Journal of Retailing 93 (1, 2017) 79–95

Journal of Retailing

The Role of Big Data and Predictive Analytics in Retailing

Eric T. Bradlow [a], Manish Gangwar [b], Praveen Kopalle [c,*,1], Sudhir Voleti [b]

[a] *Wharton Customer Analytics Initiative, University of Pennsylvania, Philadelphia, PA 19104, United States*
[b] *Indian School of Business (ISB), Gachibowli, Hyderabad 50032, India*
[c] *Tuck School of Business at Dartmouth, Dartmouth College, Hanover, NH 03104, United States*

Abstract

The paper examines the opportunities in and possibilities arising from big data in retailing, particularly along five major data dimensions—data pertaining to customers, products, time, (geo-spatial) location and channel. Much of the increase in data quality and application possibilities comes from a mix of new data sources, a smart application of statistical tools and domain knowledge combined with theoretical insights. The importance of theory in guiding any systematic search for answers to retailing questions, as well as for streamlining analysis remains undiminished, even as the role of big data and predictive analytics in retailing is set to rise in importance, aided by newer sources of data and large-scale correlational techniques. The Statistical issues discussed include a particular focus on the relevance and uses of Bayesian analysis techniques (data borrowing, updating, augmentation and hierarchical modeling), predictive analytics using big data and a field experiment, all in a retailing context. Finally, the ethical and privacy issues that may arise from the use of big data in retailing are also highlighted.

Keywords: Big data; Predictive analytics; Retailing; Pricing

Introduction

According to some estimates, Walmart collects around 2.5 petabytes (1 petabyte = 1,000,000 gigabytes) of information every hour about transactions, customer behavior, location, and devices (McAfee et al. 2012). An IT analyst firm Gartner estimates that there will be 20 billion (13.5 billion in the consumer sector) devices connected in the "Internet of Things". Imagine the amount of data that will be generated by these devices (Gartner 2015). Imagine a day where online and offline retailing data provide a complete view of customer buying behavior, and even better if the data is linked at the level of the individual customer to enable "true" customer lifetime value calculations (Gupta et al. 2006; Venkatesan and Kumar 2004). Imagine a day where data thought only to exist in online retailing, for example consumer path data (Hui, Fader, and Bradlow 2009a), exists inside the store due to RFID and other GPS tracking-based technologies. Imagine a day where integrated online/offline experiments are being run that provide exogenous variation that enables causal inference about important marketing/retailing topics such as the efficacy of email, coupons, advertising, and so forth (Anderson and Simester 2003). Imagine a day where eye-tracking data is not just collected in the laboratory from Tobii-enhanced monitors but is collected in the field due to retinal scanning devices embedded within shelves (Chandon et al. 2008; Van der Lans, Pieters, and Wedel 2008).

As futuristic as those data sources sound, all of them exist today (albeit not ubiquitously) and will soon be part of the information that marketing scientists (within and outside of retail) use for customer-level understanding and firm-level optimization. Simply and heuristically put, these data sources will be adding "columns" to our databases (and a lot of columns!) that

* Corresponding author.
E-mail addresses: ebradlow@wharton.upenn.edu (E.T. Bradlow), manish_gangwar@isb.edu (M. Gangwar), kopalle@dartmouth.edu, pkopalle@gmail.com (P. Kopalle), sudhir_voleti@isb.edu (S. Voleti).
[1] The author(s) thank the guest editors, Dhruv Grewal, Anne Roggeveen, and Jens Nordfält, two anonymous JR reviewers, participants at the 2014 Digital Summit, Indian School of Business, 2014 Nordic Wholesale and Retailing Conference, 2015 Marketing Symposium, Tuck School of Business at Dartmouth, and 2016 Wharton Retailing Conference for their thoughtful comments on earlier drafts of this manuscript. The author(s) also thank an anonymous retail chain for providing us access to their confidential data for our analysis.

http://dx.doi.org/10.1016/j.jretai.2016.12.004
0022-4359/

provide an increased ability to predict customer behavior and the implications of marketing on it. Now, add that to the technology (i.e., IP address tracking, cookie tracking, registered-user log-in, loyalty card usage, to name just a few) which enables firms to collect this from millions of customers, for each and every moment, linked to each and every transaction, linked to each and every firm-level touchpoint, and linked across distribution platforms, and we have the big data that pervades the popular press today.

While the lure (and lore) of big data is tempting, in this paper we posit that the big data revolution (McAfee et al. 2012) really is a "better data" revolution, and especially so in retailing. Our intent in this paper is to describe the newest forms of data (i.e., "new columns") that exist in retailing, the importance of experimentation and exogenous variation ("better columns"), to describe why data mining and machine learning (despite their obvious value) will never obviate the need for marketing/economic theory (i.e., "where to look in the data"), to describe how managerial knowledge and statistical methods can lead to smart data compression (i.e., "which columns" and summaries of them) that will enable researchers to shrink the data, how better data will feed into predictive models (e.g., CLV, diffusion, choice models), and how firms are likely to use these models for decision making. This framework (both the buckets and the order of them) mirrors the INFORMS (www.informs.org) definition of business analytics which includes descriptive analytics, predictive analytics, and prescriptive analytics.

Wedel and Kannan (2016) provide an excellent commentary on marketing analytics past, present, and future. Guided by one of Marketing Science Institute's (www.msi.org) top research priorities, they discuss how marketing analytics will shape future decision making by managers in the area of customer relationship management, marketing mix allocation, personalization, customer privacy, and security issues. In contrast, our aim in this paper is to highlight the challenges and opportunities facing retailers dealing with big data. The rest of the paper is organized as follows. In the next three sections, we discuss the nature of "big" data in retailing, compare it with "better" data, and describe new sources of data that also leads to better models. This is followed by a discussion of the importance of theory in the analysis of retailing and various statistical issues involved such as data compression, statistical sufficiency for modeling, and the role of Bayesian inference. Finally, we present results of a case study, that is, a field experiment that combines predictive analytics and optimization in retailing.

Big Data in Retailing

This section describes "typical" sources of big data in retailing and how there is potential to exploit the vast flows of information in a five-dimensional space: across customers, products, time, geo-spatial location, and channel. We present them in Fig. 1 and discuss each in turn.

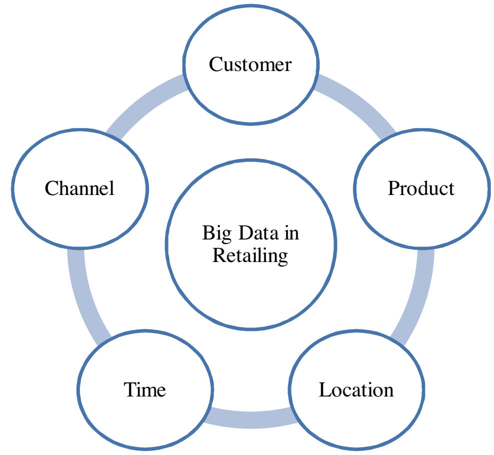

Fig. 1. Dimensions of big data in retailing.

Customers

When most people think of big data, they think of data sets with a lot of rows, and they should. Tracking technologies have enabled firms to move from aggregate data analyses which dominated marketing attribution and effectiveness studies when data was limited (Dekimpe and Hanssens 2000) to individual-level data analyses that allows for much more granular targeting (Rossi, McCulloch, and Allenby 1996). In fact, one could argue that one of the big missions of a firm is to grow the number of rows (via customer acquisition, i.e., more unique IDs) and more transactions per customer with greater monetary value (per row). In retailing, the ability to track new customers and to link transactions over time is key. Loyalty programs (Kopalle, Kannan et al. 2012; Kopalle, Sun et al. 2012; Stourm, Bradlow, and Fader 2015), widespread today, are the most common way that such tracking exists; however, credit card, IP address, and registered user log-ins are also commonplace. Besides more rows, firms also have much better measures (columns) about each row which typically, in retailing, might include a link between customer transaction data from a CRM system, demographic data from credit card or loyalty card information, survey data that is linked via email address, and in-store visitation information that can be tracked in a variety of ways. If one includes social media data and more broadly user-generated content (UGC) which can be tracked to individual-level behavior, then customer-level data becomes extremely rich and nuanced.

Products

Product information in marketing, has and likely always will be, defined by a set of attributes and levels for those attributes which define the product. However, in today's data rich environment we see an expansion of product information on two-dimensions. First, this information may be available now for hundreds of thousands of SKUs in the store, making the data set about products have a lot of rows in it. Second, the amount of information about each product need not be limited now to