

FRONTIERS OF MARKETING DATA SCIENCE JOURNAL

At the Forefront of Smart Data Marketing

**A New Perspective to Budgeting:
By Recrafting the Dorfman-Steiner Condition**

**A Novel Approach to Online Consumer
Journey Clustering Using Neural Networks**

**Correction of Sampling Bias to Produce
Unbiased Analytics**

**Effectiveness of Advertising: A Study on
the Influence of Creative Strength on
the Return of Media Use**

**Privatized Machine Learning for
Marketing Analytics**



Editor-in-Chief

Kajal Mukhopadhyay
Verizon

Executive Editors:

Joshua Koran
Criteo

Ruben Cuevas
Universidad of
Carlos III Madrid (UC3M)

Production & Design

By I-COM Global

Table of Contents

1. Preeti Mascarenhas, Vijayalakshmi Ramesh, Indranil Banerjee, Fokion Loizos / Mindshare
A New Perspective to Budgeting: By Recrafting the Dorfman-Steiner Condition 3

2. Remi Devaux, Matt Andrew / Ekimetrics
A Novel Approach to Online Consumer Journey Clustering Using Neural Networks 11

3. Shreya Jain, Swapnasarit Sahu / Zeotap
Correction of Sampling Bias to Produce Unbiased Analytics 20

4. Mark Vroegrijk / DVJ Insights
Effectiveness of Advertising: A Study on the Influence of Creative Strength on the Return of Media Use 30

5. Joao Natali, Robert Stratton / Neustar
Privatized Machine Learning for Marketing Analytics..... 40

A New Perspective to Budgeting: By Recrafting the Dorfman-Steiner Condition

Preeti Mascarenhas

Mindshare India

Vijayalakshmi Ramesh

Mindshare India

Indranil Banerjee

Mindshare India

Fokion Loizos

Mindshare UK

Classifications,

Key Words:

- Budgeting
 - Forecasting
 - Optimal advertising
 - Dorfman-Steiner condition
 - Short term & long-term budgeting
 - Brand funnel
 - Advertising to Sales Ratio for budgeting
-

Abstract

Budgeting for marketing is no longer a luxury but a necessity, if marketers are to cope with sudden changes in demand levels, price-cutting maneuvers of the competition, large swings in the economy, etc. More importantly, budgeting and the associated marketing plan are critical, if an organization must meet its business goals. While there are many ways – some very rudimentary and simple, and others long drawn and sophisticated - in which one could forecast the required budget, there is no denial that budgeting requires analytical equations that solve for the market dynamics and facilitates scenario planning. And the fundamental building block for any forecasting exercise is data – historical sales data, pricing, distribution, promotion, competition, advertising, seasonality, micro & macroeconomic factors, etc. But what if the brand is relatively new, with inadequate historical data points, or the brand is planning to enter a new market that is vastly different from the earlier one in which they have already been operating? Traditional budgeting approaches require a huge amount of historical data and a longer timeline to forecast a robust budget. And we didn't have the luxury of data and time while working with a global home furnishing retailer; we needed a model that would forecast budget for the short term (sales) as well as the long term (brand equity) KPIs. After scouring through multiple predictive techniques, we triangulated three different approaches to arrive at a robust budget. The uniqueness of this exercise was that it was a blend of data, tools, insights, and logic in varying degrees. After all, it was Albert Einstein who said: "Pure mathematics is, in its way, the poetry of logical ideas". This article details those approaches, especially the one where we recrafted the Dorfman-Steiner condition to forecast budget. These approaches are relatively simple and can help brands that are data-scarce and operating with shorter timelines.

1. Introduction

Before we get to the approach, it is critical to understand the context and the limitations we had. This approach was specifically built for a home furnishing brand which is a challenger brand in India, and this article outlines how we overcame data challenges to justify the need for an additional budget 'to build the brand

in India' at a time when every other brand was either tightening the wallet or investing minimally in short-term activation, due to COVID-19. While there were multiple challenges in carrying out the budgeting exercise, at the heart of it was the data challenge:

1. **Limited historical brand data:** The brand was only 18 months old in the city it had set up operations.
2. **Chasing two different business KPIs, that may or may not have an established relationship:** Short-term sales target and long-term brand equity.
3. **No category/competitive sales data:** Category was largely unorganized, with branded furniture accounting for only 30% of share. There were no third-party reports/estimates on either the market size or competitive sales data.
4. **Diverse country:** India resembles Europe in the sense that no two cities are homogeneous, both in terms of culture and of media landscape. Therefore, market entry cost varies across cities and cannot be generalized.
5. **No benchmarks:** There was no other comparable brand within the category, or even outside the category, to draw learnings from, as the brand is unique with respect to its offerings, range and solutions.
6. **Brand outspent competition:** Budget had to be contextualized to the brand task of driving the category growth, with a business objective 10x times the competition in terms of store visitation as well as brand saliency scores.

2. The journey

A quick feasibility study helped us to understand that none of the existing techniques would serve our dual objective. Some of them could solve only for the short-term sales target (e.g. Competitive benchmarking – SOE/SOV analysis;

Jones analysis – ESOV / SOM; Budget as a % of sales, etc.), and the ones that could solve for both the short and long term (e.g. Time series or Econometric modeling) needed rich data points. As we knew the business goals, our default approach was the objective and task-based one. However, we still needed a technique that could solve for both the long- and short-term objectives by using some of the data points that we had – brand sales data and brand health scores (Conversion funnel) – a technique that would link all three factors – Budget, Sales & Brand Equity.

Dorfman-Steiner Theorem

The Dorfman–Steiner theorem (or Dorfman–Steiner condition) is a neoclassical economics theorem that looks for the optimal level of advertising a firm should undertake. As per the theorem, firms can increase their sales by either decreasing the price of the good or persuading consumers to buy more by increasing advertising expenditure. The optimal level of advertising for a firm is found where the ratio of advertising to sales equals the price-cost margin times the advertising elasticity of demand. (Wikipedia contributors, 2018; Dorfman & Steiner, 1954)

Summarizing research from Dorfman and Steiner (1954) and Levy and Simon (1989):

The Dorfman-Steiner Theorem and increasing returns to scale answers the classical question of how much advertising is sufficient? The **Dorfman-Steiner Theorem** in economics is derived quite simply by taking a profit function where quantity sold q depends on price p and advertisement a , and goods sold are produced at constant marginal cost c . Then

$$\pi = (p - c) \cdot q(p, a) - a$$

and the corresponding two first-order conditions are

$$\frac{d\pi}{dp} = q(p,a) + (p-c) \frac{\partial q}{\partial p} = 0$$

and

$$\frac{d\pi}{da} = (p-c) \frac{\partial q}{\partial a} - 1 = 0$$

Introduce the two elasticities

$$\eta \equiv - \left(\frac{\partial q}{\partial p} \right) \left(\frac{p}{q} \right) > 0 \quad \text{and} \quad \theta \equiv \left(\frac{\partial q}{\partial a} \right) \left(\frac{a}{q} \right) > 0$$

which allows us to rewrite the two first-order conditions as

$$\eta = \frac{p}{p-c} \quad \text{and} \quad \frac{a}{p \cdot q} = \frac{p-c}{p} \theta$$

Combining the two conditions yields

$$\frac{a}{p \cdot q} = \frac{\theta}{\eta}$$

which says that the optimal level of advertisement (as a share of revenue) is equal to the ratio of advertising elasticity and price elasticity. Put another way, advertisement increases with its effectiveness but decreases with price elasticity. Advertisement expenditures tend to be larger in price-inelastic markets. (Antweiler, 2016)

The limitation of this theorem model lies in its isolated mode of decision-making where interdependence among firms with regard to decision variables has been ignored. We had to reengineer the theorem to ensure

1. Market structure must be created/simulated to capture market nuances.
2. While the theorem could solve for budgeting based on sales target, it still didn't address long-term brand equity.


Enter Logic: Recrafting the Dorfman-Steiner Theorem

Here is where our proprietary tool, along with the brand track, played a vital role.

3. Methodology

Our proprietary tool is built on the Dorfman-Steiner theorem and it recreates a market structure using sales value and volume. It analyses historical sales and media investment data for a category, generates advertising response curves, and identifies the appropriate level of media spend for a brand, or brands. (see **Table 1**)

Please enter values for the most recent period



Run ONE BRAND budget setting

Run multi-brand budget allocation

	MEDIA INV.	VALUE	VOLUME	DISTRIBUTION	MARGIN	
Brand	Required		Strongly recommended	Optional		
Brand 1	109,540	2,627,271	3,012,194	82%	6%	Competitor
Brand 2	75,687	2,812,185	1,950,648	84%	10%	Competitor
Brand 3	100,024	2,657,459	2,693,619	81%	8%	Competitor
Brand 4	78,230	884,514	259,041	83%	11%	Competitor
Brand 5	987,568	4,000,000	77,248	84%	12%	Competitor
Brand 6	131,752	2,097,378	443,039	81%	14%	Client
Brand 7	123,781	1,314,101	261,033	82%	12%	Client
Brand 8	13,229	1,190,866	452,226	80%	12%	Competitor
Brand 9	38,540	507,526	281,362	85%	8%	Competitor
Brand 10	65,093	2,125,726	411,237	81%	11%	Competitor
Brand 11	108,026	1,725,838	437,644	81%	13%	Client
Brand 12	53,870	603,195	257,186	80%	10%	Client
Brand 13	44,828	618,840	312,836	82%	8%	Competitor

-> First client brand to be chosen for Budget Setting

Table 1. Creating market structure using Sales Value, Volume & Media investment.

It offers a longer-term view of media ROI than an **MMM-based approach**, through more reliance on assumptions. (see **Table 2** and **Figure 1**)

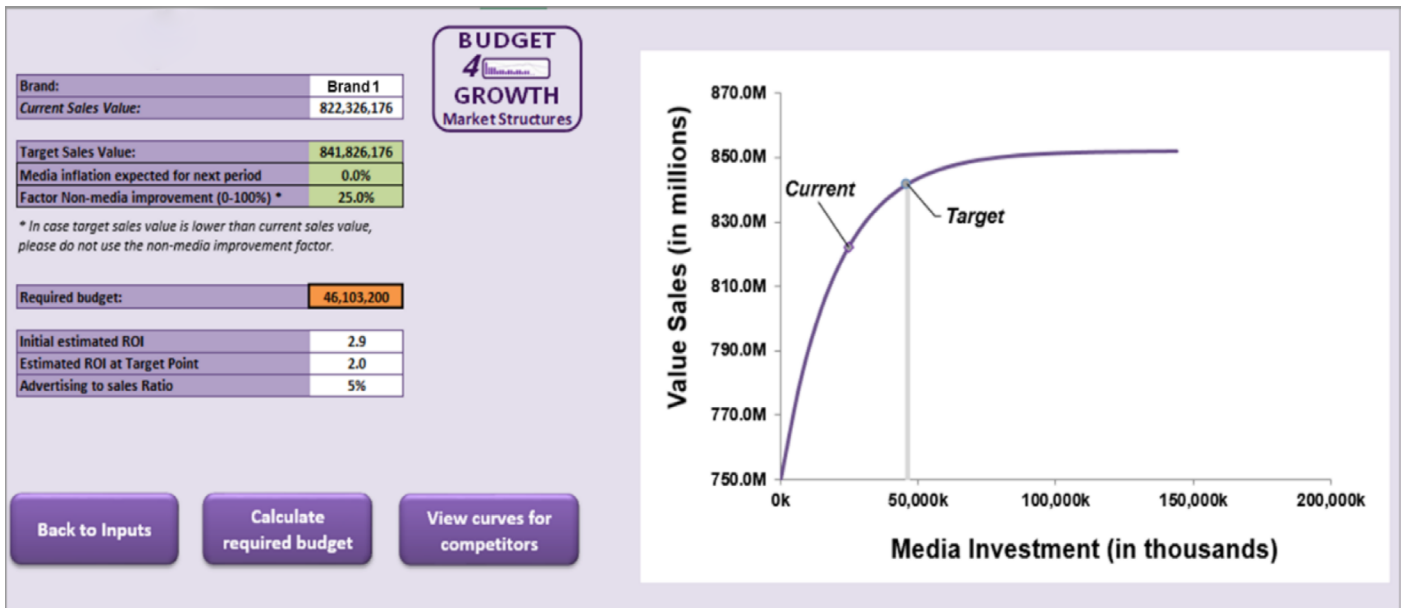


Table 2. Budget prediction for the sales scenario.

Figure 1. Sales to Investment Slope.

The tool can facilitate two things:

1. The estimation of the required media investment needed to reach desired sales for one brand, OR The sales target estimation for a planned media investment for one brand.
2. Response to advertising spend is estimated for all brands in the category ensuring that market context is easily visualised.

But the caveat was that we still didn't have the sales value/volume for competition to recreate the market structure.

3.1 Methodology 1: Inverting the Brand funnel

1. From the brand track, by equating '% purchased the brand' to the brand sales volume, we could estimate the market as well as competitors' sales volume.
2. To address point 2 from above, we inverted the brand funnel. Existing relationship between Awareness - Consideration - Purchase was flipped on its head to predict resulting brand saliency (again, by equating sales to % purchased and thereby backtracking awareness) (see **Table 3** and **4**).

Constructing the market structure

- Reconstructed the market structure using brand funnel (in the absence of competitive sales value)
- If 12% = 1,000 Mn, then the estimated organized cat. Volume is 8,333 Mn
- Competition sales value is arrived at, by using claimed purchase data from the brand track

	Current	
	Average Scores %	Conversion
TOM	12	33%
Spont Awareness	37	47%
Tot Awareness	79	
Purchase	12	
Sales Value (Mn)	1,000	

Table 3. Inverting the brand funnel and arriving at the conversion ratios

	Purchase %	Est. Value Mn
Brand	12.0	1000.0
comp1	0.09	8
comp2	0.20	15
comp3	5.5	458
comp4	2.9	244
comp5	0.36	30
comp6	0.54	46
comp7	25.0	2,130
comp8	16.0	1,318

Table 4. Estimating the sales value using the % Purchase data from Brand track

Once market structure is recreated with respect to competitors' volume, value and advertising spends, we were able to establish the response slopes for each of them. (see **Table 5** and **Figure 2**)

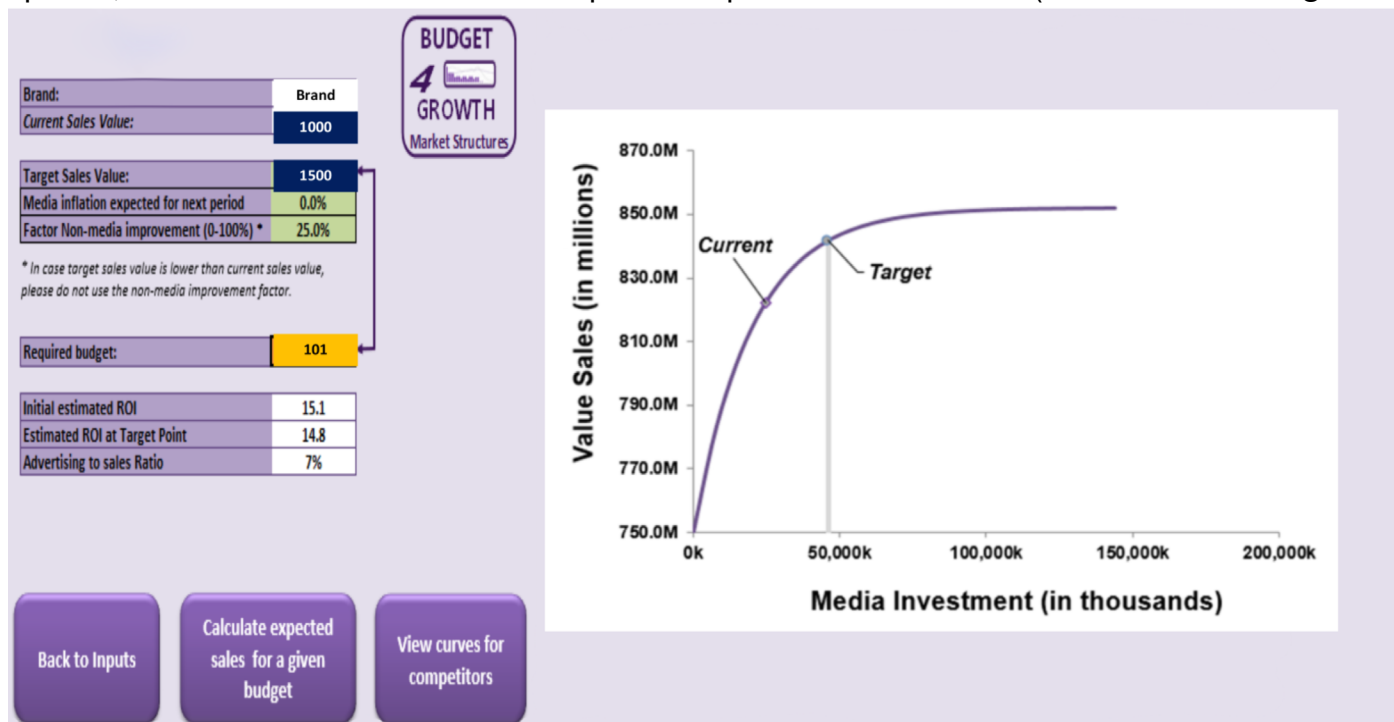


Figure 2. Graphical representation of the budget forecasting

Table 5. Budget forecasting for a sale scenario

This methodology also allowed us to look at 'what if' scenarios, such as

- if sales was entirely contributed by media (100%) as the brand was relatively new, then what should be the media investment?
- what would be the case with media contribution being 75%?
- what about 50%?
- etc.

Having arrived at the response curves (**Figure 3**), we looked at scenario planning with respect to both KPIs, sales and TOM (see **Table 6** and **7**).

FY 20 Media investment: 100Mn
Sales: 1000 Mn
Saliency (Brand Track): 12

If saliency +50% from current levels

	Budget Incr. over current investment	Resulting Sales (incr. over current sales)
Without any non-media contribution	102%	50%
With 10% non-media contribution	73%	50%
With 25% non-media contribution	53%	50%
With 50% non-media contribution	39%	50%

Table 6. Estimating budget for a target saliency level

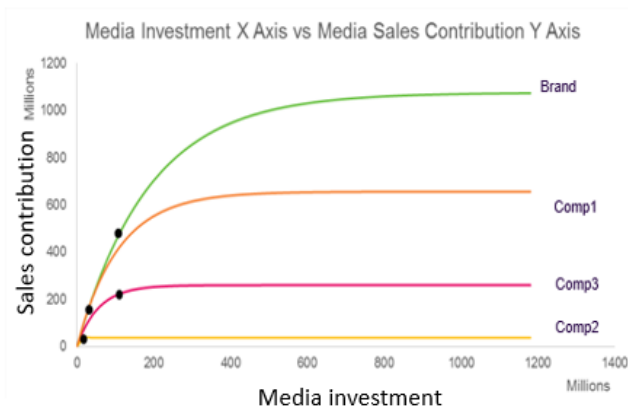


Figure 3. Media to sales response curves for competitive brands

If saliency +10,7% from current levels

	Budget Incr. over current investment	Resulting Sales (incr. over current sales)
Without any non-media contribution	17%	8%
With 10% non-media contribution	14%	8%
With 25% non-media contribution	11%	8%
With 50% non-media contribution	9%	18%

Table 7. Estimating budget for a target sales value

This method could not be deployed for the new market, as there was no brand track. We worked with Euromonitor estimates to construct the market structure and subsequently arrived at the response slope for the competition.

3.2 Methodology 2: Objective & Task Based

We built the media plan from the bottom, campaign by campaign. Here the challenge was to arrive at a unique reach across all media, as the majority of the brand's budget was being invested in outdoor advertising, for which there was no scientific way to determine the reach.

However, since we were looking at building the brand primarily, and hence impacting brand saliency score, we found a good correlation between AV (TV+Digital) medium investment and Brand health scores (TOM) using single variate correlation analysis, and therefore we increased share of investment on these 2 medium and built campaign plans.

3.3 Methodology 3: A:S Ratio benchmarking

In this approach, we looked at multiple categories, as well as brands (Figure 4) that were close to our brand in terms of offerings and life stage (Figure 5), to estimate the budget requirement.

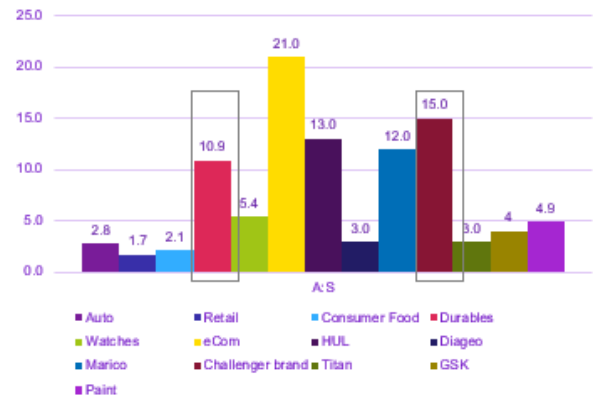


Figure 4. A:S ratio across different categories.

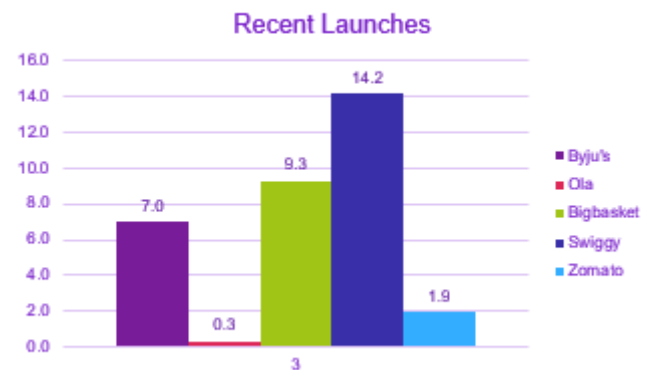


Figure 5. A:S ratio of brands, which are comparable in terms of either the offerings or at the same life stage.

4. Discussion

By exploring different methodologies, we were able to forecast a range with respect to budget and business KPIs that helped the brand to justify and procure additional budget for the long-term brand growth, which was much needed considering the life stage of the brand. And the investment paid off, as evident from the following chart, where the brand's TOM recall improved substantially (see Figure 6).

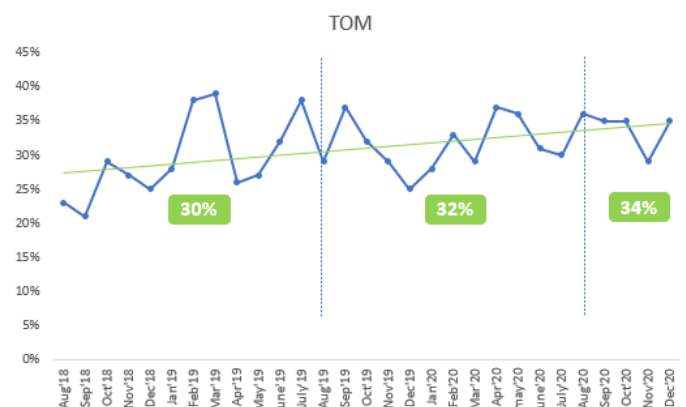


Figure 6. Upward trend in brand health scores

Referring to Queens College, CUNY (2008) insights:

Many companies set their marketing budget at what they think the company can afford. This affordability method of setting marketing budgets ignores the role of marketing as an investment and the immediate impact of marketing on sales volume. It is easier to budget for the short term, but budgeting for brand growth is essential to future-proof a business.

Kantar's BrandZ demonstrates that brands that invest in brand building and see that as an investment and not as a cost, stand to gain in the long run (Kantar, 2021). So, what if the framework is a confluence of data, tool,

technique and sound logic? To bring back Albert Einstein, "Pure mathematics is, in its way, the poetry of logical ideas".

5. Limitation of this Approach

Firstly the Dorfman-Steiner theorem assumes that the impact of advertising is positive, while this may not always be true. Secondly, the supply side is assumed to be constant, but short supply can push up prices higher than predicted if advertising ends up creating more demand than supply. The theorem deals with only one time-period and generalizes the possible lagged effect of advertising in a multi-period situation. Lastly, this approach centered around creating the market structure based on the claimed purchase behavior from the brand track, and not the actual market share.

References

1. Wikipedia contributors. (2018, August 12). *Dorfman-Steiner theorem*. Wikipedia. [Source](#)
2. Dorfman, R., & Steiner, P. (1954). Optimal Advertising and Optimal Quality. *The American Economic Review*, 44(5), 826–836. Retrieved June 18, 2021. [Source](#)
3. Levy, H., & Simon, J. (1989). A Generalization That Makes Useful the Dorfman-Steiner Theorem with Respect to Advertising. *Managerial and Decision Economics*, 10(1), 85–87. Retrieved June 18, 2021. [Source](#)
4. Antweiler, W. (2016, March 31). *The Dorfman-Steiner Theorem and increasing returns to scale*. Blog of Prof. Werner Antweiler, Ph.D. [Source](#)
5. Queens College, CUNY. (2008). *Marketing Notes* [Slides]. Course BUS 243 - Economics of Distribution and Marketing. [Source](#)
6. Kantar. (2021, April 15). *What is brand equity?* [Source](#)

Authors



Preeti Mascarenhas, VP, Head of Strategy, Mindshare, India

Preeti leads the strategy mandate across India and comes from strong research, data and strategic planning background across APAC markets. She is a transformative strategic leader & incubation partner in Mindshare for strategic planning, innovation and business intelligence. In the last 10 years in MS across APAC and India Market, she has been instrumental in building the been in harvesting strategies, Command centre (Loop) and Data at the core.

preeti.mascarenhas@mindshareworld.com



Vijayalakshmi Ramesh, Principal Partner, Strategy, Mindshare, India

As Strategy lead, Vijayalakshmi has worked with a wide range of clients across FMCG, Retail, B2B as well as new age businesses. With a background in Statistics (Operations Research), her primary interest is in communication planning and driving outcome-based planning for her set of clients.

vijayalakshmi.ramesh@mindshareworld.com



Indranil Banerjee, Principal Partner, Strategy, Mindshare, India

Indranil has about 15 years of experience in research and the media industry. He has serviced clients across various sectors in India. He is left-brained and loves anything around data. Besides research and strategy, Indranil is also experienced in advanced analytics – primarily MMM, RoI calculation & forecasting. He has worked with clients on KPI setting, deploying resources to achieve KPIs, and measuring RoI by resources deployed.

indranil.banerjee@mindshareworld.com



Fokion Loizos, Analytics Tools Manager, Mindshare, UK

Fokion is an experienced Tools Manager with a demonstrated history of working in the marketing and advertising industry. As a strong research professional with MEng in Mining Engineering from NTUA, he is highly skilled in the development of new Tools and Software that use leading-edge analytics, including MindShare's set of budget optimization and effective media planning tools.

fokion.loizos@mindshareworld.com

A Novel Approach to Online Consumer Journey Clustering Using Neural Networks

Rémi Devaux

Ekimetrics

*MINES ParisTech – PSL
University*

Matt Andrew

Ekimetrics

Classifications, Key Words:

- Customer journeys
 - Personalisation
 - Clustering
 - Neural networks
 - Online marketing
 - Purchase funnel
-

Abstract

Customer journeys are a crucial tool in the marketer's toolbox to understand how users interact with content, move through a digital ecosystem, and to understand what stage of the purchase funnel they may be in. However, quantitative analysis of the customer journey faces structural limitations, including issues with a curse of dimensionality (where the combinations of journeys grow so fast data becomes sparse). In this paper, we present a novel analytical method to cluster customers based on their online behavior, captured as they navigate an inter-connected digital journey. We have used a novel approach to image processing techniques to overcome the issues associated with traditional analysis of customer journeys. The output of this model provides brands with clusters of their customers that represent the journeys they have taken and several relevant behavioral features. These can then be used to establish the most prominent pathways through a digital ecosystem and tailor content, calls-to-action, and offers based on the most likely cluster they belong to and so increase conversion rates down the funnel and ultimately to a sale.

1. Introduction

Investment in Online Marketing continues to rise rapidly, with online advertising now accounting for more than half of all media spend (eMarketer, 2019). One reason to explain this shift is the ability of advertisers to directly track a customer's response to their marketing actions. The accumulation of these responses and actions in the digital space makes up a *customer journey*, and this is a critical part of the value seen in online activity.

The concept of a customer journey refers to a time-based, sequential mapping of customers' attitudes and behaviours toward a brand. Primarily it consists of an empirical assessment of a customer's expected utility for a good, at each stage of their journey towards purchasing it. The approach has been empowered by online marketing, which provides many proxies for a customer's utility (e.g. clicks, time spent on each page and interactions with the page).

Understanding the dynamics of customer utility is a key stake for online brands, as knowing what is important to customers enables improvement of their experience through website enhancement as well as an optimization of marketing actions at the right moment.

In this article, we present a method to explore this concept accurately by clustering similar customer journeys using image embedding algorithms. This allows a richer vision of customer behaviours in a high dimensionality space, highlighting niche behaviours that point to different stages of the purchase funnel and could otherwise be missed in traditional analysis approaches. Our model has implications for marketers' ability to understand how customers engage, convert and churn with their websites and in the actions they put in place to drive successful brand outcomes.

2. Related Work

Our research logically falls within marketing literature on the customer journey. More specifically, it contributes to an empirical assessment of the customer journey mapping (CJM) (Richardson, 2010). Such mapping is often based on theoretical, or qualitative empirical methodologies (Rosenbaum et al., 2016). The objective of CJM is to qualitatively understand how a customer's expected utility varies through their journey, and more importantly, why. In this paper, we balanced this qualitative approach with a quantitative method to determine the typical customer journeys of a business or brand online. Nevertheless, our model does not aim to replace theoretical and qualitative methodologies. Instead, it can serve as an additional input for CJMs.

Currently, quantitative analysis of customer journeys is less common in the literature. Mangiaracina et al. (2009) is an early study of customer utility through the online purchase process. Terragni and Hassani (2018) provides a model to analyze online customer journeys in order to build a page recommendation system. Closer to our approach, Temouden (2020)

clusters online customer behavior through Markov models.

Despite growing interest in the field with the increasing rise of digital, there is a lack of empirical studies, and in particular of quantitative methods, as highlighted by Halvorsrud et al. (2016). We have identified two reasons behind this:

Firstly, the quantification of the customer journey is structurally challenging due to the very high number of potential journeys. Consider a website with n pages. Even with the simplistic assumption that a user can only make one action per page, the total number of possible journeys J_n can potentially reach the following sum of arrangement:

$$J_n = \sum_{k=1}^n \frac{n!}{(n-k)!}$$

This assumes that all pages are linked to each other in a full mesh, i.e. through the sitemap, making all journeys through them possible. However, in practical terms, the observed number of journeys are lower, but importantly still inordinately complex for traditional analytics.

Additionally, a huge amount of observations are available when it comes to customer journeys (the order of actions, time spent on each action, user provenance and personal data, etc.). This raises a *curse of dimensionality* (Bellman, 1957) where the volume of the space increases so quickly that data becomes incredibly sparse. We counter these two issues by using a dimension reduction method: the convolutional autoencoder (Guo et al., 2017). More broadly, this paper applies image-processing methods (Chang et al., 2017) to counter these issues and evaluate customer journeys in a marketing context.

3. Input data

The model presented in the following section utilises customer journey data from a leading car manufacturer's digital ecosystem. We retrieved the browsing log data of ~300 users on their

website to demonstrate the complexity of the journeys encountered and how to address this. Our dataset has ~86,000 lines, each corresponding to an action, i.e. a change in a user's URL, with an average of 280 actions per user. This change can be caused by a new page visit or a script execution within a page (Table 1).

Feature	Description
Customer ID	Unique ID for each visitor
Event time	Time of an action
Page category	Identify the category of a visited page
Page name	Name of the visited page

Table 1. Description of our Data. The set contains 86.000 lines or entries representing individual consumer actions.

To ensure the relevance of the *page* variables, we grouped URLs into appropriate categories. For example, the links, *ourwebsite.com/homepage/news_1* and *ourwebsite.com/homepage/news_2* will be both classified in the *News* page which belongs to the *Homepage* category. The method holds for the scripts executed by a user on the page.

4. Methodology

As mentioned previously, we treat customer journey clustering as an image processing problem to reduce the issues of dimensionality. Encoding the information about the different stages of individual journeys into images is the necessary first step. Whilst we could use vectors or matrices to store this information, we would then be limited to focusing on one (say, pages visited) or two (say, pages visited and the time spent) customer dimensions; images have the benefit of representing three, giving us richer behavioural information to support the assessment of predicted customer utility.

The second step consists of using a neural network to reduce the dimensions of the corresponding images. Dimension reduction is crucial since it (i) reduces the computation time of our model and (ii) the process highlights

the relevant information that delivers the most variance for the clustering of journeys.

Finally, it is necessary to apply a clustering method to these images. There are many approaches to clustering, but the ultimate goal is to collect similar journeys together in a set and allow brands to view groups of typical or commonly seen customer journeys instead of being overwhelmed by focusing on individuals.

4.1 Determining the dimensions

Before proceeding with the CJC approach, one must establish the dimension(s) to clusters. We have accounted for three: the number of actions per customer, their sequential order, and the time spent on each one. We believe this mix of behaviours can give a stronger assessment of customer utility; more actions, and more time spent, represent a deeper connection than a surface level view, whilst the sequential order helps determine what is important to the customer at this time and how far down the purchase funnel they are proceeding (i.e. are they at the stage where they are focusing on a car segment, or have they chosen this and looking into the specifications available, but also which elements of the offer are most important to them, be it having a test drive, getting a promotional offer, etc.?).

4.2 Encoding customer journeys into images

Considering a website with n pages P_1, \dots, P_N , we denote $P_i^{j,k}$ as a customer journey of length l going from page j to k . We create an image of dimension $l, n+1$ cells. The l rows represent user actions and the $n+1$ columns denote the n total possible actions on the site plus when the user leaves the website.

To account for our third dimension – time – we use the content of the cell. The duration of an action is measured by the value (i.e. color) of the cell. In order to limit the contrast between cells, we use a logarithmic function to implement decreasing output:

$$M(i, j) = \begin{cases} \log(t_i \mathbb{I}_{j \in A_c} \mathbb{I}_{a_i=j}) & \text{if } t_i \geq 1 \\ \log(t_i \mathbb{I}_{j \in A_c} \mathbb{I}_{a_i=j} + t_{min}) & \text{else.} \end{cases}$$

The indicator functions \mathbb{I} ensure a user action i have been recorded on page j . t_{min} is necessary to fix the minimum duration of the action to 1.

4.3 Dimension reduction

At this stage, we have as many images as customer journeys in our dataset. Applying a clustering method on such images would overwhelm the algorithm’s capability and is thus impossible. We must first lower the dimension of the journey images we are willing to cluster. A Convolutional autoencoder (CAE) (Guo et al., 2017) is a satisfactory way to do so. We implemented the neural network model using Python’s library keras.

As an autoencoder, CAE is composed of two functions: the *encoding function* $f(x)=h$ reduces the dimensions of our input image x to obtain a reduced vector h ; then, h is processed through a decoding function $g(h)=g[f(x)]\equiv\tilde{x}$ to compute the complete image input \tilde{x} . In order to limit the information lost during the encoding-decoding process, \tilde{x} should be as close as possible in form to the original input x .

The specificity of CAE lies in the specification of its *activation functions*, CAE uses a convolutional operator between the input matrix x and the activation parameters matrix D and E :

$$\begin{aligned} f_E(x) &= \sigma(x * E) \equiv h \\ g_D(h) &= \sigma(h * D) \equiv \tilde{x}. \end{aligned}$$

Here, σ is the activation function and $*$ a convolutional operator. More specifically, our architecture (described in **Figure 1**) involves the following structure:

$$\begin{cases} f_1(x) = \sigma(x * E_1) \equiv x_1, \\ f_2(x_1) = \sigma(x * E_2) \equiv x_2, \\ f_3(x_2) = \sigma(x * E_3) \equiv h, \\ g_1(h) = \sigma(x * D_1) \equiv h_1, \\ g_2(h_1) = \sigma(x * D_2) \equiv h_2, \\ g_3(h_2) = \sigma'(x * D_3) \equiv \tilde{x}. \end{cases}$$

The stride parameter, i.e. the speed of dimensionality reduction is uniformly equal to 2 at each step. Although pooling layers are commonly implemented in image processing, we do not use these as the treated images are small.

The role of an encoder is to minimize the loss between our input journeys x and the output \tilde{x} . The encoding and decoding layers use a rectified linear unit (ReLU) activation function given by:

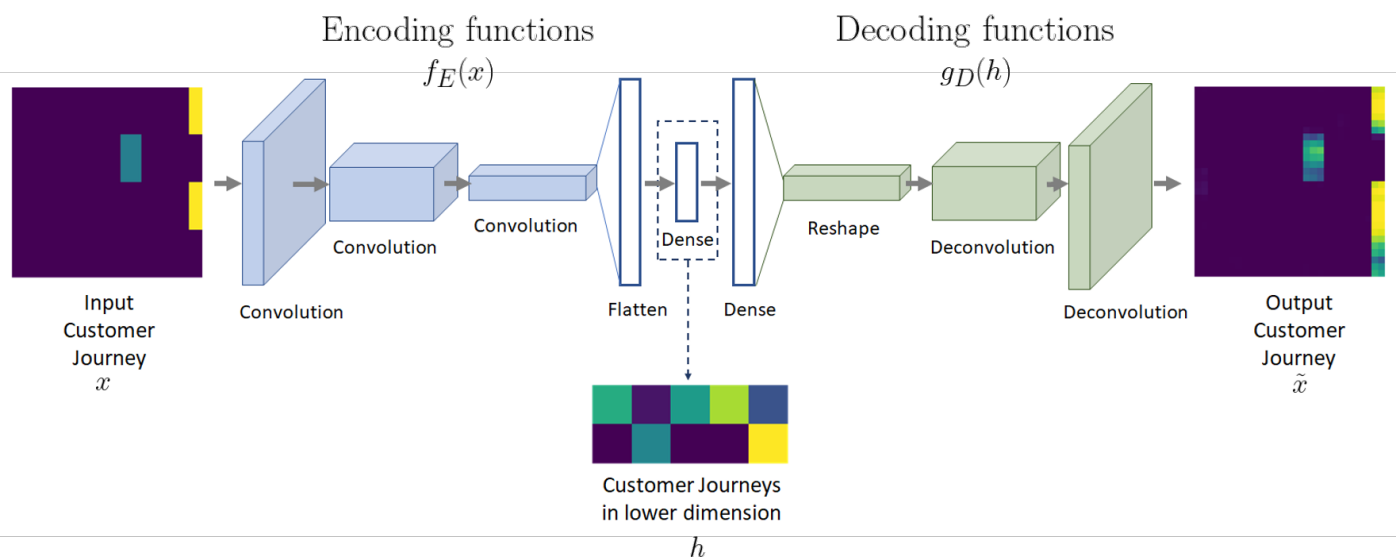


Figure 1. Architecture of our convolutional neural network. The information of the consumer journey x is summarized into a low-dimension image h which is, in turn, reconstructed into \tilde{x} . While Standards autoencoders are mainly suited for tasks implying a unique encoding and decoding layer, a convolutional neural network can perform several reduction-reconstruction tasks efficiently.

Figure based on Guo et al. (2017).

$\sigma(x) = \max\{0, x\}$. The last layer before the output \tilde{x} uses a sigmoid function $\sigma'(x) = \frac{1}{1+e^{-x}}$. In a presence of a sigmoid form, the loss function minimized by our AEC is the binary cross-entropy:

$$L(x) = - [x \log(\tilde{x}) + (1 - x) \log(1 - \tilde{x})].$$

4.4 HDBSCAN clustering

We now cluster journeys using a HDBSCAN, a density-based algorithm able to identify outliers. This method is independent of DBSCAN's hyper-parameters by using hierarchical clustering. We used Python's package hdbscan for implementation of the algorithm. It is based on the *mutual reachability distance* between two points. This metric uses (i) the distance d_k of both points to their k^{th} neighborhood and (ii) their Euclidean distance d_e :

$$d_m(x, y) = \max\{d_k(x), d_k(y), d_e(x, y)\}.$$

This distance allows to take into account the density of our graph. We then use d_m to build a *mutual reachability graph* linking our points with edges equal to d_m . The minimum spanning tree of this graph will serve to build a dendrogram. We then fix a minimum size of clusters min_{obs} such that clusters smaller than min_{obs} are dropped (they are considered as outliers). For more details on the HDBSCAN methodology, see [Cambello et al. \(2013\)](#).

We assess each cluster's performance with a silhouette coefficient ([Rousseeuw, 1987](#)). Silhouettes measure the variance within and between the clusters to ensure the relevance of the groups generated. Further details on the role of the silhouette coefficient and its interpretation are detailed in the **appendix**.

5. Results: discussion & limitations

As introduced in the last part, our algorithm clusters customers on three dimensions: actions, sequence and time. **Figure 2** demonstrates how our model accounts for the number of user actions. We then focus on two clusters and show how they contain a very different sequence of actions. **Figure 3** presents how our clustering handles the third dimension: time. This example specifically shows the diversity of pages visited and how long was spent on them across the groups of customers. In the appendix, we

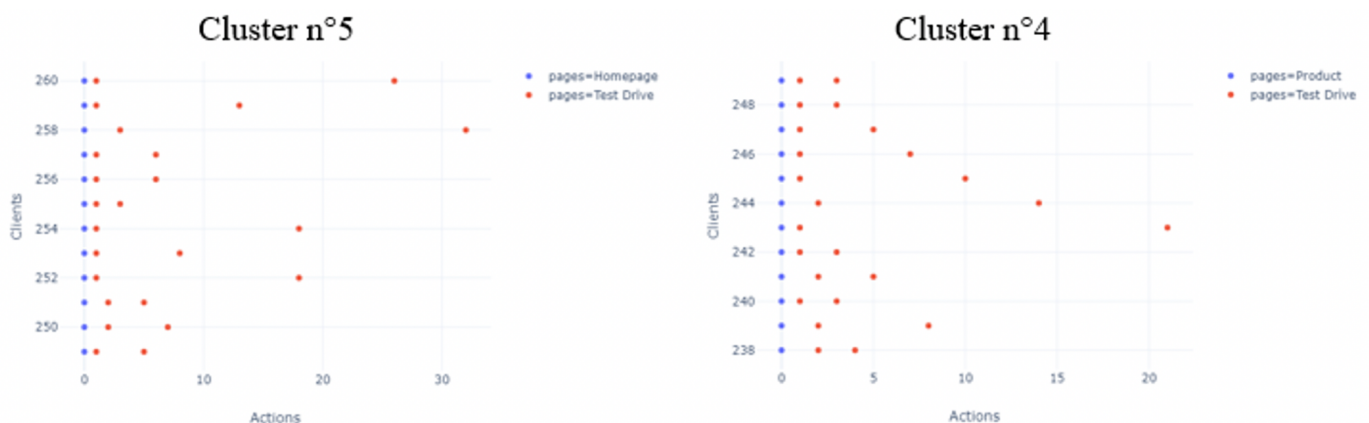
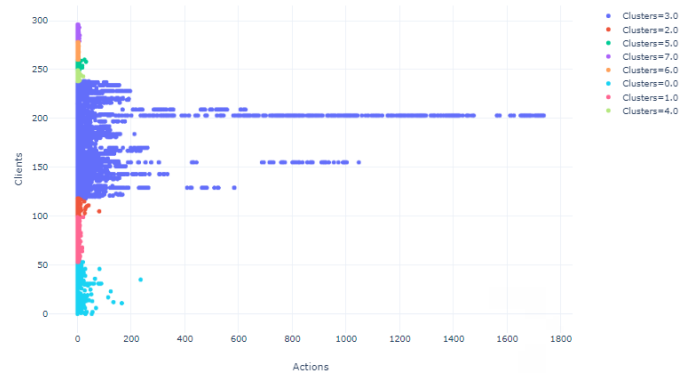


Figure 2. Number of user actions in each cluster (top) and composition of two of them by sequence of actions (bottom). “Clients” refers to an individual consumer ID. Average overall silhouette coefficient=0.85.

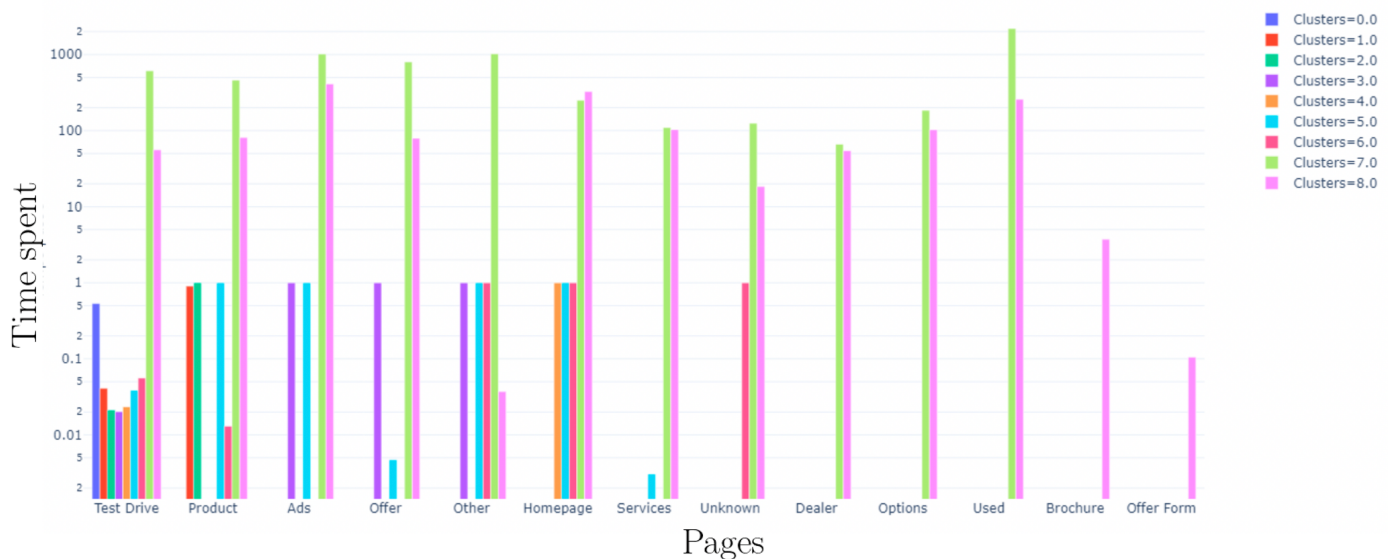


Figure 3: Composition of the clusters. For each page reported on the x-axis, the length of visits (in seconds) is reported by clusters. The scale is logarithmic.

provide the result of a model using standard autoencoders (SAE).

An evident, but important, limitation of our model is that it only focuses on online customer journeys through a brand’s tracked ecosystem.

Although disruptive, the implementation of such a model may be challenging, both technically and in terms of time.

More importantly, the method is unsupervised, meaning that we are not controlling for the relevance of the clusters generated. That is why we recommend strong business expertise is exercised alongside modelling to ensure the quality and sensibility of results. This recommendation is in line with our early idea that our model cannot replace qualitative business methodologies like CJMs. Instead, our customer journey clustering model is based and evaluated according to business assumptions. It strengthens the mapping and understanding of how customers interact with the digital ecosystem by providing empirical data on their online behaviours.

Clusters have the additional value in making this knowledge more actionable and digestible compared to focusing on individual customer journeys. In short, our customer journey clustering model should be used as a business-

driven tool and integrated into a wider range of marketing methods.

6. The implications for marketing practitioners

Our model allows a business to understand the actual typical online journey taken by their customers. The impact of this is an evolution of the power of actions that brands are already taking in the digital space; it enhances the success of actions rather than bringing completely new marketing effects into play. By bringing more dimensions of the customer journey into play, our clients have used this insight to improve the optimization their digital ecosystems in well-known ways:

6.1. Driving website inefficiencies out

Knowing how customers move effectively through your digital ecosystem is especially important since it allows brands to focus on the important areas and ensure the right content is in the right place. Alternatively, it can be used to highlight and reduce bottlenecks such as time-consuming or slow-loading pages, improving the overall customer experience.

6.2 Improving churn prediction

The results of our model contribute to the literature on churn prediction by providing firms with additional insight on how customers churn, i.e. what actions they take, in which order, and for how long. By comparing similar clusters, brands can identify patterns between journeys

and optimize clusters “at-risk” of churning. Considering two clusters $C1:\{A\rightarrow B\rightarrow C\rightarrow Convert\}$ and $C2:\{A\rightarrow B\rightarrow Exit\}$, the page C should be suggested to C2’s consumers. The same principle can be applied in order to maximize conversion rate as well.

Conclusion

This paper addresses the challenges and benefits to quantifying the online customer journey. We introduced the technical barriers of turning a sequential analysis question into an image-processing problem, a solution necessary to avoid issues with dimensionality. The method presented is innovative and presents promising results.

In particular, our model enables brands to know by what means a typical customer’s utility reaches its minimum (churn) and maximum (conversion) points. As such, the empirical assessment and prediction of those two breakpoints are a key stake for the marketing industry.

However, it is important to remember there are limitations in exclusively focusing on digital behaviours and actions. This limitation opens the door to further research on the question. Leveraging data before and after an advertising campaign to measure how typical journeys, churn and conversion paths react to media investment would be an interesting continuation of this research.

References

- Richardson, A. (2010). Using customer journey maps to improve customer experience. *Harvard business review*, 15(1), 2-5.
- Rosenbaum, M. S., Otolara, M. L., & Ramírez, G. C. (2017). How to create a realistic customer journey map. *Business Horizons*, 60(1), 143–150. [Source](#)
- Terragni, A., & Hassani, M. (2018). Analyzing Customer Journey with Process Mining: From Discovery to Recommendations. *2018 IEEE 6th International Conference on Future Internet of Things and Cloud (FiCloud)*, 224-229.
- Temouden, K. (2020, February 17). *Clustering Customer Journeys using a Mixture of Markov Models*. Medium. [Source](#)
- Halvorsrud, R., Kvale, K., & Følstad, A. (2016). Improving service quality through customer journey analysis. *Journal of Service Theory and Practice*, 26(6), 840-867.
- Guo X., Liu X., Zhu E., Yin J. (2017) Deep Clustering with Convolutional Autoencoders. In: Liu D., Xie S., Li Y., Zhao D., El-Alfy ES. (eds) *Neural Information Processing. ICONIP 2017. Lecture Notes in Computer Science*, vol 10635. Springer, Cham. [Source](#)
- Mangiaracina, R., Brugnoli, G., Aless, & Perego, R. (2009). The eCommerce Customer Journey: A Model to Assess and Compare the User Experience of the eCommerce Websites. *The Journal of Internet Banking and Commerce*, 14(3), 1-11.
- Bellman, R., & Rand Corporation. (1957). *Dynamic programming*. Princeton: Princeton University Press.
- Chang, J., Wang, L., Meng, G., Xiang, S., & Pan, C. (2017). Deep Adaptive Image Clustering. *2017 IEEE International Conference on Computer Vision (ICCV)*, 5880–5888.
- Kingma, D.P., & Welling, M. (2014). Auto-Encoding Variational Bayes. *CoRR*, [abs/1312.6114](#).
- Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. In J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (Éds.), *Advances in Knowledge Discovery and Data Mining (Vol. 7819)*, p. 160-172. Springer Berlin Heidelberg.
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65. [Source](#)

Appendix

Silhouettes coefficient

Clustering involves creating groups of observations that exhibit low intra-group variance but high inter-group variance. The silhouette coefficient evaluates a cluster's quality based on those two features. An observation x_i included in the cluster C_k has a silhouette coefficient equal to:

$$s(x_i) = \frac{b(x_i) - a(x_i)}{\max\{a(x_i), b(x_i)\}} \in [-1,1], \quad \text{where:}$$

$$a(x_i) = \frac{1}{|C_k - 1|} \sum_{j \in C_k} d(x_i, x_j) \quad \text{and} \quad b(x_i) = \min_{k' \neq k} \frac{1}{|C_{k'}|} \sum_{j \in C_{k'}} d(x_i, x_{j'}).$$

Here, $a(x_i)$ is the average distance between x_i and its neighborhoods, i.e. a proxy for intra-group variance. Whereas $b(x_i)$ measures the average distance between each cluster, i.e. a proxy for inter-group variance. A high overall silhouette coefficient can be interpreted as a performant clustering outcome. **Figure 4** provides the evolution of the silhouette coefficient according to HDBSCAN's minimum group of observation.

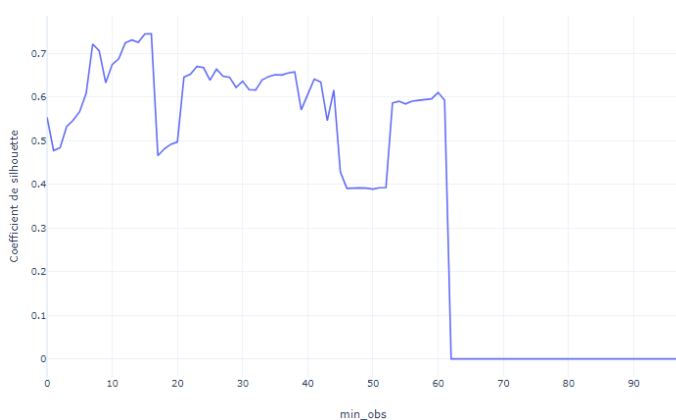


Figure 4: Overall average silhouettes coefficient evolution with HDBSCAN's.

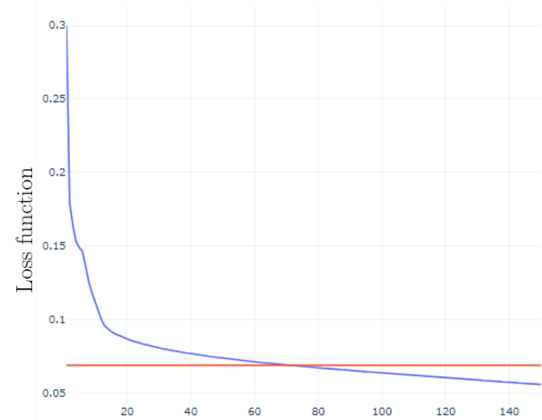


Figure 5: Sensitivity of the model's loss to the size of the training sample for the Convolutional (blue) vs Standard (red) autoencoders.

Standard autoencoders (SAE) performance benchmark

Figure 5 provides the loss function optimization gap between the two methods. It shows that the loss function of CAE is better optimized as the training sample (epoch) is large. Figure 6 shows the clusters obtained using a SAE+HDBSCAN method. The page diversity represented in the cluster and the overall silhouette coefficient suggests the configuration does not perform as well as the CAE+HDBSCAN method.

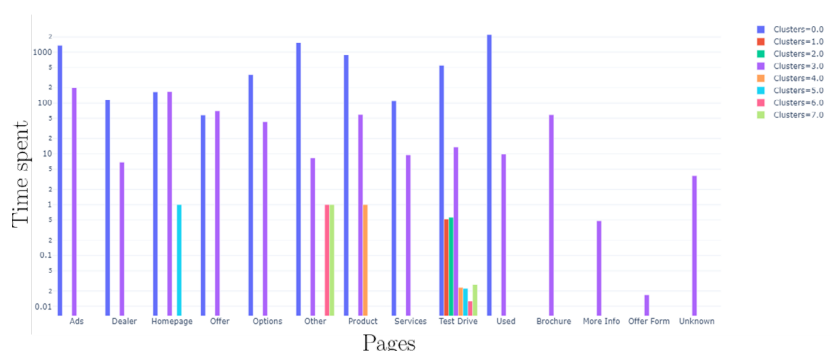


Figure 6: Clusters obtained with SAE+HDBSCAN clustering. 8 clusters with less pages diversity. Overall silhouettes coefficient = 0.71.

Authors



Rémi Devaux is a doctoral researcher at MINES ParisTech (PSL Université) and Ekimetrics, a data science company empowering marketing decisions. His thesis focuses on the targeting of consumers. Rémi masters a wide range of analytical tools on consumers behavior, as well as econometric and statistical methods to model advertising effectiveness. He also writes wider-audience articles on the marketing and media ecosystems.

remi.devaux@ekimetrics.com



Matt Andrew is a Partner at Ekimetrics and the Managing Director of the London office. Studied Natural Sciences at Cambridge University and began his career in FMCG marketing at Colgate-Palmolive, working in the UK and Europe to build brands effectively. He then worked with Clive Humby and Edwina Dunn to build out the client solutions at Starcount, before joining Ekimetrics in 2016. Now he focuses on marketing effectiveness and bringing customers to the center of brand strategies, engaging clients from multinational brands to understand the impact of their marketing efforts and how to improve them.

matt.andrew@ekimetrics.com

Correction of Sampling Bias to Produce Unbiased Analytics

Shreya Jain
Zeotap, India

Swapnasarit Sahu
Zeotap, India

Classifications, Key Words:

- Skew Correction
 - Biased dataset
 - Unbiased Analytics
 - Data Insights
-

Abstract

Due to limitations in data collection strategies, the gathered information is often incomplete or missing. If this data is aggregated from multiple sources, the gap in the depiction of the correct information is further enhanced. This eventually leads to data quality issues like degradation of model performance and analytics built on top of it. The inherent skewness present in aggregated data results in biased analytics for insights. Aiming to solve the problem of producing unbiased analytics, we first compensate for the lack of information present in the dataset. This is achieved by aligning the skewed distribution to a true distribution, gathered from external sources. In this effort, a set of generic equations is derived to address the issue of a skewed dataset. This is followed by a walkthrough of an example, which aims to correct gender bias present in the dataset. The calculations are elaborated for the specific use-case of gender bias based on some assumptions and criteria for certain selections. The paper concludes by highlighting the analogy of our approach with Heckman correction.

1. Introduction

The problem of the existence of incomplete information is prevalent across different verticals. So, it becomes imperative to correct the bias present in such data. Since this bias is introduced in our data through procurement or sampling from external sources, the problem is called sampling bias. In other words, sampling bias is a state when the sample collected from a population is not an accurate representation of it. This happens when a sampling algorithm favors the selection of certain members of the population or due to collection constraints. The resultant is the creation of a non-random sample, i.e. some members of the population are inadequately represented due to under-coverage or over-coverage.

An example would be a country's population, such as India, that roughly consists of 60% males and 40% females, but our sample corresponds to a distribution of 90% males and 10% females (see **Figure 1**).

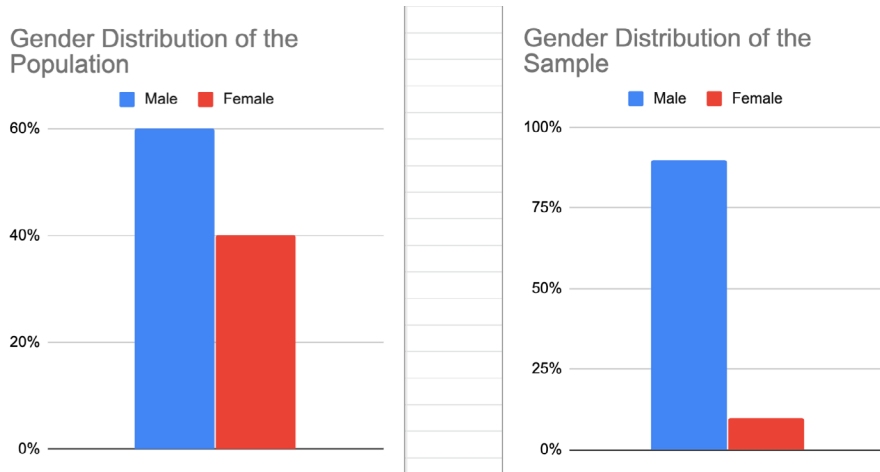


Figure 1: The sample here is not a true representation of the original population. Hence, this sample is biased and needs to be corrected.

Contribution:

The aim of this research is to derive generic mathematical equations to address the problem of lack of information with

an objective to produce unbiased analytics. This is achieved in conjunction with maintaining the variation among different sets of queries on which the analytics is produced, therefore, preserving the data property of each set. The obtained equations can be applied to various use-cases that fundamentally have the issue of biased datasets. Here, the problem of the presence of gender bias in the data is resolved by extending the results of the equations to a non-linear space.

2. Related Work

A few techniques solve the problem of selection bias correction, by using a machine learning approach that consists of reweighting the cost of an error on each training point of a biased sample to more closely reflect the unbiased distribution. This relies on weights derived by various estimation techniques based on finite samples. The effect of an error in that estimation on the accuracy of the hypothesis returned by the learning algorithm for two estimation techniques: a cluster-based estimation technique and kernel mean matching is analyzed in Cortes et al. (2008). In Huang et al. (2006), a scenario where training and test data are drawn from different distributions is considered, commonly referred to as *sample selection bias*. Most algorithms for this setting try to first recover sampling distributions and then make appropriate corrections based on the distribution estimate. Here, a nonparametric method is presented which directly produces resampling weights

without distribution estimation. The method works by matching distributions between training and testing sets in feature space.

Another category of solutions converges to making use of Heckman correction to solve for selection bias. The Heckman correction is a statistical technique to correct bias from non-randomly selected samples or otherwise incidentally truncated dependent variables, a pervasive issue in quantitative social sciences when using observational data. Conceptually, this is achieved by explicitly modeling the individual sampling probability of each observation (the so-called ‘selection equation’) together with the conditional expectation of the dependent variable (the so-called ‘outcome equation’). A specific use-case of selection bias in the criminology department is resolved in the paper Bushway et al. (2007).

3. General method for correcting bias

In order to correct the skewness present in an attribute for a biased sample, a true representation of that attribute is needed from an external source. The skewed attribute is then aligned-to/corrected-against this true distribution of the population.

All the frequency distribution curves in **Figure 2** are plotted against another attribute of choice, which helps in the alignment of our skewed distribution to its true distribution. The rectification/correction of the skewed attribute (the y-axis) is done piece by piece at the category/group level of another attribute, displayed as the x-axis.

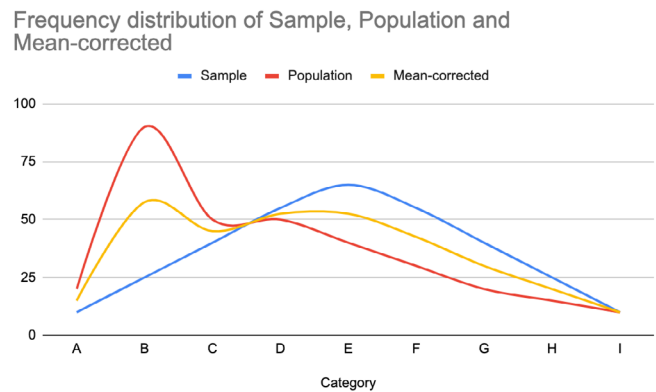


Figure 2: This diagram is for representation purposes only. Y-axis: represents the attribute to be skew corrected. X-axis: represents the attribute that is used to correct the biased attribute. Blue curve: represents the biased attribute value. Red curve: represents the true attribute value. Yellow curve: corrected attribute values, derived from the red and the blue curve.

3.1 Math behind the hypothesis

The x-axis in **Figure 2** represents the attribute along which the sample distribution needs to be corrected. This attribute first needs to be broken down into multiple strata or groups, which serve as the units on which the skewness of the attribute of the frequency distribution is corrected. Here, each stratum is correcting the bias present in the skewed-attribute in the sample progressively. This piecewise correction compensates for the lack of information present in the sample using an adjust factor:

$$adjustFactor_{Stratum} = \frac{P(population)_{Stratum}}{P(sample)_{Stratum}} \quad (1)$$

Where,
 $P(population)_{Stratum}$ = A function of the attribute to be corrected in the population

$P(sample)_{Stratum}$ = A function of the attribute to be corrected in the sample

The *adjustFactor* here is re-weighting the attribute values based on the proportion of the attribute present in the sample. If $adjustFactor > 1$, i.e.,

$P(population)_{Stratum} > P(sample)_{Stratum}$, the sample has a lower proportion of the attribute present in that particular stratum and needs to be over-weighted. On the other hand, if the sample stratum has an over-representation of an attribute value in the sample, it needs to be under-weighted with $adjustFactor < 1$.

The lack of information in the sample is corrected at a stratum level. It will be summed over all strata to arrive at the final corrected values in the sample.

3.2 Using the adjust factor at a segment-level

For a regular case, to estimate the attribute values for a subset of our sample called a *segment*, the following mean-corrected attribute equation is used at a stratum level:

$$mean_{corrected} = \sum_{i=1}^n \left(P(segmentStratum)_i \times \frac{stratumSize_i}{N} \right) \quad (2)$$

where,

$P(\text{segmentStratum})$ = A function of the attribute to be corrected at the stratum level of the segment

stratumSize_i = size of the segment stratum

N = total population size

n = number of strata

Taking the adjust factor into account to remove the bias in the attribute, the updated equation in this scenario becomes:

$$\text{mean}_{\text{corrected}} = \sum_{i=1}^n \left(P(\text{segmentStratum})_i \times \frac{\text{stratumSize}_i}{N} \times \text{adjustFactor}_i \right) \quad (3)$$

The mean-corrected attribute values maintain the *data property* of the segment. It does so by taking the attribute values at a segment stratum-level and ensuring the uniqueness of

the segment. The calculation also takes into account the adjust factor, which is responsible for the *lack of information* compensation.

3.3 Criteria behind strata selection

The selection of the attribute that is used to correct the biased attribute follows certain assumptions. This is the same attribute on which strata are defined. The objective of these criteria is to achieve a discrete distribution that does not change over time. A discrete distribution would ensure the correctness of alignment from the biased distribution to the true distribution at a stratum level.

The following points summarize the assumptions:

(A) The frequency distribution against the strata should be stable.

(B) The attribute should be categorical for the formation of a discrete frequency distribution in nature so that it can be directly used as a stratum definition. If not, the attribute values should be carefully grouped into the formation of strata.

(C) The criteria for grouping or defining a stratum can be based on the density or homogeneity of the data.

4. Skew correcting the attribute in a nominal space

The attribute to be skew-corrected in the continuous space uses **Equation (1)**. The function P in this equation accounts for the overall representation of the attribute in a segment as one floating-point value. It is straightforward to calculate the function P for continuous variables as one numerical value

is enough to comprehend the constitution of a segment. On the other hand, a nominal feature is represented by different categorical values. To account for the representation of all these cardinal values though one number, the function P is modified to represent this information in the form of a ratio.

5. Walkthrough of an example

Suppose the attribute to skew-correct is Gender here, which is a categorical variable. The generic **Equation (1)** derived above will now be modified

to serve our correction in skew for Gender attribute.

5.1 Using the generic equation to skew correct a categorical attribute

Using **Equation (1)**, and defining the function, P as the ratio of females over males for a subset of users (a segment):

$$P_{stratum}^{set} = \frac{femaleCount_{Stratum}^{Segment}}{maleCount_{Stratum}^{Segment}} \quad (4)$$

$$= \left(\frac{f}{m}\right)_{stratum}^{segment} \quad (5)$$

So, **Equation (1)** for the correction of lack in information coupled with the definition of P from **Equation (4)** is modified to:

$$adjustFactor_{Stratum} = \frac{\left(\frac{f}{m}\right)_{stratum}^{population}}{\left(\frac{f}{m}\right)_{stratum}^{sample}} \quad (6)$$

This adjust factor is incorporated to calculate the corrected attribute values for a segment across all strata:

$$mean_{corrected} = \sum_{i=1}^n \left(\left(\frac{f}{m}\right)_i^{segment} \times \frac{stratumSize_i}{N} \times adjustFactor_i \right) \quad (7)$$

$$= \sum_{i=1}^n \left(\left(\frac{f}{m}\right)_i^{segment} \times \frac{stratumSize_i}{N} \times \frac{\left(\frac{f}{m}\right)_i^{population}}{\left(\frac{f}{m}\right)_i^{sample}} \right) \quad (8)$$

$$mean_{corrected} = \left(\frac{f}{m}\right)_{stratum}^{segment} \times \frac{\left(\frac{f}{m}\right)_{stratum}^{population}}{\left(\frac{f}{m}\right)_{stratum}^{sample}} \quad (9)$$

5.2 Interplay of lack of information and data-property

Equation (9) maintains the data property of the segment by taking in the gender distribution of the segment. But at the same time, it also compensates for the lack of information present in the sample.

Lack of information in the sample is rectified using:

$$\frac{\left(\frac{f}{m}\right)_{stratum}^{population}}{\left(\frac{f}{m}\right)_{stratum}^{sample}} \quad (10)$$

Whereas data property of the segment is maintained using:

$$\left(\frac{f}{m}\right)_{stratum}^{segment}$$

So, mean-corrected value is nothing but:

$$meanCorrected = lackOfInformation \times dataProperty = l \times d \quad (12)$$

While compensating for the factor, *lackOfInformation*, the biased distribution is pulled towards the distribution of the Census data. It helps fill in the missing information but, in this effort, it fails to account for the

uniqueness in each subset of data. The other factor, *dataProperty*, mitigates this issue by preserving the data property of different subsets.

Hence, the combination of these two factors maintains the balance between removing skewness from the aggregated data as a whole and ensuring the desired skewness among subsets.

5.3 Selection of stratum attribute

For the use-case of Gender skew-correction, the attribute chosen is corrected against 'Location'. The following are the reasons for this selection:

- (A) Gender distribution is consistent with the location information, making the frequency distribution of gender against location stable.
- (B) Location data is various categorical levels, such as cities, states, etc.
- (C) Census data is readily available.

5.4 Introduction of non-linearity

The gender attribute is now skew-corrected at a stratum level for a segment using **Equation (9)**. Though the correction is made, the scale of the results is not calibrated. To find the optimal scale of the correction, we use a supervised approach. We learn a function to produce the target values from the mean-corrected values.

Segments	(f/m) Base	(f/m)Target
A	1.5	2.4
B	1.4	2.1
C	1	1
D	0.7	0.3
E	0.9	0.5

Table 1: Depiction of f/m ratios of base Mean-corrected values and target values for different segments

As it is evident from **Table 1** observations, a non-linear function is needed to approximate the target values from the mean-corrected values.

5.5 Selection of a nonlinear function

After experimentation with numerous non-linear functions, like sigmoid, exponential, etc, the Tanh function was chosen due to the following reasons:

- (A) Flexibility to cap the results as the Tanh is a closed function.
- (B) The function fitted the best and gave the least MSE among other functions.

Therefore, **Equation (9)** is updated to:

$$mean_{corrected} = \tanh \left(\left(\frac{f}{m} \right)_{stratum}^{segment} \times \frac{\left(\frac{f}{m} \right)_{stratum}^{population}}{\left(\frac{f}{m} \right)_{stratum}^{sample}} \right) \quad (13)$$

6. Results

For the specific use-case of correction of gender bias in the dataset, the plugged-in equations in Section 5 are used to attain the skew-corrected results. Three sets of results are compared here on an age query for 3 segments.

These are 1. Uncorrected results 2. Skew corrected results 3. True results. True results are obtained from online measurement companies such as Nielsen, comScore, etc.

Age Brackets	Uncorrected	Skew-corrected	True
18-24	0.278883	2.430151	2.5222555
25-34	0.2118464	1.6276676	1.726262
35-44	1.163636	2.2281639	2.42323232
45-54	0.783737	3.221235	3.4525252
55+	0.67764764	1.2311468	1.190229

Table 2: Depiction of f/m ratios of Uncorrected values, Skew-corrected values, and Target/True values for Segment 1

Age Brackets	Uncorrected	Skew-corrected	True
18-24	0.3644848	2.433598527	2.299973
25-34	0.271717	1.320830718	1.26474646
35-44	0.7363357	1.146921372	1.0864646
45-54	0.6763763	1.286387	1.178733
55+	0.0726262	1.19116379	0.9876454

Table 3: Depiction of f/m ratios of Uncorrected values, Skew-corrected values, and Target/True values for Segment 2

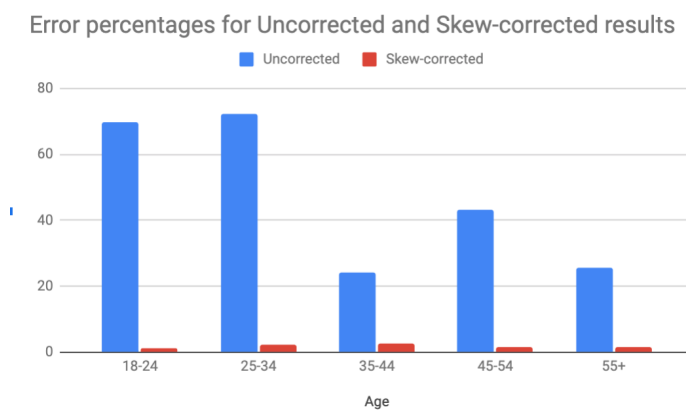


Figure 3: Pictorial representation of error percentages for gender distribution of Uncorrected values and Skew-corrected values against their true values for Segment 1

As evident from **Table 2** and **Table 3**'s results, the Uncorrected gender distribution is highly skewed towards males. The comparative results can be better depicted in the pictorial forms as error percentages as shown on **Figure 3** and

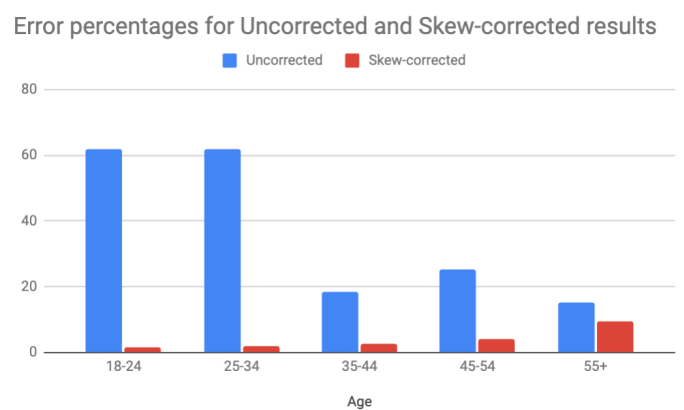


Figure 4: Pictorial representation of error percentages for gender distribution of Uncorrected values and Skew-corrected values against their true values for Segment 2. Lack of data points in the 55+ category resulted in a high error in both uncorrected and skew-corrected results.

Figure 4. The error percentages are calculated as the percentage difference between the gender percentages of the (Uncorrected and the true values) and (Skew-corrected and the true values).

Skew correction here corrects this bias to compensate for the lack of information against females in the dataset. Another point to be noted here is that the skew corrected results are different for each segment. This ensures

that the data property of each segment is still maintained while correcting the male bias in the overall dataset. Also, the skew-corrected gender distributions are much more accurate in depicting the true gender distributions.

7. Comparison with Previous Studies

The first approach, Cortes et al. (2008), of 'Sample Selection Bias Correction Theory' relies on weights derived by various estimation techniques based on finite samples. This is a learning algorithm that assumes the availability of a lot of samples in order to minimize the cost function. On the other hand, our algorithm works well with a much lesser number of samples, as samples are only used for re-scaling purposes in our approach.

In the second approach, Huang et al. (2006), of 'Correcting Sample Selection Bias by Unlabeled Data', a non-parametric method is proposed that solves the selection bias problem for different distributions for training and test data sets. The algorithm works by matching distributions between training and testing sets in feature space. This method necessitates the availability of a huge amount of training and test data sets. Whereas, our approach works well with a significantly lesser scale of samples.

The third approach, Bushway et al. (2007), of 'Is the Magic Still There? The Use of the Heckman Two-Step Correction for Selection Bias in Criminology' relies on the classic two-step Heckman correction technique. The bias is removed by modeling the missing part of the original distribution through a set of features. The

curation of these features is not straightforward and intuitive. Our approach doesn't require a hypothesis for feature generation as census data acts as a proxy for the same.

Methods related to propensity scoring algorithms, as in Dehejia and Wahba (2002), require covariates to be defined, which can explain the dependent variable (gender in this case) with good accuracy. The accuracy of the model is highly reliant on building a robust model on these covariates for all real-world scenarios. However, in cases where covariates can predict the dependent variable with reasonable accuracy, this method is superior as it doesn't require a constant update of the census truth values.

Another class of methods known as RIM weighting, which is a special form of a target weighting, provides a fair solution of how the overall dataset statistics should look like. But it can't be used as a slice and dice approach, where each user has been assigned a definitive value or a propensity score. The paper of Greenacre (2006) highlights the importance of selection bias in internet surveys and uses rim weighting as one of the weighting adjustment methods.

8. Analogy to Heckman Correction

The Heckman correction, a two-step statistical approach, offers a means of correcting the selection bias for non-randomly selected samples.

We also use a two-stage estimation process analogous to the Heckman Correction.

1st Estimation:

$$TrueAnalytics = DataProperty(BiasedSample \cap Dataset) \times LackOfInformation \quad (14)$$

Now, *Lack of Information* is estimated using:

2nd Estimation:

$$LackOfInformation = Function(DatasetSkewness \times (\frac{1}{CensusSkewness})) \quad (15)$$

The 2nd Estimation in our approach also models the missing distribution, like the Heckman. Heckman-correction relies on the

curation of features for the same and we learn a function of the census data. The 1st Estimation in our approach, finally, uses the learned information on the missing data like Heckman-correction.

The paper, by Koné et al. (2019), highlights the extent to which Heckman-type selection models can create unbiased estimates in low-income settings where key outcomes such as biomarkers or clinical assessments are often missing for a substantial proportion of the study population.

In the online retail ecosystem, the non-random assignment of marketing interventions like coupons or retention campaigns can lead to over-

or underestimating the value of the intervention. This can cause future campaigns to be directed at the wrong customers. The paper, by Walton and Sanford (2014), demonstrates how selection in the data can be modeled and shows how to apply Heckman's two-step procedure in an empirical example.

The problem of determining aggregate road crash costs is solved in the work of Giles (2001). Bias was introduced in the calculations as the collected data was a non-random sample of the true population of road crashes. Using Heckman's methodology for correcting for this selectivity bias, crash data for a region was reconciled with total (notified and not notified) crash data in the estimation of the property damage costs of road crashes.

Conclusion

With the derived set of equations, we were able to compensate for the lack of information in the gender attribute present in the dataset. This is achieved in conjugation with maintaining the variation of gender distribution among different segments, thus preserving the data property of each segment. These results were further improved with the introduction of a non-linear function to approximate these derived values close to their true values. The selection of the attribute along which the gender bias is corrected is based on the criteria of homogeneity of the attribute within strata. The choice of the non-linear function that approximates the skew-corrected results with their true results is based on supervised learning errors.

This algorithm can be directly plugged-in to correct bias in continuous features and is not only restricted to categorical variables. The key requirement is to find a supporting true distribution against an alignment attribute. This is followed by defining the adjustFactor which compensates for the lack of information and calibrates the final results using a non-linear function.

References

1. Cortes, C., Mohri, M., Riley, M., & Rostamizadeh, A. (2008). Sample Selection Bias Correction Theory. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5254 LNAI, 38–53. [Source](#)
2. Huang, J., Smola, A. J., Gretton, A., Borgwardt, K. M., & Schölkopf, B. (2006). Correcting Sample Selection Bias by Unlabeled Data. *Advances in Neural Information Processing Systems*, 19, 601–608. Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006.
3. Bushway, S., Johnson, B. D., & Slocum, L. A. (2007). Is the Magic Still There? The Use of the Heckman Two-Step Correction for Selection Bias in Criminology. *Journal of Quantitative Criminology*, 23(2), 151–178. [Source](#)
4. Dehejia, R. H., & Wahba, S. (2002). Propensity Score Matching Methods for Non-Experimental Causal Studies. *SSRN Electronic Journal*. [Source](#)
5. Greenacre, Z. A. (2016). The Importance of Selection Bias in Internet Surveys. *Open Journal of Statistics*, 06(03), 397–404. [Source](#)

6. Koné, S., Bonfoh, B., Dao, D., Koné, I., & Fink, G. (2019). Heckman-type selection models to obtain unbiased estimates with missing measures outcome: theoretical considerations and an application to missing birth weight data. BMC Medical Research Methodology, 19(1). [Source](#)
7. Walton, G.E., & Sanford, K. (2014). Did My Coupon Campaign Accomplish Anything? An Application of Selection Models to Retailing.
8. Giles, M. J. (2001). Heckman's Methodology for Correcting Selectivity Bias: An Application to Road Crash Costs. SSRN Electronic Journal. [Source](#)

Authors



Shreya Jain has 5 years of experience in the field of Data Science. She has developed a range of Machine Learning solutions in the Advertisement Tech sector in the last 3 years at Zeotap. Before this, during her stint in Samsung R&D, she's contributed to various Computer Vision projects and cloud applications. A huge round of funding was secured from the Samsung HQ for one of the demo projects.

She has been a part of several Education Tech platforms, like Springboard, UpGrad, etc, and guides Machine Learning professionals as a mentor/coach.

shreya.jain@zeotap.com



Swapnasarit Sahu has 16 years of experience in Data Science and Analytics. He spent most of his carrier in Advertising and Digital Marketing. Currently, he is the Chief Analytics Officer of Zeotap. In past, he headed Data Science and Analytics Divisions of Airpush Inc. and one of the core algorithmic developers in IBM Watson. He has numerous patents in Machine Learning and Natural Language Processing. He is also an active mentor in the Indian Data Science Community.

swapnasarit.sahu@zeotap.com

Effectiveness of Advertising: A Study on the Influence of Creative Strength on the Return of Media Use

Mark Vroegrijk
DVJ Insights

**Classifications,
Key Words:**

- Advertising
 - Creative execution
 - Recall
 - Return-on-investment
-

Abstract

In recent times, it has become increasingly important for marketers to justify what they spend on advertising. Therefore, maximizing return-on-investment in GRP's and/or impressions has become vital as well. One way of doing so is improving one's advertising in terms of creative strength. However, while several academic studies have already proven this relationship to exist, knowledge on the expected magnitude of these effects has been relatively lacking. This paper therefore presents two empirical studies aimed to provide more insight into this issue. In the first study, centered around online video advertising, we conduct a field experiment that shows that high creative quality in (skippable) ads helps them hold viewer attention for a few seconds longer – which is all that is needed in order to reap considerable gains in terms of brand and message recall. In the second study, we further build upon these findings and use tracking data to demonstrate that the in-market effectiveness of each GRP in terms of fostering ad recognition is, at least in part, a function of creative strength. For both studies, implications in terms of the most critical dimensions within creative quality and expected effect sizes are discussed.

1. Introduction

In recent months, the importance of continuing to invest in media during a crisis has frequently been brought up. At the same time, many marketers are also forced to give extra justification for every advertising expenditure. It is now more important than ever to get the most out of every GRP or impression purchased, and leave a lasting impression in the minds of consumers. Various academic studies underline that there is a clear positive influence of the use of good creatives on these efforts (Smit et al., 2006; Stewart & Furse, 2000; Lehnert et al., 2013). At the same time, relatively little is known about the magnitude of these relations – even though it is, especially now, important for marketers to understand to what extent the effectiveness of advertising increases if they optimise their creatives. Only then can one assess whether these improvements are actually worth the investment.

The above prompted us to conduct two empirical studies on the effectiveness of advertising – and the role of the quality of creatives. In the first study, we test a large number of online video advertisements to explore the extent to which an advertisement’s impact on consumer memory is driven by the amount of attention it receives from consumers and, in turn, how

this depends on the ad’s creative strength. In the second study, we go one step further and directly demonstrate how an improvement in ad quality can lead to savings in media deployment costs – by lowering the amount of GRPs that are necessary to obtain the same (long-term) impact on consumer memory.

2. Study #1

The first study was conducted among 10,000 respondents, and involved testing 100+ online video advertisements in three countries (Germany, the Netherlands, and the United Kingdom). At the beginning of this study, we let every respondent browse through a few websites. On some of these websites, a content video was shown, which was always preceded by a randomly selected advertisement video. Half of the sample was given the option to skip the ad after 5 seconds and onwards. If the respondent did, we registered (invisible to her or him) the moment where that happened. To provide a reference standard, the other half of the sample did not have the option to skip at all,

and as such were forced to watch each ad in its entirety before the content video started playing.

Then, sometime after browsing all the websites, we asked every respondent: **1)** which brands they remembered seeing an advertisement from, **2)** for which advertisements they understood the message, and **3)** which advertisements they recognised. Because we tested the same ads in both skippable- and non-skippable formats, we can analyse to which extent a longer viewing time (i.e. viewers taking longer to skip the ad) leads to uplifts in consumer memory – and how this compares to consumers who have no choice other than to watch the ad in its entirety.

2.1 How does viewing time drive what is remembered?

The results of our first study show that if a consumer is offered the option to skip an ad (albeit after a few seconds), this option is frequently used – which is consistent with the notion that consumers perceive online video advertisements as relatively intrusive (Goodrich et al., 2015). Especially in the first few seconds in which an advertisement can be skipped, there is a significant loss of attention – after only three seconds (which, in our study, implies the 8-second mark) more than one third of viewers already decided to skip it. Although the loss in viewers decreases in magnitude in the seconds thereafter, an average online video ad is viewed by less than half of the consumers in its entirety.

Given the large loss of attention for skippable advertisements, it therefore doesn’t come as a surprise that, on average, consumers are better able to remember the ad afterwards when they were forced to watch it in its entirety.

Figure 1 shows the obtained scores for unaided and aided brand recall, message recall and ad recall (averaged across ads) for both the skippable and non-skippable ("forced") conditions, with statistically significant differences (at a 95% confidence level) of 8 and 10 percentage points for unaided and aided brand recall, respectively, and 6 percentage points for message recall.

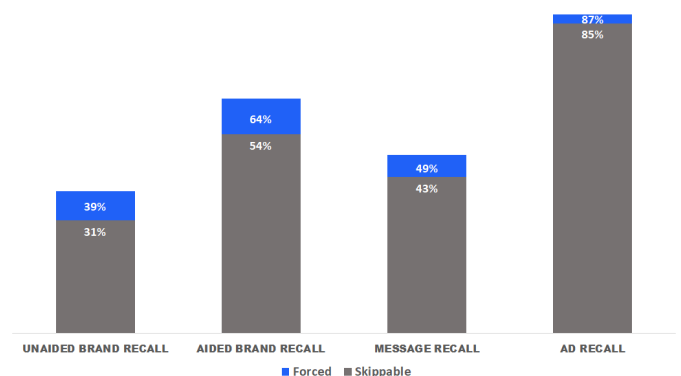


Figure 1: Average recall scores for skippable versus non-skippable advertisements

Still, when we compare this difference to the loss of attention that occurs for skippable advertisements, the ratio between effect sizes is immediately noticeable. Although more than half of the viewers don't watch skippable advertisements in their entirety (and therefore, under the corresponding "TrueView" format, doesn't constitute as costs!), the memory effect drop compared to non-skippable advertisements is of a much smaller order – 10 percentage points or less.

As such, watching an advertisement in its entirety does not seem to be a prerequisite for realizing (sufficient) impact. Instead, Figure 2 shows that it is primarily important to fascinate the viewer for a few more seconds after he or she first has the possibility to skip the advertisement (in this case: after 5 seconds). After all, each of the next few seconds where the advertisement is watched longer, ensures a significant catch-up on brand and message recall compared to non-skippable ads. At the same time, we see that from a viewing time of 8 seconds onwards (where the lag in brand and message recall is only 6 percentage points) these improvements

start to flatten. Therefore, the critical turning point seems to lie in being able to hold the viewer for 3 seconds longer after being offered the opportunity to skip.

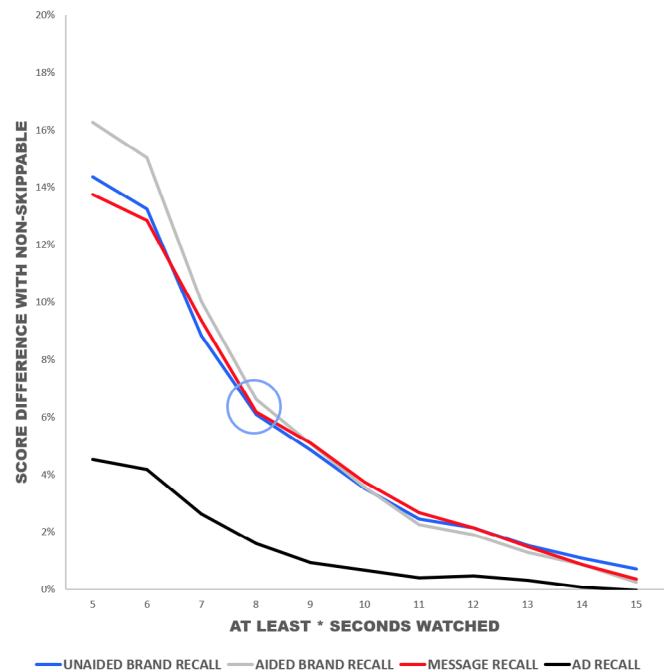


Figure 2: Difference in recall-scores compared with non-skippable advertisements at different viewing times

2.2 The impact of creative execution on viewing behaviour

But what determines how long an online video ad is actually viewed? We can distinguish external factors (which marketers cannot directly influence, but can take into account in media planning, such as the demographic profile of the viewer and the device they use to watch the ad) and internal factors, among which is the creative execution of the advertisement: the focus of this study.

After all, it seems plausible that a video with a strong creative execution will be better capable of holding the viewer's (limited) attention.

Because we have not only registered the viewing behaviour of respondents in our study, but have also asked afterwards how they would evaluate the shown creatives on different aspects, we can map out the relation between the two. We do so by conducting a respondent-level binomial logistic regression analysis, using the

respondent's decision to (not) watch an ad either for a minimum of 8 seconds or in its entirety as the dependent variable, and the respondent's ratings of the ad on a total of 9 dimensions as the independent variables.

Figure 3 indicates which variables were statistically significant (at a 95% confidence level), along with their relative effect size. Interestingly, we find that the stopping power of an advertisement (whether the ad can hold the viewer's attention for at least a few seconds), and the entire watch-until-the-end ration are not necessarily driven by the same aspects of creative strength. Still, we see that the degree to which an advertisement is perceived as distinctive, exerts a consistent positive influence on watching behaviour – both in the shorter and longer term. At the same time, we do find differences in other aspects. Where the advertisement should be seen as fun and enjoyable (but at the same time

still be credible) to not let people skip during the first few seconds, the advertisement should transfer an even stronger emotional response – a feeling of excitement – to be watched in its entirety, and also provide the viewer with new information.

As such, while marketers should take into account that (on average) more than half of viewers of skippable ads drop out prematurely, this loss of attention can be remedied by improving an ad’s creative quality. Specifically, a distinctive advertisement that is fun and/or enjoyable but also credible at the same time helps in retaining the viewer’s attention for just a few more seconds after having the opportunity to skip – which is identified as the critical tipping point that needs to be reached in order to generate sufficient impact on consumer memory.

3. Study #2

Now that we learned that ads with stronger creative execution are better able to hold the viewer’s attention and, as such, are remembered better afterwards, our second study is aimed to uncover the practical implications of this relation in terms of return-of-investment. In other words, can the same effect on (longer-term) consumer memory be achieved with a lower amount of GRPs (and thus lead to cost savings!) if an ad performs better in a creative sense?

To provide an answer to this question, we used a database of 74 Dutch TV commercials, broadcasted between 2016 - 2020. This database includes commercials of a significant number of brands (18) and product types (4 categories: electronics, fast moving consumer goods, financial services, and catering) – which enhances the generalisability of the results to other markets. Because the database includes for each of these commercials the extent to which they have been recognised from week to week (based on survey data), we have a measure of the extent to which each commercial has acquired a place in the longer-term memory of consumers – a first necessary step towards influencing knowledge, attitudes and (ultimately)

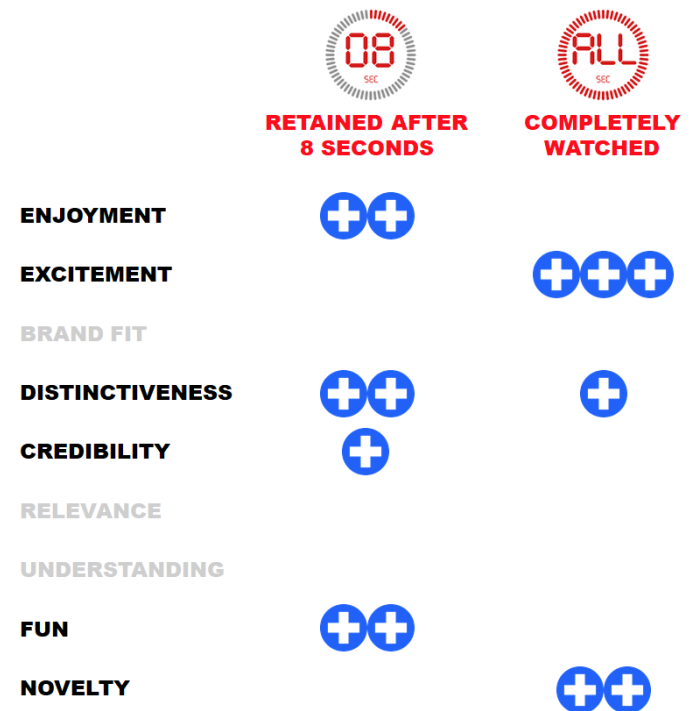


Figure 3: Main creative drivers of willingness to partially (≥ 8 sec.) and fully watch online video ads

purchasing behaviour.

As mentioned earlier, the main issue we sought to answer in this study is to understand the influence of creative strength on the effectiveness of media deployment.

To properly map this relationship (through an adapted ordinary least squares regression model with week-level ad recognition scores as the dependent variable), we approached our analysis as follows:

1. The starting point of the model is, logically, relating the actual number of GRPs deployed (in each week) for a TV commercial to the recognition score for that commercial (in that week). The corresponding parameter then thus captures the effectiveness of media deployment.
2. However, our model also "weights" the weekly GRPs for each commercial based on the creative strength of that commercial, as measured through aggregated survey responses on several evaluation statements.

If the commercial scores above average in terms of evaluation, the number of GRPs will be weighed up – the underlying idea is that with a good creative every GRP used will be worth more. The opposite applies to a commercial that is valued below average. The extent to which GRPs are weighed up or down, and the relative influence of the (interactions between) different creative aspects within the weighting, are parameters that are freely estimated in the analysis.

To estimate the effects of GRP deployment and creative strength as accurately as possible, the model also checks for several other principles:

- a. The strength of the relationship between GRP deployment and recognition can vary between categories – due to differences in the degree of involvement that consumers feel with a category, their degree of attention for (advertising) stimuli within the category may also differ (Buchholz & Smith, 1991). Therefore, category dummies are incorporated into the model as possible moderators of the GRP–recognition relationship.
- b. The strength of the relation between GRP deployment and recognition may depend on the commercial length – various authors state that a longer-lasting commercial is better recognised under the same commitment

to GRPs, due to a greater availability of 'memory hooks' that consumers can use to store the commercial in their mind (Singh & Rothschild, 1983; Zinkhan et al., 1986). Therefore, commercial duration (in seconds) is included as a moderating variable as well.

- c. Because consumers are able to remember a commercial for a longer period of time (or at least part of it) – even if it was shown less often or even no longer at all (Aravindakshan & Naik, 2010) – the model incorporates a 1-week lag of the dependent variable (recognition) as an additional predictor. Such an approach allows for the recognition of a commercial in a certain week to not only be influenced by the current GRP deployment, but also by the degree of recognition in previous weeks.

Figure 4 then displays a stylistic representation of the regression model that was estimated.

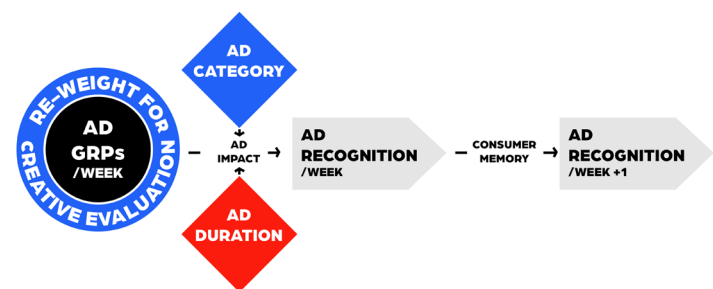


Figure 4: Schematic representation of the model.

3.1 Effectiveness in function of category & commercial length

Before covering the implications for the extent to which creative strength influences the effectiveness of media use, we will first shortly discuss the effects of the control variables included in the model. We do this by calculating (based on the analysis results) the recognition score to which the broadcasting of an average commercial, during 4 weeks with 100 GRPs per week, would lead across different combinations of categories and commercial lengths.

Table 1 shows the results of these calculations (or simulations): the longer a commercial lasts, the stronger the degree of recognition is that will follow from the use of a given number of GRPs.

These differences are particularly sizeable between commercials of short (e.g. 10 seconds) to medium (e.g. 30 seconds) duration, while at longer lengths these differences begin to flatten. For marketers, this implies that although extending a commercial can have more effect (in terms of recognition), it does not seem worthwhile to let the commercial be longer than 30 seconds. This is not too surprising, given that our first study on skipping behaviour already showed that very few ads will be able to capture the attention of their entire audience for such a prolonged time.

We also see strong differences in returns between

categories – for a high-involvement category (e.g. electronics), a total of 400 GRPs leads to a more than 60% higher recognition score than for a low-involvement category (e.g. FMCG). This can also be explained by differences in

clutter – more commercials are broadcasted in the FMCG industry, making it even more difficult to stand out. As such, when a lot of competitors advertise simultaneously, the return from each GRP decreases.

		CATEGORY				
		ELECTRONICS	FAST MOVING CONSUMER GOODS	FINANCIAL SERVICES	CATERING	AVERAGE RECOGNITION
DURATION	10 SECONDS	33%	20%	27%	27%	27%
	20 SECONDS	38%	24%	32%	32%	32%
	30 SECONDS	40%	25%	34%	34%	33%
	40 SECONDS	41%	25%	34%	34%	34%
	50 SECONDS	41%	26%	35%	35%	34%
	60 SECONDS	42%	26%	36%	35%	35%
	AVERAGE RECOGNITION	39%	24%	33%	33%	32%

Table 1: Differences in simulated recognition scores between commercial lengths and categories (based on a 4-weekly campaign with 100 GRPs per week)

3.2 Advertising returns: Which creative dimensions are decisive?

Next, we move towards the role of creative strength in determining media effectiveness. We again analyse this by using simulations (similar to our approach for the control variables): our starting point being a TV commercial (of 30 seconds) that is broadcasted with 100 GRPs per week for 4 weeks. This commercial initially

scores "middle-of-the-road" on all five creative dimensions (50th percentile / "benchmark"), but then we manipulate the scores on one evaluation dimension (by varying them between a low (30th percentile / "bottom 30") and high (80th percentile / "top 20") level). Afterwards, we look at the effects of these manipulations (and

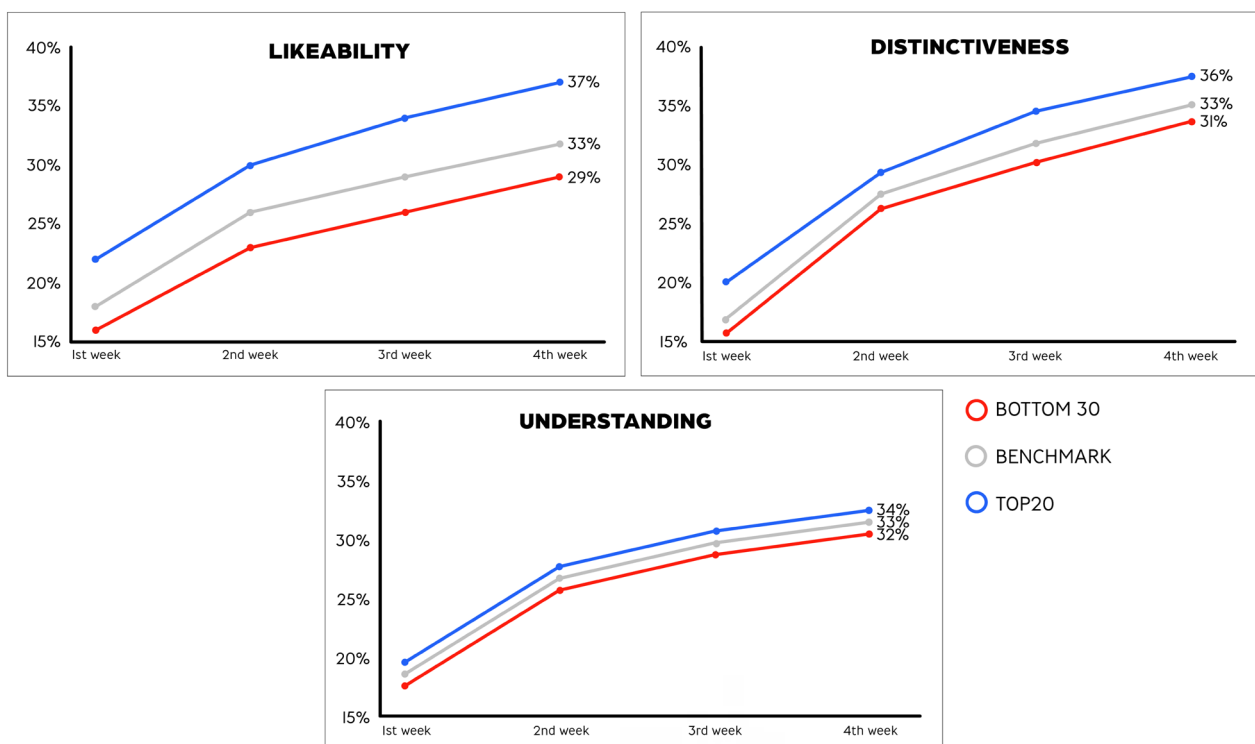


Figure 5: Differences in simulated recognition scores in function of creative strength (likeability, distinctiveness and understanding, based on a 4-weekly campaign with 100 GRPs per week)

the resulting creative re-weighting applied to the 400 GRPs) on the expected recognition scores, reporting the average effect across categories.

Figure 5 shows the results of these simulations:

1. The greatest effect is found for likeability: if a TV commercial scores at the top 20 level on this dimension, on average the same number of GRPs will achieve an 8 percentage points higher recognition score (37%) than if the TV commercial would have scored badly (bottom 30 level) in this area (29%). Not surprisingly (given the number of studies that found a link between the "liking" and effectiveness of ads (Stone et al., 2000; Ambler & Burne, 1999)), it appears that it is mainly important that a commercial is liked, to be well remembered by consumers.
2. The higher a TV commercial scores on distinctiveness or understanding, the better the commercial is recognised with the same number of GRPs. The uplift between the

bottom 30 and top 20 levels is smaller than for liking though: 5 percentage points (31% versus 36%) for distinctiveness, and 2 percentage points (32% versus 34%) for understanding.

3. At first glance, both relevance and brand fit do not seem to affect the degree to which the commercial is remembered. While the results do indeed underline this for relevance, a remark should be made for brand fit. Although this dimension does not directly determine the return that is obtained from each GRP, it does exert an interaction effect – with a good brand fit strengthening the relationship between likeability, distinctiveness and understanding on the one hand, and the GRP returns on the other. Table 2 illustrates this and shows that if a TV commercial is improved in terms of likeability, distinctiveness or understanding (from benchmark to top 20 level), the subsequent increase in recognition will be greater if the commercial also scores well on brand-fit.

UPLIFT IN RECOGNITION BY IMPROVEMENT OF BENCHMARK LEVEL TO TOP 20 LEVEL ON:				
		LIKEABILITY	DISTINCTIVENESS	UNDERSTANDING
	BOTTOM 30	Δ3.8%PT.	Δ1.9%PT.	Δ0.6%PT.
BRAND FIT	BENCHMARK	Δ4.7%PT.	Δ3.0%PT.	Δ1.0%PT.
	TOP 20	Δ5.9%PT.	Δ4.4%PT.	Δ1.5%PT.

Table 2: Uplift in recognition as a result of improvements on likeability, distinctiveness and understanding, at different levels of brand fit (based on a 4-weekly campaign with 100 GRPs per week)

3.3 A broader view on the contribution of creative strength

The interactions described above already show that the effects of improvements on specific creative dimensions can seldom be considered in isolation – they are interrelated, and improving a creative on one dimension often goes hand in hand with improvements on other dimensions.

Table 3 descriptively shows that almost all creative dimensions exhibit a positive and significant coherence. It is also interesting that this relation is strongest for the two dimensions that already have the strongest influence on GRP returns (likeability and distinctiveness).

	LIKEABILITY	RELEVANCE	BRAND FIT	DISTINCTIVENESS
RELEVANCE	,449			
BRAND FIT	,532	,628		
DISTINCTIVENESS	,798	,366	,372	
UNDERSTANDING	,380	,625	,684	-,045

Table 3: Pearson correlations between creative dimension scores based on commercial database (bold correlations: statistically significant under a 95% confidence level)]

Which is why, to give a complete picture of how the effectiveness of media use is determined by the strength of the used creatives, we once again conducted a simulation. In this case, we simultaneously manipulated the scores of the commercial on the four evaluation dimensions for which either a main and/or interaction effect was found (likeability, distinctiveness, understanding and brand fit).

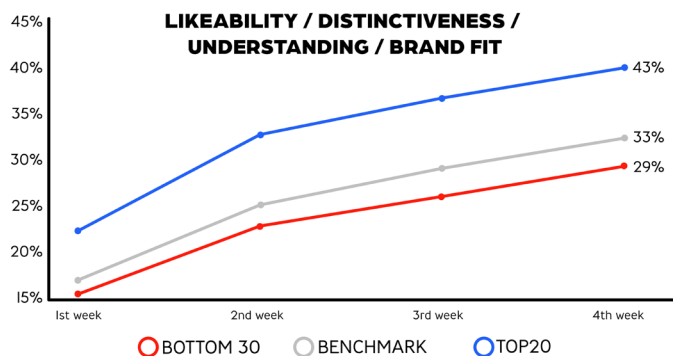


Figure 6: Differences in simulated recognition scores in function of creative strength (likeability + distinctiveness + understanding + brand fit, based on a 4-weekly campaign with 100 GRPs per week)

Figure 6 shows the results of this simulation, and shows that a TV commercial that scores at the top 20 level on likeability, distinctiveness, understanding and brand fit, achieves a 14 percentage points higher (and therefore 50% higher) recognition score (43%) with the same number of GRPs than a TV commercial that lags behind on all these aspects at bottom 30 level (29%).

Reversing this result implies that for a creatively

3.4 Creative optimisation: A valuable investment!

All in all, the second study again revealed that the extent to which the use of advertising leads to a lasting effect on the memory of consumers, depends on the strength of the creatives that were used. Firstly, similar to our first study on online video ads, we see a connection with distinctiveness (a striking and distinctive commercial attracts more attention during a commercial break, and is therefore better remembered). In addition, we again find

strong commercial, fewer GRPs need to be used than for a creatively weak commercial to still achieve the same degree of recognition. Figure 7 then shows this relation graphically (based on the simulation described above), calculating how many GRPs in total (based on a 4-week distribution) are needed to achieve different recognition scores. This happened for three scenarios: a TV commercial that scores poorly (bottom 30 level), average (benchmark level) or good (top 20 level) on likeability, distinctiveness, understanding and brand fit.

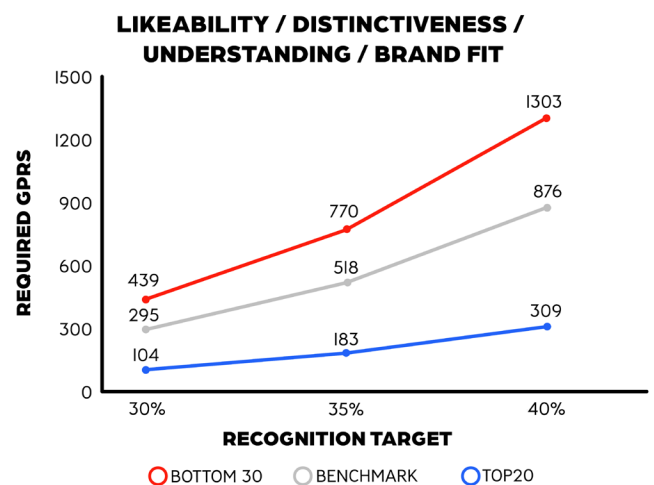


Figure 7: Differences in required GRP input in function of creative strength (likeability + distinctiveness + understanding + brand fit)

We can clearly see which savings can be achieved when a commercial is creatively strong – an improvement of a commercial from benchmark level to top 20 level on all creative dimensions, means that at least 60% fewer GRPs have to be spent to have the same recognition effect.

an (even stronger) link with the degree to which a commercial is enjoyable and fun (likeability). Other factors that play a role are related to the ease with which the consumer can place the commercial in his or her existing memory structures – understanding and, via moderating effects, brand fit. At the same time, a striking result is that our analyses showed that relevance has no influence on the degree of recognition

that results from the use of a commercial. However, it should be noted that this does not mean that the degree to which a commercial is relevant does not play any role in determining its effectiveness. Previous research has for example found that although relevance is not related to recognition, there are connections with longer-term effects – such as influencing consumers' (brand) attitudes (Smit et al., 2006).

Moreover, we can conclude that the role of creativity in determining the return on each

GRP is considerable. Various simulations showed that with a TV campaign (of average size in terms of media use), an almost 50% higher recognition score can be achieved if the commercial's creative strength scores at top 20 level (compared to bottom 30). Our analysis therefore underlines the importance of creative optimisation and the pursuit of the absolute top. After all, if, by improving overall creative strength from "average" to the Top 20 level, the same effect can be achieved with less than half of the original media expenditures, this is certainly a goal for marketers to pursue!

Conclusion

In sum, both studies revealed interesting insights with regards to increasing the effectiveness of advertising through creative execution. Within the context of online video advertising, our research shows that (the cost associated with) a full exposure to an ad is often not a prerequisite for its contents to be well-recalled afterwards. Just holding the viewer's attention for a few more seconds after he or she is able to skip the ad is generally enough – and whether an ad can actually succeed in doing so is significantly related to several dimensions of creative quality. Moreover, turning to TV advertising, we demonstrate that an ad's creative strength has a considerable role in determining the effectiveness of each GRP that is allocated to it. Specifically, one may achieve the same recognition rate for an ad that is among the best 20% rather than medium level in terms of creative quality – but with 60% less spent on GRP's. The prospect of such savings gives advertisers a clear incentive to optimize their creatives, which, following the findings across both of our studies, should mainly be centered around improvements in ad likeability and distinctiveness.

References

1. Smit, E.G., Van Meurs, L. & Neijens P.C. (2006). Effects of Advertising Likeability: A 10-Year Perspective. *Journal of Advertising Research*, 46(1), 73-83.
2. Stewart, D.W. & Furse, D.H. (2000). Analysis of the Impact of Executional Factors on Advertising Performance. *Journal of Advertising Research*, 40(6), 85-88.
3. Lehnert, K., Till, B.D. & Carlson, B.D. (2013). Advertising creativity and repetition. *International Journal of Advertising*, 32(2), 211-231.
4. Goodrich, K., Schiller, S.Z. & Galletta, D. (2015). Consumer Reactions to Intrusiveness Of Online-Video Advertisements. *Journal of Advertising Research*, 55(1), 37-50.
5. Buchholz, L.M. & Smith, R.E. (1991). The Role of Consumer Involvement in Determining Cognitive Response to Broadcast Advertising. *Journal of Advertising*, 20(1), 4-17.
6. Singh, S.N. & Rothschild, M.L. (1983). Recognition as a Measure of Learning from Television Commercials. *Journal of Marketing Research*, 20(3), 235-248.
7. Zinkhan, G.M., Locander, W.B. & Leigh, J.H. (1986). Dimensional Relationships of Aided Recall and Recognition. *Journal of Advertising*, 15(1), 38-46.
8. Aravindakshan, A. & Naik, P.A. (2010). How does awareness evolve when advertising stops? The role of memory. *Marketing Letters*, 22(3), 315-326.
9. Stone, G., Besser, D. & Lewis, L.E. (2000). Recall, Liking, and Creativity in TV Commercials: A New Approach. *Journal of Advertising Research*, 40(3), 7-18.
10. Ambler, T. & Burne, T. (1999). The Impact of Affect on Memory of Advertising. *Journal of Advertising Research*, 39(2), 25-34.

Author



Mark Vroegrijk is a Senior Methodologist at DVJ Insights, a market research company with offices in the Netherlands, the United Kingdom and Germany. Having 10 years of prior experience in academia (Ph.D. in Marketing Modeling at Tilburg University, the Netherlands, followed by a post-doc position at KU Leuven, Belgium), he now works as part of the internal “Center of Expertise” team, which is concerned with the introduction of new research techniques (along with the optimization of current ones) within the company’s workflow, and the coordination of cooperative projects with the academic world. Moreover, he frequently conducts meta-analyses on data that is collected by the company on a daily basis, in order to further advance the marketing knowledge that is shared with clients to help them grow.

mark.vroegrijk@dvjresearchgroup.com

Privatized Machine Learning for Marketing Analytics

Joao Natali

Neustar

Robert Stratton

Neustar

Classifications, Key Words:

- privacy preserving machine learning
 - differential privacy
 - homomorphic encryption
 - federated learning
-

Abstract

Two competing trends are shaping the current marketing analytics landscape. On the one hand more and more data is being generated and stored, and on the other privacy regulations and corporate policy threaten the analyst's ability to access and learn from this data. The confluence of these two forces has naturally led to technological innovations that seek to maintain the utility of the granular data for analytical purposes while at the same time offering privacy guarantees to the subjects to whom the data pertains. In this paper we evaluate several privacy preserving technologies in a workflow that reflects a typical real-world marketing analytics deployment. We study the impact of these approaches on computational cost, model fit, attribution accuracy, and the privacy of the simulated individuals, and propose some guidelines for implementing privacy preserving methods in marketing analytics.

1. Introduction

Two competing trends are shaping the current marketing analytics landscape. On the one hand, more and more data is being generated and stored, and on the other hand, privacy regulations and corporate policy threaten the analyst's ability to access and learn from this data. The confluence of these two forces has naturally led to technological innovations that seek to maintain the utility of the granular data for analytical purposes while, at the same time, offering privacy guarantees to the subjects to whom the data pertains.

Data has been called 'the oil' of the digital economy (Wedel & Kannan, 2016). Digitization has led to lower costs of data collection, storage, and transmission (Goldfarb & Tucker, 2019), while rapid growth in media channels, devices, and applications has led to a diverse range of data streams that reflect consumer behaviors, interactions, and responses. This surge in information has provided new opportunities to use data to provide enhanced experiences and satisfaction, while also providing companies with insight into how their advertising efforts are performing at a granular level. These capabilities have had a significant impact on corporate financial performance (Wedel & Kannan, 2016).

But in parallel with this growth in consumer data, questions about data privacy have become increasingly prevalent, with regulations like GDPR and CCPA — that require companies to modify their data handling practices — coming into force. In addition, Wierenga et al. argue that increased sensitivity to privacy concerns has prompted self-policing by many firms, making them reluctant to share data outside of their own firewalls (Wierenga et al., 2021).

While some observers predict that as a consequence of these growing privacy concerns first parties will retain access to their granular data while requiring those on the outside to live with data in aggregated forms, others see potential in an emerging field of privacy-preserving technologies. Although research into privacy-preserving analytics has a long history spanning multiple disciplines, in recent years there has been a significant spike in interest among academics, spawning almost 18,000 papers in 2020 alone. Given that the field is moving so quickly, it can be difficult to extract and implement useable methods from the

research, and although the practice of machine learning on granular marketing data falls within the broader scope of existing academic research into private learning, relatively little work has been done to assess the practical implications of privatizing machine learning pipelines for marketing analytics applications.

In this paper, we evaluate several privacy-preserving technologies in a workflow that reflects a typical real-world marketing analytics deployment. To achieve this, we generate a population of individuals, then learn their sensitivities to different marketing stimuli under two privacy-preserving regimes. The first regime assumes that the data itself must be protected using input perturbation. The second takes an algorithmic perturbation approach, applying privacy protection to the machine learning model itself. We study the impact of these approaches on computational cost, model fit, attribution accuracy, and the privacy of the simulated individuals, and propose some guidelines for implementing privacy-preserving methods in marketing analytics.

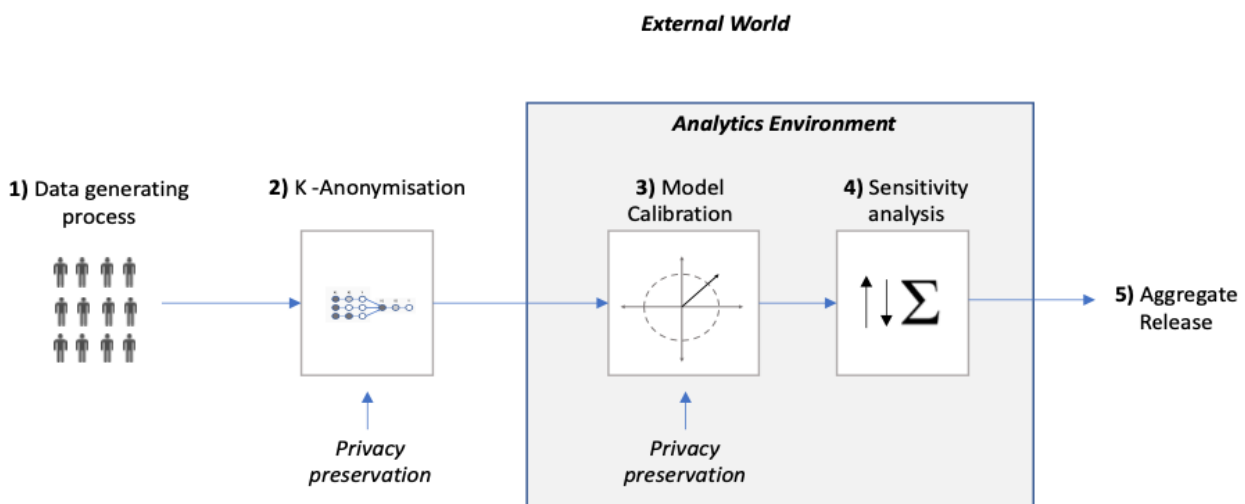


Figure 1: Overview of the simulation environment and the five steps involved in a single end to end run of the process

2. Simulation Environment

We created a simulated environment made up of an external world, in which the agents representing individuals are acting and generating data, and in which the analytics environment is contained. The data is passed from the external world into the analytics environment and progresses through a series of processes, steps 1 through 5 in **Figure 1**, and are detailed further below. At the end of the analytics process, aggregates are created and exported

back into the external world. There are two key advantages to using simulated data in this kind of exploration. Firstly, we know the real answers, so we can benchmark the accuracy of different privatized methodologies against non-privatized methods. Secondly, we can examine the impact of counter-factual scenarios, allowing us to look at what would have happened under different conditions.

3. Security Assumptions

The range of threats to which the marketing analytics pipeline is exposed depends to a large extent on the types of interface with the model and data that are available to an attacker. We assume here that the attacker does not have access to the analytics environment itself — that it is secured against intrusion — and that the vulnerabilities are limited to the points in the process that are accessible to the external world — i.e the input data and the aggregated outputs.

Attacks on machine learning pipelines are generally classified into one of three categories, differentiated in terms of the attacker's intentions (Papernot et al., 2016). Attacks on confidentiality aim to recover the model structure or parameters, or the data used to train it. Attacks on integrity

seek to induce particular outputs or behaviors of the attacker's choosing. These attacks often involve 'poisoning' the training data to mislead the model about the correct classification for a given input (Jagielski et al., 2018).

Finally, attacks on availability of the pipeline attempt to prevent access to model outputs or other features of the system.

We focused in this study on potential attacks on confidentiality. Because the analytics environment does not expose a scoring API to the external world, we exclude consideration of attacks such as reverse engineering of the training data, model weight and hyperparameter stealing, and membership inference attacks (Shokri et al., 2017).

4. Simulation Process

4.1 Simulation Step 1: Data Generating Process

The synthetic data generating process consists of:

- A heterogeneous population of simulated software agents, each of which represents an individual. Each of the individuals is given a base likelihood of making a transaction, the ability to receive advertising messages, and the ability to make a purchase when they reach a certain utility for the simulated product. The individuals have a variable level of responsiveness to advertising.

- An advertiser with the ability to use two different media channels to deliver advertising impressions to the customer agents over time. Individual agents are exposed to different levels of advertising, reflecting heterogeneity in their underlying media consumption habits.

The simulated data is generated and passed to the k-anonymization system.

4.2 Simulation Step 2: K-anonymization

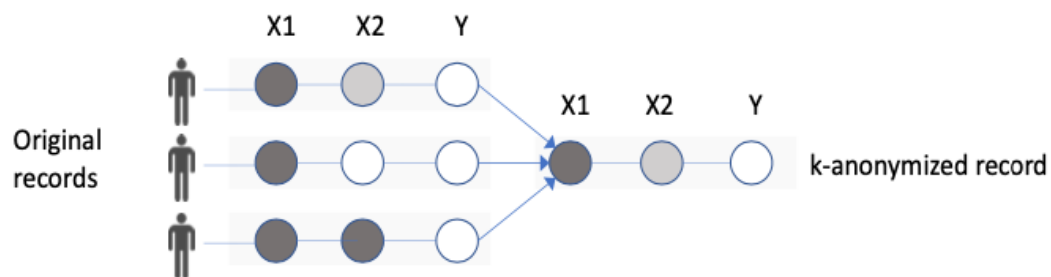


Figure 2: An example of a k-anonymization process where k=3 in which three sets of quasi-identifiers are generalized to a common set

The intuition that motivates **k-anonymization** is that attributes of an individual that uniquely identify them may remain in a record, even after traditional forms of de-identification have been performed, as illustrated in **Figure 2**. For example, in the US, zip code and date of birth create a unique combination of attributes for a large part of the population. An attacker that is unconstrained by any form of process control and has access to these attributes from another source can match them against the record in question and re-identify the individual (Sweeney, 2002).

These attributes are usually referred to in the literature as **quasi-identifiers**. To protect against this kind of re-identification, Sweeney proposed **k-anonymization** as a guarantee that each combination of quasi-identifiers appears “with at least **k** occurrences” in a given dataset. These groups of quasi-identifiers are often referred to as Equivalence Classes (EQs) (Ayala-Rivera et al., 2014).

In this exercise, we use the Mondrian algorithm, a ‘greedy’ multidimensional approach that recursively partitions the domain space into regions that contain at least **k** records that share the same EQ.

The data the agents generate, X1, X2, and Y, is collected and assembled, then passed through the Mondrian **k-anonymization** process which

may be set at any level of **k**. When **k** is set to 1, the data is unmodified by the k-anonymization process. Where **k** is > 1, the algorithm generalizes the quasi-identifiers such that EQs satisfy **k** by setting X1, X2, and Y to their mean values.

- Distances between original and k-anonymized quasi-identifiers in the dataset are computed as the L1 norm of the features’ differences. Distances between the original (f) and k-anonymized (\bar{f}) features are calculated using an averaged L1 norm of the difference between features over the entire dataset (T):

$$d(f, \bar{f}) = \frac{\|f - \bar{f}\|}{|T|} = \frac{1}{|T|} \sum_{i=1}^{|T|} |f_i - \bar{f}_i|$$

- The Discernibility metric represents how indistinguishable a record is from others in the dataset. If a record belongs to an Equivalence Class (EQ) containing $|EQ|$ members, we define the discernibility of the record as $1/|EQ|$. The discernibility of the entire dataset is defined as the sum of the discernibility of each record:

$$D(T) = \sum_{i \in \{EQ\}} \frac{1}{|EQ_i|} = |\{EQ\}|$$

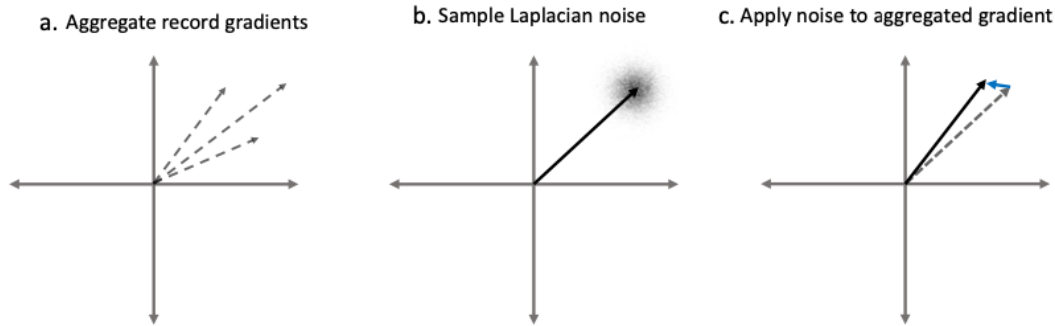


Figure 3: Steps in the computation of the differentially private gradients for the privacy-preserving model estimation. a. Record-level gradients are aggregated into coefficient update direction; b. Laplace noise is sampled with a scale proportional to gradient sensitivity and inversely proportional to ϵ ; c. Gradient vector is updated.

4.3. Simulation Step 3: Model Calibration

Some machine learning models and estimation algorithms are susceptible to leaking private information from records used during training by potentially encoding information that would otherwise not be known had an individual not been present in the training dataset. As a simple example, consider a situation in which an individual has a property unseen in other records, but which is nonetheless encoded as a feature in a model. Any association captured through the model estimation process between the individual’s property and that their outcome will allow for the identification of an individual with such property in the training dataset.

In this study, we explore the impact of privacy preservation using Logistic Regression models, which are commonly estimated using gradient-based methods. In such estimation algorithms, the encoding of individual-level information into model coefficients occurs through the effect of the gradient of the loss function with respect to a single record exerts in the direction of updates of coefficients at each iteration of the algorithm. A typical strategy to avoid privacy violations through model training is, therefore, to ensure that the loss gradients calculated at each algorithm iteration do not expose information from any single record.

To achieve this goal, we employ a Differentially Private approach to computing loss gradients in our Gradient Descent (GD) based training algorithm, as illustrated in **Figure 3**. At each

iteration of the algorithm, we calculate the Jacobian of the record-level log-likelihood loss function with respect to the model parameters:

$$g_{ij} = \frac{\partial L_i}{\partial \beta_j} \quad \forall i \in T, \forall j \in F$$

Where T is the set of all records in the training dataset, F is the set of all features in the model, L_i is the log-likelihood loss for record i , and β_j is the coefficient for feature j .

For each model feature j , we then compute the sensitivity of the loss gradient with respect to j as the smallest value S_j such that for every pair of datasets T and T' differing by a single record,

$$\frac{1}{|T|} \left| \sum_{i \in T} g_{ij} - \sum_{i \in T'} g_{ij} \right| \leq S_j$$

Finally, we compute the loss gradient with respect to each model coefficient as

$$g_j = \frac{1}{|T|} \sum_{i \in T} g_{ij} + \text{Lap} \left(\frac{S_j}{\epsilon} \right) \quad \forall j \in F$$

The above mechanism is demonstrated to be ϵ -indistinguishable by [Dwork et al. \(2006\)](#), and therefore has privacy leakage bounded by ϵ . The estimation algorithm is then:

- Set learning rate α , set tolerance τ
- Initialize model parameters: $\beta \leftarrow \beta_0$
- while iteration \leq max_iterations do:
 - Calculate Jacobian:
 $g_{ij} = \partial L_i / \partial \beta_j$ for every record i and feature j
 - Calculate Sensitivity:
 $S_j = \max(|g_{ij}|)$ for every feature j
 - Calculate gradients: $g_j = \sum_i g_{ij} + \text{Lap}(S_j / \epsilon)$ for every feature j
 - Update model parameters: $\beta_j \leftarrow \beta_j - \alpha \cdot g_j$ for every feature j
 - if $\sqrt{\sum_j g_j^2} \leq \tau$
Stop
 - Otherwise
iteration = iteration + 1

4.4 Simulation Step 4: Sensitivity Analysis and Aggregation

Once we have a calibrated model, we conduct a sensitivity analysis using the model and the dataset together to assess the contribution of each of the media channels, X1 and X2, and the base level of sales.

We then sum the contributions of each of the channels.

- Calculate the total sales available in the entire dataset
 $total_sales = \sum_i \sum_t y_{it}$
- Calculate the total sales with X1 excluded
 $attributon_to_all_except_X1 = \sum_i \sum_t \left(\frac{1}{(1 + \exp(-(\beta_1 + \sum_{t=0}^t X1_{it} * 0 + \sum_{t=0}^t X2_{it} * \beta_3)))} \right)$
- Calculate the total sales with X2 excluded
 $attributon_to_all_except_X2 = \sum_i \sum_t \left(\frac{1}{(1 + \exp(-(\beta_1 + \sum_{t=0}^t X1_{it} * \beta_2 + \sum_{t=0}^t X2_{it} * 0)))} \right)$
- Calculate sales attributable to base
 $attributon_to_base = \sum_i \sum_t \left(\frac{1}{(1 + \exp(-(\beta_1 + \sum_{t=0}^t X1_{it} * 0 + \sum_{t=0}^t X2_{it} * 0)))} \right)$
- Calculate sales attributable to X1:
 - total_sales - attribution_to_all_except_X1
- Calculate sales attributable to X2:
 - total_sales - attribution_to_all_except_X2

4.5 Simulation Step 5: Aggregate Release

The aggregate results produced in Step 4 are the output from the analytics environment into the external world.

5. Simulation Results

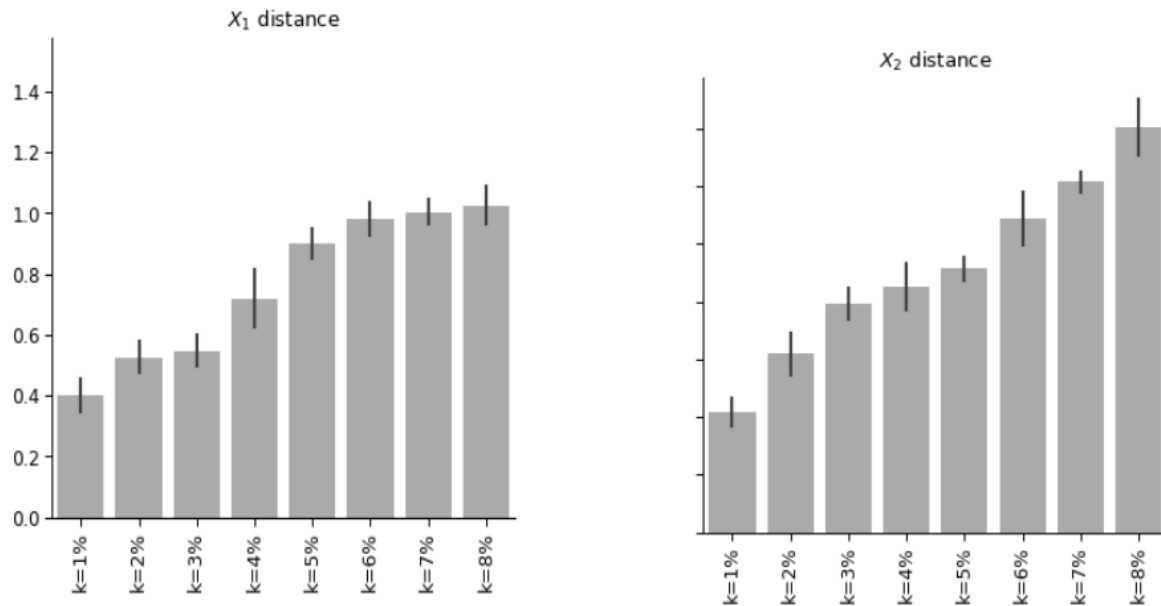


Figure 4: Distances between original and k -anonymized quasi-identifiers in the dataset, computed as the L1 norm of the feature differences. Thin bars represent the confidence interval of the measured value, defined as one standard deviation above and below the mean.

As would be expected, **Figure 4** shows that as the level of k increases, the average distance between each X_1 and X_2 value and its k -anonymized counterpart increases. Even at high levels of k though, frequently occurring combinations of values may be sufficiently common that no anonymization is required since their pre-existing EQ may already meet the k requirement. In parallel, as k increases the discernibility of any record in the dataset decreases (see **Figure 5**). As **Figure 6** shows, the computational cost of applying k -anonymization for low values of k can be substantial. As k increases, the number of partitions the Mondrian algorithm needs to apply to find EQs reduces.

For the purposes of this study, we assumed that combinations of X_1 , X_2 , and Y may act as quasi-identifiers that can potentially be used to uniquely identify the simulated individuals in the dataset. In the real world, it may be more likely that only a subset of these quasi-identifiers may be accessible to an attacker.

For the particular characteristics of the data generating process that we were using, the impact of k -anonymization on AUROC

was negligible until k approaches 8% of the total number of records in the dataset (see **Figure 7**). There may be no direct comparison of these percentages to other datasets though, since the distributions of the variables and their

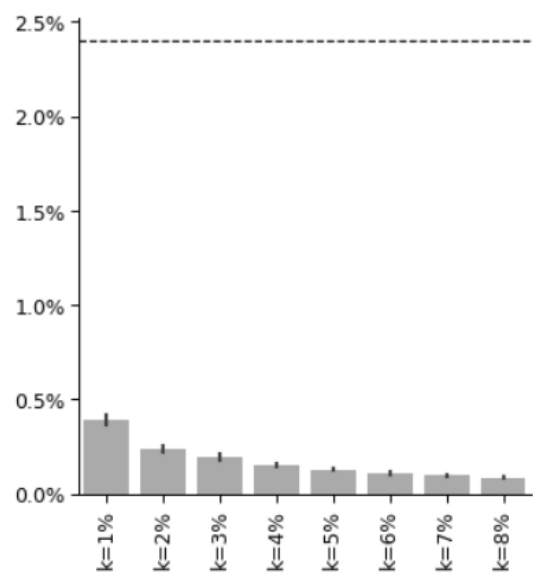


Figure 5: Ability of an attacker to single out an individual from the set of observations for different values of k as a fraction of the size of the dataset. Thin bars span one standard deviation above and below the mean.

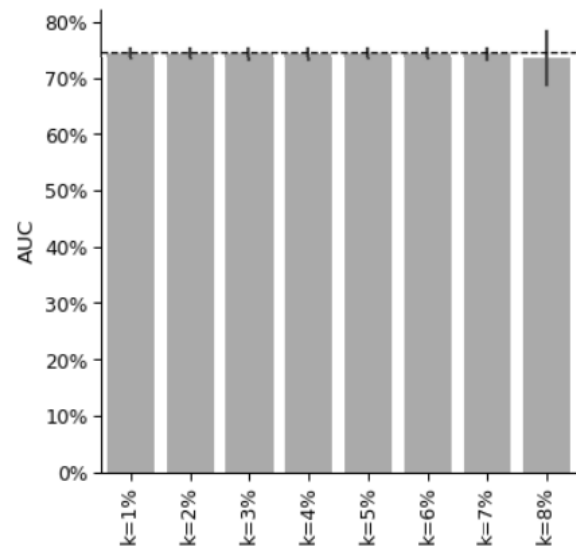
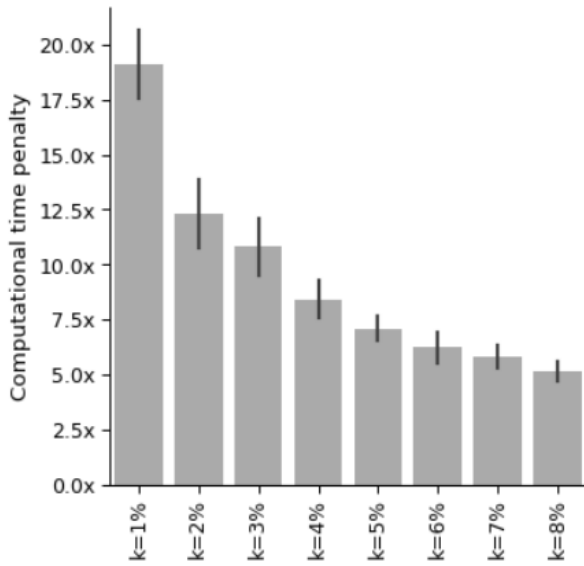


Figure 6: Variation of computational time in relation to a non-anonymized model estimation observed by applying the Mondrian k -Anonymity, for different values of k as a fraction of the size of the dataset. ϵ was set to a large value to avoid noise addition to the estimation.

Figure 7: Model predictive power measured as the area under the ROC curve (AUROC), for different values of k as a fraction of the size of the dataset. The dotted line represents the AUROC obtained for the model estimated on the non k -anonymized dataset. ϵ was set to a large value.

inter-relationships will play a major role. Also, as **Figure 8** shows, at reasonable practical levels of k we observed minimal bias in the attribution results. For example, where $k = 1\%$ of total records, e.g. a relatively large k value of 10,000 in a million-record dataset, the bias in attribution

is only 1%. As k increases, though, the level of bias trends upwards, and there would be limits on the practical usefulness of machine learning results from k -anonymized data at some level of k . In the data generating process and learning pipeline we used, attribution is generally

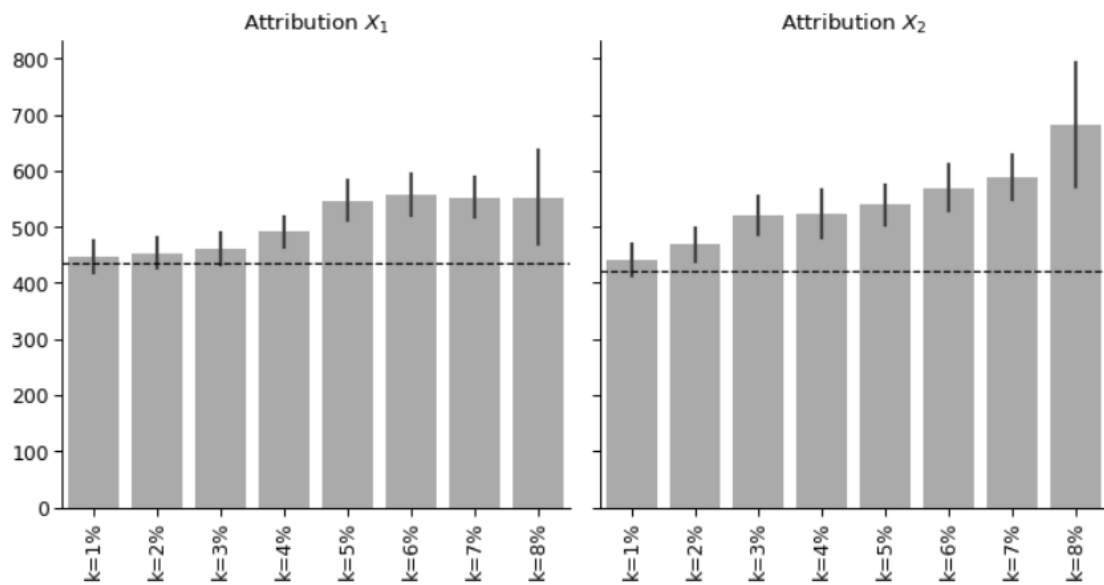


Figure 8: Outcome attributed to each model driver for different values of k as a fraction of the size of the dataset. The dotted line represents the attribution obtained without applying k -anonymity. Thin bars span one standard deviation above and below the mean. ϵ was set to a large value to avoid noise addition to the estimation.

biased upwards as k increases, but it is not clear that the bias would always be in this direction for any dataset.

In experiments that pushed k beyond the values reported here, we discovered a number of

'breaking points' that make k -anonymization above a certain level of k impractical. For example, if k exceeds the number of outcomes of a particular class in a dataset, the outcome cannot be preserved in a 1/0 form but itself becomes the average value of multiple classes.

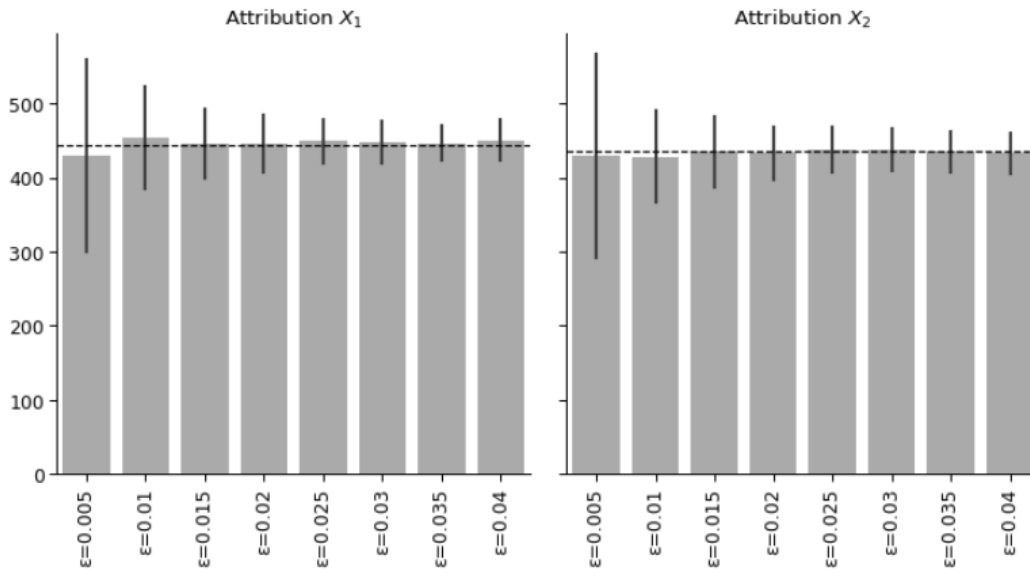


Figure 9: Outcome attributed to each model driver for different values of ϵ . Thin bars span one standard deviation above and below the mean. The dotted line represents the attribution obtained with a model estimated without privatized learning. k value was set to 1.

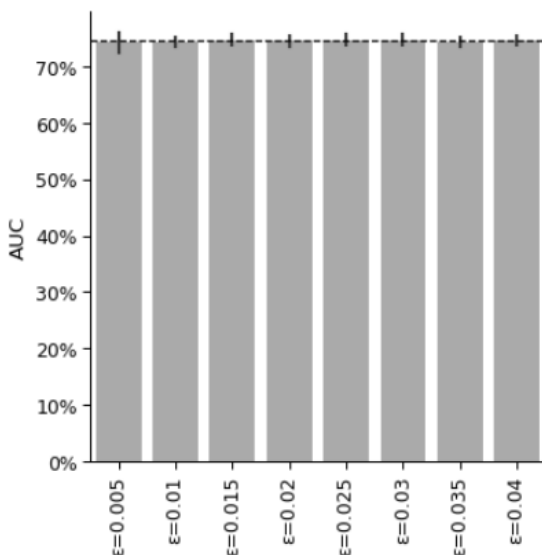


Figure 10: Distances between original and k -anonymized quasi-identifiers in the dataset, computed as the L1 norm of the features differences. k value was set to 1.

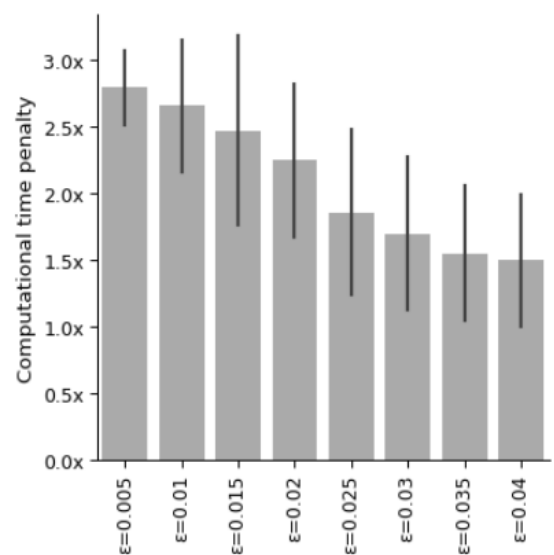


Figure 11: Variation of computational time in relation to a non-anonymized model estimation observed for different values of ϵ . Thin bars span one standard deviation above and below the mean. k value was set to 1.

In general, differential privacy can be achieved by adding a reasonable amount of noise into the output results of a computation over a dataset. The amount of noise will affect the trade-off between privacy and utility of the results. Specifically, too much noise will make the dataset useless and too little noise (Gong et al., 2020) is insufficient to provide privacy guarantees. In our model estimation approach, this trade-off is modulated by the value of ϵ selected.

The chosen value of ϵ specifies within each Gradient Descent iteration, a bound on the ratio of probabilities of calculating a specific gradient component value over two datasets that differ

only by a single record. Therefore, it limits the impact that any record can have in the gradient computation, preventing the model from encoding information particular to any single observation into the model parameters.

In practice, smaller values of ϵ yield stronger privacy protection, at the cost of higher added variation in calculated gradients and in attribution results (Figure 9). This variation may impair the effectiveness of the parameter search algorithm and result in higher computational costs for training, as observed in Figure 11, while not impacting model accuracy (Figure 10).

6. Towards an integrated view of privacy

In the previous sections, we explored the practical impact of applying two types of privacy preservation in a machine learning pipeline. To understand the role that these mechanisms play individually and their potential to work together, it is worth stepping back and reviewing the different notions of privacy that each supports. The current literature on privacy-preserving technologies does not coalesce around a universal definition of privacy. Some of the more general ideas that motivate research in the area include the suggestions that it is the “right to be let alone” (Warren & Brandeis, 1890), or “protection from being brought to the attention of others.” (Gavison, 1980). As technology has enabled the collection of increasingly detailed data about individuals, Dwork argues that “the need increases for a robust, meaningful, and mathematically rigorous definition of privacy” (Dwork & Roth, 2013). K -anonymity and differentially private gradient descent are two examples of mathematically rigorous definitions of privacy, but they are not competing approaches to achieve the same outcome, and each was developed for a different purpose with a different idea of privacy in mind.

K -anonymity was conceived as a method for protecting the identities of individuals in published data, for example, people whose details are contained in the release of medical

data from a hospital. As such, it applies privacy protection to the data itself and has an impact on any analysis or query that is done on it. It guarantees that each entity contained in the data cannot be distinguished from at least k individuals whose attributes also appear in it.

Differential privacy was designed to protect the privacy of individuals by applying noise to the results of aggregate queries in which they may be present, but does not rely on transformations on the underlying data itself. Instead, it guarantees that an algorithm’s *output* does not differ significantly statistically for two versions of the data differing by only one record.

K -anonymization and privacy-preserving learning, therefore, offer two different types of privacy protection whose applicability is determined by the constraints we impose in the measurement system. The former allows for strong guarantees around the limitations of an attacker’s ability to discern a user from others in the dataset, whereas the latter offers well-defined limitations on the impact a single user’s data have on the learning of the measurement model.

As our simulations have shown, these two approaches also have starkly distinct effects on both the increment in computational costs

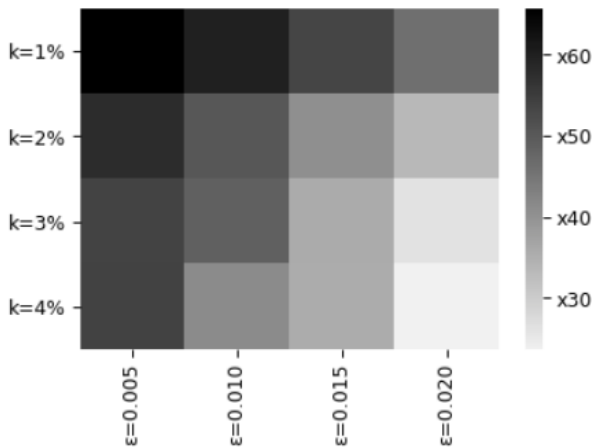


Figure 12: Computational penalty due to the sequential application of k -anonymization and privacy-preserving learning, as a factor over the baseline non-private model estimation.

and accuracy of the attribution measurement. In **Figure 12**, we can see that stronger k -anonymity constraints (larger values of k) result in lesser computational penalty added to the measurement process, since a smaller number of partitions of the dataset are required to meet the k constraint. Conversely, stronger privacy preservation at the level of model training results in higher computational costs, due to the increased difficulty of finding a solution to the estimation problem under higher amounts of added noise.

Conclusion

We have studied the application and effects of two popular privacy preservation techniques to the problem of measuring marketing impact under the controlled environment of simulated data. Our analysis shows that there are a number of considerations a marketing measurement practitioner must make in order to effectively apply privacy preservation on an attribution system. K -anonymity and private learning offer distinct sets of trade-offs related to their impact on measurement accuracy, results variance, and computational costs. We have found that k -anonymizing process inputs potentially yield measurement biases at high thresholds of k , whereas lower k values result in super-linear increases in computational costs. Conversely, privacy-preserving model estimation does not result in a significant measurement bias, even at very low values of ϵ , but will increase the variance associated with measurement, as well as training costs, at higher strengths of privacy preservation.

It is worth noting that privacy preservation techniques do not come without computational cost. As shown in **Figures 6** and **11**, which are expressed as a penalty with respect to an end-to-end model estimation process that does not include a privacy-preservation step, there is a time penalty associated with each of the methods that we evaluated.

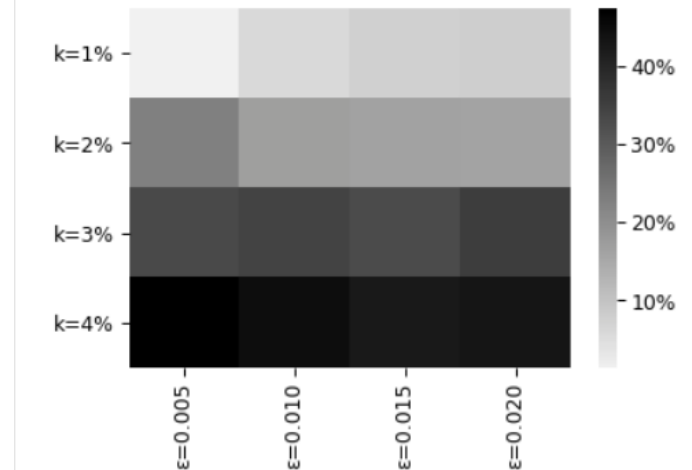


Figure 13: Mean of attribution error resulting from the application of k -anonymization and privacy-preserving learning. Higher values imply a larger bias on attribution error.

When it comes to measurement accuracy, **Figure 13** shows that k -anonymity adds a stronger bias on the mean attribution calculated under its privacy protection. This bias, however, is only observed at k values representing higher fractions of the entire dataset and, therefore, is of little practical concern when working with larger datasets. But as we saw in **Figure 9**, privacy-preserving model training does not add any significant bias to attribution measurement, although it does add variance to the results due to the noise injected into the model parameter search process.

Ultimately, the choice of which privacy protection method to employ — or combination of methods — and which protection strength lies on the balance between the measurement accuracy considered acceptable, the computational costs involved in the process, and the levels of privacy protection deemed necessary to be imposed on the system. Our results indicate that using a k below 1% of the dataset size yields very small biases and may still be computationally feasible for most use-cases. A value of k greater than 0.02, still considered small for many applications, results in a low impact in both accuracy and computational costs.

References

1. Wedel, M., & Kannan, P. K. (2016). Marketing Analytics for Data-Rich Environments. *Journal of Marketing*, 80(6), 97–121. [Source](#)
2. Goldfarb, Avi, & Tucker, C. (2019). Digital Economics. *Journal of Economic Literature*, 57(1), 3–43. [Source](#)
3. Wieringa, J., Kannan, P., Ma, X., Reutterer, T., Risselada, H., & Skiera, B. (2021). Data analytics in a privacy-concerned world. *Journal of Business Research*, 122, 915–925. [Source](#)
4. Papernot, N., Mcdaniel, P., Sinha, A., & Wellman, M.P. (2016). Towards the Science of Security and Privacy in Machine Learning. ArXiv, abs/1611.03814.
5. Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C., & Li, B. (2018). Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning. *2018 IEEE Symposium on Security and Privacy (SP)*, 19–35. IEEE (2018). [Source](#)
6. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership Inference Attacks Against Machine Learning Models. *2017 IEEE Symposium on Security and Privacy (SP)*, 3–18. IEEE (2017).
7. Sweeney, L. (2002). k -anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5), 557–570. [Source](#)
8. Ayala-Rivera, V., McDonagh, P., Cerqueus, T., & Murphy, L. (2014). A Systematic Comparison and Evaluation of k -Anonymization Algorithms for Practitioners. *Transactions on Data Privacy* 7(3), 337–370.
9. Gong, M., Xie, Y., Pan, K., Feng, K., & Qin, A. (2020). A Survey on Differentially Private Machine Learning [Review Article]. *IEEE Computational Intelligence Magazine*, 15(2), 49–64. [Source](#)
10. Warren, S., & Brandeis, L. (1890). The Right to Privacy. *Harvard Law Review*, 4(5), 193–220. [Source](#)
11. Gavison, R. (1980). Privacy and the Limits of Law. *The Yale Law Journal*, 89(3), 421–471. [Source](#)
12. Dwork, C., & Roth, A. (2013). *The Algorithmic Foundations of Differential Privacy*. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4), 211–407. [Source](#)
13. Bleier, A., Goldfarb, A., & Tucker, C. (2020). Consumer privacy and the future of data-based innovation and marketing. *International Journal of Research in Marketing*, 37(3), 466–480. [Source](#)
14. LeFevre, K., DeWitt, D., & Ramakrishnan, R. (2006). Mondrian Multidimensional K -Anonymity. *22nd International Conference on Data Engineering (ICDE'06)*. [Source](#)
15. Li, N., Qardaji, W.H., & Su, D. (2011). Provably private data anonymization: or, k -anonymity meets differential privacy. CoRR, abs/1101.2604, 49, 55 (2011) 15.
16. Song, L., & Mittal, P. (2020). Systematic Evaluation of Privacy Risks of Machine Learning Models. ArXiv, abs/2003.10595.
17. Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated Machine Learning. *ACM Transactions on Intelligent Systems and Technology*, 10(2), 1–19. [Source](#)

Authors



Joao Natali is Sr. Director, Data Science, leading analytics R&D for Neustar Marketing Solutions. Joao has over 15 years of experience in data science and marketing analytics, and stays excited by new challenges in the field of technology. He has a Ph.D in engineering, focused on optimization and machine learning applied to full genomic analysis.

joao.natali@team.neustar



Robert Stratton is SVP, Data Science, leading R&D efforts across Neustar Marketing Solutions. Robert has over 15 years of experience leading and conducting analytics projects across a wide range of industries and applications, from digital attribution and transactional analysis to process mining, marketing mix and brand equity analysis. He has a PhD from King's College in London and is an expert on computational modeling.

robert.stratton@team.neustar

