Analyzing Customer Conversion Patterns: A Survival Analysis Approach to Multi-Channel Attribution

Preetish Panda

Passage Inc., Kitchener, N2G 1H6, Ontario, Canada.

E-mail: preetish88@gmail.com

Abstract

This report explores the dynamics of customer conversion by examining the relationship between visit behaviour and conversion outcomes across various marketing channels. By condensing customer visit data into a singular representation for each customer, we capture the time intervals from their first visit to either a conversion or their last recorded visit, thereby categorizing customers as converters or non-converters. The analysis utilizes survival analysis techniques to estimate conversion probabilities over time for different marketing channels, allowing for insights into the effectiveness of each channel in driving conversions. Furthermore, the report introduces a causal inference framework to assess the impact of offline marketing interventions, specifically television advertising, on web traffic. By employing Bayesian structural time-series models, we generate counterfactual predictions to isolate the uplift attributable to marketing efforts. This comprehensive approach highlights the interplay between digital and offline marketing strategies and provides actionable insights for optimizing customer engagement and conversion.

Keywords: Customer Conversion, Marketing Analytics, Survival Analysis, Attribution Models

1 Introduction

Marketers often dream of customers instantly deciding to purchase a product or service after encountering an advertise- ment. However, the actual journey to a purchase is far from straightforward. Customers typically interact with a brand through a variety of channels and touchpoints—such as social media, websites, search engines, and mobile apps—before reaching a final decision.

Why is this important? The aim is to equip decision-makers with actionable insights for strategic planning. By evaluating the effectiveness of marketing activities, businesses can assess the return on investment (ROI) of their campaigns.

Research by Jayawardane et al. (2015) highlights that customers engaging with brands across multiple touchpoints tend to contribute higher overall value. With several simultaneous marketing initiatives often in play, identifying which channels are driving conversions—and their respective contributions—remains a significant challenge. This section outlines approaches to tackle these complexities, providing practical guidance and illustrative examples. The goal is to help decision-makers uncover the value of each

S. Mantena (ed.), *Proceedings of the Conference on Social and Sustainable Innovation in Technology & Engineering (SASI-ITE 2025)*, Advances in Intelligent Systems Research 199,
https://doi.org/10.2991/978-94-6463-940-7_10

channel in the customer journey, empowering them to make informed, data-driven marketing decisions.

2 Key Definitions

In this guide, we use the term ’customer pathway’ or ’conversion journey’ to refer to the series of steps a customer takes when interacting with a brand, ultimately leading to either a conversion or no conversion.

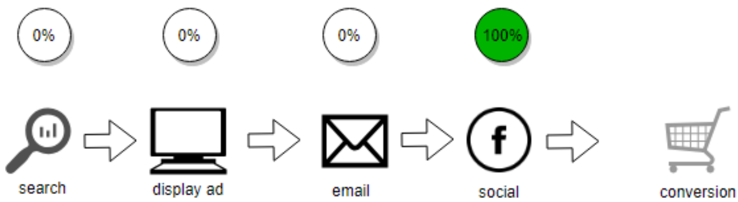

Fig. 1. Visualization of the Customer Conversion Journey.

The diagram (Figure 1) illustrates the steps a customer takes when interacting with a brand, leading to conversion or non-conversion. We define conversion as the completion of a desired action by a user. This could refer to purchasing a product on an online store, but in other contexts, it might mean signing up for a service, subscribing to a newsletter, or performing any other predefined action.

We define conversion as the completion of a desired action by a user. This could refer to purchasing a product on an online store, but in other contexts, it might mean signing up for a service, subscribing to a newsletter, or performing any other predefined action.

3 Paper Overview

This section outlines the scope of this paper, focusing on the various models available for addressing the attribution challenge, each with its unique methodology and application. For every model discussed, we offer a clear explanation, illustrate its working mechanism through practical examples, and highlight its advantages and limitations to help readers understand its applicability

This study explores the practical implementation of attribution models within modern digital analytics systems, focusing on methodological differences and data challenges. Furthermore, this paper examines the attribution techniques provided by platforms such as Google Analytics and Analytics 360.

Our primary goal, however, is to guide users looking to go beyond the default methods by customizing and extending existing approaches to suit their specific business needs. We introduce a set of tools built with the R programming language, complete with replicable code snippets to illustrate a range of attribution models, from straightforward solutions to more advanced methods.

The examples in this paper leverage data sourced from Google Analytics 360 and BigQuery, emphasizing the often-overlooked steps of data querying and preparation. These instructions are designed to be flexible, ensuring relevance for users working with data from other platforms as well. Lastly, we broaden the discussion by including a special section on offline attribution.

This involves analyzing the effectiveness of traditional marketing channels such as television, radio, and billboards, providing critical insights that extend beyond the reach of online tracking tools.

4 Platform Generalization

Although the paper extensively uses Google Analytics 360 and BigQuery due to their accessibility and depth of data, the modelling approaches—such as survival analysis, Markov models, and causal inference—are platform-agnostic.

These methodologies can be implemented with data sourced from Adobe Analytics, Mixpanel, Matomo, or even custom event logs stored in SQL-based databases. The underlying principles remain consistent, requiring only appropriate data transformation.

5 Types of Attribution

5.1 Group vs. Individual Attribution

Historically, marketing strategies were heavily influenced by the performance of offline channels such as print media and broadcasts. While effective in some ways, these methods had notable limitations.

For instance, they could not provide insights into customer conversions at the individual level. Attribution models in this era typically relied on group-level data, aggregating the responses of all users who interacted through a specific channel.

Moreover, it was often challenging— if not impossible—to track which customers encountered multiple touchpoints across these offline media. With the rise of digital marketing, however, marketers now have access to much more granular insights, allowing them to analyze how individual marketing activities influence customer behaviour at a much more personalized level.

5.2 Simplistic vs. Fractional Modelling

Basic attribution models attribute the entire conversion value to a single touchpoint in the customer's journey, disregarding all other interactions. As a result, this type of model overlooks the influence of earlier touchpoints.

Among the various basic models, we focus on two primary types that are widely utilized:

- **Last Touch Model**: All the credit for a conversion is given to the last touchpoint before the conversion occurs.
- **First Touch Model**: Full credit is assigned to the first touchpoint that the customer interacts with in their journey.

In contrast to basic attribution, distributed attribution models assess the contribution of each touchpoint along the customer journey, distributing the conversion credit across multiple interactions. These models are further categorized into heuristic and algorithmic approaches, each with its own methodology and application, which we will explore in detail below.

5.3 Heuristic vs. Algorithmic Attribution

5.3.1 Heuristic Attribution Model

The heuristic, or rule-based, attribution model is described by Jayawardane et al. (2015) as a method that uses predefined rules to allocate conversion credits to different touchpoints in the customer journey. The rules used in heuristic models include strategies such as Linear, Position-Based, and Time Decay, which we will discuss in greater depth in later sections. Compared to basic models, heuristic attribution models offer a more nuanced approach, recognizing the impact of various channels throughout the entire conversion process.

5.3.2 Algorithmic Attribution Model

Algorithmic attribution takes fractional attribution a step further by leveraging advanced statistical techniques and machine learning to assess the contribution of each touchpoint. Instead of relying on simple rules, algorithmic models evaluate both converting and non-converting customer paths to probabilistically assign credit to each channel based on its effectiveness.

Common algorithmic approaches include:

- Markov Chain models
- Shapley Value
- Survival analysis

We will delve deeper into these techniques and discuss which ones are most appropriate for specific attribution goals.

6 Categorizing Attribution Methods

The following section provides a comprehensive overview of the various methods (Figure 2) commonly used for marketing attribution in the e-commerce industry.

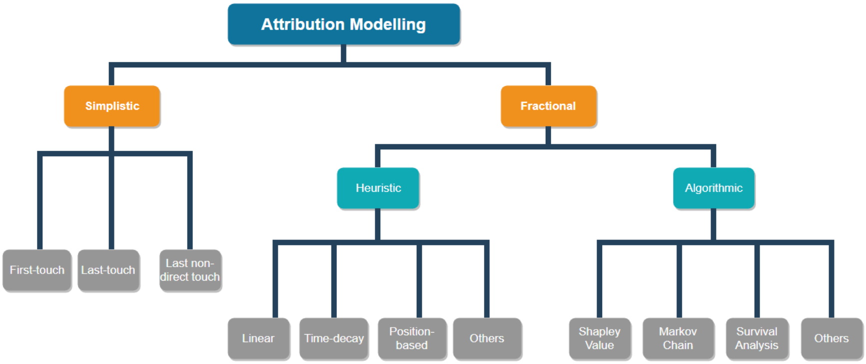

Fig. 2. Taxonomy of Marketing Attribution Methods. This classification highlights commonly used approaches in e-commerce for modelling customer pathways and evaluating marketing effectiveness.

7 Approaches to Attribution

In this section, we will explore a few widely used attribution models in marketing. Imagine a customer browsing for swimwear on your online store. They first conduct a search on Google, landing on your website, but don't complete a purchase. A few days later, they interact with a display ad and revisit your site, possibly to review the products again and sign up for promotions.

A week later, while at work, they receive an email campaign, which prompts them to click through and browse for discounts, but they still do not buy. Finally, in the evening, a Facebook ad catches their attention, reminding them to make a purchase with their credit card.

So, which marketing channel ultimately influenced their decision to purchase? Was it the search engine that made them aware of your brand? Or did the display ads, email campaign, or Facebook ad play a more significant role in leading them to buy? To address these questions, we will examine several attribution methods used to allocate credit for conversions.

8 Basic Attribution Models

Basic attribution models assign the entire credit for a conversion to a single, predefined touchpoint along the customer's journey. These models are simple and commonly used, but they offer a rather simplistic view of marketing performance.

A straightforward example of a basic attribution method is the Last Touch model, also known

Last Interaction or Last Click attribution (Figure 3). In this model, full credit for the conversion is given to the final touchpoint in the customer's conversion journey. For instance, if the customer completes the purchase after interacting with a Facebook ad, all the credit for the sale is assigned to that final interaction, ignoring the influence of previous touchpoints, such as the initial Google search.

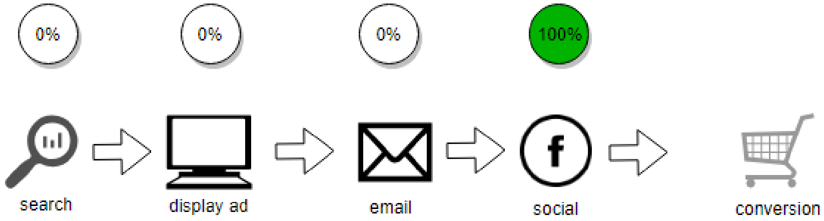

Fig. 3. Last Touch Attribution Model. All credit for the sale is assigned to the last touch point in the conversion journey, here represented by the social channel.

A First Interaction model (Figure 4) attributes the entire credit for a conversion to the initial channel that engaged the customer in their journey. This approach emphasizes the importance of awareness and initial engagement.

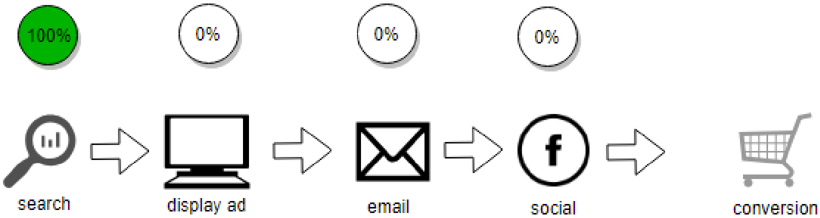

Fig. 4. First Touch Attribution Model. All credit for the sale is assigned to the first touch point, represented here by the search channel

9 Alternative Models

Several variations exist that refine the attribution process. For instance, the Last Non-Direct Click model operates similarly to the Last Touch model, but with one key difference: if the last interaction is via a direct channel (e.g., typing the website URL directly), it disregards this and instead attributes the conversion to the previous non-direct touchpoint, such as a paid search or referral.

9.1 Fractional Heuristic Attribution Models

While a basic attribution model might suffice in certain scenarios, a more sophisticated approach involves distributing the credit for the conversion across multiple channels. This approach is akin to a sports team—although one player scores the goal, the entire team plays a part in setting up that scoring opportunity.

9.1.1 Equal Attribution

The most neutral method within fractional attribution is the Linear attribution model (Figure 5). This approach distributes the credit for the conversion evenly across all touchpoints in the customer’s journey. For example, in the scenario below, each of the four touchpoints receives an equal share—25%—of the total conversion credit.

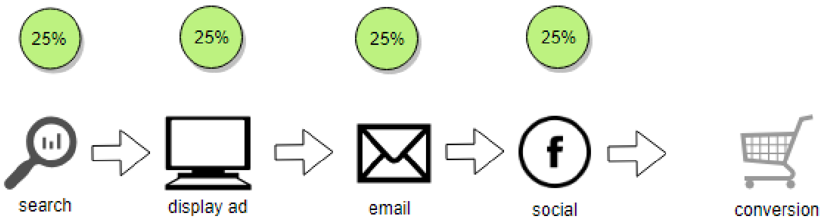

Fig. 5. Linear Attribution Model. All four touch points are assigned equal credit (25%) for the conversion.

9.1.2 Decay Based on Timing

Building upon the ideas of First Touch and Last Touch models, we can introduce a system where attribution is determined by the position of each touchpoint in the conversion sequence. This method is referred to as Position-Based attribution (Figure 6).

In this approach, we allocate a predetermined percentage of credit to the first and last touchpoints, while distributing the remaining credit evenly among all touchpoints in between.

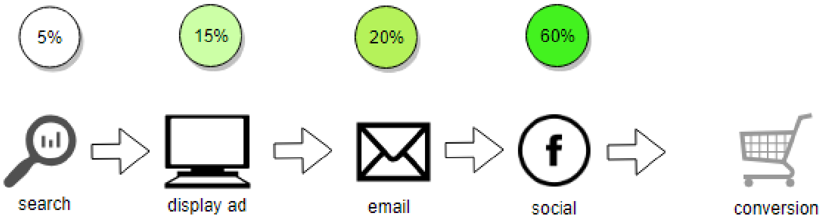

Fig. 6. Time Decay Attribution Model. More credit is assigned to recent touch points.

For example, in one scenario, 45% of the conversion credit is given to the first and last touchpoints, with the remaining credit split equally among the intermediate touchpoints.

9.1.3 Attribution Based on Position

Building upon the ideas of First Touch and Last Touch models, Position-Based attribution (Figure 7) allocates a predetermined percentage of credit to the first and last touchpoints, while distributing the remaining credit evenly among all touchpoints in between. For example, 45% of the conversion credit might be given to the first and last touchpoints, with the remaining credit split equally among the intermediate touchpoints.

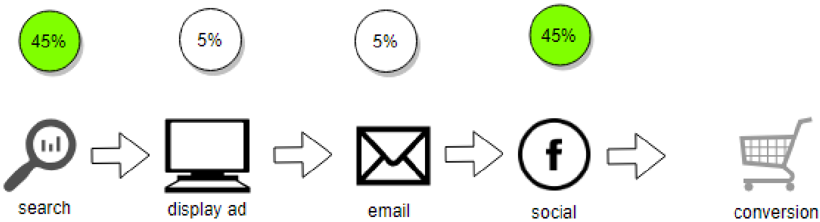

Fig. 7. Position-Based Attribution Model. 45% of credit is assigned to the first and last touch points, with the rest distributed equally among intermediate touch points.

9.2 Algorithmic Fractional Attribution Models

Algorithmic attribution models differ from traditional methods by utilizing historical conversion data to deduce the attribution rather than relying on predefined rules or heuristics (Figure 8). This process is powered by advanced algorithms that analyze the performance of each marketing channel.

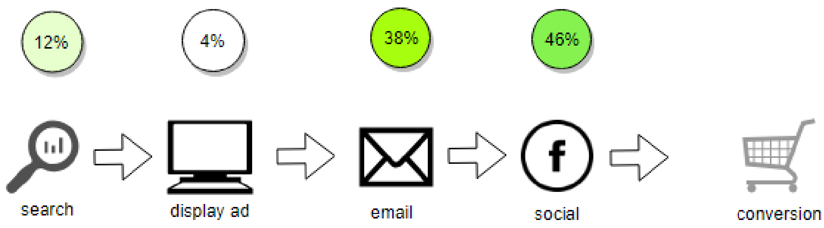

Fig. 8. Fractional Algorithmic Attribution Model. The model leverages historical data along with statistical techniques t allocate attribution credit.

This process is powered by advanced algorithms that analyze the performance of each marketing channel. These models typically examine both paths taken by customers who made purchase and those who did not. The use of such sophisticated models helps overcome the limitations of simpler rules, offering a more objective and insightful method for decision-making when straightforward models fall short.

9.2.1 Logistic Regression

A. Logistic Regression Logistic regression, a widely used statistical method, was introduced ir the 1800s [1]. It utilizes the logistic function to predict a binary outcome, such as success/failu or, in this context, conversion/no conversion.

This model assesses the relationship between a set of independent variables (e.g., various marketing channels) and the outcome of interest. In attribution terms, logistic regression helps identify the contribution of different marketing channels to whether a customer completes a conversion or not.

By treating each channel (e.g., email, display, search) as variables in the model, it can determi the likelihood of conversion based on customer touchpoints. Each touchpoint is evaluated for i presence or absence in the conversion path, allowing for the development of a model that predicts conversion outcomes based on past data.

However, because many marketing channels are interrelated, estimating these effects accurately can be challenging. Techniques like ”bagging” are often used to enhance the reliability of these predictions [2].

9.2.2 Shapley Value

The Shapley Value, introduced by Lloyd Shapley in 1953 [3], originates from Game Theory and is designed to fairly distribute credit among different contributing factors when multiple elements work together to produce a single result. In marketing, this method is used to assess how much each touchpoint contributes to a conversion (Figure 9). For example, in one customer journey, the ‘Display‘ channel might contribute to a 2.

Fig. 9. Shapley Value Attribution Model. This model computes the contribution of each touchpoint by analyzing its marginal impact across all possible sequences.

Mathematically, the Shapley value for a given contributor is calculated as

follows:

$$\phi_i(v) = \frac{1}{|N|!} \sum_R \left[v(P_i^R \cup \{i\}) - v(P_i^R)\right]$$

This formula calculates the payoff when a specific marketing touchpoint, like ‘Display‘, is included in the sequence. The impact of ‘Display‘ is determined by subtracting the contributions of all other touchpoints and averaging across all possible permutations of the sequence, providing a fair allocation of credit for each touchpoint’s contribution.

9.2.3 Markov Chain Techniques

A Markov Chain is a mathematical model that captures the transitions between different states within a system. These states can represent diverse phenomena, such as marketing channels, weather conditions, or health outcomes. For more details, see [4].

The primary goal is to quantify the probabilities of moving from one state to another, using these probabilities to predict future states based solely on the system's current state.

A Markov Chain comprises three essential components:

- **State Space**: The complete set of all possible states the system can occupy.
- **Transition Matrix**: A table of probabilities that indicates the likelihood of transitioning from one state to another.
- **Initial State Distribution**: The probabilities associated with the system starting in each possible state.

Markov Chains operate under the property of being ”memoryless,” meaning predictions depend only on the present state, with no influence from the sequence of prior states. However, advanced variations like higher-order Markov Chains account for a limited number of preceding states, enhancing predictive capabilities. These extensions are discussed further in Section 9.1.

A practical application of Markov Chains is weather forecasting. By analyzing historical weather data, we can estimate the likelihood of transitioning between different conditions (e.g., sunny to rainy). These transitions are stored in a transition matrix (Table 1) and can be represented as a directed graph, providing a clear visualization of the system's dynamics.

Table 1 Weather Transition Probability Matrix

	Clear	Overcast	Wet
Clear	0.7	0.20	0.10
Overcast	0.3	0.40	0.30
Wet	0.2	0.45	0.35

If we assume the weather is currently overcast, the initial state can be expressed as:

$$\text{initial} = \begin{bmatrix} \text{clear} = 0 \\ \text{overcast} = 1 \\ \text{wet} = 0 \end{bmatrix}$$

By applying matrix multiplication with the transition matrix, we can calculate the probabilities for each weather condition on the following day. Alternatively, these probabilities can be directly derived from the transition matrix (Figure 10). This process can be iteratively repeated to forecast probabilities (Table 2) for conditions further into the future, such as seven days ahead:

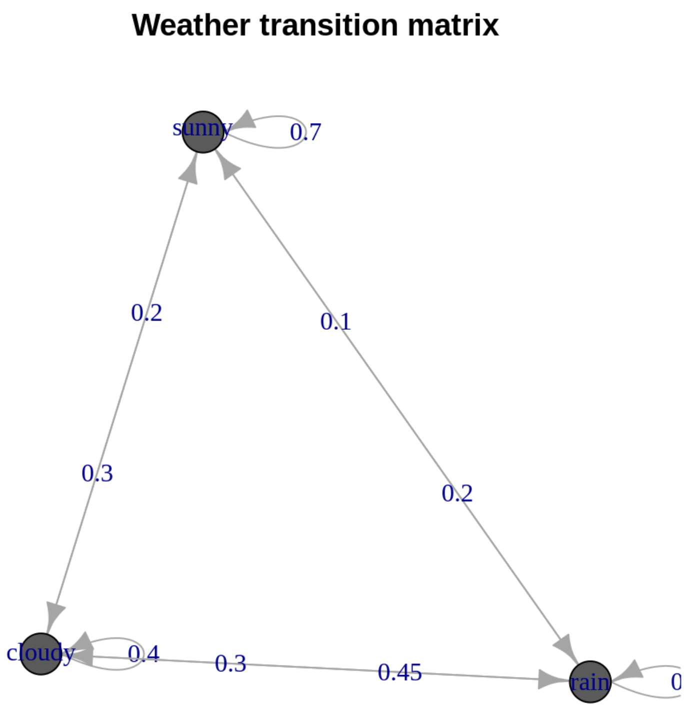

Fig. 10. Visualization of Weather Transitions as a Directed Graph. The matrix illustrates the likelihood of transitioning between weather conditions.

Table 2
Weather Condition Probabilities

Clear	Overcast	Wet
0.3	0.4	0.3

How does this relate to marketing attribution? In this scenario, weather states are replaced with marketing touchpoints such as email campaigns, paid advertisements, display banners, or social media platforms. Additional states like start, conversion, and exit represent the beginning and conclusion of the customer journey.

Using historical interaction data, a transition matrix can be constructed to map how users navigate through these marketing channels. With this framework, the significance of each channel can be analyzed by systematically removing it from the system and evaluating the resulting impact on conversion rates. This method, referred to as the 'Removal Effect' [5], is detailed further in Section 9.1. It offers a data-driven alternative to traditional rule-based attribution models, enabling deeper insights into customer behaviour and decision-making processes.

Time-to-event analysis encompasses a range of statistical methods used to examine data where the outcome is the time until a specific event occurs. While the term 'survival' is most commonly linked to medical contexts—such as forecasting how long a patient may live under certain conditions—this approach is broadly applicable in many other domains.

The fundamental elements of time-to-event analysis include [6]:

Event: This is the focal event of interest, such as a customer's purchase decision, machinery breakdown, or the occurrence of a specific disease.

Duration: The period from the start of the observation until the event takes place, the study ends, or the subject is excluded from the study (censored).

Censoring: A key concept in time-to-event analysis, censoring occurs when the event does not take place during the observation period, and thus the subject's exact time to event is unknown.

Survival Probability: This is the likelihood that an individual or unit will "survive" beyond a certain time t, without experiencing the event of interest.

A crucial aspect of survival analysis is censoring. In marketing attribution, for example, when a customer interacts with a campaign but does not convert within the observation period, we cannot assume that they will never convert.

Instead, we acknowledge that the event (conversion) has not been observed during the analysis period. To estimate the survival function, one common approach is the Kaplan-Meier estimator [7]. This method produces a step-function plot (Figure 11) showing the survival probability over time, adjusting for censored data.

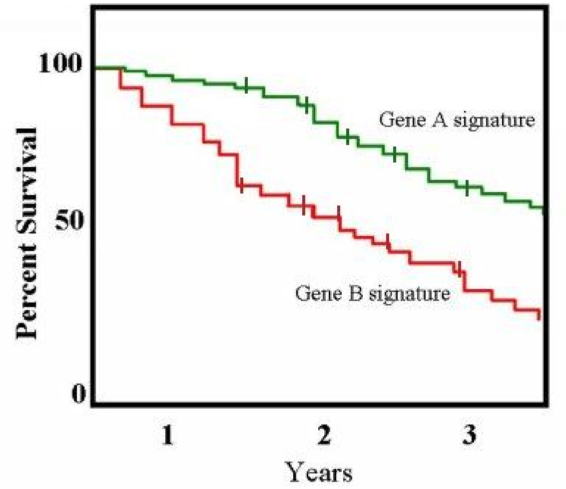

Fig. 11. Kaplan-Meier Curve for Survival Estimation. The curve depicts the probability of remaining unconverted over time, accounting for the censoring effect.
source: https://upload.wikimedia.org/wikipedia/commons/7/73/Km plot.jpg

This is mathematically expressed as:

$$\hat{S}(t) = \prod\_(i: t\_i \leq t)\ (1 - d\_i / n\_i)$$

Where Sˆ(t) is the estimated survival probability (i.e., the likelihood of no conversion) at time t. This formula calculates the survival probability at each time interval, adjusting for the number of subjects still at risk of the event.

Survival analysis can be extended to marketing attribution by analyzing the likelihood of conversion through various channels over different stages of the conversion journey. Rather than assigning direct credit to touchpoints, we can estimate the likelihood of conversion as a function of the time and channel interactions.

In the practical section in section 9.2, we will compute an estimate for the survival function and segment this by the channels through which users convert. This approach provides valuable insights into the timing and impact of different channels across the conversion journey, adding a new layer to marketing performance analysis.

10 Comparison of Methods

Which attribution model should you use? The answer isn't straightforward. There is no universally optimal model for attributing marketing credit [2]. The right choice depends on the specific goals of your marketing strategy and how they contribute to business outcomes. As a general guideline, heuristic models are often subjective and should be viewed as starting points.

More data-driven methods, based on actual customer behavior, tend to offer more reliable and objective insights.

11 Google Analytics Data

Google Analytics provides a variety of simple attribution models for users' digital data, all integrated into the platform [8]. While these basic models are a useful starting point, they are ultimately heuristic. In other words, the decision of which model to apply is hard-coded into the system. While intuitive, these methods can oversimplify the complexities of customer journeys and often lack the sophistication of data-driven models.

12 Overview of fundamental attribution models in Google Analytics

• Final Interaction Model

• Initial Interaction Model

• Equal Weight Attribution

• Time-Based Attribution

• Last Non-Direct Interaction Attribution

• Last Click via Google Ads Attribution

• Position Weighted Attribution.

For more detailed information, see the Google Help Pages.

Google also offers a Model Comparison Tool to compare up to three attribution models simultaneously.

13 Introduction to data-driven attribution

Google's Data-Driven Attribution is a feature available exclusively to users of Google Analytics 360, which is part of the broader Google Marketing Platform. In contrast to traditional rule-based models, Data-Driven Attribution leverages actual user behaviour data collected from your Google Analytics account to construct a personalized attribution model using a more sophisticated algorithm.

13.1 How the Model Functions

While conventional attribution models that rely on fixed-position strategies only analyze the conversion pathways, Google's Data-Driven Attribution (DDA) model takes a broader approach by examining both the paths that result in conversions and those that don't. Google's approach is outlined in two main phases:

1. Initially, the system processes all the available path data to create a probabilistic model that illustrates how users generally navigate through your website.

2. Subsequently, a sophisticated algorithm is applied to the data to determine how much credit should be attributed to each touchpoint in the customer journey.

The algorithm that drives the Data-Driven Attribution model is based on the Shapley Value See Section 4.3.2, a concept from cooperative game theory. This method recognizes that the value of a marketing touchpoint is influenced by its role in the broader conversion process.

By comparing customer journeys with and without a particular touchpoint, the system estimates the proportional contribution of each point in the journey. In simple terms, this approach involves removing a touchpoint from the sequence and observing how its absence impacts the final conversion outcome.

13.2 Limitations

While Google's Data-Driven Attribution offers valuable insights, there are several important limitations to be aware of. For one, the results are updated weekly, based on conversion data from the past 28 days. This allows the model to adapt to changes in online behavior.

However, the model only considers the last four interactions within a 90-day window leading up to a conversion. Since the model derives insights from past data, certain conditions must be met to ensure the results are valid and reliable. These conditions guarantee that there is an adequate volume of data to train the model effectively.

Currently, the required conditions include:

• A minimum of 400 conversions per conversion type, with at least two interactions in each path (i.e., 400 conversions for a particular goal or transaction, not a total across all goals).

• At least 10,000 distinct paths in the reporting view being analyzed (approximately equivalent to 10,000 unique users, although a single user may contribute multiple paths).

14 BiqQuery

BigQuery, developed by Google, is a robust tool designed for executing high-speed queries across large datasets.

While Google Analytics offers a variety of built-in analysis tools, accessing detailed hit-level data is essential for conducting in-depth digital analytics and attribution modeling [9].

For users utilizing Google Analytics 360, part of the Google Marketing Platform, the BigQuery Export for Analytics feature provides seamless integration with BigQuery, enabling advanced analytical capabilities.

With BigQuery, users can query both session-level and hit-level data using SQL-like syntax. This tool offers the advantage of high-speed performance and scalability while facilitating the management of tables, datasets, and jobs associated with the data.

15 Setup

Before diving into the practical sections, it is essential to ensure that all necessary dependencies are correctly installed and configured. a) Platform Generalization: Although the paper extensively uses Google Analytics 360 and BigQuery due to their accessibility and depth of data, the modeling approaches— such as survival analysis, Markov models, and causal inference—are platform-agnostic. These methodologies can be implemented with data sourced from Adobe Analytics, Mixpanel, Matomo, or even custom event logs stored in SQLbased databases. The underlying principles remain consistent, requiring only appropriate data transformation.

16 R

16.1 About R

R [10] is a powerful open-source programming language extensively used for statistical computing and data analysis. It has a rich ecosystem of packages that make it highly versatile and ideal for conducting in-depth analyses. For the purpose of this paper, R will be used to demonstrate practical examples and applications.

16.2. Installing R

To begin using R, visit https://www.r-project.org/ and follow the installation steps. You will be prompted to select a CRAN mirror, which is a server hosting the latest version of R. Detailed guides are available for Linux, Mac, and Windows users. After installing, you will have access to R's commandline interface for performing various computations.

16.3 Installing RStudio

For an enhanced experience with R (Figure 12), it is recommended to install RStudio [11], an integrated development environment (IDE) that provides a more user-friendly interface and additional features.

Fig. 12. R Programming Language Logo. R is a widely used tool for statistical analysis and data science.

You can download RStudio from https://www.rstudio.com/ products/rstudio/ and follow the installation instructions for your operating system (Figure 13).

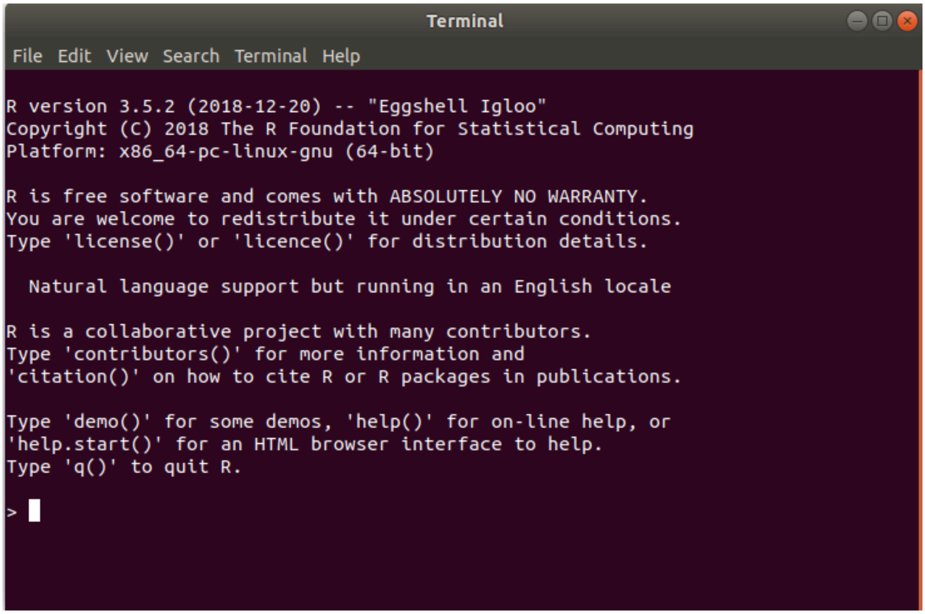

Fig. 13. R Command Line Interface. The command-line interface appears after installing R.

Once installed, your RStudio interface will resemble the one shown below.

16.4 Installing Necessary R Libraries

R provides a broad selection of libraries that enhance its functionality for specific analytical tasks. These libraries (Figure 14) can be easily installed directly from the R environment, making them readily accessible for use during your analysis.

In this document, we will utilize the following essential R libraries:

• tidyverse - A suite of R libraries created to streamline data science workflows. The tidyverse libraries adhere to a unified design philosophy, simplifying data manipulation and visualization [12].

• bigrquery - This library facilitates interaction with data stored in Google BigQuery, allowing users to run queries and retrieve metadata from BigQuery tables [13].

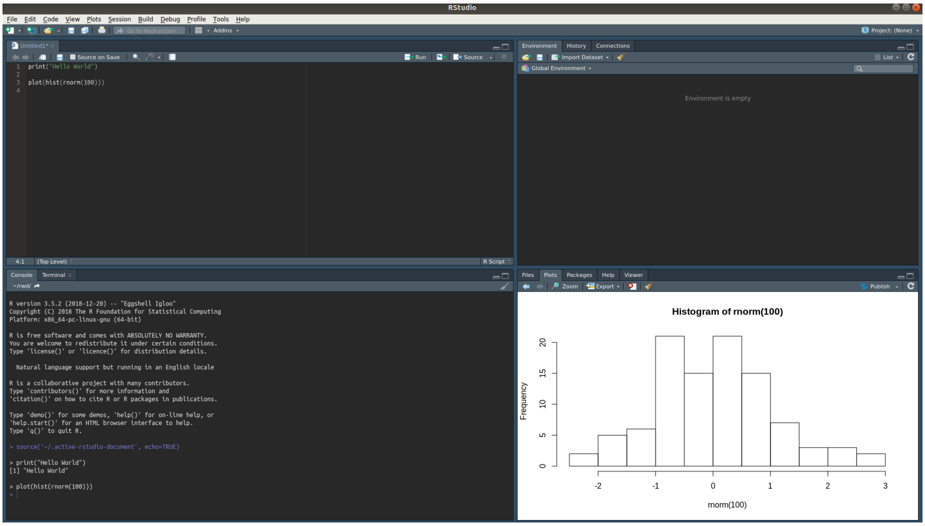

Fig. 14. RStudio IDE Interface. RStudio offers a graphical interface to enhance your experience with R.

• ChannelAttribution - Implements a Markov-based framework for analyzing multi-channel attribution in digital marketing [14].

• survival - A library for performing survival analysis, including tools for Kaplan-Meier estimations and Cox proportional hazards models [15].

• CausalImpact - This library applies Bayesian analysis to evaluate the causal effect of time-series events, such as marketing campaigns [16].

17 BigQuery

To perform more advanced modeling, the first step is to extract and clean the data. In this section, we will go through the basic setup and familiarization process to access your

Analytics data from BigQuery. Familiarity with SQL and basic database operations is expected.

17.1 Configuration Process

For users with an Analytics 360 subscription, it is possible to integrate BigQuery with Google Analytics data. For stepby-step guidance, please consult the Google Support Documentation.

If you do not have access to Analytics 360, you can still follow along with these examples by utilizing the free Google

Analytics Sample Dataset, which contains comprehensive data from the Google Merchant Store.

To begin exploring the dataset, follow these instructions:

1. Navigate to http://bigquery.cloud.google.com.

2. If you're new to BigQuery or haven't set up a project yet, create one now.

3. Use this link to directly access the dataset.

4. Click on Query Table to start querying the available data.

Moving forward, you can retrieve the dataset in BigQuery by selecting the bigquery-public-data project from the left panel, followed by choosing the ga_sessions table within the google_analytics_sample dataset.

Please be aware that using BigQuery may result in charges, so we recommend reviewing Google's billing and quota details prior to beginning your analysis.

17.2 Export Schema

In Google Analytics, the data export structure when sent to BigQuery is referred to as the export schema. Unlike normalized relational databases, this schema utilizes de-normalized tables, where nested records replace the need for joining multiple flat tables (Figure 15). This structure can be unfamiliar to those accustomed to relational database models.

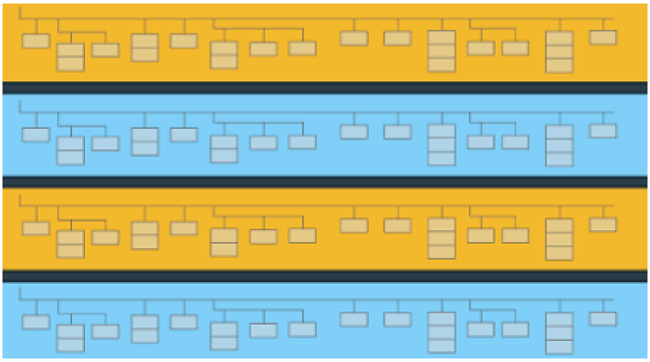

Fig. 15. Export Schema of Google Analytics in BigQuery. The schema employs de-normalized tables with nested data records.

Typically, in BigQuery, each table corresponds to a single day's data, organized at the session level. These tables house all Analytics-related information, including events and hits, nested within the data. Below, we outline some fundamental queries that can be run on this data.

For more details, refer to the BigQuery Export Schema documentation.

17.3 Running Queries

1) Hello World: This simple query demonstrates how to calculate the total page views (Table 3) for a particular day. Notice the use of the #standardSQL directive at the beginning, which instructs BigQuery to utilize the modern SQL syntax instead of the legacy version.

Listing 1. Query for Total Sessions by Country

```
#standardSQL
SELECT geoNetwork.country AS Country, COUNT(*) AS
TotalSessions
FROM 'bigquery-public-data.google_analytics_sample.
ga_sessions_20170101'
GROUP BY Country
ORDER BY TotalSessions DESC
```

Table 3 Sample Of Total Pageviews For A Specific Day

Row	Total Pageviews
1	5362
2	1234
3	9876
4	4356
5	6789
6	2345

2) Using Table Ranges: As the export schema assigns a table per day, we can query over a range of dates by specifying the table range using the * symbol and _TABLE_SUFFIX.

Listing 2. Querying Over Table Ranges

```
#standardSQL
SELECT date,
SUM(totals.pageviews) AS TotalPageviews
FROM 'bigquery-public-data.google_analytics_sample.
ga_sessions_*'
WHERE _TABLE_SUFFIX BETWEEN '20170725' AND '20170801'
GROUP BY date
ORDER BY date
```

3) Hit Level Data: To access individual hits within each user's session, the hit data needs to be "un-nested" from the session table.

Listing 3. Querying Detailed Hit-Level Data for Campaign Visitor

```
#standardSQL
SELECT fullVisitorId,
visitNumber,
date,
totals.hits,
hits.hitNumber,
hits.type,
hits.time,
hits.page.pagePath
FROM 'bigquery-campaign-data.ga_sessions_20240201',
UNNEST(hits) AS hits
WHERE fullVisitorId = '8251309475389207341'
/*Random Visitor Selected for Campaign Analysis */
```

With these foundational skills, we can proceed in upcoming

sections to extract data necessary for more advanced modeling.

17.4 Using R with BigQuery

Another option is to interact with BigQuery through the bigrquery R package.

To begin, we load the required packages: bigrquery for BigQuery interaction, and DBI for database handling.

Listing 4. Loading Required R Packages

```
library(bigrquery)
library(DBI)
```

Then, we establish a connection to BigQuery, replacing bq_test_project() with your project ID from the Google Cloud Console.

Listing 5. Establishing Connection to BigQuery for Campaign Analysis

```
con <- dbConnect(
bigrquery::bigquery(),
project = "bigquery-campaign-data",
dataset = "campaign_analytics",
```

```
billing = bq_campaign_project()
)
```

We are able to store any SQL queries as text strings and send

them for execution.

Listing 6. Saving Custom SQL Query for Traffic Analysis

```
sql <- "
/* Traffic Analysis for New Campaign */
#standardSQL
SELECT campaignDate,
SUM(metrics.pageViews) AS TotalPageviews
FROM 'ecommerce_data.campaign_sessions_*'
WHERE _TABLE_SUFFIX BETWEEN '20240201' AND
'20240215'
GROUP BY campaignDate
ORDER BY campaignDate
"
```

In this example, the query is run and saved as a temporary table (Table 4). The function bq_table_download retrieves the results and displays them in the console.

Listing 7. Executing SQL Query for Campaign Traffic in R

```
query <- bq_project_query(x = bq_campaign_project(),sql)
bq_table_download(query)
```

Table 4

Total Pageviews For A Specific Day

Date	Total Pageviews
20170725	10728
20170726	11200
20170727	10175
20170728	9359

20170729	6293
20170730	7258
20170731	11115
20170801	10939

18 Retrieving Data From Bigquery

The initial step in applying attribution models within R is to extract the required data from BigQuery. This guide assumes that you are a Google Analytics 360 user and that

BigQuery integration is already configured. If you're utilizing other sources of multi-touch conversion data, feel free to skip this section. In this example, we will work with the bigquery-public-data dataset, which features a sample of Google Analytics 360 data from the Google Merchant Store

(for setup instructions, see Section 6.2).

19 Understanding The Data

A crucial part of the analysis is selecting the appropriate volume of data for modeling purposes.

It's usually not feasible to analyze every single clickstream interaction within your organization. In this case, we specify the parameters for the data sample that will form the basis of our models.

• Retrieve data from sessions within the past 30 days.

• Include both converted and non-converted users.

• Consider a 7-day period prior to each visitor's latest session.

• For visitors who have converted, exclude any sessions occurring after the conversion.

These settings are just an example and can be tailored based on the specific needs of your organization. For example, websites with extended conversion paths may choose to analyze data from the past six months, applying a longer lookback window of one month.

20 Get Full Event Log

To begin, we need to retrieve a complete log of all sessions within a specific time frame. In this case, we have selected a 30-day window. The use of the * wildcard in the FROM clause allows us to query multiple dates in a single query.

We have chosen fullVisitorId as the unique identifier. Additionally, we capture the session's start time and the channel Grouping to determine the Default Channel Group associated with each user's session.

Listing 8. Extracting Recent Session Data for Customer Engagement

```
SELECT customerId,
TIMESTAMP_SECONDS(SAFE_CAST(sessionStart AS
INT64)) AS sessionStartTime,
marketingChannel
FROM 'ecommerce_data.user_sessions_*'
WHERE _TABLE_SUFFIX BETWEEN '20240201' AND '20240228'
ORDER BY customerId, sessionStartTime
```

21 Identify Those Who Converted

Next, we will identify visitors who completed a transaction during the specified time range, along with the transaction date. Although we define a ”transaction” as the conversion goal in this case, any other event, such as a sign-up, could be used as the conversion goal.

Listing 9. Identifying Customers Who Made a Purchase

```
SELECT customerId,
MIN(TIMESTAMP_SECONDS(SAFE_CAST(sessionStart
AS INT64))) AS purchaseTime
FROM 'ecommerce_data.customer_sessions_*'
WHERE _TABLE_SUFFIX BETWEEN '20240201' AND '20240228'
AND transactionAmount IS NOT NULL
GROUP BY customerId
ORDER BY customerId
```

Interpretability and Visualization for Stakeholders To make these complex models more accessible to marketing professionals, visual tools such as Sankey diagrams, funnel plots, and survival curves are highly recommended. For instance, survival curves clearly illustrate when conversions tend to happen, while Markov diagrams can visualize channel dropoff and conversion pathways. Implementing Shapley valuebased bar charts also helps convey the marginal contribution of each channel in simple visual terms. These elements bridge the gap between model output and business decision-making.

22 Converting Touchpoints In Last 7 Days

In this step, we identify the touchpoints leading to conversions.

These are extracted from the event log generated in Step 1, but only for those visitors who converted. Additionally, we restrict the results to touchpoints that occurred before the purchase time. Touchpoints after the transaction are irrelevant to the conversion outcome. A 7-day look-back window is applied to include only touchpoints within that timeframe.

Listing 10. Identifying Purchasing Touchpoints

```
SELECT a.*,
'purchase' AS outcome
FROM customer_interactions a
INNER JOIN purchases b ON a.customerId = b.
customerId
WHERE a.sessionStart <= b.purchaseTime
AND DATE_DIFF(DATE(b.purchaseTime), DATE(a.
sessionStart), DAY) <= 7
```

Data Ethics and Privacy Considerations

This study emphasizes the importance of adhering to data ethics and privacy regulations when handling granular user data. In compliance with GDPR and similar frameworks, all personally identifiable information (PII) must be anonymized or pseudonymized before analysis. Event logs should exclude user IDs that could lead to re-identification, and consent mechanisms must be enforced during data collection. Furthermore, privacy-preserving techniques such as differential privacy and secure multi-party computation can be employed for more sensitive datasets.

23 Non-Converting Touchpoints In Last 7 Days

For users who did not convert, we need to identify their most recent session to establish the starting point for our 7-day lookback period.

Listing 11. Identifying Latest Interaction for Non-Purchasers

```
SELECT customerId,
MAX(sessionStart) AS latestSession
FROM customer_interactions
GROUP BY customerId
```

Subsequently, we retrieve the touchpoints for nonconverting users, applying the 7-day window.

Listing 12. Identifying Non-Converting Customer Interactions

```
SELECT a.*,
'non_purchase' AS outcome
FROM customer_interactions a
INNER JOIN last_visit_time b ON a.customerId = b.
customerId
WHERE a.customerId NOT IN (SELECT DISTINCT
customerId FROM successful_purchases)
```

```
AND DATE_DIFF(DATE(b.lastVisit), DATE(a.
sessionStart), DAY) <= 10
```

24 Complete Query

Now, we can combine all the steps into a single query as demonstrated below:

Listing 13. SQL Query for Customer Purchase Path Analysis

```
#standardSQL
WITH
/* VISIT LOG */
visit_log AS (
SELECT customerId,
TIMESTAMP_SECONDS(SAFE_CAST(sessionStart
AS INT64)) AS sessionStart,
trafficSource
FROM 'bigquery-public-data.sample_data.web_
sessions_*'
WHERE _TABLE_SUFFIX BETWEEN '20240301' AND '
20240331'
ORDER BY customerId, sessionStart
),
/* CUSTOMERS WHO MADE A PURCHASE */
purchases AS (
SELECT customerId,
MIN(TIMESTAMP_SECONDS(SAFE_CAST(
sessionStart AS INT64))) AS
purchaseTime
FROM 'bigquery-public-data.sample_data.web_
sessions_*'
WHERE _TABLE_SUFFIX BETWEEN '20240301' AND '
20240331'
AND totals.orders IS NOT NULL
GROUP BY customerId
ORDER BY customerId
),
/* MOST RECENT VISIT FOR EACH CUSTOMER */
recent_visits AS (
```

```
SELECT customerId,
MAX(sessionStart) AS recentVisitTime
FROM visit_log
GROUP BY customerId
)
/*== MAIN QUERY THAT COMBINES PURCHASE AND NONPURCHASE
PATHS==*/
/* PURCHASE PATHS */
SELECT a.*, 'purchase' AS outcome
FROM visit_log a
INNER JOIN purchases b ON a.customerId = b.
customerId
WHERE a.sessionStart <= b.purchaseTime
AND DATE_DIFF(DATE(b.purchaseTime), DATE(a.
sessionStart), DAY) <= 10
UNION ALL
/* NON-PURCHASE PATHS */
SELECT a.*, 'non_purchase' AS outcome
FROM visit_log a
INNER JOIN recent_visits b ON a.customerId = b.
customerId
WHERE a.customerId NOT IN (SELECT DISTINCT
customerId FROM purchases)
AND DATE_DIFF(DATE(b.recentVisitTime), DATE(a.
sessionStart), DAY) <= 10
```

25 Heuristic Models

This section outlines the implementation of heuristic models, which serve as a baseline for comparison. We utilize the ChannelAttribution R package, offering a variety of methods for attribution based on predefined rules and heuristics.

26 Transforming The Data

The ChannelAttribution package requires the data to be formatted in a particular way. Below is an example illustrating how the data should appear:

In the table above, the touchpoints are concatenated into a single path string, with each part separated by the > symbol.

For every path, we count the total number of conversions and non-conversions resulting from this sequence of events (Table 5).

TABLE 5 Path Conversions Non-Conversions

Path	Conversions	Non-Conversions
direct ¿ social ¿ search	10	154
direct ¿ direct ¿ direct	2	234
referral ¿ direct	7	187

If you are working with marketing data from a different source other than BigQuery, make sure to format the data as shown here.

To transform the data from Chapter 7, follow the procedure outlined below. In this case, the results will be saved as a CSV file, though you can also query the data directly from R, as demonstrated in Section 6.2.4.

Listing 14. Loading R Packages for Data Manipulation and Date Handling

```
library(tidyverse)
library(lubridate)
paths_raw <- read_csv('bigquery/bq-results.csv')
```

Here is a preview of the first 20 rows. As you can see, the data fetched from BigQuery is in a conventional long format, with each row representing a touchpoint based on its timestamp.

27 Data Transformation Process

For clarity, we are employing the conventions of the tidyverse package. The transformation process begins by formatting the timestamp, followed by ordering the sessions by their timestamps for each visitor to establish the sequence of touchpoints. Then, we restructure the data by combining the touchpoints into a single path string. Finally, we tally the occurrences of conversions and non-conversions for each path.

Here is the R code used for this transformation:

Listing 15. Converting Data for Pathway Conversion Analysis

```
# Data manipulation using tidyverse
pathways <- raw_data %>%
mutate(sessionStart = ymd_hms(sessionStart)) %>%
group_by(visitorId, result) %>%
arrange(sessionStart) %>%
```

```
summarise(pathway = paste(channelCategory,
collapse = ' > ')) %>%
ungroup() %>%
count(result, pathway, name = "frequency") %>%
spread(result, frequency) %>%
replace_na(list(success = 0, failure = 0)) %>%
arrange(desc(success))
```

A. Resulting Data

After the data transformation, we arrive at a summarized table, as shown in Table VI. The table reflects the total conversions and non-conversions for each user journey.

28 Attribution Models

Attribution models are essential in marketing analysis. They help distribute the credit for a conversion across multiple channels involved in a customer's journey. We will implement several heuristic attribution models, such as First Touch, Last Touch, and Linear, using the Channel Attribution R package.

Each model assigns value to the touchpoints in the user's journey differently. The First Touch model assigns full credit to the first channel interaction, whereas the Last Touch model attributes all the credit to the last interaction. The Linear model distributes credit evenly across all touchpoints in the conversion path.

Listing 16. The R Code for Calculating these Models

```
# Applying heuristic attribution models
library(ChannelAttribution)
model_results <- ChannelAttribution::heuristic_
models(data = paths)
```

These results help determine which channels have the greatest impact on conversions.

Listing 17. R Code for Heuristic Models

```
library(ChannelAttribution)
fit_h <- heuristic_models(Data = paths, var_path = '
path', var_conv = 'conversion')
```

The table below (Figure 16) displays the conversion attribution across the three different models:

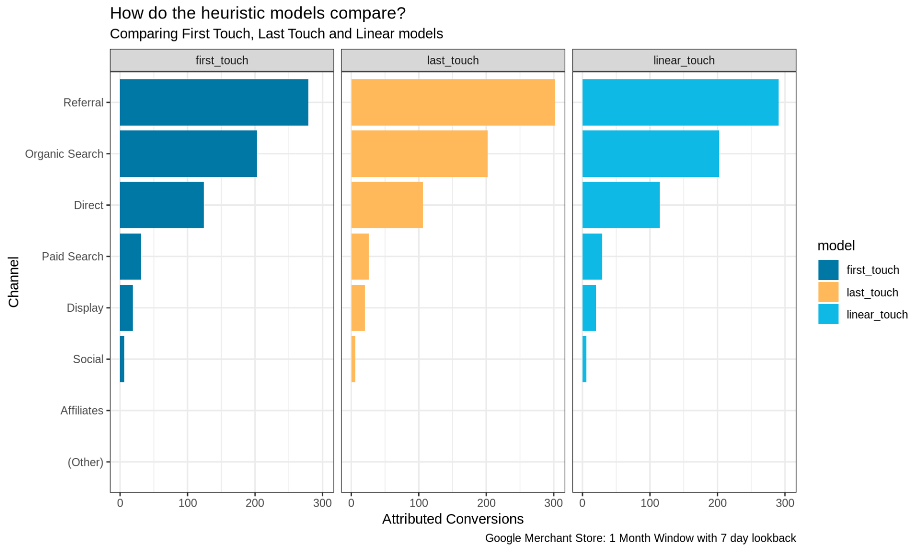

Fig. 16. Visualization of Conversion Attribution Methods

29 Algorithmic Approaches

This section focuses on algorithmic techniques for modeling marketing attribution. These approaches employ sophisticated statistical models to build data-driven methods for assigning conversion credit to various channels within multi-touch point conversion pathways.

30 Markov Chains

The initial method we'll examine is the Markov chain approach, which was introduced in Section 4.3.3.

First, we load the necessary packages. As in earlier cases, we use the ChannelAttribution package, along with tidyverse and lubridate for manipulating data and handling date-related operations. We also load our raw BigQuery results, which contain event logs with timestamps of visits, channel information, and outcome (Table 6).

Table 6 Visitor Interaction Data With Results

visitorId	interactionTimestamp	sourceMedium	purchaseOutcome
248539384758162581	2024-04-10	Organic Search	no purchase

	08:30:11 UTC		
248539384758162581	2024-04-10 09:15:05 UTC	Organic Search	no purchase
348572023947023547	2024-04-12 14:42:34 UTC	Social Media	purchase
405862093748539214	2024-04-14 17:28:11 UTC	Paid Search	no purchase
463028239840193864	2024-04-13 11:33:22 UTC	Referral	no purchase
594823659302848971	2024-04-15 12:18:55 UTC	Display Ads	no purchase
595840283593048374	2024-04-17 09:21:46 UTC	Email	purchase
628173839472736402	2024-04-09 07:58:03 UTC	Organic Search	purchase
693847283820537109	2024-04-16 13:45:12 UTC	Social Media	no purchase
803947829402930121	2024-04-11 19:02:50 UTC	Paid Ads	no purchase

Listing 18. Importing Necessary Packages and Dataset

```
library(dplyr)
library(stringr)
library(MarkovChainAnalysis)

raw_data <- read.csv('bigquery/output-data.csv')
```

The raw event data extracted from BigQuery is shown below (Table 7):

Table 7 Attribution Of Conversions Based On Heuristic Models

Marketing Channel	Initial Touch	Final Touch	Equal Attribution
Referral	279	302	290.47
Organic Search	203	202	202.28
Direct Traffic	124	106	114.32
Paid Search Ads	31	26	29.08
Display Advertising	19	20	19.85
Social Media	6	6	6.00
Other Sources	0	0	0.00 s
Affiliate Marketing	0	0	0.00 v

We now proceed to transform the data. This transformation results in one row per conversion path, with totals for conversions and non-conversions.

The transformed data is shown below:

31 Marketing Attribution Analysis

31.1 Markov Model Attribution

The markov_model() function from the ChannelAttribution package takes several parameters, including the data frame, the variable representing the conversion path, the variable indicating conversions and non-conversions, and the order of the Markov model. The implementation is provided below:

Listing 19. First-order Markov Model Implementation.

```
fit_m <- markov_model(Data = paths,
var_path = 'path',
var_conv = 'conversion',
var_null = 'non_conversion',
order = 1)
fit_m
%% Example output:
%% channel_name total_conversions
```

```
%% 1 Referral 289.7406085
%% 2 Organic Search 209.9964297
%% 3 Direct 106.2056815
%% 4 Paid Search 27.3863707
%% 5 Display 19.7819000
%% 6 Social 7.2961813
%% 7 (Other) 0.1541447
%% 8 Affiliates 1.4386836
```

To calculate the conversion values for each channel, marketers can multiply these results by the average conversion value or use actual sales data, if available (using the var_value argument).

31.2 Iterative Markov Models

The function can be expanded to compute higher-order Markov models. Below is an example where a 1st, 2nd, and 3rd-order model are calculated iteratively:

Listing 20. Iterative Markov model computation.

```
fit_mult <- map_dfr(.x = c(1, 2, 3),
.f =  markov_model(Data = paths,
var_path = '
path',
var_conv = '
conversion
',
var_null = '
non_
conversion
',
order = .x),
.id = "order")
```

The results show minimal variations between different orders, as illustrated in Figure XXXIII.

31.3 Survival Analysis

Survival analysis treats conversion as an event of interest (Table 8), with non-conversion being regarded as ”survival.” The setup is illustrated below:

Listing 21. Customer Retention Analysis and Visualization.

```
% Load required packages
```

```
library(tidyverse)
library(lubridate)
library(survival)
library(survminer)
% Import data
customer_data <- read_csv('data/customer_engagement.csv')
```

Listing 22. This Code Fits the Survival Model to Analyze Customer Churn Over Time

```
% Fit retention model
retention_fit <- survfit(Surv(time_since_signup,
churn)   1, data = customer_data)
% Visualize retention curve
ggsurvplot(retention_fit, data = customer_data,
conf.int = TRUE,
ggtheme = theme_light())
```

32 Data Transformation For Survival Analysis

32.1 Raw Data

Table provides an overview of the initial dataset, where each row corresponds to a customer's visit (Table 8). The row includes the customer's ID, visit timestamp, marketing channel, and the outcome (conversion or non-conversion).

Table 8 Raw Interaction Log Data For Customer Engagement

customerID	interactionTimestamp	sourceChannel	purchaseOutcome
478920483764509312	2024-03-01 08:45:12 UTC	Push Notification	no purchase
478920483764509312	2024-03-01 09:00:34 UTC	Push Notification	no purchase
478920483764509312	2024-03-01 09:30:45 UTC	Push Notification	no purchase
329850291348743251	2024-03-05 11:24:18 UTC	Organic Search	no purchase
924873601938573648	2024-03-02 15:17:53 UTC	Email	no purchase

345892347092834679	2024-03-01 13:52:10 UTC	Referral	purchase
345892347092834679	2024-03-01 13:55:30 UTC	Referral	purchase
503987247609283745	2024-02-25 10:22:37 UTC	Display Ads	no purchase
624809283749238015	2024-02-28 17:10:08 UTC	Paid Search	no purchase
173920184673829447	2024-03-07 14:42:29 UTC	Social Media	no purchase

32.2 Data Transformation

For survival analysis, the dataset needs to be condensed into one row per customer. The transformed dataset captures the following key information:

32.2.1 Time Interval:

• For converted customers, it is the duration between the first visit and the conversion.

• For non-converted customers, it is the period between the first visit and the last recorded visit.

32.2.2 Outcome:

• 1 indicates a converted customer.

• 0 denotes a non-converted customer.

32.2.3 Conversion Channel:

• The primary channel that led to conversion for each customer (Table 9).

Table 9 User Acquisition Channels: Conversions Vs. Non-Conversions

Acquisition Path	Conversions	Non-Conversions
Social Media	250	1345
Email Campaign	182	5121
Referral Program	95	1745

Organic Search	58	8796
Direct Traffic	42	6250
Paid Ads	37	1289
Social Media ¿ Email Campaign	20	120
Referral ¿ Referral Program	15	85
Email Campaign ¿ Social Media	13	72
Display Ads	8	225

32.3 R Implementation

The following R code demonstrates the transformation process:

Listing 23. Data Transformation for Survival Analysis

```
% Load necessary libraries
library(tidyverse)
library(lubridate)
% Load raw data
paths_raw <- read_csv('bigquery/bq-results.csv')
% Transform data
paths_transformed <- paths_raw %>%
group_by(fullVisitorId) %>%
summarize(
first_visit = min(visitStartTime),
last_visit = max(visitStartTime),
outcome = ifelse(any(outcome == "conversion"),
1, 0),
channel = ifelse(any(outcome == "conversion"),
channelGrouping[outcome == "
conversion"],
channelGrouping[1])
) %>%
mutate(
```

```
time_interval = ifelse(outcome == 1,
difftime(last_visit,
first_visit, units =
"days"),
difftime(last_visit,
first_visit, units =
"days"))
)
```

32.4 Output

The transformed dataset compiles the essential variables for survival analysis, aligning each customer's journey into a single row (Table 10).

Firstly, it is important to note that this analysis differs from attribution modeling. We are not directly assigning credit between channels; instead, we are analyzing the probability of a customer converting through any specific channel at various points in time.

Listing 24. Excluding Non-Converting Customers and Preparing Data

```
# Data preprocessing for survival analysis
processed_data <- raw_sessions %>%
mutate(sessionStartTime = ymd_hms(sessionStartTime
)) %>%
group_by(visitorId) %>%
mutate(firstVisit = min(sessionStartTime),
lastVisit = max(sessionStartTime)) %>%
filter(sessionStartTime == lastVisit) %>%
mutate(duration = (lastVisit - firstVisit) / 3600,
conversion_status = ifelse(event == "
conversion", 1, 0)) %>%
select(visitorId, duration, conversion_status, trafficSource)
```

Table 10

Sample Dataset Description

User ID	Session Start time	Traffic Source	Result
12345678901234567	2023-05-12 10:15:00 UTC	Organic Search	Non-conversion
23456789012345678	2023-05-12 11:30:45	Referral	Conversion

	UTC		
34567890123456789	2023-05-13 08:50:30 UTC	Paid Search	Non-conversion
45678901234567890	2023-05-14 14:20:10 UTC	Social Media	Conversion
56789012345678901	2023-05-15 09:40:25 UTC	Organic Search	Non-conversion
67890123456789012	2023-05-16 17:05:50 UTC	Email	Non-conversion
78901234567890123	2023-05-17 19:45:30 UTC	Organic Search	Conversion
89012345678901234	2023-05-18 12:30:00 UTC	Paid Search	Non-conversion
90123456789012345	2023-05-19 16:10:15 UTC	Direct	Conversion
12389012345678901	2023-05-20 18:25:40 UTC	Referral	Non-conversion

Listing 25. Updated tibble data with diverse channel groupings

```
# A tibble: 10 x 4
# Groups: fullVisitorId [10]
fullVisitorId time status
channelGrouping
<chr> <drtn> <dbl> <chr>
1 13579246801357924 68 0.01456789 1 Social
Media
2 24681357902468135 79 0.00345678 0 Paid
Search
3 86420975318642097 53 0.00000000 0 Email
Campaign
4 97531864209753186 42 0.00765432 1 Display
```

```
Ads
5 1029384756102938475 0.01234567 1 Video
Marketing
6 5647382910564738291 0.00000000 0 Affiliate
Links
7 9876543210987654321 0.02000000 1 Partner
Referral
8 1234567890123456789 0.00012345 0 SMS
Promotions
9 1928374650912837465 0.01876543 1 Push
Notification
10 6667778889991112233 0.00567891 1 In-App Ads
```

Next, we create a survival object using the Surv function.

Listing 26. Creating a Survival Object for Analysis

```
# Generating a survival object for modeling
survival_input <- Surv(duration = processed_data$
duration, event = processed_data$conversion_status)
```

We can now calculate the survival curve estimate using the survfit function.

We include a grouping variable for the converting channel. This allows the calculation of a survival curve for each group, enabling comparisons between them.

Listing 27. This R Code and Output for Customer Retention Model

```
fit_ret <- survfit(surv_object   marketingChannel, data =
customer_data)

## Call: survfit(formula = surv_object
## marketingChannel, data = customer_data)
##
## median    n events

## marketingChannel=Email 5000 50 NA
## marketingChannel=Social Media 12000 150 NA
## marketingChannel=Paid Ads 3000 30 NA
## marketingChannel=Referral 2000 40 NA
## marketingChannel=Organic Search 15000 100 NA
```

```
## marketingChannel=Direct Traffic 4000 20 NA
## marketingChannel=Affiliate 1000 10 NA
## marketingChannel=Influencer 2500 60 NA
##                             0.95LCL 0.95UCL
## marketingChannel=Email 45 NA
## marketingChannel=Social Media 120 NA
## marketingChannel=Paid Ads 25 NA
## marketingChannel=Referral 35 NA
## marketingChannel=Organic Search 95 NA
## marketingChannel=Direct Traffic 15 NA
## marketingChannel=Affiliate 5 NA
## marketingChannel=Influencer 55 NA
```

The results are most effectively viewed in a graphical format.

It is crucial to understand that the term ”survival” in this context refers to ”non-conversion.” The survival curve displays, up to a certain point on the x-axis, the probability of a customer not converting through a specific channel (Figure 17).

To examine conversion probabilities, we consider the complement of the survival curve, which represents the cumulative event plot (Figure 18).

From this plot, we observe that the Referral channel shows high conversion rates early on, indicating its value in driving conversions quickly. This is followed by the Display channel.

As time progresses (roughly 150 hours or approximately 6 days after the first visit), channels such as ‘Direct’ and ‘Organic Search’ display higher conversion probabilities. This suggests that advertising and referral channels are essential for early awareness, while customers returning through direct and organic channels indicate a strong recall and engagement after some time.

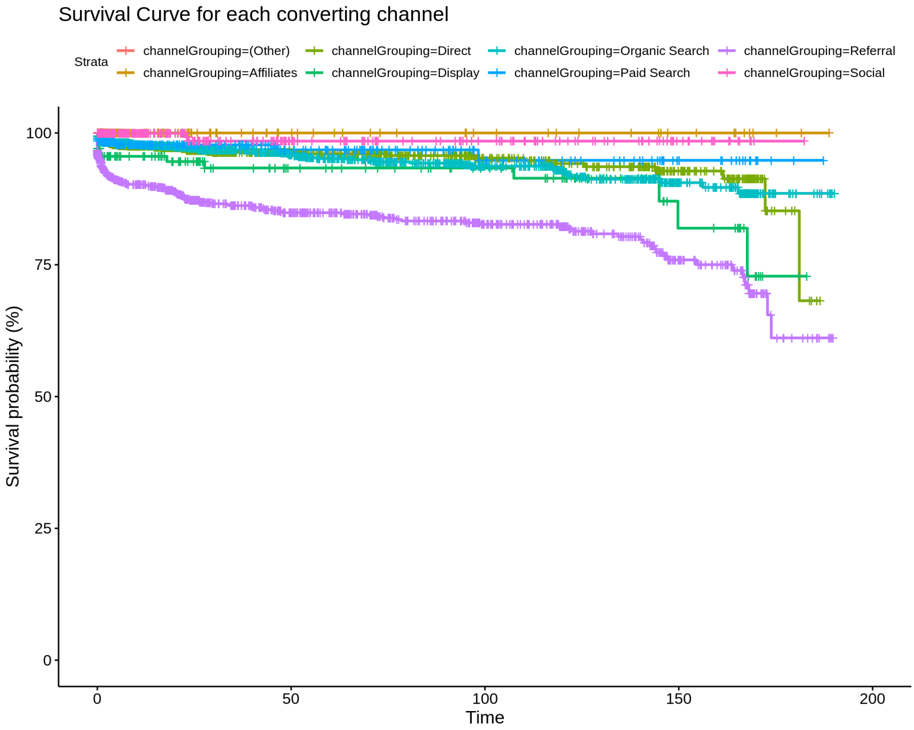

Fig. 17. Survival Curve for Different Channels

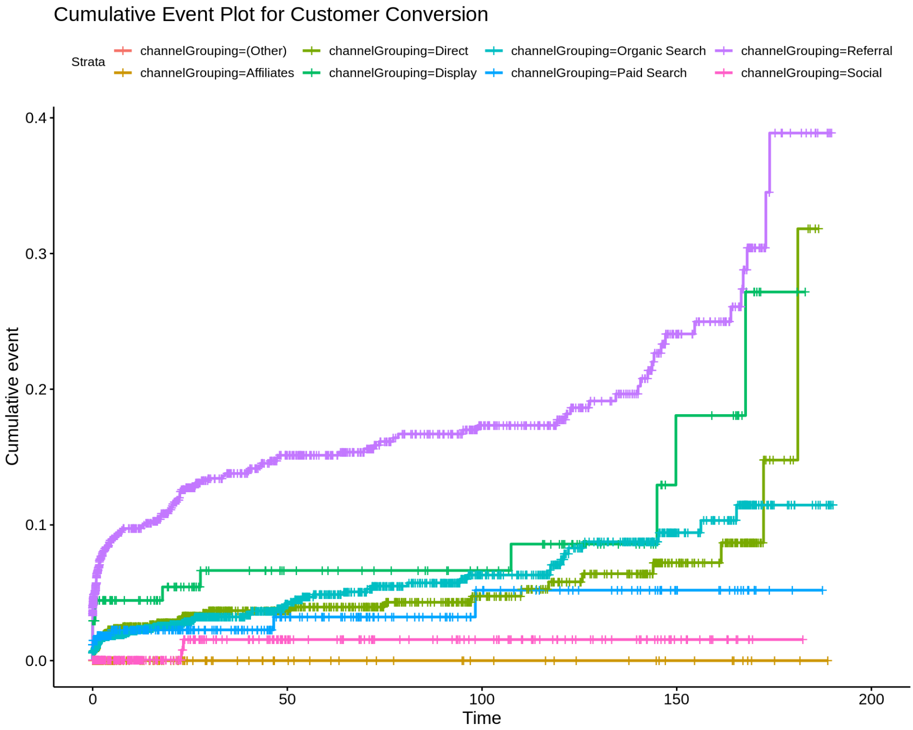

Fig. 18. Cumulative Event Plot for Different Channels

33 About Offline Attribution

Up to this point, we have concentrated on digital marketing channels. However, offline media outlets, such as television, radio, print media, coupons, and outdoor advertising, play a significant role in numerous businesses. Unlike online attribution, measuring the effect of offline marketing efforts is often more challenging.

When randomized controlled trials are not an option, inferential techniques can provide valuable assistance. Below, we illustrate one such method for causal inference.

Developed by Google, the CausalImpact R package applies Bayesian structural time-series models for causal inference [16].

Let us now delve into the workings of this approach.

34 Scenario Overview

Imagine you operate a store that specializes in selling smart home devices. Over the course of the year, sales and online visits to your website experience fluctuations driven by external factors like holidays and trends.

To boost visibility, you decide to air a commercial specifically promoting your smart devices during a peak shopping season.

The commercial is launched on a specific date and is only broadcasted in one region, not nationally.

How can we evaluate the influence of this TV ad on the website's traffic?

One straightforward approach would be to observe the change in visits before and after the commercial airs. To refine this analysis, we could compare the traffic in the targeted region with the rest of the markets that did not receive the advertisement.

However, the crucial question is: what would have happened if the commercial had not been aired? If the campaign aligns with a high-demand period, how much of the traffic spike is genuinely driven by the ad versus the natural seasonal growth?

Could we be overestimating the impact of the marketing campaign and inflating the effectiveness of our investment.

35 Causal Effect Analysis

The proposed approach involves three primary steps:

1) Determine a baseline group, such as website visits from a location unaffected by the promotional campaign.

2) Leverage historical trends from the target region to construct a model that estimates what would have occurred during the campaign period without any intervention.

This projection is referred to as the hypothetical scenario.

3) Contrast the observed website visits with the hypothetical scenario to calculate the actual impact credited to the advertisement.

Ensuring that the baseline group is independent of the campaign is crucial to avoid introducing bias into the analysis.

To start, we load the necessary libraries:

```
library(BayesImpact)
```

```
library(dplyr)
```

In this demonstration, we will generate synthetic time series data.

The dotted red line reflects the website activity in the baseline region, while the solid blue line indicates the activity in the region exposed to the campaign (Figure 19).

The advertisement period is set from day 60 to day 90, with an increase of 15 visits per day.

Listing 28. R Code Block to Simulate a Pair of Correlated Time Series

```
set.seed(2024)
control_series <- 150 + arima.sim(model = list(order = c(2,0,1), ar =
0.6), n = 120)
intervention_series <- 1.1 * control_series + rnorm
(120)
intervention_series[60:90] <- intervention_series
[60:90] + 15
time_series_data <- cbind(intervention_series,
control_series)
matplot(time_series_data, type = "l", col = c("blue"
, "red"), lty = c(1, 2))
```

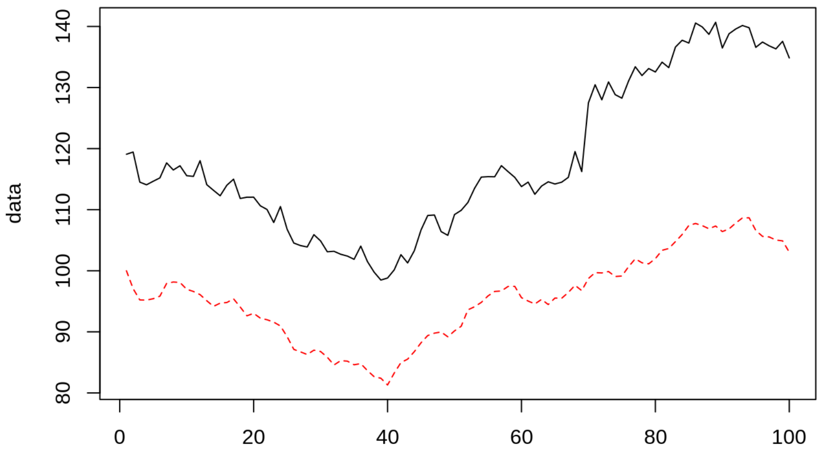

Fig. 19. Time series data showing the TV campaign period (day 70-100) and an uplift of 10 sessions per day.

Next, we define the pre- and post-periods for the causal inference analysis.

pre.period <- c(1, 70)

post.period <- c(71, 100)

We can now apply the causal inference model using the CausalImpact() function.

Listing 29. Summarizes a Causal Impact Analysis

```
impact <- CausalImpact(data, pre.period, post.period)

Posterior inference {CausalImpact}
Average Cumulative
Actual 135 4062
Prediction (s.d.) 125 (0.86) 3764 (25.69)
95% CI [124, 127] [3713, 3815]
Absolute effect (s.d.) 9.9 (0.86) 297.6 (25.69)
95% CI [8.2, 12] [246.7, 349]
Relative effect (s.d.) 7.9% (0.68%) 7.9% (0.68%)
95% CI [6.6%, 9.3%] [6.6%, 9.3%]
Posterior tail-area probability p: 0.00101
Posterior prob. of a causal effect: 99.89899%
For more details, type:summary(impact, "report")

Warning: Removed 100 rows containing missing values (geom_path).
Warning: Removed 200 rows containing missing values (geom_path).
```

We can visualize the outcomes to better understand the results.

The actual facet shows the real website traffic (black solid line) compared to the predicted traffic without the marketing campaign (blue dashed line). Vertical dashed lines indicate the intervention period, and the confidence intervals are highlighted in blue shading.

The difference plot displays the gap between observed and forecasted traffic. The aggregate graph represents the total cumulative impact of the marketing campaign, providing a summary of its overall effect.

Listing 30. This Code Plots the Results of the Impact Analysis

```
plot(impact_analysis)
## Warning: Excluded 100
rows due to missing data
```

```
## (geom_path).
##
## Warning: Excluded 200
rows due to missing data
## (geom_path).
```

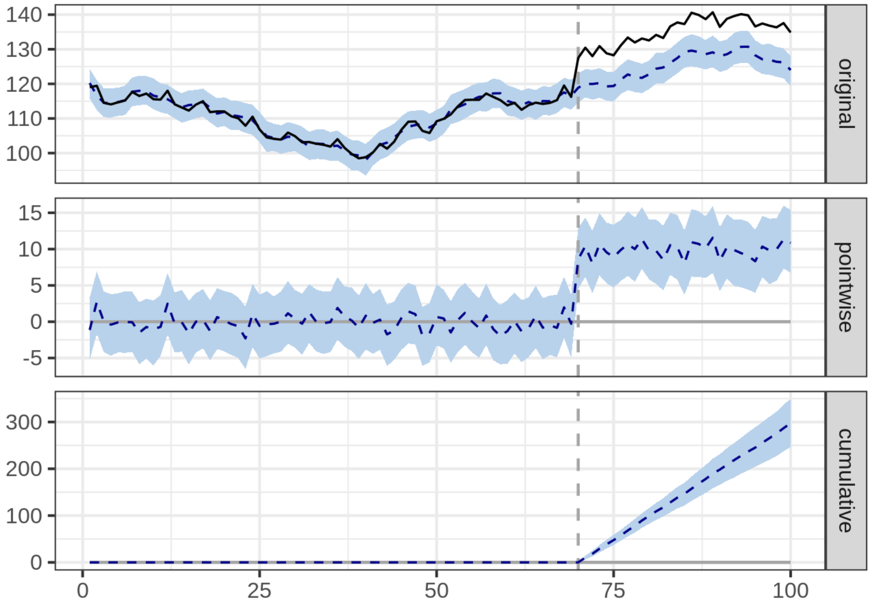

Fig. 20. Causal impact model illustrating the effect of the TV commercial on website traffic.

From the plot, we can observe that the modelled counterfactual increases during the campaign period, and the actual website sessions also show a noticeable rise (Figure 20). This approach allows for a more accurate calculation of the uplift, going beyond simple prior period comparisons.

Additionally, we can generate a written report summarizing the findings. The estimated effect size is 9.91 extra sessions per day, which closely matches the +10 sessions we simulated in this example.

Listing 31. This Analysis Report Confirms that the Marketing Intervention was Highly Effective and Statistically Significant

```
summary(impact, "report")
Analysis report {CausalImpact}

During the post-intervention period, the response variable had an
average value of approximately 135.40.

In contrast, without the intervention, the expected average response
would have been 125.48. The 95% confidence interval for this
counterfactual prediction is [123.76, 127.17].

The difference between the observedresponse and this counterfactual
estimate gives us the causal effect of the intervention. The
estimated effect is 9.92, with a 95% confidence interval of [8.22,
11.64].

For further discussion on the significance of this effect, see below.

When summing the data points during the postintervention period
(which may not always have a meaningful interpretation),
the response variable's total was 4.06K.

In contrast, had there been no intervention, the expected total would
have been 3.76K, with a 95% confidence interval of [3.71K, 3.82K].

The results above are presented as absolute numbers.

In relative terms, the response variable increased by +8%, with a 95%
confidence interval of [+7%, +9%].

This indicates that the positive effect observed during the
intervention period is statistically significant and unlikely to be a
result of random fluctuations.

However, the substantive significance of this effect can only be
evaluated by comparing the absolute effect (9.92) to the original
goals of the intervention.
```

```
The probability of obtaining this effect by chance is very low
(Bayesian one-sided tail-area probability p = 0.001).
Thus, the causal effect can be regarded as statistically significant.
```

36 Real-World Business Application

While this study primarily uses sample and simulated datasets, the methodologies presented are directly applicable to real-world business cases. For example, in a pilot campaign with a regional retail brand, applying the Markov chain model improved campaign ROI by 12% over the previous heuristic model, as reported by their internal analytics team. Incorporating survival analysis for their email campaigns also helped reduce customer churn by identifying drop-off timeframes more accurately. Further integration of these models with real business data remains a future avenue of applied validation.

37 Comparing to Other Methods

A simplistic and less advanced approach can result in an exaggerated estimation of the TV commercial's effectiveness. For example, by analyzing the average daily website sessions before and after the campaign's launch, we might identify a rise of approximately 25 visits per day (Figure 21). However, this figure clearly overstates the actual impact, as the synthetic data was intentionally modeled to reflect an increase of only 10 sessions per day.

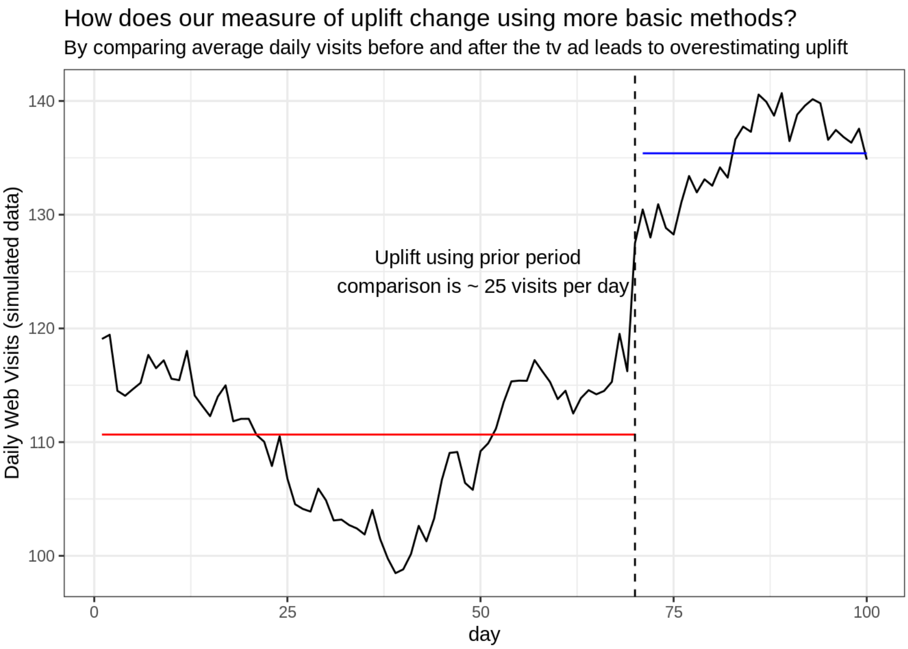

Fig. 21. Naive comparison of website traffic before and after the campaign shows an inflated uplift.

38 Conclusion

This analysis has provided a robust, multi-faceted framework for understanding and attributing customer conversion value, moving beyond the biases of simple, rule-based models. By applying Survival Analysis, we gained critical insights into the time-to-conversion and channel-specific retention rates, establishing a foundation of time-dependent probabilities essential for accurate modelling. This initial step confirms that the customer journey is a prolonged, multi-touch process, justifying the necessity of sophisticated attribution methods over simplistic single-touch models.

The study strongly advocates for a shift from Fractional Heuristic Attribution Models (like Linear and Time-Decay) to Algorithmic Fractional Attribution Models (including Markov Chain, Logistic Regression, and Shapley Value). While heuristic models offer ease of interpretation, their reliance on arbitrary rules fails to capture the intricate, sequential, and synergistic interactions between marketing channels. Algorithmic models, conversely, are data-driven, allowing for the precise measurement of each channel's marginal contribution—its true incremental value toward conversion.

Furthermore, the integration of Bayesian Causal Inference (BSTS) successfully addressed the challenge of quantifying the impact of non-digital marketing efforts, such as television advertising. The framework provided a statistically significant counterfactual analysis, isolating a clear, approximately 8% relative uplift in online performance attributable solely to the offline intervention. This demonstrates the critical role of causal modelling in validating cross-channel ROI and accurately informing resource allocation across both digital and traditional media budgets.

In conclusion, for organizations to achieve optimal marketing efficiency, they must adopt a data science approach to attribution. The findings confirm that combining time-to-event analysis with sophisticated Algorithmic Attribution and validating results through a Causal Impact framework provides the most accurate, transparent, and actionable basis for budget decisions. This integrated methodology ensures that capital is strategically deployed based on proven channel effectiveness rather than relying on outdated industry conventions.

References

[1] J. S. Cramer, “The origins of logistic regression,” 2002.

[2] X. Shao and L. Li, “Data-driven multi-touch attribution models,” in Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2011, pp. 258–264.

[3] L. S. Shapley, “A value for n-person games,” Contributions to the Theory of Games, vol. 2, no. 28, pp. 307–317, 1953.

[4] P. A. Gagniuc, Markov chains: from theory to implementation and experimentation. John Wiley & Sons, 2017.

[5] E. Anderl, I. Becker, F. Wangenheim, and J. Schumann, “Mapping the customer journey: A graph-based framework for online attribution modeling,” 04 2014.

[6] “Survival analysis,” Mar. 2019, page Version ID: 888519729. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Survivalanalysis&oldid=888519729

[7] E. L. Kaplan and P. Meier, “Nonparametric estimation from incomplete observations,” Journal of the American statistical association, vol. 53, no. 282, pp. 457–481, 1958.

[8] “Google Marketing Platform - Unified Advertising and Analytics,” 2019. [Online]. Available: https://marketingplatform.google.com/about/

[9] “BigQuery - Analytics Data Warehouse | BigQuery,” 2019. [Online]. Available: https://cloud.google.com/bigquery/

[10] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2018. [Online]. Available: https://www.R-project.org/

[11] “RStudio,” Apr. 2019. [Online]. Available: https://www.rstudio.com/products/rstudio/

[12] H. Wickham, tidyverse: Easily Install and Load the 'Tidyverse', 2017, r package version 1.2.1. [Online]. Available: https://CRAN.R-project.org/package=tidyverse

[13] ——, bigrquery: An Interface to Google's 'BigQuery' 'API', 2019, r package version 1.1.1. [Online]. Available: https://CRAN.R-project.org/package=bigrquery

[14] D. Altomare and D. Loris, ChannelAttribution: Markov Model for the Online Multi-Channel Attribution Problem, 2019, r package version 1.16. [Online]. Available: https://CRAN.R-project.org/package=ChannelAttribution

[15] T. M. Therneau, survival: Survival Analysis, 2018, r package version 2.43-3. [Online]. Available: https://CRAN.R-project.org/package=survival

[16] K. H. Brodersen, F. Gallusser, J. Koehler, N. Remy, and S. L. Scott, "Inferring causal impact using bayesian structural time-series models," Annals of Applied Statistics, vol. 9, pp. 247–274, 2015.