not directly lead to a conversion. Both metrics acknowledge the value of contributing to the awareness of a product. In contrast, LTA only assigns credit when the interaction represents the final touchpoint leading to a conversion, what Facebook differentiates.

In order to visualize the attribution of each channel under several attribution metrics, Figure 4 is presented. The first thing to note is that LTA and UWA show almost identical results and follow the exact same pattern as the total number of interactions, displayed in Figure 1. This emphasizes the inadequacy of these metrics to provide insights on how effective an ad-campaign is, apart from the number of users it reaches. In general, all metrics follow a roughly similar pattern as the distribution of total interactions. This observation could also be attributed to the AUC results regarding the model fit, which revealed that the predictive accuracy of the models was only slightly better than that of a random predictor. It could suggest that the media channels had limited influence on the likelihood of conversion, implying that the effectiveness of the channels may have been relatively low. Consequently, it can be assumed that the marketing channels did not exert a significant impact on the conversion process, and therefore, none of the channels show attribution scores that significantly exceed their occurrence rates.

The computation of SV and UWA is very comparable, only SV corrects for the fact that some channels appear multiple times in one path. Compared to UWA, SV assigns significantly more value to Instagram (3.9%), and less credit to Facebook (3.2%). The SV computes the marginal contribution of each channel to evaluate its contribution to conversion. Hence, this indicates the impact of both Instagram on the probability of conversion is high, and the impact of Facebook is low, relative to their occurrence rate. IVA is the most conflicting measure when comparing it to the total number of interactions (Figure 1), even more than the SV results. One evident outlier is the value (20.6%) IVA assigns to Instagram, and the value (22.9 %) it allocates to Paid Search. Singal et al. (2019) find that IVA gives higher credit to channels that have a higher appearance rate, due to the fact that it's scaled by action intensity. This is exactly opposite to the results presented in this research. However, Singal et al. (2019) also claim IVA is an unjustifiable method as it generates more value than the system can distribute. Hence, this may be the reason for the conflicting results compared to other metrics.

5.2.2 Extension

Table 5.2.2 provides the Shapley Value outcomes obtained from the traditional Random Forest model. However, the results for the SHAPAA method, which eliminates the ordering of the paths by averaging over the all the possible orders of the SV, are not included in the Table. The computation time required for this method, exceeded several days, even when only using a small number of test observations. The Shapley value is essentially the average marginal contribution

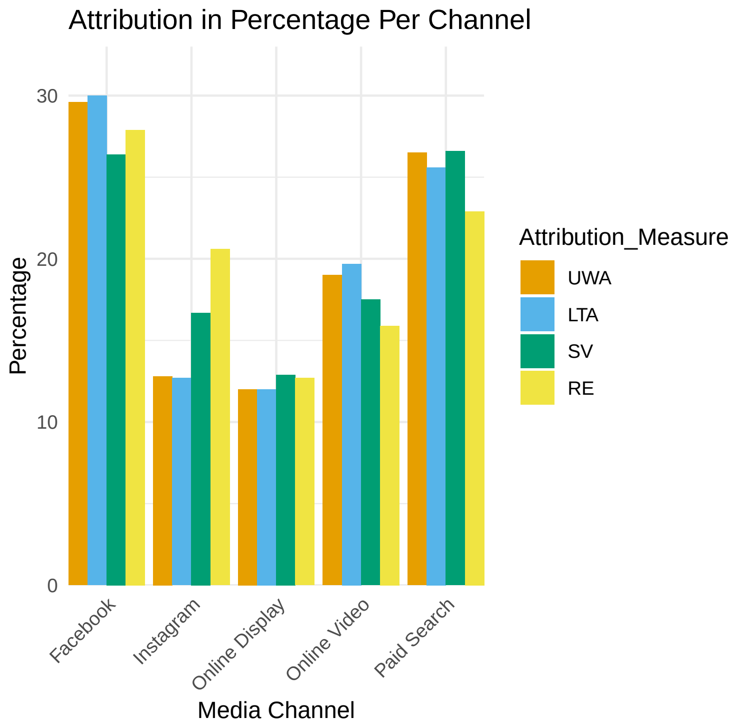

Figure 4: Histogram of attribution across 5 channels in percentages: a comparison of 4 attribution metrics

of a variable when all feasible combinations are taken into account. This already results in a lot of computations. When the order also has an impact on the outcome, this estimation procedure becomes even more tedious. Besides estimating all possible feature combinations, averaging over all possible orderings of these combinations is added to the analysis. For a considerably large dataset, this execution becomes extremely time consuming, and unrealistic. Consequently, due to time constraints and the unavailability of a GPU processor, the implementation of the SHAPAA method was not feasible. Moreover, it can be concluded that the SHAPAA method poses significant computational challenges, and is thus not a suitable method for the attribution framework. Despite its ability to address the path sequence issues, it does not outperform the effectiveness of the Markov Chain combined with the Shapley Value. Even though the RF is able to handle a high dimensional covariate space, the SHAPAA used to provide interpretation to this black box, becomes the bottle neck.

As the traditional RF model outperformed the moving RF and the MC, it will be utilized for further research. Subsequently, the Asymmetric Shapley Values (ASV), as described in Section 4.3, and also the conventional Shapley Value, which will be denoted as the Symmetric Shapley Value (SSV), are computed based on the traditional RF. The ASV and SSV are displayed in Table 5.2.2. As proposed by Molnar (2022) Shapley Values based on tree based models, can be viewed as 'forces'. Each variable acts as a force to either increase or decrease the prediction. Hence, a positive SV indicates that a feature influences the model to predict conversion=1, and a negative value indicates that a feature influences the model to predict conversion=0. The SSV and the ASV yield significantly different results. Specifically, the SSV assigns a negative value to Instagram, suggesting a negative impact of interactions with Instagram advertisements on users'

Table 4: Average Symmetric Shapley Value (SSV) feature effect, average Asymmetric Shapley Value (ASV) feature effect, and the ASV in percentages across 5 media channels

Media Channel	SSV	ASV	ASV %
Facebook	0.0131	0.135	22.5
Instagram	-0.0161	0.0916	15.2
Online Display	0.0228	0.107	17.7
Online Video	0.0265	0.0695	11.5
Paid Search	0.00404	0.200	33.1

conversion decisions. This finding contradicts the results obtained from other attribution metrics discussed previously. However, it is important to note that the negative value does not imply inadequate performance of the SSV. According to the mathematical definition of Shapley Values, they can be either positive or negative. On the other hand, the ASV consistently produces positive values, indicating that the negative results of the SSV are a result of the symmetry. While negative values are mathematically allowed for Shapley Values, they are generally not desirable from a marketing perspective. Therefore, in the context of marketing, the ASV would be a more suitable metric to consider. Moreover, the SSV values exhibit significant differences when considering the overall number of interactions shown in Figure 1 and the attribution patterns presented in Figure 5.2.1. These findings indicate that the use of the traditional RF, despite its computational advantages over a Markov Chain, in combination with the SSV for interpretability, is not consistent with previous results. It appears that this discrepancy might stem from not adequately accounting for the sequential nature of consumer paths in both the RF and the SSV.

The ASV and SSV feature effects are visualized in Figure 5. The ASV algorithm allows for causal relations computing the Shapley Values, relaxing the *Symmetry* axiom of the Symmetric SV framework. This way, the sequential time structure within consumer paths can be preserved. Also, as proposed by Frye et al. (2020), a sparser explanation can be obtained. This means that this framework can prioritize the most influential features, while assigning less importance to less significant features. The histogram provides insights into the attribution assigned by the ASV and the SSV metric to different marketing channels. It should be noted that even though Figure 5 is not depicted in percentages, it can be inspected in the same manner as Figure 4. If the ASV values were scaled by the total attribution value and calculated as percentages, an exact similar Figure would have been displayed. However, since the SSV metric also includes negative values, it is valuable to illustrate these in absolute values. As discussed earlier, the SSV illustrates very different attribution patterns compared the number of interactions in Figure 1 and the attribution percentages in Figure 4. Observing the values of the ASV in Figure 5, Paid Search is considered the most valuable channel, followed by Facebook. Upon examining Figure 5, several observations can be made. Firstly, the attribution pattern of the ASV does not completely align with the distribution of total interactions per channel, as shown in Figure 1. Paid Search receives a significantly higher credit allocation (0.2) relative to its occurrence, while Facebook receives less credit than expected. This suggests that the Paid Search advertisement has a substantial impact on increasing the conversion probability by 0.2, which holds significant implications for marketers. Additionally, Instagram and Online Display are attributed slightly more credit than anticipated based on their interaction frequencies. Where Online Display has the lowest occurrence rate, it's the third most valuable channel according to the ASV. This result highlights the effectiveness of an ad-campaign based on Online Video, where the probability of conversion is increased a lot relative to its occurrence rate. Surprisingly, Online Video, which has an interaction rate of 19.3%, receives a notably low attribution value according to the ASV method. These findings emphasize the opposite as the aforementioned results. An advertising strategy using Online Video seems to be ineffective, and has a lower impact on conversion relative to its occurrence rate. This demonstrates the importance of considering the ASV results and its deviations from the expected attribution patterns when evaluating the effectiveness of marketing channels.

The allocation of attribution by the ASV method aligns more closely with the results obtained from applying the Shapley Value (SV) to the Markov Chain, which serves as a benchmark model. Both approaches rank Paid Search as the most preferred channel, however, ASV more extensively than SV. As briefly mentioned above, according to Frye et al. (2020) the ASV provides a more sparse framework. This would indicate that Paid Search is a more influential feature than considered by the SV. Furthermore, both methods assign notably more credit to Instagram, as compared to the occurrence rate of the channel. As mentioned above, this emphasizes the effectiveness of online advertising campaigns displayed on Instagram. Notably, where the SV mitigates the feature importance of Facebook compared to other attribution metrics, this result becomes even more visible through the ASV method. Where the SV allocates 26.4% of the total value to Facebook, ASV only assigns 22.5% (Table 5.2.2). Moreover, there is a discrepancy in the rankings of the least effective channels. While the SV method identified Online Display as the least effective, the ASV method indicates that Online Video shows the worst performance. Previous research by Frye et al. (2020) suggests that ASV tends to concentrate the importance of features at the beginning of a sequence, whereas SV distributes value more evenly across the entire sequence. This finding may explain why Paid Search and Online Display receive higher attribution percentages according to the ASV method. It suggests that Online Display and Paid Search are often found as an initial interaction in consumer paths that lead to conversion.

The attribution percentages allocated by the ASV method exhibit closer alignment with the results obtained from applying the Shapley Value (SV) to the Markov Chain, which serves as a benchmark model. Both approaches consistently rank Paid Search as the most preferred channel, with the ASV method emphasizing its influence even more than the SV method. This finding supports the notion proposed by Frye et al. (2020) that the ASV provides a more sparse framework, indicating a higher level of importance for Paid Search than suggested by the SV. Additionally, both methods assign significant credit to Instagram, exceeding its occurrence rate, highlighting the effectiveness of advertising campaigns on this platform. Notably, the SV method downplays the importance of Facebook compared to other attribution metrics, and this difference becomes even more pronounced with the ASV method. Table 5.2.2 shows that while the SV assigns 26.4% of the total value to Facebook, the ASV only assigns 22.5%. However, there is a discrepancy in the rankings of the least effective channels. While the SV method identifies

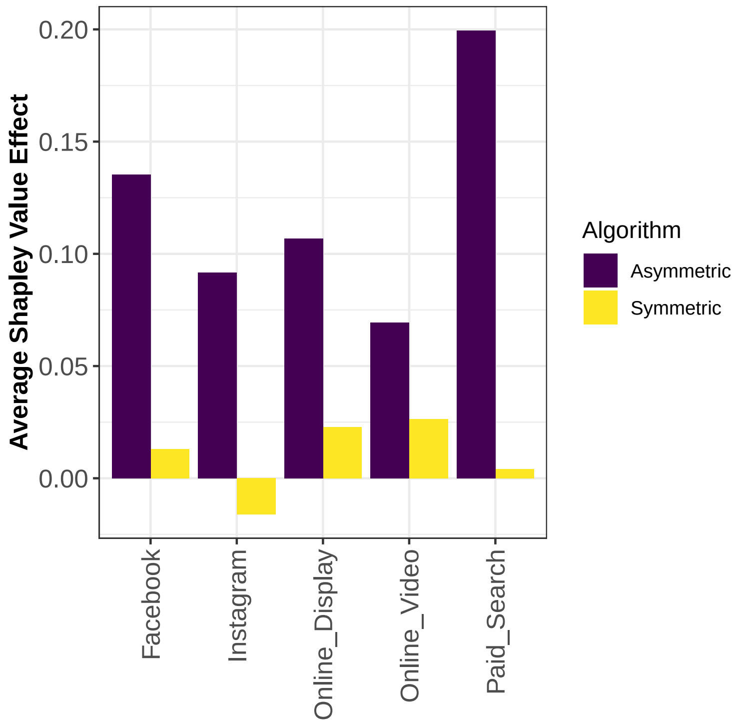

Figure 5: Histogram of the attribution value across 5 channels based on the Asymmetric Shapley Value (ASV) metric and the Symmetric Shapley Value (SSV) metric

Online Display as the least effective, the ASV method indicates that Online Video performs the worst. Previous research by Frye et al. (2020) suggests that the ASV tends to concentrate the importance of features at the beginning of a sequence, whereas the SV distributes value more evenly throughout the sequence. This finding may explain why Paid Search and Online Display receive higher attribution percentages according to the ASV method. It suggests that Online Display and Paid Search are often found as an initial interaction in consumer paths that lead to conversion.

In summary, the ASV framework provides causal and sequential explanations for the effects of features in advertising. Unlike the SSV, the ASV assigns only positive attribution values to media channels. This positive-only attribution aligns better with marketing perspectives, making the ASV a preferred method in the field. Furthermore, the ASV method assigns a significantly higher percentage of influence to Paid Search, while Online Display and Instagram also receive slightly higher credit compared to other attribution metrics. However, Online Video performs poorly in terms of attribution according to the ASV. When comparing ASV with the Shapley Value on the Markov Chain, Paid Search consistently emerges as the most preferred channel. The influence of Facebook, on the other hand, is less pronounced in the ASV results, diverging from patterns based solely on occurrence rates. The use of the ASV in conjunction with RF is computationally feasible and advantageous for companies handling multiple advertising channels. However, the variations between the ASV and the Shapley Values derived from Markov Chains raise concerns regarding the accuracy of one or both of these methods.

6 Conclusion and Discussion

This research attempted to find suitable methods to model the attribution of media channels in the online advertising industry. The commonly used methods, as demonstrated in the reference paper by Singal et al. (2019)), are inadequate for modeling channel attribution as the number of channels increases. Hence, the need to investigate the possibilities of applying Machine Learning techniques to model this framework arises. However, computational complexities arise when applying the Shapley Value as a post-hoc-interpretability method in attribution modeling. Consequently, this study investigated the following research question:

What is the most effective approach for incorporating the Shapley Value as an interpretability measure in Random Forests for attribution in online advertising, considering the inclusion of the sequential order of channels in the consumer path, and how does this approach compare to applying the Shapley Value to a Markov Chain in terms of prediction performance and attribution value?

In order to do this, two predictive models were employed and compared to the Markov Model, which served as a benchmark model. These predictive models include the traditional Random Forest and the moving Random Forest. The moving Random Forest uses an alternative bootstrapping method, the moving block-bootstrap, to secure the sequential structure of the consumer paths in the data. In order to compare the models the Area Under the Curve (AUC) value was computed for each model, as this value can be estimated for both RF models, and for the Markov Chain. The AUC values for the moving RF and the traditional RF were 0.52 and 0.60, respectively. Hence, the traditional Random Forest is found to have a better model fit compared to the moving Random Forest. The moving RF is expected to perform much better than the traditional RF, because it should represent the consumer path much better. However, the low AUC score of the moving RF can be explained by the inadequacy of the model to adjust the block-length based on the length of each consumer path. The traditional RF also outperformed the Markov Chain in conversion prediction, which presented an AUC value of 0.52. The Markov Chain's poor performance can be explained by the low order of the model. As the order of the Markov model is one, the transition probabilities are estimated only using information regarding the current state. This simplifies the already complex computation of the model, which increases the feasibility of the model. However, this results in low performance due to a lack of information. Even though the performance of the traditional RF is not outstanding, it has a representative performance compared to the Markov Chain, the moving RF, and to a random classifier.

After the evaluation of predictive models, the attribution metrics were computed. The heuristic metrics Last Touch Attribution (LTA) and Uniform Weight Attribution (UWA), and the algorithmic metrics Incremental Value Attribution (IVA) and Shapley Value (SV) were computed for the Markov model. Among these metrics, it was consistently observed that Facebook is considered the most influential media channel in terms of influencing conversions, except for the SV metric. This result aligns with the high occurrence rate of Facebook, which was the most frequently encountered channel in the data with a rate of 30%. Furthermore, this highlights the essential difference between the SV and the other metrics. The SV only accounts for the "unique" contribution of channels, which results in a lower attribution assigned to Facebook. Notably, the heuristic metrics produced similar results, that were also comparable to the occurrence rates of the media channels. This suggests that these metrics do not account for the varying contributions of the channels based on their individual properties. Interestingly, both the IVA and SV metrics assigned less value to Facebook but attributed higher value to Instagram. The algorithmic approaches, which consider the marginal contribution of each channel, emphasize the added value of Instagram and indicated the relatively ineffective influence of Facebook on conversions. However, it is worth noting that the results deviated from the results reported by Singal et al. (2019), who demonstrated that IVA assigns more value to channels with higher occurrence rates compared to SV. As mentioned earlier this finding can be explained by the fact that SV only considers the "unique" channels within each consumer path. However, in this study IVA assigns substantial value to Instagram, despite its low occurrence rate, and relatively less value to Paid Search, which has a significantly higher occurrence rate. This questions the generalizability of the findings of Singal et al. (2019), and suggests the need for further investigation of the justification of the IVA metric.

Finally, to interpret the preferred traditional RF model, two methods for computing the Shapley Value were examined in this study. Since the traditional RF model cannot account for the sequence of consumer paths, an adjusted version of the Shapley Value was employed to incorporate the time dependency. Initially, the SHAPAA method was implemented, which considers all possible channel combinations within each user path and averages out the ordering. However, this method proved to be computationally intractable, with estimation times exceeding several days for only a few observations. Hence, it's concluded the method is not suitable to model the attribution framework. Consequently, the Asymmetric Shapley Value (ASV) method was applied to interpret the Random Forest. The ASV method relaxes the *Symmetry* axiom of the conventional Shapley Value to maintain the causal structure in the data. The ASV was compared to the results obtained using the conventional Symmetric Shapley Value (SSV) applied to a Random Forest. The SSV yielded inconsistent results when compared to our benchmark model. The metric even assigned a negative attribution value to Instagram. This could question the choice to use the RF model in combination with SSV for modeling online advertising attribution, and suggests that the SSV method may not be optimal from a marketing perspective. The inconsistency compared to other attribution metrics could be explained by the fact that the order of the consumer path is not considered by this framework, which leads to a high loss of information. In contrast, the ASV method yielded more consistent results for channel attribution. It ranked the Paid Search channel as the most valuable, followed by Facebook, Online Display, Instagram, and Online Video, in that order. These rankings align with the findings of the benchmark model, using Shapley Values to interpret Markov Chains, which also preferred Paid Search followed by Facebook. However, the two frameworks also exhibited some varying results. The ASV attributed higher value to Paid Search, and less credit to Online Video. Frye et al. (2020) suggest that this variation can be explained by the fact that Paid Search and Online Display are more frequently observed at the beginning of a consumer path compared to other channels. They find that the ASV considers features at the beginning of the sequence to be more important, while the SV distributes the value more evenly. However, it is worth noting that our data does not exhibit typical time-series properties such as seasonality, trends, and cyclical fluctuations. This may limit the application of the model and lead to inaccurate results of the ASV framework. Additionally, the computation of the SV for the Markov model was based on 100,000 path simulations using Markov transition probabilities. This is different to the synthetically adjusted data sample employed for estimating the ASV. Despite both samples containing a substantial number of observations, which enhances generalizability, this variation in methodology could potentially account for the divergent results between the two metrics.

Returning to the research question, the most effective approach to incorporate SVs as an interpretability metric to Random Forests, is to use the Asymmetric Shapley Value in conjunction with the traditional RF. The ASV can determine feature importance for the traditional RF algorithm, while preserving the sequence in the data. This approach offers the advantage of accommodating a high-dimensional covariate space, unlike the combination of Markov Chain and Shapley Values. Additionally, the ASV method preserves the order of paths by eliminating the *Symmetry* property of the Shapley Value. As a result, it produces solely positive attribution values, which cannot be guaranteed by Symmetric Shapley Values. Therefore, from a marketing perspective, the ASV is a desirable and preferred metric. In terms of prediction performance, RF outperforms the Markov Chain of order one by 0.08 (AUC). For higher-order Markov Chains, this difference is expected to be smaller. However, the attribution values assigned to the five media channels by the ASV metric differ from those of the SV benchmark model. This difference can be explained by the generation of synthetic data used to train the model. However, it cannot be validated that the ASV framework accurately determines the attribution value for online advertising channels.

There are various limitations to consider on this study. First of all, the choice of an attribution framework heavily relies on the specific data and the objective of a companies' analysis. Since only the Random Forest and the Markov Model were investigated, this study does not provide a comprehensive framework that applies to all contexts where different models may be more suitable. Examples of other data-driven models available for attribution tasks are the Neural Network (as discussed by Du et al. (2019)), and gradient boosting (proposed by Kadyrov and Ignatov (2019)). Both ensemble learning models have the ability to handle a high covariate space, including a high number of channels. Furthermore, Du et al. (2019) apply the SV method that incorporates sequence-dependence of exposures to ads. Further research is needed to explore the potential of all possible predictive models and to determine the appropriate model for specific contexts. And in specific, models that are able to remain the path dependent structure.

A second limitation of this study is that the framework has been tested solely on a single dataset, resulting in company specific findings. This reduces the generalizability of research and raises concerns about endogeneity. To address these concerns, future research could explore multiple datasets and investigate pairwise correlations between variables to mitigate endogeneity issues, as demonstrated in a study by Anderl et al. (2016). Additionally, the current dataset included only 5 media channels, limiting the opportunity to test the feasibility of the framework with a larger number of channels. This presents an interesting area for future research, as the Random Forest is known for its capability to handle high-dimensional covariate spaces.

A third shortcoming of this study is the poor performance of the predictive models used for attribution, namely the Markov Chain, the Random Forest and the moving Random Forest. All models exhibited low AUC values, only slightly better than those of a random predictor. This poor performance could be attributed to either inadequate model fitting, as discussed earlier, or the possibility that the ad campaigns presented on the media channels had minimal impact on conversions. Consequently, the ability of a post-hoc interpretability method to accurately allocate attribution to the different channels is limited. To address this limitation, future research could utilize a dataset where the success of the ad campaigns has been verified.

Another limitation of this study is the lack of verification regarding the accuracy of the ASV method. While the ASV framework appears to be suitable, it produces different results compared to the benchmark model. This variance could be attributed to the use of a different sample dataset, where the sample is synthetically generated for both predictive models. Furthermore, previous research by Frye et al. (2020) suggests that the ASV provides a sparser framework compared to the Symmetric SV. However, without further verification, it is challenging to confirm the accuracy of the ASV. Future research should focus on validating the ASV framework as an interpretability method within the attribution framework. Additionally, exploring other interpretability metrics, such as the Local Interpretable Model-agnostic Explanations (LIME) method, can offer valuable insights, particularly for understanding channel attribution at the individual observation level. This approach significantly reduces computation time as it focuses on a single observation.

Lastly, the information pertaining to the dataset used in this study had certain limitations that need to be acknowledged. Firstly, the construction of consumer paths relied on the use of *cookieIDs* as identifiers. However, users may have multiple *cookieIDs* if they use various phones and computers. Secondly, the dataset did not include data on offline advertisement channels. Consequently, it remains unclear whether the ad campaign also had exposure on offline channels, which could potentially have had an impact on conversion. The incompleteness of the constructed consumer paths limits the accuracy of the attribution estimates. Furthermore, although the dataset contained information about the conversion value for each conversion, this information was not utilized in the analysis of the binary classification models. Considering that insights regarding revenues and profits resulting from conversions are valuable for companies, exploring attribution allocation based on revenue, rather than solely on the conversion event, presents an interesting area for further research.

Finally, this study has made a significant contribution to the field of attribution in online advertising. Its main objective was to identify a suitable and feasible model that offers companies a framework for determining advertising attribution. The insights gained from this research can assist marketers and advertising companies in allocating their budgets effectively and with accountability, ultimately leading to increased revenues. Based on the analysis of this specific dataset, several recommendations can be made to companies, advertising platforms, and publishers. Firstly, based on three attribution metrics (UWA, LTA, and IVA) it is recommended to prioritize Facebook, having the most impact on potential conversion. Additionally, heuristic methods tend to underestimate the value of channels that contribute to product awareness (i.e. future conversion), rather than direct conversion. The Shapley Value, on the other hand, recognizes these channels and has indicated that an ad-campaign presented on the media channels

Paid Search and Instagram marginally contribute to conversion relative to their occurrence rates. Therefore, it is advised that besides Facebook, all stakeholders allocate a higher budget to these two channels.

References

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

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.

Arava, S. K., Dong, C., Yan, Z., Pani, A. et al. (2018). Deep neural net with attention for multi-channel multi-touch attribution. *arXiv preprint arXiv:1809.02230*.

Archak, N., Mirrokni, V. S. & Muthukrishnan, S. (2010). Mining advertiser-specific user behavior using adfactors. , 31–40.

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

Biau, G. & Scornet, E. (2016). A random forest guided tour. *Test*, *25*, 197–227.

Biecek, P. & Burzykowski, T. (2021). *Explanatory Model Analysis*. Chapman and Hall/CRC, New York. Retrieved from `https://pbiecek.github.io/ema/`

Brajnik, G. & Gabrielli, S. (2010). A review of online advertising effects on the user experience. *International Journal of Human-Computer Interaction*, *26*(10), 971–997.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Cao, W., Wang, X., Ming, Z. & Gao, J. (2018). A review on neural networks with random weights. *Neurocomputing*, *275*, 278–287.

Dalessandro, B., Perlich, C., Stitelman, O. & Provost, F. (2012). Causally motivated attribution for online advertising. , 1–9.

Davis, R. A. & Nielsen, M. S. (2020). Modeling of time series using random forests: Theoretical developments.

Department, S. R. (2023). *Digital advertising spending worldwide 2021-2026.* Retrieved from `https://www.statista.com/statistics/237974/online-advertising-spendingworldwide/`

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

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

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

Frye, C., Rowat, C. & Feige, I. (2020). Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in Neural Information Processing Systems*, *33*, 1229–1239.

Geyik, S. C., Saxena, A. & Dasdan, A. (2014). Multi-touch attribution based budget allocation in online advertising. In *Proceedings of the eighth international workshop on data mining for online advertising* (pp. 1–9).

Goehry, B., Yan, H., Goude, Y., Massart, P. & Poggi, J.-M. (2021). Random forests for time series.

Gonçalves, L., Subtil, A., Oliveira, M. R. & de Zea Bermudez, P. (2014). Roc curve estimation: An overview. *REVSTAT-Statistical journal*, *12*(1), 1–20.

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

Huang, J. & Ling, C. X. (2005). Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, *17*(3), 299–310.

Kadyrov, T. & Ignatov, D. I. (2019). Attribution of customers' actions based on machine learning approach.

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

Lechner, M. & Okasa, G. (2019). Random forest estimation of the ordered choice model. *arXiv preprint arXiv:1907.02436*.

Lundberg, S. M. & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Mahima Kaushiva. (2020). *Attribution for online advertising using markov chains.* `https://github.com/MahimaKaushiva/Building-an-attribution-model-using-Markov-Chain`. (Accessed: 2023-23-06)

Mishra, S. (2017). Handling imbalanced data: Smote vs. random undersampling. *Int. Res. J. Eng. Technol*, *4*(8), 317–320.

Molnar, C. (2022). *Interpretable machine learning* (2nd ed.). Retrieved from `https://christophm.github.io/interpretable-ml-book`

Nisar, T. & Yeung, M. (2015). Purchase conversions and attribution modeling in online advertising: an empirical investigation.

Singal, R., Besbes, O., Desir, A., Goyal, V. & Iyengar, G. (2019). Shapley meets uniform: An axiomatic framework for attribution in online advertising. , 1713–1723.

Thomaidou, S. & Vazirgiannis, M. (2011). Multiword keyword recommendation system for online advertising. In *2011 international conference on advances in social networks analysis and mining* (pp. 423–427).