International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
Article Received: *25 February 2024* ***Revised:*** *12 March 2024* ***Accepted:*** *30 April 2024*

Causal Inference Methods for Understanding Attribution in Marketing Analytics Pipelines

Suresh Sankara Palli

Independent Researcher, USA.

Abstract

In strategic decision-making, the limits of conventional predictive analytics have become more apparent as businesses negotiate more complex, data-rich settings. Because predictive models often just reveal correlations rather than the fundamental causes of change, they expose organisations to misunderstandings and inefficient responses. A revolutionary development is provided by causal machine learning models, which separate cause-and-effect correlations from big, multidimensional datasets. By simulating the possible effects of business decisions prior to execution, these models help decision-makers close the gap between insight and consequence. In these kinds of campaigns, many channels often provide ads to specific individuals. The industry is very interested in "attribution," which is the process of allocating conversion credit to the different channels. Marketing researchers have a plethora of options to better forecast and maybe explain customer behaviour because to the massive amount of data. In this work, a causally justified approach to conversion attribution in online advertising campaigns is presented. However, as this article will argue, academics studying marketing should not hastily forsake methodological and cognitive processes that have been honed over centuries of scientific and philosophical contemplation. By combining the literature from many hard sciences, we talk about the importance of machine learning in causal inference as well as current issues with data management and measurement in the age of digital data.

Keywords: - Causal Machine Learning, Offers Marketing, Predictive Models, Strategic Decision-Making, Attribution, Online Advertising, Causal Inference.

I.INTRODUCTION

Over a century has passed since the development of marketing research strategies and analytical techniques [1]. Numerous causes, including changes in consumer demands and expectations, the necessity for scholarly publications, technology advancements, and the expansion of knowledge and skills, have contributed to these developments. Perhaps the most significant factor behind these developments in recent years has been the rise of digital (big) data and the techniques for analysing it [1, 2]. Some examples include;

(1) User-generated material on social media platforms that facilitates the exchange of ideas and information,
(2) Transactional data recorded in CRM platforms that improve performance evaluation, and [1],
(3) The Internet of Things (IoT), a network of geographically dispersed objects having integrated sensing, actuation, and/or identification capabilities that interact constantly.

Numerous chances to forecast and maybe explain consumer behaviour arise when machine learning algorithms are combined with this recently accessible data [2, 3].

Data analytics pipelines and other modern computer systems are usually made up of many components, each of which has a wide range of configuration choices that may be used alone or in combination with other components on various hardware platforms [3, 6]. With hundreds or even thousands of software and hardware configuration choices that interact with one another in a nontrivial way, such highly adjustable systems have a combinatorially vast configuration space. The performance behaviour of the systems that make up the components is usually only partially understood by individual component developers, who tend to have a more localised perspective [3, 4]. As a result, the complexity of creating and configuring components often overwhelms developers and end users of the finished system, making it difficult and prone to mistakes to set these systems to meet intended performance objectives.

In contrast to common software defects, these non-functional flaws do not result in system crashes or overt abnormalities. On-device machine learning (ML)

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
Article Received: *25 February 2024* ***Revised:*** *12 March 2024* ***Accepted:*** *30 April 2024*

systems, cloud infrastructure, and internet-scale systems may have significant problems as a result of misconfigured systems that continue to function but perform worse [3, 4]. For instance, a developer expressed displeasure because,

"I have a complex system with various parts that uses several sensors and an NVIDIA Nano, and I've noticed a number of performance problems."

Another time, a developer queries,

"I'm very annoyed with how much CPU the Jetson TX2 uses while the TCP/IP upload test software is running" [5].

The developer comes to this conclusion after battling for many days to address the problems,

"To properly assess CPU load and optimise the network stack, a great deal of information is needed [4, 5]. I made an effort to experiment with each configuration option described in the kernel documentation."

They also want to know how configuration choices affect things and how they work together, for example.,

"What impact does swap memory have on throughput growth?"

Using a browser to visit websites, where multimedia items and information are presented in an easily comprehensible manner, is how the great majority of computer users engage with the internet [6, 7]. This is only the beginning of the web's potential, however, since there is a wealth of valuable raw data that is concealed from view [8]. Most of this information can be readily accessed across several websites via APIs, but this is up to the site owner, and it's simple not to make this information available via APIs if they so desire. On the other hand, web scraping or web crawling is much quicker and more efficient. It may be used to collect and aggregate data from hundreds of millions of pages for processing and information extraction. This method is quite useful in many different applications, but it is especially useful in the area of business intelligence.

Web scraping is a crucial technique for establishing a strong online presence while staying competitive for any business hoping to stay relevant in the twenty-first century. Businesses and organisations rely heavily on data to help them make decisions, and much of that data is presently available online [5, 9]. The initial stage of any data science research and development is data acquisition, which is the process of gathering data from either public or commercial sources, such as journals, open data, websites, and financial reports, or by buying data. The three main, interconnected steps of online scraping are website analysis, website crawling, and data organisation.

Data mining is different from web scraping in that the later involves data analysis, while the former does not care about data collection [11, 12]. Advanced statistical methods are also required for data mining. In general, web scraping is a very easy procedure since there are many readily available tools and libraries that provide effective implementations of much of the necessary functionality. Most web scraping software comes with the capability to send bespoke HTTP requests with various headers and payloads.

By the end of 2012, it's expected that the US will have about $40 billion in total income from online advertising. Despite making approximately 25% of all media time, estimates indicate that the internet only contributes 13–19% of overall media ad expenditure. This disparity points to a sizable chance for online ad expenditure to increase. Two categories account for the majority of online ad spend: Paid Search, which accounts for the largest portion of expenditures [12], and Display, which comprises banner advertisements and newer video forms. The general consensus is that the absence of suitable measuring methods that fully account for the category's worth is a hindrance to Display's expansion [13].

Conversion attribution is the most contested measurement standard in the industry.It is generally understood to be the assignment of conversion credit when a particular online user is reached by multiple advertising channels, where a "advertising channel" is any publisher or vendor that displays advertisements on behalf of an advertiser [13]. The industry's default conversion attribution approach at the moment is often known as "last-touch" attribution. This rule-based attribution methodology gives full conversion credit to the channel that last showed an ad to a user that converted [13].

A significant change has occurred in the contemporary business environment, with data-driven strategies replacing intuition-based management. How businesses produce insights, assess performance, and distribute resources has changed as a result of the widespread use of digital systems, cloud computing, and analytics tools [14]. Data is becoming a key strategic asset in today's economy, not just a support role. Nowadays, businesses from many industries use data to find market

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
Article Received: *25 February 2024* ***Revised:*** *12 March 2024* ***Accepted:*** *30 April 2024*

possibilities, streamline processes, customise client interactions, and reduce risk. The capacity to instantly aggregate, clean, and analyse large datasets is essential to this change. Businesses of all sizes may now incorporate quantitative reasoning into strategic planning thanks to platforms for business intelligence, machine learning, and big data infrastructure that have reduced the hurdles to advanced analytics [11].

For example, banking firms utilise data models to evaluate credit risk and identify fraud trends, while retail businesses use transactional and behavioural data to inform assortment and price choices [14]. Competitive advantage has also been redefined as a result of this change. Businesses that successfully use their data assets do better than their counterparts in terms of creativity, effectiveness, and responsiveness.

With the use of feedback loops that continually adjust strategy in response to changing measurements, they create more flexible business models. As a result, data-driven businesses are more robust to upheaval and can react to change more quickly [16]. In the end, the emergence of data-driven strategy represents a more general paradigm change in which executive choices are guided by algorithmic insights and evidence takes the role of intuition. This study examines how this development is influencing the next phase of corporate decision-making, namely by combining explanatory modelling with causal inference [11].

Although multi-touch attribution systems for internet advertising have advanced, there is still no widely recognised structure or set of guidelines that serve as the foundation for attribution measurement [16]. Two common obstacles to the expansion of online ad spending are said to be a lack of standardisation and openness [16]. The lack of uniformity and openness in online attribution measurement is especially concerning. Making progress in this approach is our aim with this research.

Our first addition is a list of general characteristics of a successful attribution system that encourage the use of a causal framework to establish attribution [15]. We establish attribution criteria that are intended to distribute credit according to each advertisement's causal impact. Secondly, we provide a strategy for approximating the whole causal framework that is flexible enough to accommodate various data assumptions. In situations when the causal interpretation is not supported, we recast these approximation techniques as variable significance measures, which are derived from earlier attribution work in cooperative game theory [16]. Lastly, we demonstrate that earlier research in the area is consistent with our guiding principles.

The structure of this document is as follows: In addition to offering a concise overview of the variable significance and causal procedures that guide our attribution system, Section 2 suggests three characteristics of an effective attribution system. In Section 3, the usefulness of this method is initially discussed and attribution parameters are defined within a well-known causal framework. In order to accommodate various data assumptions, it then expands this into an approximation of the complete technique. We make the connection between this and previous work to wrap out section 3. Estimation and an empirical examination of both simulated and actual campaign data are covered in Section 4.

The process of web scraping transforms unstructured online data into organised data that can be stored and examined in a central database or spreadsheet. Because big data is always changing and updating, this allows the bot to obtain vast amounts of data quickly, which is beneficial in today's society. Let's say you own a clothes store and you want to monitor the prices of items from your competitors. Although it would take a lot of effort, you may visit the websites of your competitors and collect data every day to compare their prices with yours for each product [18]. It will also be quite hard to maintain track of the prices if you have thousands of rivals. This is where web scraping may help. An approach known as "web scraping" transforms unstructured online data into organised data that can be kept and examined in a spreadsheet or central database [19].

There are many businesses that employ web scraping, including cyber security. By evaluating the collected data, it may be utilised to calculate pricing and manufacturing costs. Depending on the user's location and cookie settings, advertisements may be made to promote items to various people [15]. One may characterise it as a novel approach to online data collecting [10]. Because web scraping makes it possible to create large, personalised data sets at cheap prices, it is already being utilised for both commercial and scientific purposes. It is intended to take the role of antiquated data collecting techniques as more companies try to keep up with emerging consumer preferences.

This study discusses new problems in machine learning applications and the usage of digital data in general by combining the literature from several hard science

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
Article Received: *25 February 2024* ***Revised:*** *12 March 2024* ***Accepted:*** *30 April 2024*

domains [11]. We specifically discuss the difficulties in measuring and managing digital data, the ongoing importance of survey research, and the effectiveness of machine learning in drawing conclusions about causality. By doing this, we provide potential new avenues for study design and methodology throughout the course of the next ten years.

II.CHALLENGES IN DATA MANAGEMENT AND MEASUREMENT

For almost 60 years, survey data has been the primary source of research for marketing experts. Up until ten years ago, surveys and in-person interviews, or some mix of the two, accounted for over 60% of academic empirical research in marketing. Data gathering techniques have evolved throughout time, moving from human interviews and postal surveys to phone surveys (including mobile ones) and, most recently, online data collection using software platforms like Survey Monkey, Qualtrics, Google Survey, and Mechanical Turk [10]. Approximately 15% of marketing survey data is now gathered using mobile phone polls, while 70% of data is being gathered via web platforms. Online survey platforms make it easy to get samples of respondents, however some studies have expressed doubts regarding the integrity of their data [19].

2.1 Data quality

The granularity, volume, velocity, and diversity of data gathered in the consumer and business-to-business sectors have all altered as a result of recent advancements in data collecting [19]. As data becomes more difficult to handle and analyse using conventional statistical analysis techniques, these data features provide significant problems for both marketing academics and practitioners, even while they hold great potential for producing unique consumer insights [18]. Consequently, marketing researchers are starting to focus on computer and data science research methodologies.

2.2 Data sparsity

In machine learning applications, data sparsity is a prevalent issue [19]. Data are considered scant when just a tiny portion of them have essential information. Though it usually arises as part of the data-generating process, data sparsity may be caused by missing values [11, 18]. The vast majority of the millions of data entries in Netflix movie ratings, for instance, are missing or unseen since each user has a restricted behaviour budget in comparison to the vast number of films available. Consider an online retailer's recommender system in a similar manner, where customers usually buy for themselves, but sometimes also for other people [10]. Insufficient examples of each sort of purchase event in the dataset will make it impossible to discern between them, making it difficult to define what is "normal" and what is not [23]. Data sparsity, however, is troublesome not just for transactional data but also for rich data kinds like music, video, and photographs.

2.3 Measuring unobserved phenomena

Digital data usually takes the form of observations, including tracking data from sensors, consumer wearables, Internet cookies, or other IoT devices [21, 23]. Because of this, these data are especially suitable for monitoring the behaviour of customers or businesses and evaluating their performance. Predicting future behaviour requires an awareness of previous behaviour, but understanding the psychological processes that underlie behaviour is a key component of the scientific endeavour [24]. Theoretical notions pertaining to customers' attitudes, beliefs, or intentions must be measurable in order for researchers to comprehend these psychological dynamics. Since these ideas cannot be seen directly, they are often quantified using scales made up of groups of things.

III.CONVERTING UNSTRUCTURED INTO STRUCTURED DATA

It is estimated by experts that between 80 and 90 percent of information is unstructured data. Unstructured data includes things like audio, [23], blogs, images, text, and video. A large number of these data are,

"Organisations often do not utilise the information assets they gather, analyse, and store during routine business operations for other reasons (such as analytics, business relationships, [22], and direct monetisation)".

These data have grown to be a valuable resource with the advent of machine learning techniques, leading to the development of new business models [22]. However, information extraction must be used to transform this unstructured data into structured data in order to extract their value.

IV.MACHINE LEARNING AND CAUSAL INFERENCE

4.1 Are correlations enough?

Marketing researchers have been examining known ideas in consumer and business-to-business behaviour [24], or just much any other area of marketing concentration, for

decade. In order to accomplish this goal, marketers have used ideas from a variety of fields, including biology, sociology, and psychology [21, 25]. Qualitative research or a panel of marketing specialists were usually the basis for adapting these ideas to a marketing setting. This strategy has proven beneficial to the marketing field, and significant strides have been achieved in shifting the field's focus from descriptive research methods to testing theoretical models.

4.2 Turing the black box into a white box

Scholars from a wide range of disciplines have recognised that machine learning techniques are a mystery. The models often lack an explicit declarative knowledge representation, making it challenging to identify the underlying explanatory structures, even when the fundamental mathematical concepts are well-defined [10]. In response to this issue, computer scientists have called for machine learning models to provide both an outcome and a human-understandable description of the process. Explainable artificial intelligence, another name for this expansion, would assist decision makers accept it more and provide information about the causes of the outcomes [18].

4.3 Implications for marketing research

Although machine learning techniques may not adhere to the traditional scientific methodology of hypothesis, modelling, and testing, this does not imply that marketing academics cannot benefit from them [20].

Through the discovery of significant patterns in huge data sets, predictive modelling may reveal connections that qualitative research methods would not have acknowledged and suggested. When done well, predictive modelling may provide useful direction for theory development, opening the door to explanations that are driven by predictions [19], which have not been problematic in the physical sciences [20]. The effectiveness of machine learning techniques for testing theories, however, and induction have an intrinsic limit that marketing researchers must recognise. Furthermore, it is difficult to distinguish the basic distinctions between the target functions of explanatory-causal models and predictive models as they are used in machine learning [11].

V.CONCLUSION

Like many other fields, marketing research is undergoing a revolution. As life is captured at ever-higher levels, technology has made data even more accessible to marketing researchers. The volume of data will continue to grow dramatically with the emergence of the Internet of Things and the estimated 75 billion linked devices by 2025. These data will provide distinctive insights into consumer and business-to-business behaviour when combined with contemporary machine learning techniques, progressively facilitating the active control of corporate operations in real-time.

Although there are numerous potentials associated with the seemingly limitless quantity of digital data, researchers and practitioners alike still find it difficult and tiresome to extract valuable insights from various data sources. Major quality problems that impair the information extraction and analysis process include missing, incomplete, inconsistent, erroneous, duplicate, and outdated data. Researchers in computer science and management information systems have made significant strides in addressing these issues by identifying bias patterns, handling missing numbers, and creating accuracy measurements.

Similar to this, methodologists in various domains have created sophisticated methods for extracting information from audio, video, pictures, and text. The large volume, dimensionality, variety, dynamicity, and heterogeneity of the data to be analysed continue to pose major obstacles to these methodologies, even as they push the limits of data analysis. Large-scale parallel computing systems that are dispersed are necessary to facilitate the extraction and analysis of information. In fact, most businesses are still unable to use these approaches in real time, despite having access to enormous amounts of computer power.

Lastly, it is important to note that we do not want to imply that machine learning techniques for large-scale data analysis are useless. On the contrary, we think the reverse will be true. In order to make sure that their analyses are up to date and represent the current state of affairs, marketing researchers and practitioners must be aware of the limits of these approaches and take into account advancements in computer science and related disciplines. Researchers and practitioners will probably be tempted to support a strictly data-driven approach to scientific and commercial decision-making due to the volume of data. But the value of data is only as great as the information we can draw from it.

VI.REFERENCES

[1] Ahmed F, Ahmed MR, Kabir MA, Islam MM. Revolutionizing Business Analytics: The Impact of Artificial Intelligence and Machine Learning. American Journal of Advanced Technology and

Engineering Solutions. 2025Feb 14;1(01):147-73.

[2] Vercellis C. Business intelligence: data mining and optimization for decision making. John Wiley &Sons; 2011Aug10.

[3] Mehrotra P. Applications of Artificial Intelligence in the Realm of Business Intelligence. InResearch AnthologyonArtificial Intelligence Applications in Security 2021 (pp. 358-386). IGI Global.

[4] Wang J, Omar AH, Alotaibi FM, Daradkeh YI, Althubiti SA. Business intelligence ability to enhance organizational performance and performance evaluation capabilities by improving data mining systems for competitive advantage. Information Processing & Management. 2022 Nov 1;59(6):103075.

[5] Noah GU. Interdisciplinary strategies for integrating oral health in national immune and inflammatorydiseasecontrol programs. Int J Comput Appl Technol Res. 2022;11(12):483-498. doi:10.7753/IJCATR1112.1016.

[6] Conboy K, Mikalef P, Dennehy D, Krogstie J. Using business analytics to enhance dynamic capabilitiesinoperations research: A case analysis and research agenda. European Journal of Operational Research. 2020Mar16;281(3):656-72.

[7] Nagy M, Lăzăroiu G, Valaskova K. Machine intelligence and autonomous robotic technologies in the corporatecontext of SMEs: Deep learning and virtual simulation algorithms, cyber-physical production networks, andIndustry4.0-based manufacturing systems. Applied Sciences. 2023 Jan 28;13(3):1681.

[8] J. Pearl. Causality: models, reasoning, and inference, volume 47. Cambridge Univ Press, 2000.

[9] F. Provost, B. Dalessandro, R. Hook, X. Zhang, and A. Murray. Audience selection for on-line brand advertising: privacy-friendly social network targeting. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 707–716. ACM, 2009.

[10] D. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of educational Psychology, 66(5):688, 1974.

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

[12] L. Shapley. A value for n-person games. The Shapley value, pages 31–40, 1953.

[13] Ferrell, O. C., Hair, J. F., Marshall, G., & Tamilia, R. D. (2015). Understanding the history of marketing education to improve classroom instruction. Marketing Education Review, 25(2), 159–175.

[14] Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. Science, 349(6245), 273–278.

[15] Ghasemy, M., Teeroovengadum, V., Becker, J. M., & Ringle, C. (2020). This fast car can move faster: A review of PLS-SEM application in higher education research. Higher Education, 80(6), 1121–1152.

[16] Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. Journal of Behavioral Decision Making, 26(3), 213–224.

[17] Gudivada, V., Apon, A., & Ding, J. (2017). Data quality considerations for big data and machine-learning: Going beyond data cleaning and transformations. International Journal on Advances in Software, 10(1), 1–20.

[18] Hair, J. F., Howard, M., & Nitzl, C. (2020). Assessing measurement model quality in PLS-SEM using confirmatory composite analysis. Journal of Business Research, 109, 101–110.

[19] Hair, J. F., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM – Indeed a silver bullet. Journal of Marketing Theory & Practice, 19(2), 139–151.

[20] Bich, W. (2013). Error, uncertainty and probability. In E. Bava, M. Kühne, & A. M. Rossi (Eds.), Proceedings of the International School of Physics “Enrico Fermi”, Course 185: Metrology and physical constants (pp. 47–73). IOS; SIF.

[21] Bollen, K., & Pearl, J. (2011). Eight myths about causality and structural equation models. In S. L. Morgan (Ed.), Handbook of causal analysis for social research (pp. 301–328).

[22] Bosu, M. F., & MacDonell, S. G. (2019). Experience: Quality benchmarking of datasets used in software effort estimation. ACM Journal

of Data and Information Quality, 11(4), Article No. 19.

[23] Bound, J., Jaeger, D. A., & Baker, R. (1993). The cure can be worse than the disease: A cautionary tale regarding instrumental variables (NBER Working Paper No. t0137). National Bureau of Economic Research.

[24] Boyd, D., & Crawford, K. (2012). Critical questions for big data. Information, Communication & Society, 15(5), 662–679.

[25] Brick, T., & Bailey, D. H. (2020). Rock the MIC: The matrix of implied causation for design and model checking. Advances in Methods and Practices in Psychological Science, 3(3), 286–299. forthcoming.