POWER MARKETING ATTRIBUTION **WITH MACHINE LEARNING**

A WHITEPAPER BY DATAIKU

www.dataiku.com

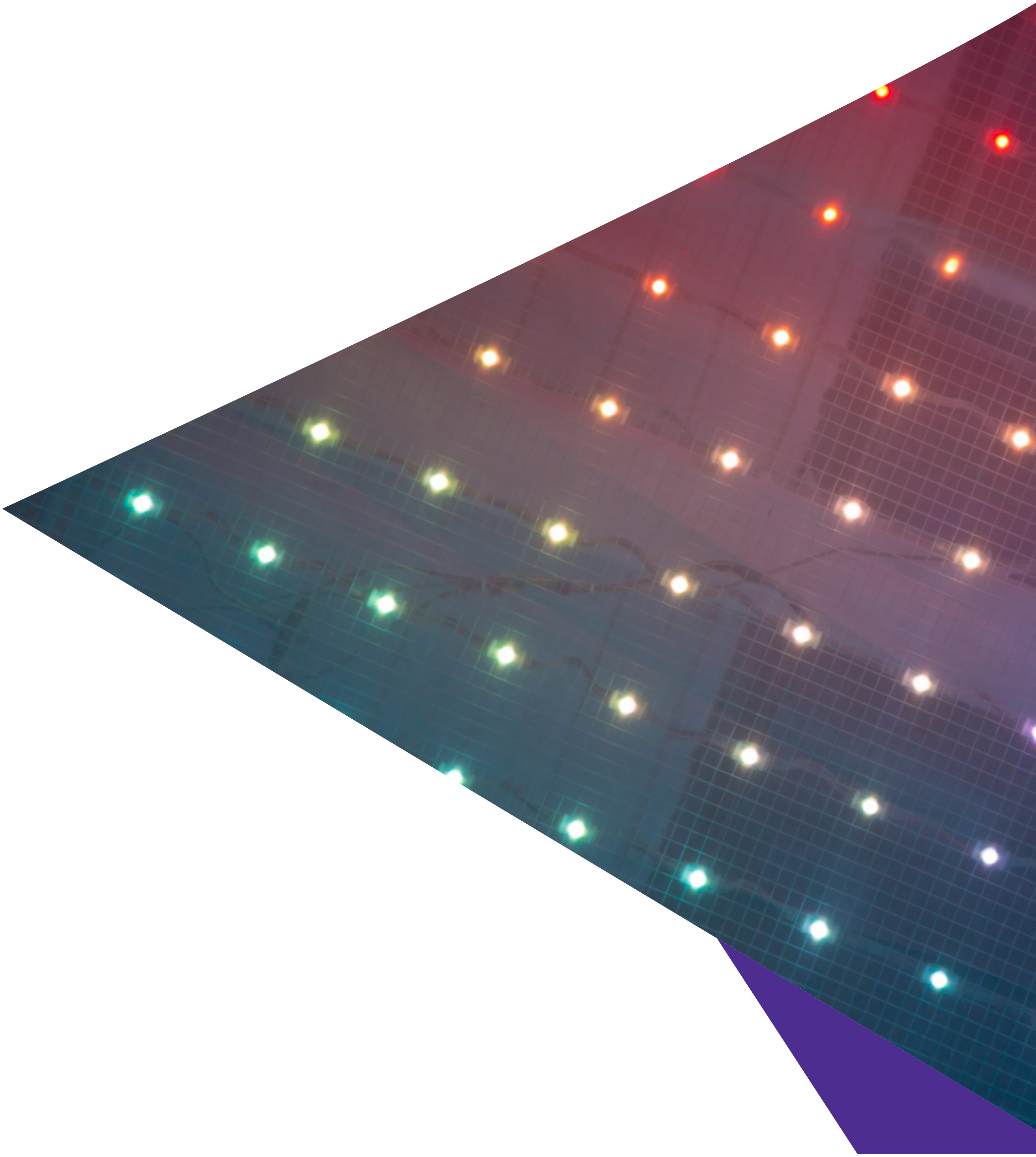
INTRODUCTION

Marketing attribution has been around for many years, and as the number of available advertising channels continues to shift and expand, so do the strategies employed by teams to leverage those channels. This guidebook is not intended to be an overview of every possible strategy, but a deep dive specifically into using machine learning models as opposed to heuristic models for marketing attribution across digital channels (both the why and the how).

By the end, readers should have an understanding of:

- **What it means to use data science for marketing attribution.**
- **How it can make the difference in scaling efforts to reach customers with more customized targeting (whether in a business-to-consumer or business-to-business enterprise).**

For those completely unfamiliar with data science in the context of marketing attribution, this guide will provide a short introduction to the topic and walk through the core aspects. But on top of that, for those already familiar, the guide includes some code and practical examples for execution.



DATA SCIENCE IN MARKETING ATTRIBUTION

Marketing attribution is the process of measuring campaign effectiveness by quantifying the influence those campaigns have on a desired outcome (e.g., starting a free trial, making a purchase, etc.). By understanding which channels or what content leads to a higher conversion rate to these desired outcomes, marketing teams can better optimize spend and messaging.

Today, in addition to there being more channels available for marketers on which to advertise, there is also more data than ever before on not only the channels, but the customers themselves and their specific habits. SalesForce did a study a few years back that said, on average, it takes six to eight touches to generate just one viable sales lead.

This drive to find a better way to solve one of the biggest challenges facing marketers today has turned what traditionally was a business question into a data science problem, fundamentally changing the core question: "How can I get more pepople to buy my product using advertising? to "How can I quantify the influence an advertisement has on a customer decision to make a purchase?" or "How can we measure the effectiveness of each advertising channel?"

Today, ML and AI allow marketing teams to go far beyond the methods of attribution introduced 10 (or more) years ago to:

Marketing attribution is, therefore, the perfect space for data science, which can incorporate vast amounts of data from various sources to help marketers understand in a scalable way and down to a granular level where the best (and worst) conversions are coming from. From there, marketers can adjust spend (either manually or automatically) accordingly.

- Get Personal: Teams can build ideal customer journeys down to granular user segments or, in some cases, down to the individual level for hyper-personalization (which generally translates to more desired actions).
- Scale: With algorithms handling data from multiple sources and giving near real-time feedback on the most effective channels, marketing teams can scale their efforts to reach more people more effectively (ideally by spending less money).
- Automate: Tight integration with customer relationship manager (CRM) or ad platforms can reduce manual processes and introduce more automation.
- Get Creative: With all of the hard work being done by machine learning (ML) models, marketing teams are more free to get creative and experiment when it comes to channels and messaging (especially with real-time feedback on effectiveness to pivot if needed).

MARKETING ATTRIBUTION DEEP DIVE

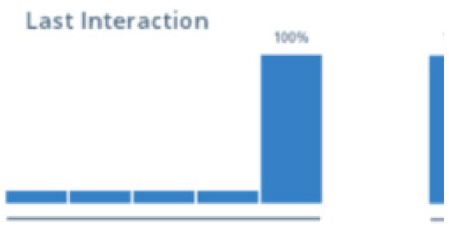
It's worth noting that the term marketing attribution is often used to encompass three distinctly different processes:

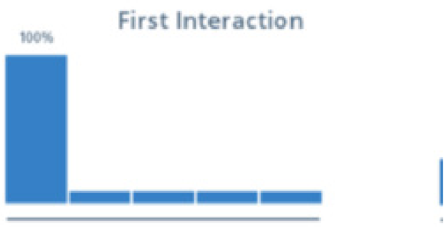
- **Attributing offline outcomes** (e.g., brick-and-mortar store purchases) to a particular campaign.
- **Tracking campaigns across different user devices** or different media (e.g., knowing that a particular user first saw an ad on television, then visited the website from their phone, then used a tablet to actually make the final purchase).
- **Measuring the relative effectiveness of different strategies** as part of a digital-only campaign across one specific device (e.g., one user on a certain device sees three different types of ads - which was the most effective?)

All three flavors of marketing attribution are quite complex (and are only becoming more so as customer experience becomes more fragmented across channels and devices), but the approach to each is different. This guidebook focuses exclusively on the third type of marketing attribution, where data science and ML have the most direct and effective application. Get suggested resources on tackling the first two types of marketing attribution.

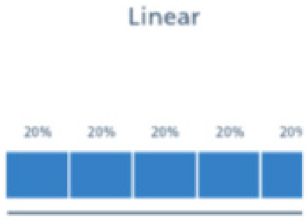
HEURISTIC VS. ALGORITHMIC APPROACHES

Historically, marketing attribution has been a painstakingly manual process that often turns out to be more difficult (and less effective) than desired. And unfortunately, due to their relative simplicity, many marketing teams turn to single-source /single-touch attribution or other heuristic models, which are based on simple rules (like tying desired outcomes to a single source along the customer's journey or assigning equal credit to all channels across a journey).

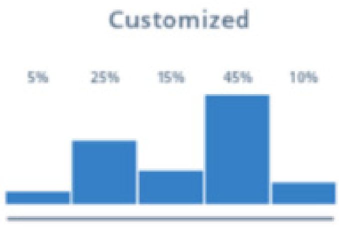




With rare exceptions, heuristic models for marketing attribution are a gross oversimplification and generally come with inaccuracy, especially for products and services with long sales cycles and many touches along the way since more often than not, a combination of messages could have led to the desired behavior.

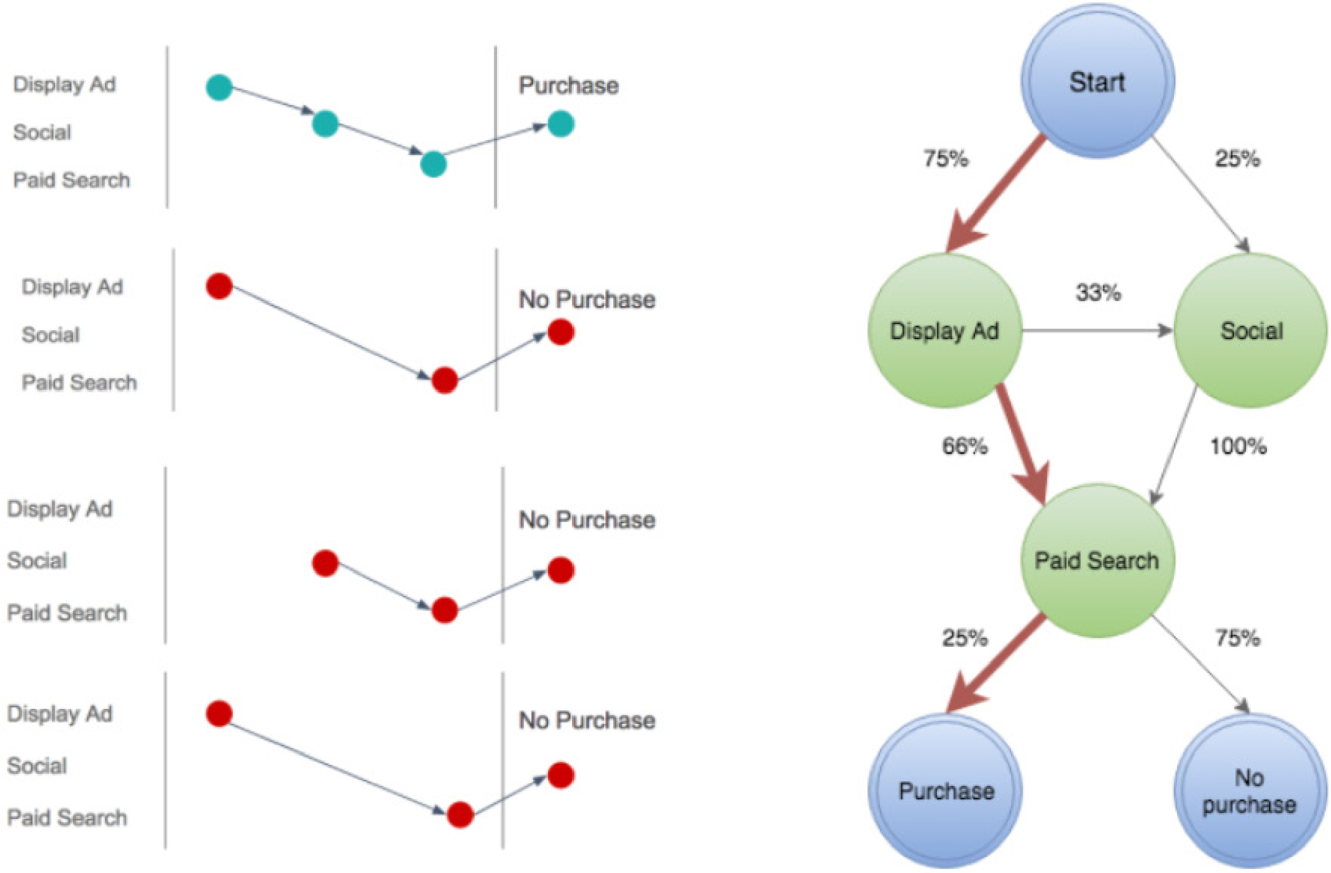


Heuristic models also introduce a great deal of bias; for example, last- or first-click models can place unwarranted emphasis on retargeting or Google search as effective ad targeting platforms. OK, so if heuristic models are ineffective, what is effective? Again, this guidebook doesn't cover every possible approach and model, but the most popular and effective options that data scientists at Dataiku have tested with real-life customers.



MARKOV CHAIN MODELING

The output of a Markov model is the probability that a user will move from one step in the customer journey to another. Essentially, it models the customer journey, and from there, it allows marketers to answer the question: “If channel X were not present in my marketing strategy, what would be the effect on the probability of conversion?” This will ultimately give a “removal effect” for each channel, and through that, marketing teams can decide which channels are the most important.



The image on the left would be the first step in building the Markov chain (the sequences in red and green correspond the customer journeys); these sequences are then aggregated to form the image on the right.

GAME THEORY & SHAPLEY VALUE

In using game theory for marketing attribution, one can actually model the interactions that customers have with the marketing channels as a cooperative game where each marketing channel can be seen as a player in the game, and the set of all players/channels can be thought of as working together in order to drive the conversions.

So in other words, game theory in marketing attribution assigns each touchpoint credit for a conversion based on its contribution. The Shapley value stipulates that if two players (or in this case, channels) are interchangeable, they should get the same payments (in this case, credit for conversion). And if a channel doesn't add any value to all the coalitions (in marketing attribution, combinations of actions in a user journey), that channel should get the conversion credit that it generates alone. For a more in-depth explanation behind the theory, this is a great post (with visuals).

BUILDING YOUR MODEL: STEP 1

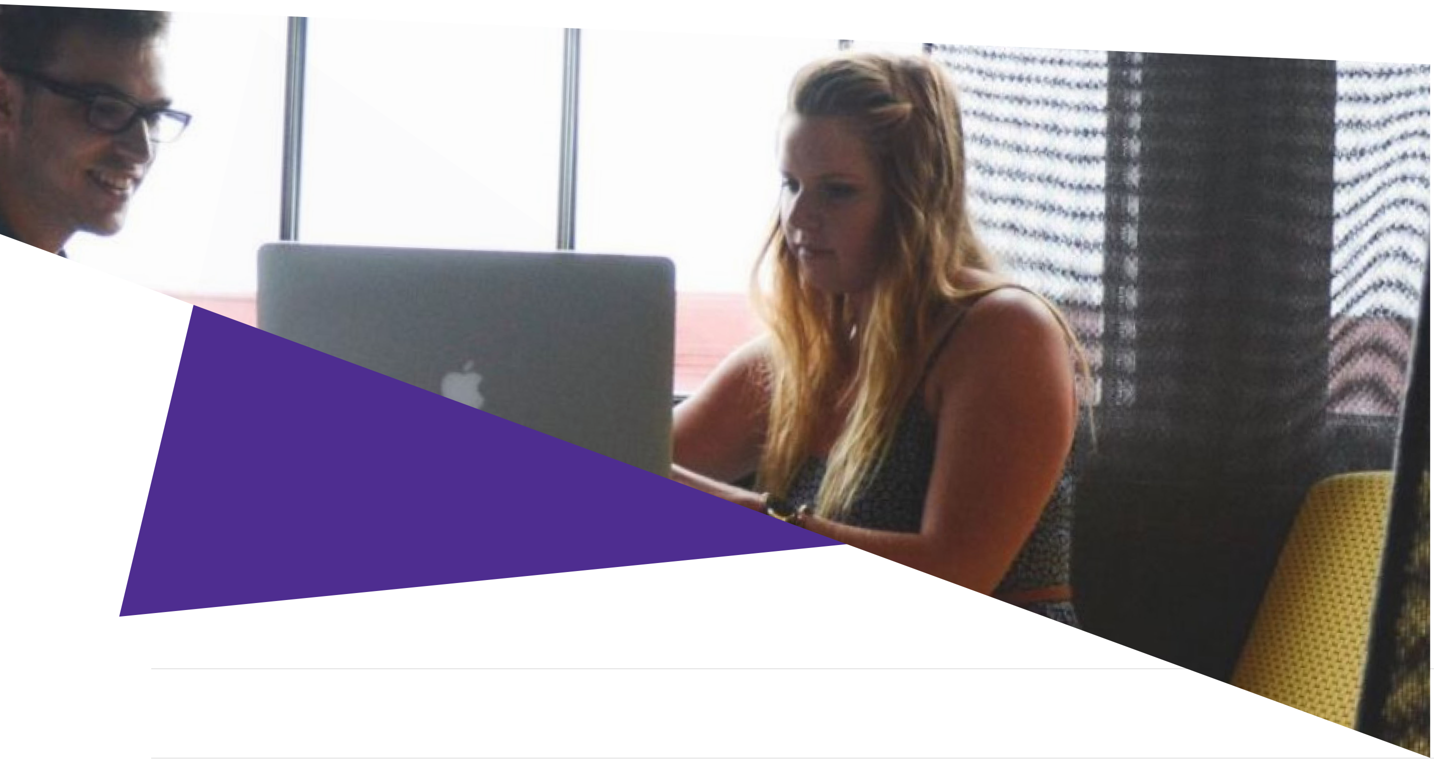
UNDERSTAND THE BUSINESS

In using game theory for marketing attribution, one can actually If you're familiar with the seven fundamental steps to building a data project, then you already know the basics for how to get started using ML to the benefit of your marketing team. But there are also several particularities to bear in mind when working with marketing attribution. As with any data science project, marketing attribution must begin on the business side. Before diving into the data, the team needs to take a step back and answer the following questions (preferably with business/marketing and data teams together):

How are we currently doing marketing attribution?
Of course, before starting a new project, it's important to understand what teams are already doing (or have already tried to do) to address the question of channel attribution. Every member of the team tackling marketing attribution should know how it's being done right now, why it's being done that way, how it works, the results it's delivering, and who is using those results (as well as how they're using them). This will provide a more clear picture of needs.

How many different types of campaigns do we have, and what is the desired action for each campaign (or campaign type)?
For some businesses, or for some particular campaigns, the desired action might be making a purchase. For others, it might be more awareness-based, so a potential customer simply visiting the website would be considered the goal action. In any case, the desired action - or goal - for each marketing campaign must be defined, and it should be specific. Different attribution models might work better or worse with certain campaigns, so mapping this out clearly before getting started is critical.

What is the ideal way to deliver results that will have real business impact? In other words, what is the deliverable?
Whether it's a dashboard or real-time, automated campaign spend allocation, failing to define deliverables before kicking off a marketing attribution project sets the stage for failure (especially when data scientists and marketing teams aren't aligned and the result is something the marketing team can't make use of).



BUILDING YOUR MODEL: STEP 2

GET YOUR DATA

Coming up with a good data science solution for a business question starts with properly scoping out the business needs, but once that's finished, the second most essential component is good data.

The first step is to map out all channels and touchpoints along the customer journey to be sure that no channels are forgotten From there, good data means, of course, the prerequisite tracking of all user actions on each targeted channel.

But moreover, it means understanding exactly what data is attached to each touchpoint and where the data comes from as well as what limitations (e.g., missing data) might exist. Understanding attribution data is not only fundamental to the accuracy of models, but it's also essential for business teams and leaders to trust model outcomes.

GO FURTHER

The "Go Further" sections of this guidebook will walk through a marketing attribution example going from raw data to a working model, assuming that there is a history of recorded customer interactions from a website in the following format:

```
log_data: user_id | touchpoint | timestamp
```

Though this guidebook has introduced both Markov chains and game theory as viable models, this example will walk through just one - game theory - since between the two, data scientists at Dataiku have found it to be the most effective in clients' businesses.

BUILDING YOUR MODEL: STEP 3

PREPARE DATA

After identifying all the right data sources, no matter what algorithm is ultimately chosen for the attribution, the next step in all cases is to ensure the data is clean and in the right format. This requires, among other things, that the user sessions be constructed and well defined. It is at this point in the process that one may discover channels where data is missing altogether.

Should this process uncover holes in the data (like missing tags for certain channels, for example) the best approach is to stop and address the problem. It's not possible to build an accurate attribution model with missing data, so taking the time to fix the issue to ensure data is attributed properly before moving forward is critical.

GO FURTHER

The goal here is to recreate the entire journey for each customer that led him to purchase an item on the website (i.e., put together all the marketing channels that the customer visited prior to the purchase, in the right order).

1

The first step is to construct the micro-sessions, which are a series of clusters of events that belong to a unique user navigation experience (read more in depth about sessionization). The goal is to create a new table with the following schema, where user_micro_session would be a unique identifier for each user's micro-session:

```
log_data_sessions: user_id | user_micro_session | touchpoint | timestamp
```

Obviously it's necessary to distinguish between micro-sessions (that is, when one stops and the next begins), so the rule is to stop the micro-session when the time between two clicks is greater than T_inactivity (usually, a value of 30 minutes is used - this is, for example, the default value in Google Analytics). Here, compute the difference of time between each two clicks/visits for each user using SQL language (Postgres) and window functions in particular:

```
SELECT
    *,
     extract(epoch FROM "timestamp")
    - lag(
        extract(epoch from "timestamp")
        ) over (
        PARTITION BY "user_id"
        order by "timestamp"
                  ) as "time_difference"
FROM log_data
```

2

The next step is to compare every value of the column time_difference with the value of T_window (using the standard 30 minutes, in this example), flagging the cases where the value of time_difference is greater than T_window with a value of 1, and give all the other cases a value of 0. Building on the previous code:

```
SELECT
  *,
  CASE WHEN
  EXTRACT(epoch FROM "timestamp")
          - LAG(
              EXTRACT(epoch FROM "timestamp")
              ) OVER (
          PARTITION BY "user_id"
          ORDER BY "timestamp"
                      ) >= 30 * 60 THEN 1 ELSE 0 END as new_micro_session
  FROM log_data
```

3

Following this, we'll create a new session_id for a user every time the column new_micro_session is equal to 1. One possible solution for doing this is to concatenate the user_id value with the cumulative sum of the column new_micro_session:

```
SELECT *,
       "user_id" || '_' || SUM("new_micro_session")
       OVER (PARTITION BY "user_id"
             ORDER BY "timestamp") AS "user_micro_session"
FROM (
  SELECT
      *,
      CASE WHEN
          EXTRACT(epoch FROM "timestamp")
          - LAG(
              EXTRACT(epoch FROM "timestamp")
              ) OVER (
          PARTITION BY "user_id"
          ORDER BY "timestamp"
                      ) >= 30 * 60 THEN 1 ELSE 0 END as new_micro_session
  FROM log_data
  )t
```

4

Of course, the customer's navigation on the website will have generated additional records (touchpoints) that aren't useful in the context of attribution analysis. So use the following code to retain only the first touchpoint of each user session (the portion useful for attribution):

```
SELECT "user_id",
       "channel",
       "timestamp"
FROM(
    SELECT *,
            FIRST_VALUE("touchpoint") over(
                PARTITION BY "user_id", "user_micro_session"
                ORDER BY "timestamp") as "channel"
    FROM(
        SELECT *,
               "user_id" || '_' || SUM("new_micro_session")
               OVER (PARTITION BY "user_id"
                     ORDER BY "timestamp") AS "user_micro_session"
        FROM (
          SELECT
              *,
              CASE WHEN
                  EXTRACT(epoch FROM "timestamp")
                  - LAG(
                      EXTRACT(epoch FROM "timestamp")
                      ) OVER (
                  PARTITION BY "user_id"
                  ORDER BY "timestamp"
                              ) >= 180 * 60 THEN 1 ELSE 0 END as new_micro_session
          FROM "log_data_sessions"
          )t1
      )t2
  )t3
GROUP BY 1,2,3
```

5

Taking a step back, it’s important to redefine that a customer journey is a succession of touchpoints that may or may not end in a conversion. In other words, a user could make several purchases or conversions along the way, which means that more than one journey can be linked to one customer. These journeys are called macro-sessions, and the final step is to create the macro-sessions using the previously computed table as the input:

```
user_touchpoints: user_id | channel | timestamp
```

With the goal of creating the following table:

```
user_macro_sessions: user_id |user_macro_session | channel | timestamp | conversion
```

Where user_macro_session is a unique identifier of each user’s macro-session. For this, we will also need the table of conversions/purchases:

```
conversions : user_id | conversion_timestamp
```

Each record in this table corresponds to a purchase that the user user_id completed and when.

First, we are going to perform a left join on the conversions table with user_touchpoints as the user identifier. Since each conversion event will define a unique macro-session, we will also create a new column user_macro_session by concatenating the user_id and the conversion_timestamp.

```
SELECT *,
       CASE WHEN
           "timestamp" >= "conversion_timestamp" - interval '30 days'
           AND "timestamp" <= "conversion_timestamp"
           THEN 1 else 0 END as "keep_touchpoint"
FROM(
  SELECT "conversions".user_id,
         "conversions".conversion_timestamp,
          "touchpoints".channel,
          "touchpoints".timestamp,
          "conversions".user_id || extract(epoch from "conversions"."conversion_timestamp") AS
"macro_user_session"
  FROM "conversions" "conversions"
  LEFT JOIN "user_touchpoints" "touchpoints"
  ON "conversions"."user_id" = "touchpoints"."user_id"
```

Next, we need to filter out in each macro-session, the touchpoints that happened after the conversion, as well as the

6

touchpoints that happened prior to the conversion with a time greater than T_window = 30 days.

```
SELECT *,
       CASE WHEN
           "timestamp" >= "conversion_timestamp" - interval '30 days'
           AND "timestamp" <= "conversion_timestamp"
           THEN 1 else 0 END as "keep_touchpoint"
FROM(
  SELECT "conversions".user_id,
         "conversions".conversion_timestamp,
          "touchpoints".channel,
          "touchpoints".timestamp,
          "conversions".user_id || extract(epoch from "conversions"."conversion_timestamp") AS
"macro_user_session"
  FROM "conversions" "conversions"
  LEFT JOIN "user_touchpoints" "touchpoints"
  ON "conversions"."user_id" = "touchpoints"."user_id"
```

Let's name this intermediate table: intermediate_table. We need to extract from this table two types of macro-sessions:

Macro-sessions which end up in a conversion: those are the rows that we flagged with a 1. We only need to filter on those. Macro-sessions which don't end in a conversion: those are the rows we flagged with a 0. Let's focus on those for

7

the rest of this part. For those we need to compute the time difference between each touchpoint and flag the rows where this value is greater than 30 days.

```
SELECT *,
           CASE WHEN
               "time_difference" >= INTERVAL '30 days' THEN 1  ELSE 0 END AS "flag_macro_session_
limits"
    FROM(
        SELECT *,
               LAG("timestamp",1)  OVER (
                   PARTITION BY "user_id"
                   ORDER BY "timestamp" DESC
                    )
               - "timestamp" as "time_difference"
        FROM(
            SELECT "user_id",
                    "channel",
                    "timestamp"
            FROM(
                SELECT  "user_id",
                        "channel",
                        "timestamp",
                        SUM("keep_touchpoint") as "keep_touchpoint"
                FROM "non_converting_macro_sessions"
                GROUP BY "user_id", "channel", "timestamp"
                ) t1
            WHERE "keep_touchpoint" = 0
            ) t2
        ) t3
```

And now for the final trick. We are going to use this flag to create the user_macro_session column.

8

```
SELECT "user_id",
       "user_id" || SUM("flag_macro_session_limits") OVER (
            PARTITION BY "user_id"
            ORDER BY "timestamp" ASC
           )
           - "flag_macro_session_limits" AS "user_macro_session",
       "channel",
       "timestamp"
FROM(
    SELECT *,
           CASE WHEN
               "time_difference" >= INTERVAL '30 days' THEN 1  ELSE 0 END AS "flag_macro_session_
limits"
    FROM(
        SELECT *,
               LAG("timestamp",1)  OVER (
                   PARTITION BY "user_id"
                   ORDER BY "timestamp" DESC
                    )
               - "timestamp" as "time_difference"
        FROM(
            SELECT "user_id",
                    "channel",
                    "timestamp"
            FROM(
                SELECT  "user_id",
                        "channel",
                        "timestamp",
                        SUM("keep_touchpoint") as "keep_touchpoint"
                FROM "non_converting_macro_sessions"
                GROUP BY "user_id", "channel", "timestamp"
                ) t1
            WHERE "keep_touchpoint" = 0
            ) t2
        ) t3
    ) t4
```

BUILDING YOUR MODEL: STEP 4

EXPLORE AND ENRICH DATA

It is at this point that it's necessary to define the model that will be used for the project, as all subsequent steps of working with the data depend on which model is being used for attribution.

The "Iterate" section dives in further, but with marketing attribution, trying multiple approaches in parallel is not practical or recommended (though if you're currently not doing any marketing attribution, it's a good idea to use one of the simple heuristic models first to get a baseline idea of what the start of the customer funnel looks like).

Unlike other types of machine learning models (like, for example, churn, predictive maintenance, or anomaly detection) where it's possible to split data into train and test sets to compare the model's predictions to actual outcomes, the only way to actually test a marketing attribution model is to use it. Unlike these other models, marketing attribution isn't a true predictive model, so there are no "actual" outcomes with which to compare before making the model live.

GO FURTHER

The schema of our nicely cleaned and prepared data from the previous step now looks like this:

```
user_logs : user_id | channel | timestamp | conversion
```

Where:

user_id = A unique identifier of each user
Note: This can be anything that uniquely identifies each customer. If the customer is logged in on the website, then his interactions with marketing channels can be precisely tracked. If not, cookies are usually used to record these interactions.
channel = The marketing channel visited by the user user_id
timestamp = The time of the visit
conversion = A binary variable that is 1 if the user converted after visiting channel and 0 otherwise.
As a first step toward applying game theory, we need to compute the sum of conversions that each subset of channels has yielded. The

following SQL query (still using Postgres) does this, and it produces a table with the following schema:

```
subsets_conversions : channels_subset | conversions_sum

SELECT  "channels_subset",
        sum("conversion") as "conversions_sum"
FROM
    (
        SELECT  "user_id",
                string_agg(DISTINCT("channel"), ',') as "channels_subset",
                max("conversion") as "conversion"
        FROM
            (
                SELECT  "user_id",
                        "channel",
                        "conversion"
                FROM "MARKETINGATTRIBUTIONSIMULATIONS_simulated_data"
                where "sample_id" = 0
                ORDER BY
                    "user_id",
                    "channel"
            )a
        group by "user_id"
    )b
GROUP BY "channels_subset"
```

BUILDING YOUR MODEL: STEP 5

GET "PREDICTIVE"

Traditionally in a data project, once data is clean and prepared, predictive models can be applied. In the case of marketing attribution, nothing is actually being predicted (hence the title of this section as get “predictive”). Instead, the outcome of the model will be a percentage or score for each channel.

GO FURTHER

Now, we can start to work toward computing the Shapley value for each channel. First, compute the worth of each coalition by summing the number of conversions of each of its subsets.

1 Here, subsets is a function that returns all the subsets that each coalition contains (for example, if coalition A={a,b,c}, then the function would return the list [{a},{b},{c},{ab},{ac},{abc}]).

```
def v_function(A,C_values):
    '''
    This function computes the worth of each coalition.
    inputs:
            - A : a coalition of channels.
            - C_values : A dictionnary containing the number of conversions that each subset of
channels has yielded.
    '''

    subsets_of_A = subsets(A.split(","))
    worth_of_A=0
    for subset in subsets_of_A:
        if subset in C_values:
            worth_of_A += C_values[subset]
    return worth_of_A
```

2 Next, the following Python script applies the above-defined function to all the possible coalitions that can be formed:

```
# First, let's convert the dataframe "subsets_conversions" into a dictionnary
C_values = user_logs_aggr.set_index("channels").to_dict()["conversions"]

#For each possible combination of channels A, we compute the total number of conversions yielded by
every subset of A.
# Example : if A = {c1,c2}, then v(A) = C({c1}) + C({c2}) + C({c1,c2})
v_values = {}
for A in subsets(channels):
    v_values[A] = v_function(A,C_values)
```

3

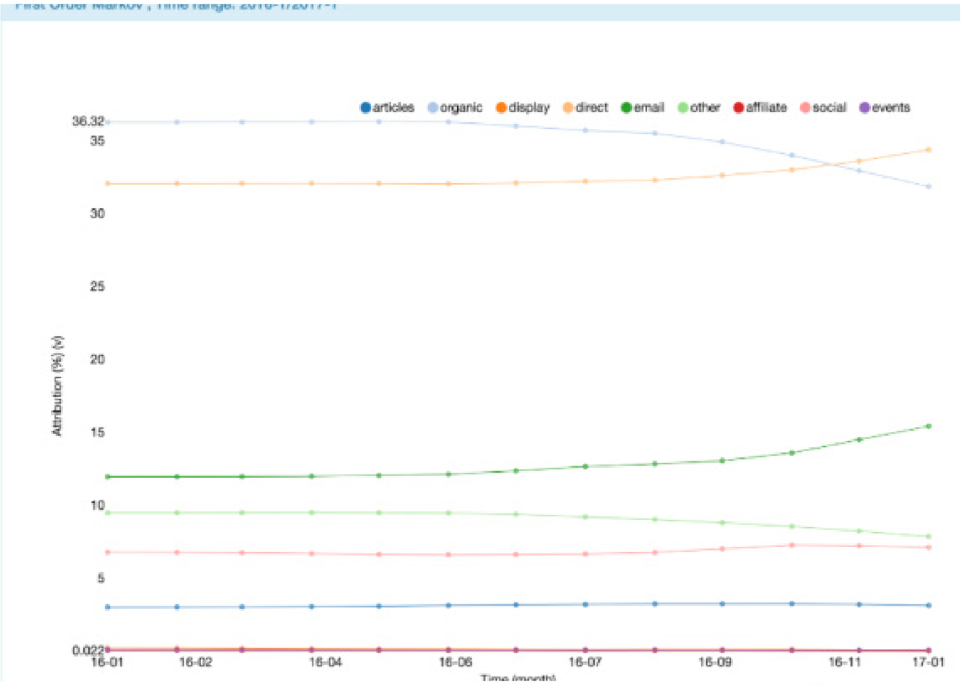
Finally, compute the Shapley value for each channel. The following Python script implements the Shapley equation:

```
from collections import defaultdict

n=len(channels)
shapley_values = defaultdict(int)

for channel in channels:
    for A in v_values.keys():
        if channel not in A.split(","):
            cardinal_A=len(A.split(","))
            A_with_channel = A.split(",")
            A_with_channel.append(channel)
            A_with_channel=",".join(sorted(A_with_channel))
            shapley_values[channel] += (v_values[A_with_channel]-v_values[A])*(factorial(cardinal_A
)*factorial(n-cardinal_A-1)/factorial(n))
    # Add the term corresponding to the empty set
    shapley_values[channel]+= v_values[channel]/n
```

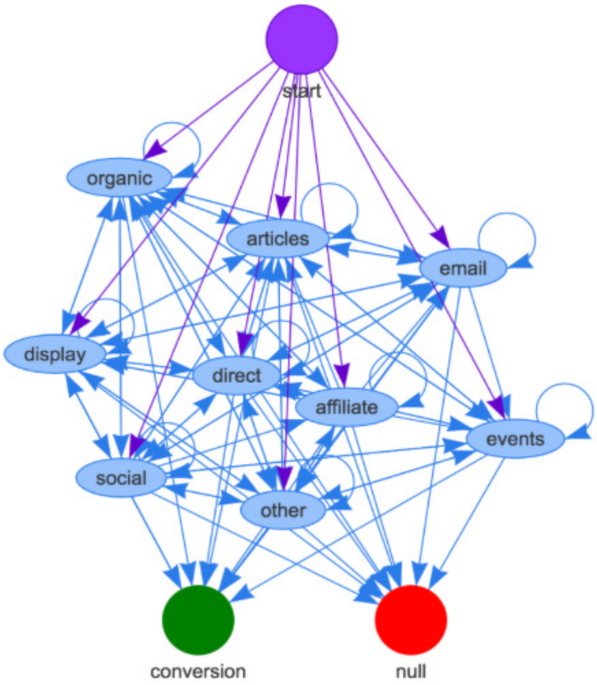
BUILDING YOUR MODEL: STEP 6

VISUALIZE

Of course, visualizations can be useful when it comes to marketing attribution to illustrate the distribution of the conversions for the channels themselves. This might be a bar chart showing conversions (or percentage of conversions) per channel for all time. Or it could be a line chart showing conversions per channel over time, which can be useful to see if there is fluctuation. Fluctuation could either indicate seasonality or, more likely, that the algorithm is unstable, which is a good sign that iteration is necessary:



Visualization can also be useful when leveraging a Markov chain model for plotting the channels and the percentage of conversions between them. For example, this is an interactive chart (clicking on the arrows displays the percentage):



BUILDING YOUR MODEL: STEP 7

DEPLOY AND ITERATE

Deploying a marketing attribution project can mean any number of things depending on the predefined deliverables with the business and marketing teams. But at a very minimum, it means having a model working on actual data and updating regularly based on current data (again, this should have been pre-defined in the deliverables agreed up with marketing - depending on their needs and the nature of the business, it could be daily, weekly, monthly, etc.).

Marketing attribution is unique as a data science project in that the only way to see its effects is to deploy the model, update marketing spend accordingly, and observe the change on the business side. In other words, based on the model and adjusting spend, look at the number of conversions - how did allocating less budget to a specific channel effect those conversions overall?

By repeating this process for different channels and measuring the resulting business outcome, marketing teams will be able to identify the optimal balance.

FEATURED GUEST CONTENT

6 PRODUCTION CONSIDERATIONS FOR ATTRIBUTION MODELS

BENJAMIN BAUSILI

PRINCIPAL | ANALYTICS PRACTICE LEAD @ INTERWORKSW

You've been in your data lab working to create the perfect model. It's been hard work gathering, cleaning, and analyzing all that data, but you're sure you have produced something of value. Yet most data science projects fail to make it all the way through production. Unfortunately, like the original Netflix Prize, it's too easy to create a winning solution that can never be used. Let's tackle the six top things to consider to ensure you create a lasting impact:

#1 - Premature Optimization is the Root of All Evil

This is a common rule in computer science, and yet I find it so often ignored in the real world. It's easy to get excited about deploying APIs and creating custom web apps right out of the gate, especially when you have a tool like Dataiku enabling you. However, consider the endpoint(s) that delivers the most initial value. In market attribution, it often means delivering a visualization or dashboard to end users. Don't over-complicate the process, which in this case should be creating a simple, scheduled workflow that populates a clear and actionable report. Save the APIs and coding for later iterations.

#2 - Deploy Early and Iterate

The longer your model stays in development, the longer it takes to see ROI. Set a reasonable threshold for your model that represents a small but a real improvement over your current process. Once your model has reached that threshold, deploy it. Then, focus on delivering incremental updates to your process that provide additional value. For the absolute best results, begin following a regular sprint interval for updates to the model and release regularly.

#3 - Don't Stop Evaluating Your Production Models

ABT: Always Be Testing. A common mistake in the industry is improper monitoring of a deployed model. Just because you put in all the hard work to develop and test the model's predictive power doesn't mean it will continue to deliver valid results. Markets and goals change, so your models have to change too. Today's prediction is tomorrow's test set. How well did your models do this month predicting user engagement for a given level of ad spend? Collect all of your new actuals to see how your model is doing with real data, and make sure you're still delivering the value you think.

#4 - One Model Doesn't Rule Them All

To build an appropriate attribution model for your business, you need to understand the model effects that are important. Each of the marketing attribution models you could choose has different benefits. So you have to really think about what question you're trying to answer. It may be that different campaigns need different models to answer the key questions of the team responsible, while yet another model is built to help executives understand everything in the aggregate. Be flexible with your selection, keep a focus on the question your answering, and be transparent to the end-users about the strengths and weaknesses of each approach.

#5 - Communication and Design Is Key

If you want people to use your model, give them confidence in it. Getting your model(s) to run is really just the first step and often can be the easiest. Don't skimp on the delivery and presentation elements. All models have performance metrics; don't be afraid to show them! Work on creating the right data visualizations to give users a before and after view of dollars spent to advertising ROI. This will help end users trust your models and get excited about the results. Remember: The goal is to make better decisions. Easy-to-understand charts and pleasant graphics are more than just window dressing; they are key to user acceptance.

#6 - Disaggregate

Beware aggregated data! You've built a model that you're proud of, but if you feed summarized data, your results will be distorted. When you aggregate daily data to the monthly levels, it hides the immediate effect of your advertising. We understand that an ad seen on the first day of the month likely deserves a small share of attribution for a sale made on the last day of the month. However, when you aggregate your data to the monthly level, there is no way to make that distinction. You should make sure the data you use are as temporally disaggregate as possible.

The Bottom Line

With these six things in mind, you've given your project the best chance at impacting the business. Remember that as you progress, you should continue to document your unique learnings, helping to build upon each winning solution and ensuring that what you develop in your lab is able to reach the light of day. With each successful iteration, you're helping increase the organizations trust in analytic projects, helping to establish a data-driven culture and paving the way for new and exciting business problems to tackle.

About InterWorks: InterWorks is a people-focused tech consultancy, delivering premier service and expertise. For over 20 years, they've empowered clients around the globe by combining the right mix of people, knowledge and partners into solutions that propel their businesses forward. This comprehensive, client-centric approach has helped them forge strong relationships with Fortune 500 companies and small businesses alike. From foundation to vision and every step in-between, they're committed to helping others get further, faster with their IT and data strategy.

About the Author: Ben is on a mission to understand more of everything and help others along the way. His knack for research and experimentation helps him see solutions to problems that others often miss – a vital skill when helping clients overcome their data challenges. Based out of Tulsa, Oklahoma, he leads InterWorks' analytics, UX, and data science teams.

When not instigating positive changes for clients, Ben is most at home on a hiking trail with his family. His wife, Rachel, shares his love of adventure and travel. So, when they're not in a new city to try the local food or the great outdoors to escape it all, they are likely dreaming and planning their next family destination over (locally-roasted) coffee.

PITFALLS & SOLUTIONS

1. Poor Quality Data

The single biggest issue companies face when beginning a marketing attribution project is actually the quality of the data they start with. Essentially, they have not been tracking one (or more) channels as they should have been, so it turns out they don't have the information they need to start the project.

Solution

Unfortunately, there is no magic solution here aside from pausing the marketing attribution project, going back to fix the issues with data collection, and resuming once they are addressed. There is little sense in moving forward if data from certain channels is missing or incomplete.

2. Not Properly Tracking Users Across Devices

One of the factors further complicating marketing attribution efforts today is the difficulty in tracking users as they are exposed to advertising content on their mobile phone, tablet, laptop, and any number of other devices. The fact is that this is an issue nearly every marketing team faces, and unfortunately, it's almost impossible to solve.

Solution

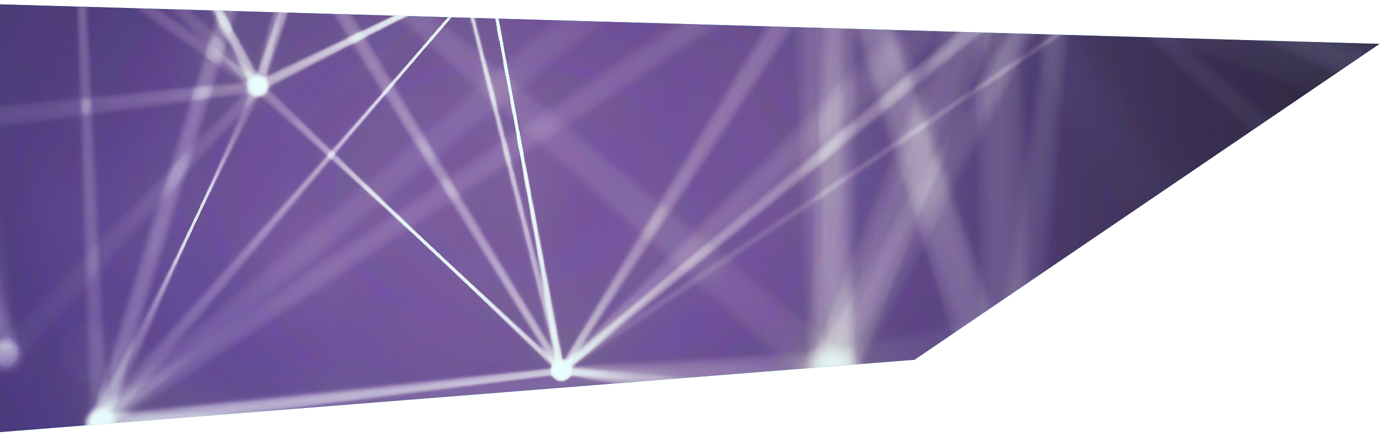
The only reprieve is for businesses that have a sign-in functionality, which makes tracking single users on many devices much, much, more simple (but by no means easy). If possible, forcing login as early as possible is best. For businesses where this is not possible, correctly setting expectations and being transparent about this reality when talking about marketing attribution is the best way forward.

3. Struggling with online vs. physical store conversions

Entirely online businesses that don't have any offline campaigns or opportunity for conversion have it relatively easy; marketing attribution becomes even more challenging when there's one (or both) of these offline factors. Perhaps the largest and best example is e-commerce with brick-and-mortar stores as well.

Solution

There are a number of techniques that well-known companies have used to address the offline advertising/offline purchase struggle. This is a great resource that looks at several case studies on how offline advertising contributes to online purchases, and here's one for the reverse - offline conversions from online campaigns.



CONCLUSION

TRENDS & WHAT'S NEXT

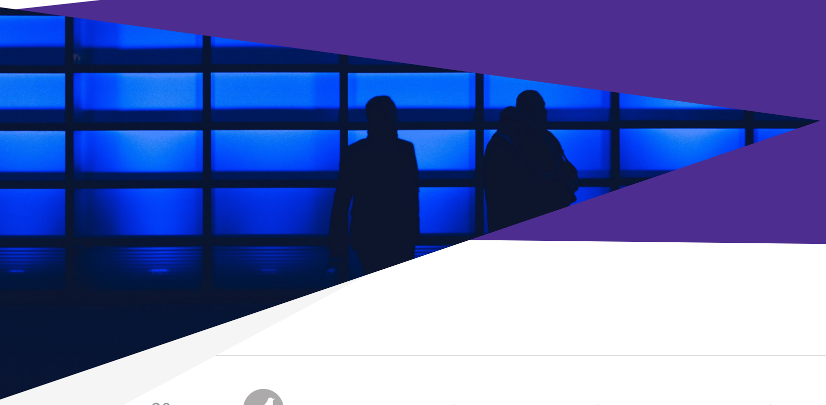
It is true that, as previously mentioned, there is traditionally no way to test marketing attribution models before using them in a real business context, seeing as it's not possible to compare the algorithm's output to some source-of-truth-data.

However, that being said, marketing attribution is still an evolving discipline, and data scientists are exploring possible ways of testing these models for a possible look into performance before applying them to live data and doing a sort of real-life testing by moving marketing budget around.

For example, the data scientists at Dataiku have developed a way to create data simulating users going through the customer journey. From there, they choose one channel that will result in conversion 100 percent of the time to see if the model can also accurately attribute 100 percent credit to that channel. This has proved to be an effective way to weed out algorithms that don't perform well (and, of course, identify the ones that do). If you're interested in starting a marketing attribution project, you can give Dataiku a try for free or get a demo.

Attributing advertising channel conversions is perhaps the biggest - yet also most complex - challenge that today's marketing teams face. And there is no magic bullet solution; though employing data science and ML techniques can significantly lower the time spent and deliver better results than traditional heuristic models, it's still not a one-and-done deal. Marketing teams must continuously evaluate channels, and the use of those channels, at regular intervals to understand and address shifts in consumer behavior over time.

What's more, this landscape will continue to grow more complex over time as new avenues for reaching potential customers emerge. Taking an algorithmic approach to attribution is just the beginning in driving change by moving toward a more detailed, data-driven approach in marketing.



data
iku

Your Path to Enterprise AI

Dataiku is the centralized data platform that moves businesses along their data journey from analytics at scale to enterprise AI. Data-powered businesses use Dataiku to power self-service analytics while also ensuring the operationalization of machine learning models in production.

20,000+

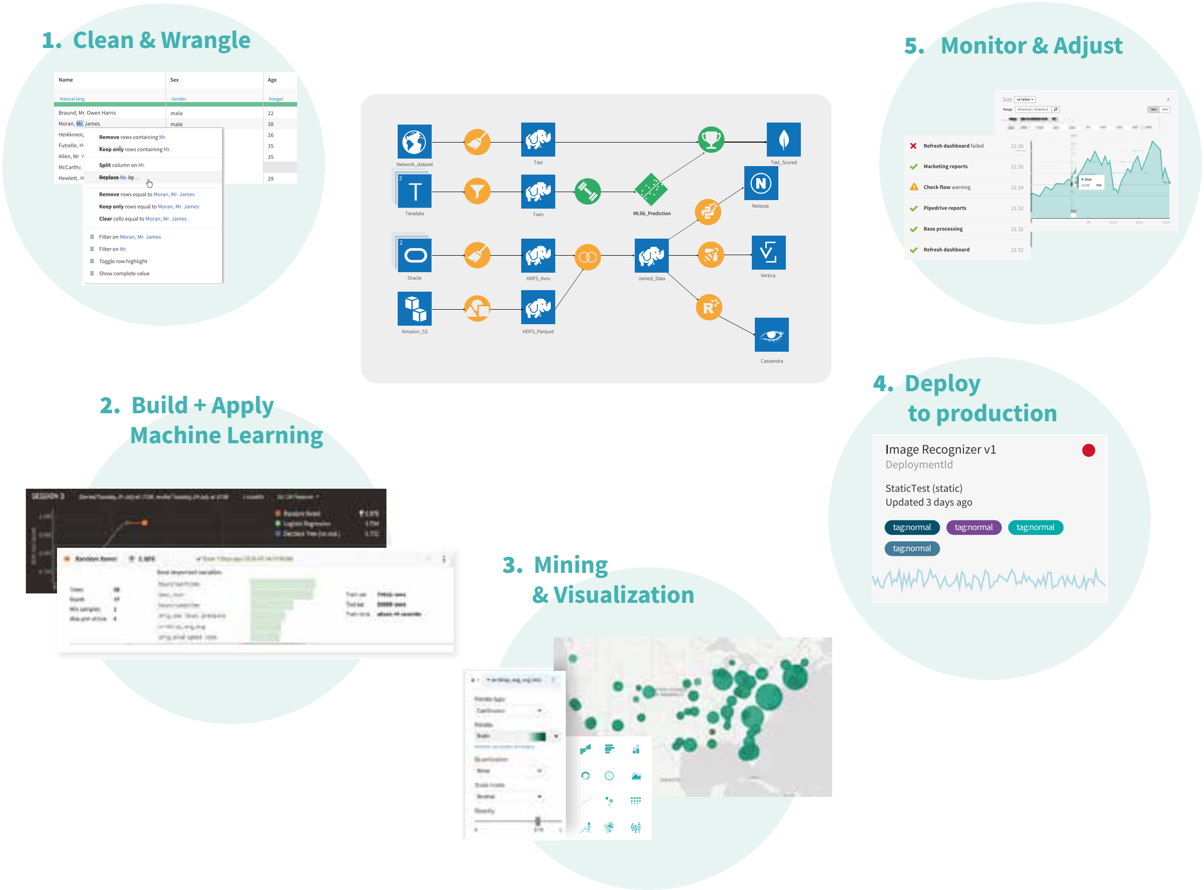
ACTIVE-USERS

*data scientists, analysts, engineers, & more

200+

CUSTOMERS

BNP PARIBAS
Santander
paloalto NETWORKS
PREMERA BLUE CROSS
FOX NETWORKS GROUP
Unilever
NXP
SEPHORA



1. Clean & Wrangle

2. Build + Apply Machine Learning

3. Mining & Visualization

4. Deploy to production

5. Monitor & Adjust

data
iku

WHITEPAPER

www.dataiku.com