Comparative Study of Classification Models and the Markov Model in Attribution Modeling for Digital Advertising

Bachelor Thesis

Midas Goos 509612

June 28, 2023

Abstract

Given the substantial growth in the digital advertising market, understanding the effectiveness of different channel combinations that drive product purchases is of utmost importance. This study compares the Shapley Value results of the Gradient Boosting and Random Forest classification models with the Markovian model in the context of attribution modeling. The analysis was conducted on a subset of data from a Fortune 500 company, which included user paths and a conversion indicator. One of the surprising factors is that the top three converted paths only contained of paid search advertisements and was responsible for almost one third of all the converted paths. The Shapley Value results suggest that the Markov model is most suitable, given its alignment with the high ratio of paid search ads in converted paths, reasonable computation time, and model complexity. Among the classification models, Random Forest showed slightly better performance, primarily due to its higher attribution value for paid search ads. Evaluation metrics demonstrated similar performance in predictability for both classification models.

Contents

1 Introduction

As of 2018, digital advertising is the most dominant advertising medium, accounting for more than half of global advertising spend. The global digital advertising market is growing fast as people's behaviour increasingly shifts to the online world. According to Statista [2022], the number of internet users doubled within a decade from 2.3M in 2012 to 5M in 2022. The global digital advertising revenue of $616bn in 2022 is expected to grow to $1tn up to 2027, with an expected annual growth rate of more than 10%. As companies increasingly invest in digital channels, it becomes crucial to determine the effectiveness of their spending.

One way to measure the effectiveness of marketing spend is called attribution. Attribution involves assessing the contribution of individual advertiser actions such as emails, display ads and search ads to ultimate conversion. Attribution models play a crucial role in the realm of digital advertising by providing frameworks and rules that govern the allocation of value to specific marketing channels. These models assist advertisers in understanding the contribution of the channels towards achieving their objectives. Given the growing importance of online advertising, possessing the best-performing attribution model is of great value. Enhanced models can result in more accurate assessment of the contribution of a certain advertiser action, eventually leading to more efficient and valuable advertising methods. Therefore, it is highly beneficial to identify the most effective attribution model within the online advertising industry.

This paper is build upon previous research from Singal et al. [2019] where they examined the performance of different attribution schemes using the Markovian model. Under the several attribution schemes as the Last Touch Attribution (LTA), Uniform, the (Normalised) Incremental Value Heuristic (IVH), Shapley Value (SV) and the Counterfactual Adjusted Shapley Value, they discovered that the Counterfactual Adjusted Shapley Value outperformed the other schemes. This paper endeavours to examine how the classification models: Gradient Boosting (GB) and Random Forest (RF) perform compared to the Markovian model. In therms of classification models, existing literature suggests that the GB and RF models are the most suitable option compared to other classification models due the fact that these models use decision trees which we can interpret as user paths. However, limited research exists regarding the performance of these two models as attribution models.

To examine how the GB and RF models perform compared to the Markovian model, the results of the SV of these models will be applied on a subset of the same dataset as in aforementioned article. Besides these models, the heuristic models: LTA, IVH and Uniform attribution will also be included to provide understanding regarding the user paths and the ratios of the channels.

The SV results of the classification models and the Markovian model will be assessed for their complexity, the computation time and the attribution ratios of the dataset. Given the fact that the we deal with an imbalanced dataset, the most appropriate evaluation metrics for predictability of the classification models are precision, recall and F1-score. Therefore, in this research paper the following research question will be considered:

"How do classification models perform compared to the Markovian model in the context of Attribution modeling for Online Advertising?"

To give an answer to this research question, we first took a deep dive into the dataset where we discovered that the top three user paths only consists of paid search ads which consists of almost one third of the total converted paths. This implies that the attribution of paid search ads should be noticeably significant for all the models. Furthermore we checked the results for the heuristic models for some insights on the user paths. These were interesting due to the fact that the attribution results of the email and paid search ads where the highest. And at last, we gained the Shapley Value results for the advanced attribution models where the processing time for obtaining the results from the models was reasonable. The conclusion of this research is that the Markovian Model outperformed the classification models as the least complex model with an acceptable computation time and a sufficient attribution value for the paid search advertisements.

The outline of the paper is as follows, we will begin with describing the relevance of this paper, based on literature research. In this section we will also explain the choices of the attribution models and the evaluation methods. In the third section, an extensive analysis of the dataset will be provided, encompassing the same dataset as discussed in Singal et al. [2019]. Subsequently, the methodology will be discussed upon in fourth section, containing a comprehensive examination of the various models and the evaluation metrics. The fifth section will contain the results of the performances of both models. Finally, in the sixth section, conclusive findings will be drawn and potential options for further research will be discussed.

2 Literature

In this section, we will discuss some attribution models used in previous literature for attribution modeling in digital advertising to find which model is the most appropriate for this research. We consider the best performing models in relevant literature and argue which one suit our data best. At last, we will discuss what evaluation metrics are the most appropriate for the assessment of the performances of the models.

2.1 Attribution Models

As mentioned before, this paper is a further research, namely on the article of Singal et al. [2019] where they examined how several attribution models performed in digital advertising. These attribution models were the LTA, Uniform, IVH, SV and CASV. As mentioned in Xuhui Shao [2011], the LTA is a simple heuristic attribution model where only the last channel receives 100% of the credit. They concluded that Multi-Touch Attribution (MTA) models outperform the LTA due to the fact that they discovered that a mix of channels can lead to a conversion and not only the last one. However, the LTA as the Uniform attribution and IVH are suitable for gaining more information regarding the dataset. So, for this research, we will apply the heuristic models LTA, Uniform and IVH for comparison of the results and giving the reader some clarification about the user paths. For the most preferred and high-performing attribution models, one advanced attribution models will be chosen. For the results of these models, The SV will applied. Furthermore, in the previous paper they discussed that the attribution scheme the Counterfactual Adjusted Shapley Value (CASV) was considered as one of the best performing schemes. Unfortunately, due to the time frame constraints of this paper and the challenges associated with implementing the CASV, it is not feasible to employ this attribution scheme.

Moving on to the relevance of our paper. As suggested in the previous article, it is recommended to investigate other attribution models that may outperform the Markovian Model, as the Markovian Model is one of the traditional models that is commonly used but has his limitations, as described in DelSole [2000]. One of them is the assumption of independence from past events and another its the stationary transition probabilities, disregarding potential variations over time. By comparing and evaluating these models, a deeper understanding of the performances and effectiveness of different attribution models can be gained and potentially improved models can be identified. An important aspect of this research is that we will use the same dataset containing several components. The user path consists of the channels where eventually the consumer either quits or converts. An important fact for the literature research for attribution models and especially evaluation metrics is to mention is that the dataset is imbalanced. This means that most of the customers eventually leave the system (quit) than buy the product (convert). The total is ratio is 2.41% of converted user paths.

Given the time and size constraints of this paper, we have chosen to compare the results of the Markovian Model with two other models. We will analyse the two classification models: Gradient Boosting and Random Forests and evaluate the Deep Neural Networks.

Gradient Boosting

The first machine learning technique we will cover is the Gradient Boosting method. As described in Natekin and Knoll [2013], this method combines several 'decision trees' that collectively lead to accurate predictions. A decision tree is a tree in which each node represents an individual decision and where the path taken ultimately leads to a final decision or class label. In our case is it the user path containing the channels and where it eventually quits or converts. The prediction accuracy is gradually improved as each new tree corrects the 'errors' of the previous tree. Through gradients, the process targets areas where the model is not yet performing well. Ultimately, a strong algorithm models complex relationships and patterns in data.

There has been limited research on the application of Gradient Boosting in attribution modeling for digital advertising. For example, in the article of Kadyrov and Ignatov [2019] they compared the GB with the Markovian Model and applied it on an imbalanced dataset and used the AUC ROC curve as evaluation metric. They concluded that the GB outperformed the other models. However, as mentioned in Hand [2001] and Weng and Pong [2006] the mentioned evaluation metric is highly infeasible for imbalanced datasets so we cannot conclude if these conclusion are correctly drawn. Nevertheless, Gradient Boosting is an intriguing model to explore in attribution modeling due to its utilisation of decision trees and the optimisation process it employs. This technique takes into account the sequence of user paths and evaluates the contributions of different touchpoints. In contrast to the other models, with Gradients Boosting it is possible to assign varying attribution values to the different channels.

Random Forest

The Random Forest Model as described in Breiman [2001] is a model based upon many individual decision trees, as the Gradient Boosting. The idea of the RF model is that out of the ensemble of many decision trees the most popular, more precisely the most 'voted', outcome is chosen as the class label. By using Random Forests, significant improvements can be made on the performance metrics, due to its wide variety of uncorrelated decision trees.

In the article written by Leguina et al. [2020] were different attribution models compared, including the Random Forest model. The article showed that the model performed less than other models and we think to know what the reason is. It is mainly because their RF model did not hold the sequence of the user path into account where the ordinality plays a crucial role for attribution modeling. The (Ordinal) Random Forest Model as described in Janitza et al. [2016] may be more suitable due to the ordinal ordered nature of the data. However, there has not been much research about ordinal random forest for attribution modeling in digital advertising. It can estimate the relative contribution of different channels or touchpoints to the conversion on an ordinal scale level. We believe that it can be valuable in assigning the right value to different channels in the customer journey.

Other classification models

We have also analysed some other classification models and we discovered that the Deep Neural Networks (DNN) could also be a model that would fit for this research. In the last decade, classification models, as described in Fawaz et al. [2019], have experienced a surge in popularity. The DNN consists of many layers, in which each layer is a representation of an input domain. The advantage of DNN is that high complex problems can be solved and DNN has the ability to automatically learn abstract attribute representations from available data. Nevertheless, the DNN will not be chosen for this study. One of the reasons is that DNN lacks transparency and interpretation. DNN is often difficult to interpret (blackbox modeling) while within attribution modeling it is crucial to understand what the contribution of specific channels are to the final conversion. In addition, our problem is also not modelled on a high complexity level so the added value of solving complex problems is somewhat redundant.

Concluding from the findings of Scott M. Lundberg and Lee [2019], it was determined that RF and GB, as classification models, are the most suitable for attribution models due to their applications with tree ensembles as user paths. These models offer the ability to model interactions and patterns between different channels, allowing it to better handle complex attribution modeling scenarios. It could model the impact of different combinations of channels and the order of interactions in a flexible way.

2.2 Heuristic Models

Since this is a follow-up study to Singal et al. [2019], where we compare the Random Forest, Gradient Boosting and Markovian Model using the same dataset, we will also verify the results for the heuristic attribution models. We are working with a subset of the same dataset. Below is a brief description of the different models:

- Last-Touch Attribution (LTA): This metric assigns all the value to the last touchpoint or interaction that occurred before the consumer purchased the product (conversion). It ignores other touchpoints that may have contributed to the conversion.

- Uniform Attribution: Uniform Attribution is an attribution model that evenly distributes the credit or value of a conversion across all the channels in a customer's journey. It assumes that each channel contributes equally to the conversion.

- Incremental Value Heuristic (IVH): The Incremental Value Heuristic (IVH) is an attribution approach that aims to allocate credit based on the incremental contribution of each touchpoint. It considers the difference in performance when a channel is present versus when it is removed.

- Shapley Value (SV): The Shapley Value is an attribution concept borrowed from cooperative game theory. It calculates the contribution of each touchpoint by evaluating its marginal impact on the overall outcome, considering all possible combinations of touchpoints. It provides a fair distribution of credit among the touchpoints.

These are the different heuristic attribution methods used in digital advertising to attribute value to the channels in a customer's journey. Each method has its own assumptions and calculation techniques for determining the contribution of channels.

2.3 Evaluation Metrics

For attribution models, there are several known performance metrics, including accuracy, recall, precision, F1 score, the ROC curve, and the AUC value, as described in Koyejo [2014]. In our case we are dealing with a dataset that is not evenly distributed. The models we have chosen produce a confusion matrix as a result, which shows how many correct and incorrect predictions we have made. Using the confusion matrix, we can calculate various performance indicators.

The simple performance indicators are accuracy, recall, precision and F1 score. Accuracy shows the percentage of correctly predicted predictions. Recall shows what percentage of "converted" people is correctly predicted to convert. Precision is the percentage of people predicted by the model as converted, out of those who actually have converted. Finally, we have the F1 score, which is a combination of precision and recall. These simple performance indicators are suitable for an imbalanced datasets because they focus on the performance of the minority class.

The Receiver Operating Characteristic (ROC) curve as described in Bradley [1997] is a well-known performance metric that graphically represents the performance of a binary classification model. It plots the true positive ratio (sensitivity) versus the false positive ratio (1-specificity). The Area Under the Curve (AUC) value is a numerical measure that summarises the overall performance of the ROC curve.

When evaluating a model on an imbalanced dataset, where the number of samples in one class is significantly higher than the others, the ROC curve and AUC value may not provide an accurate assessment. As mentioned in Hand [2001] where they compare the performance metrics ROC and recall, the ROC-curve are insensitive to class imbalance and recall not. The ROC and AUC treat each class equally and do not consider the actual distribution of the classes in the dataset. As mentioned in Weng and Pong [2006] the ROC curve and AUC value can give a less accurate assessment when evaluating a model on an imbalanced dataset with a substantial class imbalance, such as having many more ”quit” cases than ”conversion” cases. With a binary distribution for conversion below 3%, these metrics are not most appropriate.

In case of this dataset, where the minority class (”conversion”) holds particular importance, the evaluation metrics such as precision, recall, and the F1 score are the most appropriate. These metrics focus on the performance of the minority class and provide a better understanding of how well the model identifies positive cases. That is the reason why we choose for these metrics to evaluate the performances of the models.

3 Data

In this section we will discuss the insights and features of the dataset. A subset of the same dataset as Singal et al. [2019] will be used in order to make the best comparison. There has been direct contact with the one of the authors, namely R. Singal from the Dartmouth College, and fortunately he made it available. This section will consist of a brief explanation of the overview of the dataset, such as size and information about the channels. In addition, we will discuss some characteristics of the users of all the paths, converted paths and the top-10 most used paths.

We discovered relevant information regarding the lengths of these paths and the concentration of the different channels which are of high relevance for the conclusion of this paper.

3.1 Overview of the dataset

This real-world dataset corresponds to a single product (software) promoted and sold on the Internet by a Fortune 500 company. Given anonymity reasons, the name and specific statistics related to real data of this company will not be disclosed. The company is the advertiser as opposed to a website that serves ads for many advertisers. The dataset consists originally of three elements where each row of these three elements is part of a user path consisting of states, channels and conversion indicator. Given the time and the complexity of working with both the states and channels, only the channel matrix and the conversion vector will be applied for this paper. So please consider the user paths as channel - channel - .... - channel, with eventually a conversion or quit. We deal with a subset of the complete dataset, comprising 31,168 user paths. The length of the channel matrix is 31.168 x 125 and the conversion indicator is a binary 31.168 x 1 vector consisting of 1's (conversion) or 0's (quits). Note that every type of advertisement (also no-ad) are channels.

Each path start with a "sign-up", i.e., the user creating an account on the company's website. From the date of the sign-up, we have access to the user's interaction with the company for a period of 8 months[1]. The type of channels are the following: no-ad, email-ad, display-ad and paid search ad.

The first channel no-ad consists of no touchpoints (no type of advertising) between two states in a user path. The second channel email-ad consists of three related touchpoints: advertiser sends an email (ad action), user opens the email (user action) and a user clicks on a link in the email (user action). Thirdly, the display-ad contains two related touchpoints: advertiser shows a display ad (ad action) and a user clicks on the display ad (user action). Finally the paid search only contains one touchpoint: user clicks on the paid search ad (user action). The paid search will only be registered if the user clicks on the advertisement it. The last two touchpoints are when a user creates an account (sign-up) or buys the product (conversion).

[1] Our user path data is prone to common issues, such as losing track of a user when they clear cookies from their web browser.

3.2 Dataset characteristics

The first characteristics we will discuss are the total amount of users, converted users and the top 10 converted users and their average length of the user paths. As shown in Table 1 the total amount of users are 31,168 and converted users are equal to 751. This means that our dataset is imbalanced where only 2.41% are converted paths. The number of the top-10 most common converted user paths is equal to 297. Furthermore, we can observe that the average length of the converted paths are almost half of the average length of all the paths. And finally, it is shown that the average length of the top 10 converted paths is equal to 2.40, which is less than a fifth of the average length of the converted paths. We just observed that the share of the top-10 most common converted user paths is 39.5% of the total converted paths where the top three is responsible for 29.7%, as shown in Table 2. An interesting observation is that the top three most common user paths only consists of paid search advertisements, where the first one is significantly high compared to others. This could be interpreted as highly interested customers when they first signed up. Seen the average length of the common user paths compared to the average length of all the converted paths, we could say that these 297 customers were already interested in the product. Furthermore, we can observe that the display and no-ads are less common in this table and that the email and paid search ads are more popular.

	Total	Converted	Top-10 Conv.
Users	31,168	751	297
Avg. length	21.84	12.03	2.40

Table 1: Conversion Data

In the last characteristic Table 3 are the ratios of the channels in the total number of user paths and the converted user paths shown. We see that the ratio differs significantly, where the display and paid search ads are more commonly involved in the converted user paths. Furthermore, the no-ad halves in size and the ratio of the email ad of all the ads is half in the converted user paths compared to more than 80% in the total user paths.

4 Methodology

In this section, we will first briefly examine and explain the heuristic models: LTA, Uniform, IVH and SV. Then, we will discuss the methodology of the advanced attribution models: the

Nr.	Path	Conversion rate	Ratio
1.	4	168	22.4%
2.	4→4	39	5.2%
3.	4→4→4	16	2.1%
4.	2	14	1.9%
5.	2→2→2→2	14	1.9%
6.	2→2→2	12	1.6%
7.	3	10	1.3%
8.	2→2→2→2→2	9	1.2%
9.	2→2	8	1.1%
10.	1→4	7	0.9%
	Total	297	39.5%

Table 2: Top 10 converted paths

Channel	Total	Total (%)	Converted	Converted (%)
No-ad	90,774	13.3%	607	6.7%
Email ad	552,255	81.1%	4,565	50.5%
Display ad	32,448	4.8%	3,086	34.1%
PS ad	5,237	0.8%	789	8.7%
Total	680,714	100.0%	9,047	100.0%

Table 3: Conversion Statistics by Channel

Markovian Model, the Random Forest Model and the Gradient Boos Model. Finally, we will provide a theoretical explanation of the evaluation metrics that we will employ to compare the performance of the models.

4.1 Heuristic Attribution Models

In this subsection, we will explain the heuristic models: LTA, Uniform, IVH and SV. Note that these schemes are the same as those in the paper Singal et al. [2019]. We refer to that paper for more theoretical information.

4.1.1 Last Touch Attribution (LTA)

In LTA, all value generated by a purchase is attributed to the last channel in the path. Let $P = \{a_1, a_2, \ldots\}$ be a sequence of the channels before the example path ends at one of the absorbing states $\{q, c\}$. For $a \in A$ we define:

$$\delta_{(a)} = \begin{cases} 1, & \text{if } P \text{ converts and } a \text{ is the last channel in } P \\ 0, & \text{other} \end{cases}$$

where a gets full credit if it is the last channel before the conversion state c on the P path. Therefore, LTA is clearly budget balanced and easy to implement. Within the LTA, the channels before the last channel are not rewarded for promoting the user's progress in the conversion funnel while this can be of high relevance.

4.1.2 Uniform attribution

With the Uniform attribution model, the total generated credit is divided equally among all channels that occur in the path. For mathematical notation, we now take the path $P = \{a_1, a_2, \ldots\}$ representing a series of channels before the example path ends at one of the absorbing states $\{q, c\}$. For $a \in A$ we define:

$$w_{a,\text{uni}}(P) = \begin{cases} \frac{n_a}{|P|} & \text{if } P \text{ converts and } a \in P \\ 0 & \text{other} \end{cases}$$

where n_a is the number of times a) occurs in P and $|P|$ is the number of (not necessarily unique) channels in P indicates. This means that each channel receives an "equal share" of each path it contributed to. Such a scheme is budget-balanced and scaleable.

4.1.3 Incremental Value Heuristic (IVH)

The (Normalised) IVH assigns to each channel a the increase in final conversion probability by looking at the channels: email, display and paid search ads, compared to the "no ad" action. For each $a \in A$, we define a helper variable $z_{a,\text{IVH}}$, which captures the forward-looking increase conditioned on the user being in a certain state:

$$z_{a,\text{IVH}} = P_a h\beta + p_a - (P_1 h\beta + p_1) = (P_a - P_1)h\beta + (p_a - p_1),$$

where $z_{a,\text{IVH}} = [z_{a,\text{IVH}}] \in \mathbb{R}^m$. This value represents the assignment at the level of a "track". In other words, if the advertiser decides to take action a, the corresponding assignment $z_{a,\text{IVH}}$ is used. To interpret this metric as attribution across the entire population, it must be scaled. The values of z_a for the channels are the same as in the paper where the no-ad action receives no value and the other channels receive the value of 1. For each converted user path, the total attribution value for each channel will be normalised.

4.1.4 Shapley Value

The SV will be applied for the advanced attribution models and is a concept from cooperative game theory. It provides a fair way to allocate the value created by the channels to each individual channel. In our research, we utilise the SV to attribute the contribution of each marketing channel in the customer journey. The SV considers all possible combinations of channels and calculates their marginal contributions to the overall conversion value $v(\bullet)$. The formula for calculating the Shapley Value in the context of attribution modeling is as follows:

$$\pi^{a,\text{Shap}}(v) = \sum_{X \subseteq A \setminus \{a\}} \frac{|X|!(|A| - |X| - 1)!}{(|A|)!} \cdot (v(X \cup \{a\}) - v(X))$$

In this formula, the first parameter is defined as $\pi^{a,\text{Shap}}(v)$ representing the Shapley Value of the marketing channel a in the attribution model, where the characteristic function v represents the total conversion value. The set A is the set of all the channels in the attribution model. The set X is a subset of the channels, excluding channel a. The algorithm for computing the Shapley Value can be found in the appendices of this paper. For a more detailed explanation of the Shapley Value and its application in attribution modeling, including mathematical proofs and further insights, you can refer to Singal et al. [2019].

4.2 Advanced Attribution Models

In this subsection we will discuss the methodology of the three advanced attribution models. We will start with the Markovian Model, followed by the classification models the Random Forest and the Gradient Boosting.

4.2.1 Markovian Model

Within our Markovian model, transitions in the user state are stochastic and depend only on the state they are currently in and the advertiser's action. We denote the Markov chain by M and the process ends when it stops or converts. The states are omitted and only the channels and conversion indicator are on the path. The channels configure as states in therms of the Markovian model. We can split the methodology regarding the Markovian model into six parts. The first part is the state space, then the arrival process, followed by the action space, and finally the transition probabilities. Afterwards, we will shortly discuss some mathematical features regarding our attribution problem.

The state space $A := \{a\}_{a=1}^{m}$ is the set of all the channels excluding the two absorbing channels: quit q and conversion c. The notation of the total set of states is equal to $\mathbb{A}^{+} := \mathbb{A} \cup q, c$. The two absorbing channels come into play when the product is purchased or when the consumer leaves the system. Then the arrival process, where external traffic arrival is defined as channel $a \in \mathbb{A}$ w.p. λ_a (*initial channel probability*). We define the vector $\boldsymbol{\lambda} \in \mathbb{R}^m$ as $[\lambda_a]_{a \in \mathbb{A}}$. No external traffic is able to arrive at c and q.

The transition probabilities represent the probabilities of switching from one channel to another in the user path. These probabilities can be derived from the historical data of the dataset, looking at the order of the ad channels in the paths. We denote $p'_{aa} = \mathrm{P}(\text{user moves from } a \text{ to } a' \text{ in one transition})$ with $a \cup a' \in A$. Also, for $\forall p_{aa'} \in \mathbf{P}$. The first assumption is the absorption function where any Markov Chain corresponding to $(\boldsymbol{\lambda}, \mathbf{P})$ is absorbing, so the probability is equal to one that each user will eventually convert of quit.

Now we have come to the absorbing channels q and c, with the notation p_{aq} and p_{ac} are the probabilities that the user quits $q \in \mathbb{A}^{+}$ or converts $c \in \mathbb{A}^{+}$. Due to the fact that q and c are absorbing channels, the transitions for them are self-loops w.p. 1. So, the Markov Chain M is defined as $(\boldsymbol{\lambda}, \mathbf{P}, \mathbf{p_{aq}}, \mathbf{p_{ac}})$.

The last two elements are the expected number of visits and the eventual conversion probability. First, the expected number of visits where $\mathbf{F} \in \mathbb{R}$ denote the fundamental matrix of the Markov Chain M. Where the (x, y)-th entry of F is equal to the expected number of visits to channel x when the initial channel is y. As stated in Grinstead and Snell [2012], the initial assumption guarantees the existence of $\mathbf{F}$, where $\mathbf{F}$ is equal to $(I - P)^{-1}$. Additionally, let us define $\boldsymbol{\mu} := [\mu_a]_{a \in \mathbb{A}} \in \mathbb{R}$, where μ_a represents the expected number of visits to channel a. Demonstrating $(\boldsymbol{\mu})^T = \boldsymbol{\lambda}^T \mathbf{F}$ is straightforward. To establish the eventual conversion probability, we introduce h_a as the probability of ultimately being absorbed in channel c from channel $a \in \mathbb{A}$. Furthermore, we define the vector h as $\mathbf{h} := [h_a]_{a \in \mathbb{A}} \in \mathbb{R}$. So, we can state that $h = Ph + p_c$. Thus, $h = Fp_c$. We set $h_q = 0$ and $h_c = 1$.

4.2.2 Random Forest Model

The (Ordinal) Random Forest model is built by combining multiple independent random decision trees. Each individual tree depends on a random vector of observations. This vector is drawn independently from the original dataset, these observations have the same distribution across trees as mentioned in Breiman [2001]. As described in Janitza et al. [2016], that process is called the bootstrapping phase. Bootstrapping involves the process of randomly selecting a subset with replacement from the original data to obtain the training set. As mentioned in Hastie [2009] a well-known technique is also called bagging or bootstrap aggregating. In this technique, the original sample of size N is converted into a random sample of exactly the same size N, but with replacement. This subset is then used to construct a decision tree. Given the randomness of the sample, not all observations are involved and those that are not are called the out-of-bag (OOB) sample which are then used for performance evaluation.

The RF model uses the training set to build a classification and regression tree (CART). A CART is a model that uses binary splits of variables to ultimately provide predictions. With CART, it is a model that uses splits of variables such that a prediction is ultimately made with them. Based on a randomly selected subset of variables, these splits are made. Then on to the final step and that is the construction of the RF model consisting of individual decision trees, with tree defined as k. This tree is generated by the random vector k with the same distribution as the other random vectors. Then the tree is constructed by using the training set and Φ_k. Thus, this tree results as the classifier $h(\mathbf{x}, \Phi_k)$ where $\mathbf{x}$ is the input vector. The Random Forest is defined as a model or classifier consisting of a collection of decision tree classifiers $h(ß\mathbf{x}, ß\Phi_k), k = 1, ...,$ where the classifiers are independent identically distributed random vectors and each tree casts a vote for the most common class. Ultimately, the class that receives the most "votes" in the aggregation is chosen as the final class label.

Thus, for an individual decision tree, let us say tree k, an independent random vector Φ_k is generated with the same distribution as the other random vectors. The tree is then created using the training set and Φ_k. The resulting tree is a classifier $h(\mathbf{x},\ \Phi_k)$ where $\mathbf{x}$ is an input vector. Therefore, we define a Random Forest as a classification model or classifier consisting of a collection of decision tree classifiers $h(\mathbf{x}, \Phi_k), k = 1, ...$ where the trees are independent identically distributed random vectors and each tree casts a vote for the most common class. The class that receives the most votes in the aggregation is chosen as the final class label. This definition was introduced by Breiman [2001]. A nice and important own of Random Forests is that it cannot overfit because of the generalization error that has a limit of:

$$P_{X,Y}\left(P_{\Phi}[h(X,\Phi) = Y] - \max_{j \neq Y} P_{\Phi}[(X,\Phi) = j]\right) < 0.$$

Given that the generalisation error cannot go above its stated limit, it is impossible to be below 0 and so overfitting will not be a problem.

4.2.3 Gradient Boosting

The second classification model we will cover is the Gradient Boosting model. We will briefly describe this method and we please refer to the articles Natekin and Knoll [2013] and Chen and Guestrin [2016] and the algorithm in the appendices for more information. This model is a powerful method that uses the user paths as decision trees and here assigns credits to the different channels in the customer journey. Through the decision trees it creates a strong ensemble model. The first step within this technique, after defining the decision trees, is to find the best step size rho_t that minimizes the loss function:

$$\rho_t = \arg\min_{\rho} \sum_{i=1}^{N} \mathcal{L}\left(y_i, \hat{f}_{t-1}(x_i) + \rho h(x_i, \Phi_t)\right)$$

where ρ_t represents the optimal step size for the current iteration t. The parameter $\hat{f}_{t-1}(x_i)$ is the prediction of the ensemble model up to the previous iteration and $h(x_i, \Phi_t)$ is the prediction of the base model with its associated parameters Φ_t. The parameter $\mathcal{L}$ is the chosen loss function, which measures the deviation between the predicted values and the actual values. The model will be adjusted by the following method:

$$F_M(x) = F_{M-1}(x) + \sum_{m=1}^{M} \gamma_m h_m(x)$$

where $F_M(x)$ represents the final ensemble model, which estimates the contribution of each marketing channel to the desired outcome. $F_{M-1}(x)$ is the ensemble model of the previous iteration and h_m is the multiplier or step size determined through optimization, and indicates the contribution of the weak learner $h_m(x)$ at each iteration. Moreover, the parameter $h_m(x)$ represents the basic model, such as a decision tree, trained on the pseudo-residuals to capture the relationships between the features x and the target variable. Finally, M indicates the total number of boost-iterations or weak learners added to the ensemble. Using the gradient boost formula and step size optimisation, by applying the Shapley Value method we can effectively capture the complex interactions between marketing channels and assign appropriate credits based on their contribution to the desired results.

4.3 Evaluation metrics

There are several performance metrics available for evaluating the models, which are derived from the Confusion Matrix depicted in Figure 2 in the Appendix A. Each performance measure serves its own purpose and can be applied to different types of data. As stated in the Literature review, we will utilise the recall, precision and F1 score statistics as performance metrics. In this subsection, we will delve into the mathematical notation of these metrics.

The first metrics we will consider are precision and recall. To recap, precision measures the number of correctly predicted positives, while recall indicates how many positives we missed. In our case, recall will be seen as important as the precision for the results. The mathematical definitions of precision and recall are as follows:

$$\text{Precision (\%)} = \frac{TP}{FP + TP} * 100\%,$$

$$\text{Recall (\%)} = \frac{TP}{FN + TP} * 100\%.$$

In simple terms, recall represents the ratio of correctly predicted cases of conversion to the sum of correct predictions and false negatives. Precision is defined as the ratio of true positives to the total number of positive predictions. A higher precision value indicates a lower rate of false positives. The false positives should be minimised. But... it can be misleading, therefore: maximising this ratio is also significant. The F1 score combines both recall and precision, providing a balanced measure, and is mathematically defined as follows:

$$\frac{2 * Precision * Recall}{Precision + Recall}.$$

The F1 score is advantageous because it serves as an indicator when either precision or recall is significantly low. If the F1 score is low, it suggests that something is amiss, and the desired balance between precision and recall has not been achieved with the model. In other words, the F1 score helps identify situations where the model is not performing well in terms of both precision and recall.

5 Results

In this section, we will discuss the results of the attribution models. First, we will discuss the heuristic results. Then, we will delve into the Shapley Value results of the Markov model and the classification models RF and GB. Lastly, we will evaluate the predictive performances of the models RF and GB.

5.1 Heuristic models

In this subsection, we will discuss the results of the heuristic models for the LTA, IVH, and Uniform attribution. which are shown in Table 4. When comparing our results with the findings presented in the article by Singal et al. [2019], we observe a discrepancy. This difference comes from how we conducted our study using a different method. We focused only on the channels and the conversion indicator, without including the states. We have shared our results and codes with the author R. Singhal for information and verification. Unfortunately, we have not received a response.

Moving on, we started by receiving the LTA results where the paid search ad and the email ad obtained the most attribution with percentages of 44.9% and 32.1%. These attributions were followed by the display and no ads, with values of 14.8% and 44.9%, respectively. Secondly, if we examined the Uniform results, we can observe that the email and paid search ads received the most attribution with 38.5% and 37.7%. The Uniform results for the display and no ad were equal to 18.2% and 6.1%. At last, the email and paid search ad for the IVH model were equal to 42.2% and 38.6%, followed by 19.2% for the display ad and 0.0% for the no ad. The no-ad have not received attribution because all attribution values for the no-ad were equal to 0. As discussed in the Data section, the top three converted paths (4, 4 → 4, and 4 → 4 → 4) contribute for 29.7% of all the attribution. As expected, the results of the paid search ad for all of the three heuristic models are higher than 29.7%.

Channel	LTA	Uniform	IVH
No-ad	8.3%	6.1%	0.0%
Email ad	32.1%	38.5%	42.2%
Display ad	14.8%	18.2%	19.2%
PS ad	44.9%	37.3%	38.6%

Table 4: Heuristic results

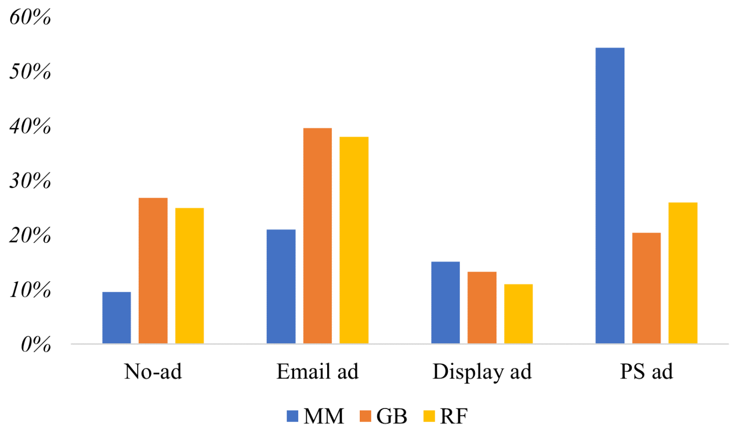


Figure 1: Shapley Value results

5.2 Advanced Attribution Models

First of all, the computation time for receiving the results of the advanced models were reasonable and did not require a significant amount of time. The Shapley Value results of the advanced attribution models are depicted in Figure 1. A notable disparity is observed between the results of our Markovian model and the findings of the previous research. We recognise that the results show differences, and this can be attributed to the different methods we used. In our approach, we excluded the states. This difference in methodology explains the variations in the results.

First of all, the no-advertisement channel received 10% attribution for the Markovian model and obtained 27% attribution for the GB and 25% for the RF models. Secondly, the email advertisement acquired 21% attribution in Markovian model and it received the highest attribution from both the GB and RF models, with 40% and 38% respectively. Moving on to the display advertisement, our Markovian model attributed 15% to the display advertisement. Interestingly, the display advertisement received the least attribution among all channels for the GB model with 13% and for the RF model with 11%. Lastly, in our model, the paid search ad gained the most attribution with 54% attribution. Surprisingly, for the classification models, the paid search ad obtained 20% attribution in the GB model and 26% in the RF model. We can state that both of the classification under performed in therms of the paid search ad due to the fact that their attribution levels are below 29.7%. For this case, we can state that the Markovian model performed the best, based on the paid search ad results.

5.3 Performances classification models

In this subsection, we will discuss the performance of our classification models. These models are predictive models that are trained using a portion of the dataset (in this case: 64%) and then evaluated on a separate test set (36%).

The evaluation metrics results are presented in Table 5, where the scores are quite similar. In terms of precision, the RF model demonstrates better performance than the GB model, achieving a score of 93.4% compared to 90.5%. This suggests that the RF model achieved a 93.4% accuracy rate in predicting conversions, correctly identifying them in the majority of cases. These results indicate a high probability of conversion when the model makes such predictions. On the other hand, recall measures the percentage of actual converted paths that were correctly predicted by the models. In this aspect, the GB model slightly outperformed the RF model, achieving a recall rate of 69.6% compared to the RF model's 68.4%. This means that both models accurately predicted approximately 7 out of 10 converted paths, successfully. As described in the Methodology section, the F1-score combines both precision and recall scores. In this regard, the RF model performs slightly better with a score of 79.0% compared to the GB model's 78.7%. The difference between the two models in terms of F1-score is relatively small and can be considered negligible. We can conclude that both of these models demonstrated similar performance in terms of their predictive abilities. Furthermore, the results of the confusion matrices are shown in Tables 6 and 7 in the Appendices.

	RF	GB
Precision	93,4%	90,5%
Recall	68,4%	69,6%
F1-score	79,0%	78,7%

Table 5: Perf. Metrics

6 Conclusion

The digital advertising market has experienced significant growth in the past decade and is projected to continue expanding. Companies strive to sell products to people of all ages through digital platforms in an intelligent manner. Therefore, it has become crucial to determine the value of different channel combinations that ultimately lead to consumer purchases. Attribution modeling is employed to analyse this attribution process. In a study conducted by Singal et al.

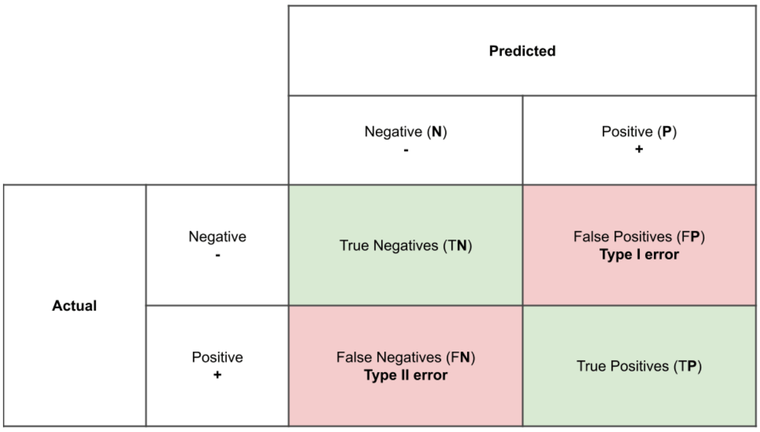
[2019], the relationship between heuristic attribution models and the Shapley Value of the Markovian Model was investigated. This research serves as a follow-up study, comparing the Shapley Value results of two classification models, Gradient Boosting and Random Forest, with those of the Markovian Model. Following our research question: *How do classification models perform compared to the Markovian model in the context of Attribution modeling for Online Advertising?*

For this study, a subset of the same dataset was utilised. This dataset is from a Fortune 500 company and consists of user paths containing the channels: no-ad, email ad, display ad and paid search ad and a conversion indicator whether the customer bought the product (conversion) or left the system (quit). The total size of the dataset was 31,168 user paths with a ratio of 2.41% converted paths which implies an imbalanced dataset. An interesting observation of the dataset is that 29.7% of the converted paths are the top 3 common user paths only containing paid search ads. Given this ratio, the attribution of the paid search ad should be significant for our models. Furthermore, the classification models were tested on their predictability using the F1 score, due to the imbalance dataset.

First of all, the Shapley Value results of our models where the computation time for receiving the results of the models was reasonable and did not entail significant time consumption. In terms of complexity for applying these models to different datasets, our suggestion is that the Markovian Model stands out the most while the classification models required additional assistance. And at last, we mentioned the high ratio of paid search ads in the converted paths which should match with the Shapley Value results. Therefore, with the Shapley Value results of the classification models for paid search below 29.7%, we suggest that our Markovian Model is the most suitable model. Finally, both classification models demonstrated similar performance in terms of the evaluation metrics. So, for the classification models, we suggest the Random Forest model as the preferred option due to the higher attribution value for the paid search advertisement.

Finally, for the last part of the paper there are some suggestions for further research. One of them is that the used models will be applied for the whole dataset to find extract more information about the functions of our models and theirs. Furthermore, it would be interesting to apply the Counterfactual Adjusted Shapley Value to the models and compare these results. Lastly, the states of the individual users could be applied to the user path to receive more detailed information about the users.

A Figures



		Predicted	
		Negative (**N**) -	Positive (**P**) +
Actual	Negative -	True Negatives (T**N**)	False Positives (F**P**) **Type I error**
	Positive +	False Negatives (F**N**) **Type II error**	True Positives (T**P**)

Figure 2: Confusion Matrix

B Tables

		Predictions	
		0	1
Actuals	0	10909	12
	1	78	169

Table 6: Conf. Matrix (RF)

		Predictions	
		0	1
Actuals	0	10903	18
	1	75	172

Table 7: Conf. Matrix (GB)

C Algorithms

Algorithm 1 Friedman's Gradient Boos Algorithm

```
Require: Input data $(x, y)_N$
Require: Number of iterations $M$
Require: Choice of the loss-function $\mathcal{L}(y, f)$
Require: Choice of the base-learner model $h(x, \Phi)$
Initialize $f_0$ with a constant
for $t = 1$ to $M$ do
    Compute the negative gradient $g_t(x)$
    Fit a new base-learner function $h(x, \Phi_t)$
    Find the best gradient descent step-size $\rho_t$:
    $\rho_t = \arg\min_\rho \sum_{i=1}^{N} \mathcal{L}\left(y_i, \hat{f}_{t-1}(x_i) + \rho h(x_i, \Phi_t)\right)$
    Update the function estimate:
    $\hat{f}_t \leftarrow \hat{f}_{t-1} + \rho_t h(x, \Phi_t)$
end for
```

Algorithm 2 Shapley Value Algorithm

```
Require: Set of channels $A$
Require: Conversion increments $v$ for each subset of channels
Ensure: Shapley Value $\pi^{a,\text{Shap}}(v)$ for each channel $a$
$\pi^{a,\text{Shap}}(v) \leftarrow$ empty dictionary
$n \leftarrow |A|$
for $a$ in $A$ do
    $\pi^{a,\text{Shap}}(v) \leftarrow 0$
end for
for $k$ in 1 to $n$ do
    for all subsets $X$ of $A$ with size $k$ do
        $m \leftarrow |X|$
        for all $a$ in $X$ do
            $\pi^{a,\text{Shap}}(v) \leftarrow \pi^{a,\text{Shap}}(v) + \frac{M\cdot(N-M-1)!}{N!} \cdot (v(X \cup \{a\}) - v(X))$
        end for
    end for
end for
return $\pi^{a,\text{Shap}}(v)$ for each $a \in A$
```

References

A. P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, pages 1145–1159, 1997.

L. Breiman. Random forests. 2001.

T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data*, pages 785–794, 2016.

T. DelSole. A fundamental limitation of markov models. pages 2158–2168, 2000.

H. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4):917–963, 2019.

C. Grinstead and J. Snell. Introduction to probability. 2012.

R. J. T. Hand. A simple generalisation of the area under the roc curve for multiple class classification problems. *Kluwer Academic Publishers*, 2001.

F. Hastie, Tibshirani. *The Elements of Statistical Learning.* Springer, 2009.

S. Janitza, G. Tutz, and A.-L. Boulesteix. Random forest for ordinal responses: Prediction and variable selection. pages 57–73, 2016.

T. Kadyrov and D. I. Ignatov. Attribution of customers' actions based on machine learning approach. *National Research University Higher School of Economics, Russian Federation, St. Petersburg Department of Steklov Mathematical Institute of Russian Academyof Sciences, Russia*, 2019.

P. K. R. I. S. D. Koyejo, Nagarajan Natarajan. Consistent binary classification with generalized performance metrics. *Department of Computer Science, University of Texas at Austin*, 2014.

J. R. Leguina, C. Rumín, and R. C. Rumín. Digital marketing attribution: Understanding the user path. 2020.

A. Natekin and A. Knoll. Gradient boosting machines, a tutorial. *frontiers in NEURO-ROBOTICS*, 2013.

G. G. E. Scott M. Lundberg and S.-I. Lee. Consistent individualized feature attribution for tree ensembles. 2019.

R. Singal, O. Besbes, A. Desir, V. Goyal, and G. Iyengar. Shapley meets uniform: An axiomatic framework for attribution in online advertising. pages 385–393, 2019.

Statista. Digital advertising report 2022, 2022.

C. G. Weng and J. Pong. A new evaluation measure for imbalanced datasets. *School of Inofmration Technologies*, 2006.

L. L. Xuhui Shao. Data-driven multi-touch attribution models. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, page 258–264, 2011.