


Predicting Purchase Behavior of Website Audiences

Saar Kagan & Ron Bekkerman

To cite this article: Saar Kagan & Ron Bekkerman (2018) Predicting Purchase Behavior of Website Audiences, International Journal of Electronic Commerce, 22:4, 510-539, DOI: [10.1080/10864415.2018.1485084](https://doi.org/10.1080/10864415.2018.1485084)

To link to this article: <https://doi.org/10.1080/10864415.2018.1485084>




View supplementary material 



Published online: 17 Sep 2018.



Submit your article to this journal 



View Crossmark data 

Predicting Purchase Behavior of Website Audiences

Saar Kagan, and Ron Bekkerman

ABSTRACT: This paper proposes a methodological framework that extends the advantages of behavioral targeting while preserving the privacy of the individual. Instead of profiling individual users according to their general interests, we profile website *audiences* according to their online *purchase behavior*. This presents a trade-off between looser, aggregate audience profiling and deeper understanding of actual purchase behavior, the holy grail of online advertising. Our framework is based on the analysis of raw clickstream data of Web users who explicitly agreed to participate in an online audience panel. We experiment with data collected by an online analytics company, SimilarWeb, which consists of 3,463,796 records of online purchases and 1.1 billion records of Website visits. We train a multilabeled classification model on the clickstream of panel members with distinctive online purchase profiles to predict the purchase potential of the entire panel. We aggregate the individual purchase behavior profiles (both ground-truth and predicted) into purchase behavior profiles of Web domain audiences and test the resulting methodology on 3,408 Web domains, with very promising results. If privacy-related regulation tightens up in the near future, the proposed panel-based, purchase-focused ad targeting mechanism might be the panacea for online advertisers.

KEY WORDS AND PHRASES: Online targeting, online profiling, panel profiling, privacy preservation, behavioral targeting, online advertising, online behavior, online purchasing.

Introduction

The field of ad targeting has expanded significantly in the last few decades. From traditional media advertising (e.g., on print, radio, and television), where the advertiser struggles with a dearth of information on the consumer, to the realm of the Internet, where the advertiser is bombarded with an overabundance of information, the industry now encompasses a wide range of ad targeting models and related technologies. In traditional media advertising, most ad targeting models were and are still based on audience demographics. Statistics on markets and audiences are compiled from census data, customer surveys, psychographics, and other forms of consumer polling [14]. Using these statistics, segmentation models are developed to allow advertisers to understand their audiences and to target them [23].

Although demographic factors can often indicate an affinity to certain product categories and can be used as a proxy for purchase habits, they are very limited in their ability to accurately predict a person's actual product preferences and selections [12]. This is why in modern advertising, namely, online advertising, other methods are being used to target consumers. Two of the most popular methods currently employed are behavioral targeting [10] and re-marketing based targeting, otherwise known as

retargeting [19]. Behavioral targeting is the monitoring of a person's behavior online in order to build an interest profile, which is then used to advertise a range of products to them. Retargeting follows a person around the Internet, notes when he or she visits a certain site, and then advertises that same site to the person while he or she visits other sites. Both of these involve the use of third-party tracking technologies and have invoked questions in the public about the ethics and legality of these practices. Regardless of where one falls on the ethical and legal debate, it is hard to ignore the effectiveness of these targeting models in understanding consumer patterns and matching ads to them that are more relevant and personalized [2].

In this paper, we propose a methodology for predicting the purchase behavior of Website audiences. With this methodology, we create Website audience profiles, which can effectively be used by advertisers, marketers, analysts, and executives to understand the makeup of any Website's audience. Like the segmentation methodologies used traditionally by offline advertisers (e.g., Nielsen PRIZM, CACI Acorn, etc.) that help them understand the audience makeup of physical geographical markets, our methodology allows advertisers to understand the audience makeup of Websites. The segments are based on actual product purchases and therefore do not suffer from the deficiencies of relying on demographics alone. Instead, they leverage precise data on user browsing habits and purchase behavior in order to effectively understand Website audiences, all while respecting and protecting the privacy of any consumer who did not explicitly agree to be tracked.¹

Our model is similar to behavioral targeting in that it leverages information on the browsing behavior of individuals but is also different in the following ways:

1. Our model predicts product purchases while taking the ad targeting a step further by basing it on actual purchases rather than just on expressed interests.
2. Our model does not require tracking every individual user in order to serve them ads. Instead, it is based on data generated by a representative panel of users who agreed to share their browsing behavior.

The panel of users studied in this research (and the data they provide) is maintained by a company called SimilarWeb, which is a competitive intelligence and traffic analytics tool used by marketers, analysts, and managers to perform research and generate insights on digital properties (i.e., Websites and apps). SimilarWeb relies on a large panel of Web users (more than 50 million) to generate the statistics that the company provides in its tools and reports. The panel participants consciously give up their private data in exchange for a variety of benefits they receive online (mostly in the form of free tools and add-ons). We are very lucky and thankful to have been given access to this vast database (~1.3 petabytes) of raw clickstream data on which this research was based.

We applied Hadoop MapReduce methodology to sift through this massive database and extracted records of product purchases. The product purchases that we captured were made on Amazon.com² over the course

of one year (2013). After extracting all the purchases, we matched them to product categories using Amazon's Product Advertising API. We then mapped users in the panel to categories of their purchases, which served as labels for training our predictive model. Browsing behavior (in the form of clickstream data) of all panel users was processed to create feature vectors. The Gradient Boosted Decision Trees (GBDT) classifier was trained on the labeled feature vectors and then applied to predict purchase categories of users whose purchases were unknown. We combined the predicted purchases with the known purchases into the aggregated audience profiles. The audience profiles were created on a per-domain basis, with each profile consisting of purchase categories (known and predicted) of all the users who were observed visiting a particular domain.

A sample website audience profile for the domain gamepur.com, a gaming website, is given in Table A1 of the online appendix.

It can be seen that 35.8 percent of the visitors to the Website, as estimated by our model, are people who purchase products in the Video Games category. This is expected for a gaming Website. However, it is also interesting to see among the top 10 product categories of the Website's audience profile the categories Health & Beauty, Sports, and Apparel. This insight could lead advertisers—who would have otherwise overlooked this audience—to consider them a viable target for their products.

It is important to remember that advertising online is not just about detailed cost-per-click analysis and conversion optimization. A big swath of the online advertising world is concerned with brand exposure and more broad-stroked digital marketing campaigns. This is why it is important for advertisers, especially the larger ones, to understand the makeup of website audiences. It can be gleaned from the Website profile just presented, for example, that the visitors of gamepur.com do not just sit sedentary at home playing video games but also have an interest in purchasing health, sports, and apparel products. This could prompt a sports equipment/fashion brand such as Under Armour to engage with the gamepur audience, perhaps sponsoring content on the site, hosting a competition, or simply running banner ads.

In addition to ad targeting, another application for audience purchase profiles is within the reports provided by competitive intelligence and analytics tools such as SimilarWeb, ComScore, Quantcast, and Alexa. Their customers pay a subscription fee to access data on Websites, industries, and apps. These data are then used by marketing managers, media buyers, brand managers, business analysts, product managers, and executives to understand the audiences of these digital properties. Professionals on the marketing/advertising side of the spectrum use the data for selecting targets for campaigns, researching potential partnerships, and generating ideas for expanding their marketing reach, while product managers, analysts, and executives use the tools to understand their competition, their customers' use of these digital properties, and general industry trends. Of the four tools just mentioned, only two include some form of audience profiling *by product* (i.e., not just demographics). Quantcast includes a feature called Shopping Interests, which is based on offline purchases and covers only seven product categories. ComScore, on the other hand, offers a more comprehensive solution

called Segment Metrix 2.0, which integrates with 10 leading segmentation schemes and offers behavioral data based on its panel of users.³ By integrating the Website profiles generated by our methodology, these companies can provide segmentation data that are more relevant to their customers: not only the age and gender of audience members but also their lifestyle categories and, more important, their purchase behavior.

Let us emphasize that we can assess the audience purchase behavior for any Website as soon as users from the panel visit it, regardless of how many online purchases those panel participants made so far. In fact, they do not have to make any purchases at all; we predict their potential purchase behavior based on purchase data of other panel participants. Since the historical data of online purchases is hard to obtain, while the general clickstream data are plentiful, our purchase behavior prediction model is an enabler of large-scale purchase-focused ad targeting.

This paper's contributions are threefold:

1. We introduced a novel panel-based, purchase-focused ad targeting paradigm that preserves privacy of all Web users but those who explicitly agreed to have their clickstream analyzed.
2. We developed a methodology that implements the new paradigm. Our methodology employs machine learning to predict purchase behavior of all panel participants based on known purchase behavior of some participants. It aggregates purchase behaviors (known and predicted) of Website audiences to create Web domain purchase-focused profiles. The profiles may be used for audience-specific, purchase-focused ad targeting.
3. We applied the proposed methodology to a large-scale data set of millions of purchases (more than 1 billion clickstream events) and thousands of Web domains. We went through all the pitfalls of developing a large-scale machine learning system⁴ and empirically demonstrated its high quality and utility.

Literature Review

Relevant literature splits into two large bodies: (1) research on privacy in ad targeting and (2) Website audience characterization, with or without focusing on online purchases.

Ad Targeting and Privacy

Ad targeting plays a pivotal role in the online advertising industry. In an ever-increasing effort to raise revenues, the ad targeting technologies have been honed over time. As these technologies advance, however, concerns have been raised over privacy. A nationally representative survey conducted in [38] found that, on the whole, Americans reject tailored advertising and

the methods used to enable it. Of the respondents surveyed, 66 percent were “not OK” with advertising that was based on information about Websites they previously visited. Ninety-two percent agree that a law should exist requiring “Websites and advertising companies to delete all stored information about an individual, if requested to do so.” Seventy percent suggest that “if a company purchases or uses someone’s information illegally,” that company should be fined. Finally the study shows that the view commonly held by marketers that “young people don’t care” is not true, with 86 percent of young adults (ages 18–24) rejecting tailored advertising if it involves having their activity tracked on Websites other than the one they are visiting (i.e., third-party tracking).

These findings are corroborated by other studies that were conducted later. In 2010, McDonald and Cranor [22] conducted a survey that revealed that 68 percent of respondents find the idea of behaviorally targeted advertising invasive. Later that year, *USA Today* and Gallup conducted a poll asking if advertisers should be allowed to target ads based on specific interests garnered from Websites visited. Sixty-seven percent of respondents answered “No” [17]. In 2011, TRUSTe and Harris Interactive conducted interviews with 1,004 consumers and found that only 15 percent of them “probably would consent” to having their browsing behavior tracked to increase the relevancy of the ads that are shown to them [36]. Finally, a Pew Research Center study conducted on search engine use in 2012 found that 68 percent of adult Internet users surveyed did not like their online behavior tracked/analyzed and view online targeted advertising negatively [29].

Given the current state of affairs, it is understandable that privacy concerns are being raised over the use of certain ad targeting technologies in the online advertising industry. These concerns include a lack of transparency, a lack of control, and a general sense of being taken advantage of. Turow et al. [38, P. 24] speculated that much of the concern that surfaced in their study revolves around a worry that these practices “can lead to hidden forms of social discrimination.” For example, retail policies offer certain products and discounts to some people and not others. Conscious of this environment, our research has developed an ad targeting model that protects consumer privacy while still maintaining relevance.

The literature offers technological solutions for allowing privacy-preserving ad targeting. The majority of those solutions are based on ideas of local, client-side aggregation of private data, as opposed to server-side or cloud-based aggregation, which is common practice nowadays. Examples of client-based tracking systems are RePriv developed by Fredrikson and Livshits [13] and Privad by Guha et al. [15]. Bilenko and Richardson [5] focused on privacy-preserving keyword advertising and store user search profiles on the client side. Adnostic [35] went further into downloading and storing locally a large repository of ads such that the targeting is executed purely on the client side.

All these solutions are technologically challenging, expensive, and therefore hard to deploy in real-world systems. Moreover, client-side ad targeting cannot deal with user purchases; this would have been a severe privacy violation. We propose a solution that does not lead to client-side computations, encryption overheads, or high network communication burden. We

predict user purchase behavior on the server side, in an aggregative—and transparent—manner.

Website Audience Characterization

Over the past two decades, much research interest has been concentrated on online purchases. The theory distinguishes between purchase *intentions* and purchase *behavior* [25], as between the *will* to purchase a product and the actual purchasing *process*. Marketing literature proposes theoretical models of understanding [40] and predicting [16] online purchase intentions, neither of which is in the focus of our work: We do not study the way the users establish a new purchasing habit but rather observe existing habits.

Many previous works deal with modeling the actual purchasing behavior as a sequence of actions taken while the user is engaged with a (single) e-commerce Website, in a session that may or may not culminate with an online purchase. The motivation of those studies is clear: suggesting ways to increase the transaction convergence. Prominent works study Website usability [41], model the convergence as task completion [33], or predict the convergence [39]. Olbrich and Holsing [28] measured the effects of consumer search behavior, as well as social gestures (tagging, product rating, etc.), on the purchasing behavior of individuals. They investigate one shopping community site, “similar to Polyvore.” The connection to our work is that they also base their study on clickstream data. The difference is in the goal: We predict the categories of products that audiences of a variety of Websites are likely to purchase, for the purpose of better ad targeting and competitive intelligence.

All the aforementioned works employ controlled experiments in which hundreds of subjects are engaged in a purchasing activity, actual or intended. We, however, observe (but do not control) the online behavior of a panel consisting of millions of users, some of whom are involved in online purchases, while the others are not (or, more precisely, their online shopping behavior is unknown). We need to predict their purchase behavior based on the available purchase records of some users, with the goal of revealing the existing purchase habits of the entire panel—assuming that most of contemporary Web users purchase products online, or at the very least are *likely* to shop online for certain product categories.

To the best of our knowledge, the following two studies are closest to ours. In the first, Tsai and Chiu [37] developed a purchase-based segmentation methodology that uses product specific data on items purchased, along with their monetary value, to cluster together customers with similar purchasing behavior. They build on work done by Wedel and Kamakura [42] that distinguishes between “general variables” (such as demographics) and “product specific variables” and argues that characteristics of purchased products are more reliable for user segmentation than the characteristics of the users themselves. The main difference between our work and the work of Tsai and Chiu’s is that we infer the purchase behavior of users whose purchase behavior is unknown, while Tsai and Chiu segment only users

whose purchase behavior is known. Our approach is more useful in real-world settings, where ground truth on purchase behavior cannot be obtained for a large pool of online users.

In the second most relevant study, Wu et al. [43] use Probabilistic Latent Semantic Analysis, taken from the field of document clustering, to process the content of search logs and infer the purchase intent of search engine users. The ground truth of actual purchases is unavailable to Wu et al., so they have to stick to an unsupervised methodology, which is clearly inferior to classification. Also, their inference is based on search query logs that are much more sparse than the clickstream we use in our study. Finally, Wu et al.'s goal is to improve search engine ad targeting while we focus on content page advertising.

It is worth mentioning the work of De Bock and Van den Poel [11], who predicted demographics of website audiences. Their ground truth data are obtained through online surveys. The authors argued that the demographics split of the audience is an important factor in ad targeting. While not disagreeing with them, we go far beyond demographics (which is still limited in its effectiveness [12]) into the purchase behavior, with the goal to offer products that the audiences are actually likely to purchase.

Some Recent Developments

Surprisingly, recent e-commerce literature is less relevant to our work. The majority of recent studies of user purchase behavior are conducted over a single Website, being an online retailer [1, 26, 34] or a gaming platform [32]. A strong focus is put on the social aspect of online shopping [21] and on branding of e-commerce websites [18].

Su and Chen [34] clustered clickstream data of users of an e-commerce Website to infer the users' interest in specific product categories. We infer users' interests in product categories given the users' clickstream all over the Web. In fact, we can infer purchase interests of a user who is yet to access any e-commerce Website. Nishimura et al. [26] proposed a method for estimating product purchase probabilities from the clickstream of users of an e-commerce Website. We go further to estimating product purchase probabilities of entire audiences of any Website, even those that are not related to e-commerce. Baumann et al. [1] used a graph-theoretic methodology to predict whether a user browsing session on an e-commerce Website leads to a purchase. We, in contrast, focus on predicting user purchases in the long run, to improve the relevance of banner ads on any Website, without compromising the user's privacy.

Cross-platform studies of user behavior are rarer and—obviously—heavier on resources. To perform a cross-platform study, a researcher would need to recruit a user panel or team up with a company that maintains a user panel. An example of a cross-platform study is by Nottorf [27], who aims to find out which types of online advertisement are most effective in driving consumer clicks. Nottorf analyzed the data produced by a panel of 5,000 users. Our panel, in contrast, consists of more than 770,000 users. Moreover,

we focus on predicting online *purchasing* behavior, which is in the heart of the e-commerce industry.

Recent literature proposes methods for audience extension [31]/audience expansion [20] in online advertising as a way of leveraging existing audience selection criteria to effectively target larger audiences. In a sense, our approach is similar to audience extension, as we start with an audience of online shoppers and extend it into the global online audience. However, our methodology is fundamentally different, as we do not bound our audience with specific selection criteria in the first place so that the proposed audience extension methods are not applicable in our case.

Methodology

To predict the purchase behavior of Website audiences, we developed the methodology that takes clickstream data generated by members of a panel of online users, preprocesses the data to understand the purchase history and browsing behavior of those users, creates a classifier that can predict the purchase behavior of users whose purchase behavior is unknown, and aggregates all the purchase behavior (both known and predicted) into a Website audience profile. For any Website visited by members of an audience panel—regardless of their purchase data availability—our methodology allows to predict the categories of the products that the Website’s audience would potentially purchase online.

This chapter explains the methodology developed and is organized into the following sections:

1. Raw Data Acquisition
2. Data Pre-Processing
 - a. Purchase Data
 - b. User Selection
 - c. Clickstream Data
3. Feature Vector Construction
4. Classifier Training
5. Classification of Users
6. Aggregation of User Behavior

Each section represents a step taken in the methodological process. This process represents the bulk of the work performed in this research and can be replicated to profile the audiences of Websites.

The following is a high-level summary of each of the steps taken in the methodological process:

- The raw clickstream data are acquired and processed in two data sets: purchase data and domain visit data.
- The purchase data set is used to generate the labels (or the purchase history) for a particular set of users, and the domain visit data set is

used to generate the feature vector (or the browsing behavior) for that set of users.

- The two data sets are joined into a combined data set, which is used to train a machine learning classifier.
- The classifier, given a user with a browsing history, is trained to predict a probability distribution over the labels (i.e. the categories), which is the purchase behavior of the user.
- For a given domain, all users who visited that domain are categorized and aggregated so that a product category distribution (PCD) for that domain emerges.
- The PCD determines the audience profile for the given domain. The share of a category relative to all categories determines the probability with which the website audience is likely to purchase products of that category. This probability distribution can be used for ad placement decision making.

Table A2 of the online appendix summarizes the scope of the data used in this research. It shows the progression from raw purchase data to the final data set, which was used to train the classifier.

Raw Data Acquisition

The raw data that serve as the foundation of this research are a collection of clickstream URLs (plus associated metadata) that are gathered by the online analytics company SimilarWeb. SimilarWeb's product is a research and analytics platform that provides traffic statistics, analytics reports, and other Web usage trends that are used by its clients to perform market research, create competitive intelligence reports, prepare financial analyses, and investigate general Web industry trends. SimilarWeb gathers its clickstream data from a large panel of online users who share their browsing behavior via add-ons installed on their browsers.

SimilarWeb has a team of mathematicians and data scientists who process the raw clickstream data using statistics and machine learning algorithms in order to estimate traffic volume and other Web traffic metrics (such as bounce rate, time on site, etc.) for more than 30 million Websites on the Internet.

Clickstream data are "the electronic record of a user's activity on the Internet" [6, P. 36]. These data serve as the foundation for many research projects, including those that apply the data to Web advertising and marketing [8, 11, 35] and those that investigate the use of Websites [24, 28, 30]. In this research, clickstream data also serve as the foundation and are used to extract and analyze the browsing and purchase behavior of known users. These behaviors are used to predict the purchase behavior of unknown users and then to generate Website audience profiles. The Website audience profiles can be applied in online marketing and advertising, as well as in understanding Website usage.

In this study, each record of raw clickstream data represents a visit to a single page by a user who is part of the SimilarWeb panel. Each record is made up of several elements, including the URL of the page, the time that the page was accessed, the type of browser that was used to access the page, and so on.

The following is a definition of the data elements contained within each record of clickstream data:

TimeStamp: In UNIX epoch time

UserID: Unique identifier of panel participant

IP: Internet Protocol address of participant

CountryID: Country of participant

ClientUserAgent: Browser and OS of participant

RequestedSite: URL of the page that was requested by the browser

The elements used in this research are Timestamp, UserID, and RequestedSite. All personally identifiable information is removed during the first step of preprocessing. In later stages of this analysis, UserID is ignored since it is the aggregation of user behaviors into domain profiles that is interesting to us. No privacy of any person is violated during this work, and all data collection methods comply with SimilarWeb's terms of use and privacy policy.

Data Preprocessing

Purchase Data

The first set of data that are collected by SimilarWeb and processed in this research consists of all of the clickstream records in the SimilarWeb database that represent purchases made on Amazon.com over the course of one year.

To locate these records in the SimilarWeb database, we constructed a query that retrieves the clickstream records of Amazon's "thank-you pages." A thank-you page is a page that the user sees after having paid and confirmed a purchase on Amazon.com. It was found in a preliminary experiment (that we performed by purchasing a \$0.01 gift card) that Amazon's URL structure for thank-you pages is made up of two main elements: `amazon.com/gp/buy/thankyou` and `asins=`

The first element is the domain and path that indicates that the page is an Amazon thank-you page. The second element indicates a product included in the purchase, encoded via an ASIN (Amazon Standard Identification Number)—a 10-character alphanumeric value used by Amazon (and its

partners) to identify all items that are sold on Amazon.com, including Amazon Marketplace items. Note that a thank-you page URL can contain multiple asins= parameters, each indicating a separate product that was purchased.

As a result of querying the SimilarWeb database, approximately 1.2 million clickstream records of thank-you page visits were retrieved. These records represent purchases made by 770,155 users in the SimilarWeb panel.

For each row in the purchase data, the ASIN is extracted from the URL (from the RequestedSite parameter of the clickstream record) and is used to look up the product category of the product. The product category is queried using the Amazon Product Advertising API. This API is available to developers for free and is accessible on Amazon's affiliate program website. The API takes the ASIN as an input parameter and returns the Amazon category, which the product belongs to. Amazon maintains a list of more than 70 product categories (e.g., Home & Garden, Music, Sports, etc.) to which all of its items are assigned, including Amazon Marketplace products.

Once the category of a product is retrieved, it is associated with the purchase data record. If there are multiple products in the same purchase data record, then that row is duplicated once for each product. For example, if a user purchased four products in a single purchase, then four rows will be created each containing one ASIN and one product category, as well as the identical timestamp and other metadata.

After processing the purchase data to extract the ASINs from each clickstream record, separating them into distinct records, and looking up each one's product category, we end up with a data set comprising 3,463,796 records. Each record represents a single product. Table A3 of the online appendix shows an example of some products

User Selection

To select the users that are most appropriate for training the classifier, we group all the purchases in the purchase data by unique user and count the number of purchases each user made. This is done per category. For example, we find that user X purchased seven items in category A, five items in category B, and two items in category C. We then remove all instances from the data set where a user purchased less than three items in a category. In the preceding example, we remove the instance where the user purchased only two items in category C, keeping the two instances where the user made three or more purchases in categories A and B.

We do this because, when training a classifier, we wish to feed the prediction model only with cases in which the link between a user and their category is strong. When the classifier is trained, it is important that the users whose browsing behavior it is learning are those that "represent" the particular labels that it is learning about. For example, if the classifier is learning to identify people who purchase photography products, then it is important to provide it with examples of people whose number of photography purchases is not negligible. If the classifier was also trained using

instances where the number of purchases was small, then the behavior learned by observing those individuals may introduce noise into the prediction model and reduce its capability to predict accurately.⁵

Clickstream Data

As a result of user selection, a set of 215,867 unique users is created. For each one of these users, browsing behavior is retrieved. The browsing behavior of a user is represented as a collection of clickstream URLs. Since the raw data collected and used in this research span an entire year and are for a significant number of users (more than 1.1 billion records), the data have to be stored in the cloud. We choose S3, Amazon's Simple Storage Service, to store the clickstream data in a way that is convenient and accessible for processing.

To process the data, we use another Amazon cloud service known as Elastic MapReduce. Processing the data involves, for each user, running through the entire data set and counting the number of times the user visited each domain from a constructed list (see next). Elastic MapReduce, a big-data-processing framework, allows one to break up this task into multiple small tasks (that are run separately on commodity machines) and process it in parallel.

Feature Vector Construction

To construct the feature vector for each user, the entire user data set is scanned and the number of visits that the user made to a list of domains is counted. (The list of domains is a list of *top 10,000 domains* detailed next in the next subsection, which has the same name.) The feature vector for user X is a list of the 10,000 domains on which the number of times the user visited each domain is indicated. An example of a user's feature vector is in Table A4 of the online appendix.

The resulting feature vectors are processed using a development framework called GraphLab.⁶ GraphLab is a project (now owned by Apple) that allows one to process large data sets and perform data mining analyses. GraphLab uses a Python programming object called the SFrame, which is a dataframe (similar to the ones used in Pandas and R). The SFrame is employed to load and manipulate large quantities of data using a combination of memory and disk. GraphLab is also known for providing multiple machine learning algorithms to build predictive models such as the one developed in this research.

We transformed the feature vectors into their sparse representation, in the format of, for example, {'X47': 1, 'X187': 3, 'X890': 1, 'X7585': 1, ...}, which implies that the user visited domain X47 once, domain X187 three times, domain X7585 once, and so on. This sparse representation is much more efficient in terms of storage and memory and allows the GraphLab toolkit, when training the classifier, to process the feature vectors much faster.

Top 10,000 Domains

To construct a feature vector that can be relevant to all users in the research, it was decided to use a list of the top 10,000 domains on the Web. Every person browses a different set of Websites, and there is a subset of domains that many people visit in common. Therefore, a *standard set* of sites is chosen to act as a “common denominator” for all the users. All user visits to Websites outside the standard set were ignored.

Ten thousand Websites is a quantity that is large enough to encompass most people’s browsing habits while not venturing too far into the fringe and esoteric. The list of top 10,000 domains was obtained from SimilarWeb. The company offers a tool to its customers called Industry Analysis, which marketers, advertisers, analysts, researchers, and journalists use to investigate high-level industry trends. The Industry Analysis tool aggregates large amounts of Websites into categories (or industries), which can then be analyzed as a whole. For example, if looking at the Travel category (i.e. all sites that are categorized as Travel), it can be seen that traffic to that industry as a whole has risen by 5 percent, relative to the previous month.

One of the features of Industry Analysis is Top Websites. Top Websites ranks all the domains in a category by traffic: The domain that receives the most traffic in the category is ranked first, then the domain that receives the second-most traffic is ranked second, and so on. It is also possible to filter the list by geography, such as a list of the top 10,000 domains in the United States (i.e. visited by users who are browsing from the United States), a list of top 10,000 domains in Germany, or even a list of all the top 10,000 domains in the world. A design aspect of our methodology was to decide whether to use a list of the top 10,000 domains in the United States or a list of the top 10,000 domains in the world. This is not obvious, because the raw data collected for this research were not limited to a certain geography.

To facilitate this choice, and to support a data-driven decision, we performed a comparative analysis of the two lists with the top 10,000 domains obtained from the clickstream data of the selected users. It was found that the U.S. list had a significantly wider overlap (with 5,189 domains in common) with the raw data, while the worldwide list had only 2,846 domains in common. This is probably because the majority of people making purchases on Amazon.com are based in the United States.

Classifier Training

To train the classifier, the feature vectors are combined with the labels to create a combined dataset (see Table A5 of the online appendix). The combined data set consists of four columns: UserID, ProductCategory, Purchases, and Features. The only two columns that are used for training the classifier are the ProductCategory (label) and the Features (feature vector) columns. Before training the classifier, however, the data set is filtered to include only *selected product categories*.

Selected Product Categories

In the combined data set, 55 of the 71 product categories are present. Of these, 24 are selected. The 24 categories are selected by taking the top 31 categories⁷ (by number of purchases in the purchase data) and removing the following categories:

- Book
- Wireless
- Not Available
- Shoes
- Watch
- Single Detail Page Misc

The Book category is removed because it is overrepresented in the data. In the purchase data set, 434,661 of 3,463,713 purchases were books. This represents 12.5 percent of the data set, a disproportionately large number of purchases for a single category. Indeed, as Amazon's origins are as a bookseller and the company continues to be a very large seller of books. The remaining categories are removed because they are not relevant to any site in the list of *selected domains*.

Selected Domains

The list of selected domains is a list of domains (out of the top 10,000 domains) that are matched to a product category, which we call the *subject category*. The domains that are not on the list are the domains that could not be matched to a product category, and therefore the domain does not have a subject category. Out of the 10,000 domains in the top domain list, 3,408 are successfully matched to a product category. The reason for matching domains to product categories is for the process of *evaluating the effectiveness of our model*, which is discussed in greater detail next.

The process of assigning a product category to a domain consists of two steps, automatic labeling and manual matching.

1. **Automatic labeling.** The domain category of the domain is looked up using the SimilarWeb Website Categorization API. The API receives the domain as a parameter (e.g., cnn.com) and returns the domain's category. SimilarWeb maintains a list of categories.⁸ The list of categories may have changed as of the time of publishing. SimilarWeb categorizes domains using a machine learning algorithm that analyzes domains' content and relation to other.
2. **Manual matching.** Once a domain is assigned a domain category, a manual process of matching the domain category to a product category is employed. There are 220 domain categories, which we tried to assign to one of the 24 selected product categories. For example, the

domain category Sports/Winter Sports was assigned the product category Sports. If a domain category could not be assigned to a product category, for example, Travel/Tourism, then it was ignored.

Training Set and Test Set

After the combined data set is filtered to include only records that are labeled with one of the 24 selected product categories, the data set is split into a training set and a test set.⁹ The split operation is performed multiple times, each time the ratio of train to test is varied. We vary the train-to-test ratio because we wish to test the prediction model under varying conditions. Since in a production environment the amount of labeled data is unlikely to correlate with the amount of unlabeled data, it is important to measure the effectiveness of the model under varying conditions to see at which point the performance degrades too far to be useful, as well as how the performance varies over the various train-to-test ratios.

The following ratios of train to test are created:

- 40 percent train/60 percent test
- 20 percent train/80 percent test
- 10 percent train/90 percent test
- 5 percent train/95 percent test
- 2.5 percent train/97.5 percent test

In addition to creating five variants of train to test proportions, each variant is randomly restarted five times to achieve statistical significance of the results. Twenty-five training set and test set pairs are created.

Classification of Users

Classification of users is done using the GBDT classifier in GraphLab. The GBDT classifier is considered one of the strongest classification algorithms to date [7, 9].

The classifier is controlled by multiple parameters, which dictate how the model treats the data and how the algorithm performs the training task. Each parameter has a range of values, which impacts the performance of the classifier on the test data. To determine which parameter values yield the best results, a parameter tuning exercise was performed.

Parameter turning was implemented via a built-in function in GraphLab called `model_parameter_search()`. For each train/test ratio, a parameter search was performed over the training set only, where the classifier was trained and tested 50 times, each time slightly varying the parameter values. The values were varied for the following parameters:

- `column_subsample`
- `early_stopping_rounds`

max_depth
 max_iterations
 min_child_weight
 min_loss_reduction
 row_subsample
 step_size

For each train/test ratio, a set of best parameter values was determined and fixed. The classifier was then trained on the entire training set using the fixed parameters and evaluated over the test set, which was held out during the parameter tuning process.

In addition to the GBDT classifier, the Random Forest (RF) classifier was used as a baseline against which we compared the GBDT classifier's performance. Parameter tuning of the baseline classifier was conducted in the same way as it was for the GBDT classifier. The parameters tuned for the RF classifier were as follows:

column_subsample
 max_depth
 max_iterations
 min_child_weight
 min_loss_reduction
 row_subsample

Aggregation of User Behavior

To establish the audience profile of a domain, all of the product categories of the domain's visitors are aggregated into a PCD for that domain: The PCD of a domain is a normalized sum of the PCDs of all users who visited that domain. The PCD of a user is created by determining the *weighted purchases* of the user in each category of the 24 *selected product categories*.

Weighted Purchases

The weighted purchases W_{ijk} made in category i by user j who visits domain k are determined by multiplying the *probability of purchase* P_{ij} (for that user in that category) by the amount of times V_{jk} the user has visited the domain: $W_{ijk} = P_{ij} \times V_{jk}$. When aggregating users of a particular domain, only those n_k users who visited the particular domain are aggregated: $W_{ik} = \sum_{j=1}^{n_k} W_{ijk}$, and the weighted purchases are calculated for each of the categories in the *selected product categories*. The motivation behind this design choice is that a typical advertiser of a Web domain cares about the purchase behavior of frequent visitors to the domain. A user who visited a domain once is less representative for the domain than the user who visits the domain on the daily basis.

Probability of Purchase

The probability that a user will make a purchase in a given category is determined by one of two methods. The first method is for those users whose label is predicted, and the second method is for those users whose label is given.

- **Predicted Labels.** Users whose labels are predicted are those users whose labels are not known a priori and therefore determined by our classification model. When the classifier determines the label of the user, it calculates the probability within which it is correct. Each user is given a probability for the label that is assigned. A user can be assigned multiple labels, the probabilities of which compose a distribution.
- **Given Labels.** Users whose labels are given are those users whose label is already known (from the purchase data). These are the users who are included in the training set. There is no calculation necessary to determine the labels of these users. To determine the probability of purchase for these users, in a given category, the total number of purchases made by the user in the category is divided by the total number of purchases made by the user in all categories.

The PCD of a user is illustrated in table A6 of the online appendix.

The PCD of a domain is illustrated in Table A7 of the online appendix.

The PCD of a domain is simply a normalized sum of all of the values of all of the users in each of the product categories.

Measuring the Effectiveness of our Model

The effectiveness of the model is measured by observing the performance of the model in its ability to correctly determine the audience profile of a Website. The audience profile of a Website is represented by a distribution of product categories, where each product category represents a segment of the Website's audience. For example, the Health & Beauty category represents a segment of the audience that is most strongly associated with purchasing health and beauty products. A domain will have a PCD, generated by our model, which shows the relative size of each audience segment. For example, CNN.com has a PCD, which shows that 16.5 percent of the Website's audience tends to purchase Health & Beauty products, 10.9 percent purchase Toy & Hobby products, and 7.7 percent tend to purchase Sports & Outdoor products. The entire PCD is the audience profile of the Website.

To determine whether the model is actually able to discern this profile with reasonable accuracy, we select one of the product categories of the domain as a test category. We call this the subject category. The subject category for photo.net, for example, is Photography. (For a discussion on

how we determine the subject category of each domain, and which domains were selected, see the preceding Selected Domains section.) We examine the rank of the subject category, within the category distribution determined by our model. In the photo.net example, we find that the rank of Photography in the domain's PCD is 1. In the other words, the model determined, based on user purchases (both known and predicted), that Photography is the most popular category of products purchased by visitors to photo.net.

We compare this rank to the rank of Photography in the general product category distribution (GPCD), which is the PCD of all the domains in the data set combined (see Table A8 of the online appendix). The rank of Photography in the GPCD is 18: Of all the products that were purchased by all the visitors to all the domains in our data set, Photography was the 18th most popular category. This implies that if our model was not employed, a generic choice of a product category to be advertised on photo.net would highly unlikely be Photography, while our model makes Photography the most probable choice. The difference in rank of Photography in photo.net's PCD from the GPCD equals 17; we use this *difference in rank* to evaluate our model.

Despite that inferring the most probable product category for photo.net appears fairly straightforward, this is not trivial for domains without a strong product category affiliation, such as, for example, news feeds, blogs, and social networks. Which products should be advertised on, say, linkedin.com, the professional social network? Our model helps to make this decision.

Illustration of Performance Evaluation on a Small Subset of Domains

To get an initial understanding of the model's effectiveness we hand-selected 10 domains to evaluate. Each domain represents a different product category and was chosen just for the sake of an illustration. For example, bluefly.com is an online fashion/clothing store, and its subject category is Apparel. For each domain, we generated a PCD using the model, noted the rank of the subject category in the PCD, noted the rank of the subject category in the GPCD, and calculated the difference in rank. Table A9 of the online appendix shows the results of this preliminary analysis.

The initial investigation shows positive signals. In all of the examples in Table A9, the difference in rank is positive, with the exception of bluefly.com and apple.com; their difference in rank is 0. The rank of Bluefly's subject category cannot be positive since Apparel is the top rank in the GPCD. The difference in rank of Apple is 0 because the rank of its subject category remained unchanged in the domain's PCD.

While a large difference in rank can indicate that the model is more effective in determining the audience profile of a domain, it is not necessary for a large difference to exist. It is more important to emphasize the direction of the difference (i.e., positive or negative) rather than the magnitude. If the model determines that a product category associated with the domain (i.e., the subject category) is less represented in the PCD of that domain, then we

cannot conclude that the model was effective. For example, if the model shows that the relative size of the Photography segment of the audience of photo.net is smaller than it is in the general distribution, we must conclude that the model was ineffective in determining the audience profile for photo.net.

Evaluating the Performance of the Model

To test the overall performance of the model, we compiled a list of 3,408 selected domains (described in the Methodology section) that are evaluated individually and collectively. For each domain, a probability distribution over product categories is constructed, and the categories are ranked according to their probabilities. The subject category is examined, and its corresponding *difference in rank* is determined as the delta between its rank in the domain's PCD and GPCD. We then build the distribution of domains by the difference in rank to see how many domains have a positive difference in rank and how many have a negative difference in rank after application of our methodology. In addition, the *Average Difference in Rank* (ADR) was calculated as a measure of the performance of the model for all of the domains combined.¹⁰

The entire process is repeated for five variations of the prediction classifier. Each variation of the classifier corresponds to a different ratio of training set to test set (as described in the preceding Training Set and Test Set section).

The performance of the prediction classifier—and therefore the entire model—is expected to improve as the size of the training set increases. The charts in [Figure 1](#) show the distribution of domains by difference in rank for each of the five train/test ratios. With a training set of 40 percent, which is the highest of the classifier variations, a positive distribution is observed. The vast majority of domains have a positive difference in rank. Out of 3,408 domains tested, 2,308 have a positive difference in rank. Only 856 of the domains have a negative difference in rank. The average difference in rank for the entire data set is 2.61.

It is noticeable in this and in all the subsequent graphs that there is a significant peak at 9. Our investigation shows that the vast majority of these are domains with a subject category of Video Games. Since the rank of the product category Video Games is 10 in the GPCD, and our classifier was particularly effective on that category, many Video Games' domains moved nine spots to the number one position.

Also, we noticed that the peak at 2 is somewhat surprisingly high. Upon further investigation, we found that the Health & Beauty category is responsible for this result. The Health & Beauty category is ranked third in the GPCD, and for every domain with this subject category, our model determined Health & Beauty at rank 1; therefore the difference in rank of 2 was observed (see [Figure 2](#)).

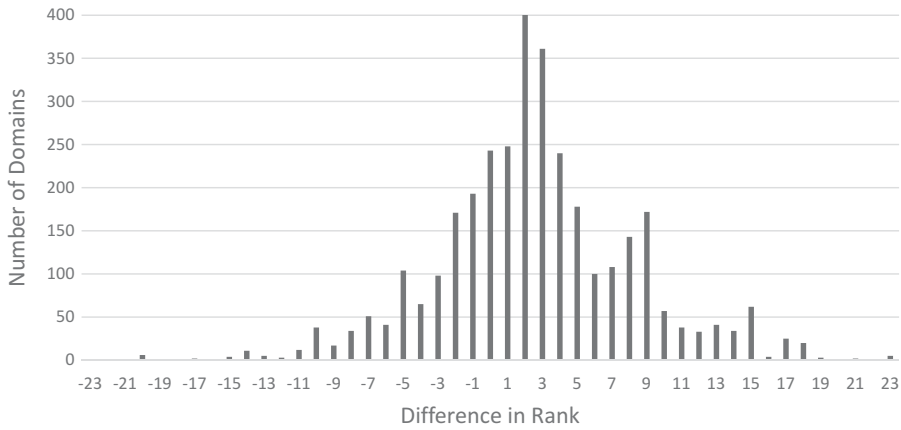


Figure 1. Distribution of Domains by Difference in Rank (40% train, 60% test)

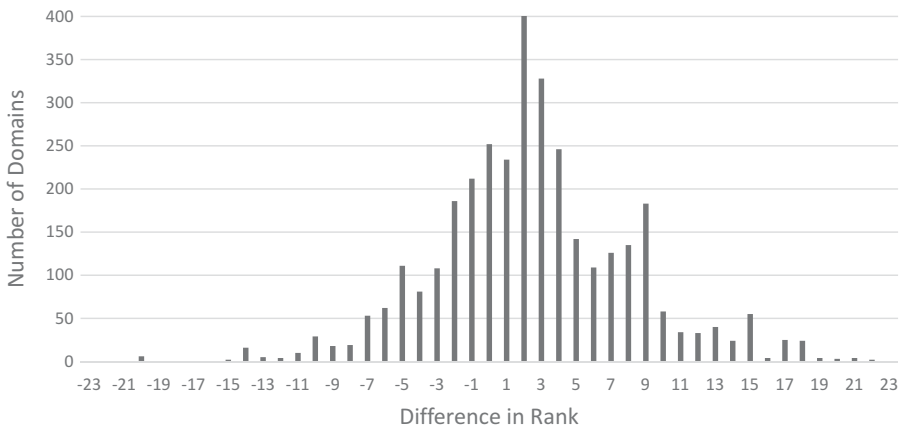


Figure 2. Distribution of Domains by Difference in Rank (20% train, 80% test)

At 20 percent training set the performance of the model degrades only slightly while still maintaining a large majority of domains with a positive difference in rank. Of the 3,408 domains in the data set, 2,233 domains result with a positive difference in rank, while 922 of the domains result with a negative difference in rank. The average difference in rank for the entire data set is 2.49 (see [Figure 3](#)).

The results for 10 percent training set show a significant degradation in performance from the 20 percent training set. The average difference in rank drops from 2.49 to 2.22. This is a larger drop in performance than was experienced when the training set reduced from 40 percent to 20 percent. Of interest, however, the proportion of domains with a positive difference in rank to domains with a negative difference in rank did not change very

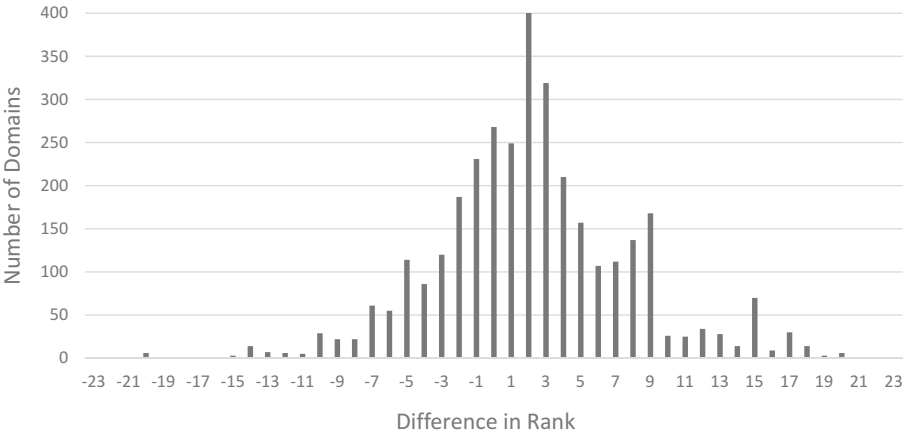


Figure 3. Distribution of Domains by Difference in Rank (10% train, 90% test)

much. The amount of domains with a positive difference in rank went down only by 62 to 2,171 domains, and the number of domains with a negative difference in rank increased only by 46 to 968 (see [Figure 4](#)).

The degradation in performance at 5 percent training set is not significantly large. The total number of domains with a positive difference in rank is 2,105. The total number of domains with a negative difference in rank is 1,005. This is not far from the results for the 10 percent training set. The average difference in rank for the 5 percent training set is 2.03. This is a surprisingly good result for such a small training set size (see [Figure 5](#)).

The test in which the performance degrades the most is at a training set size of 2.5 percent. Here the number of domains with a positive difference in

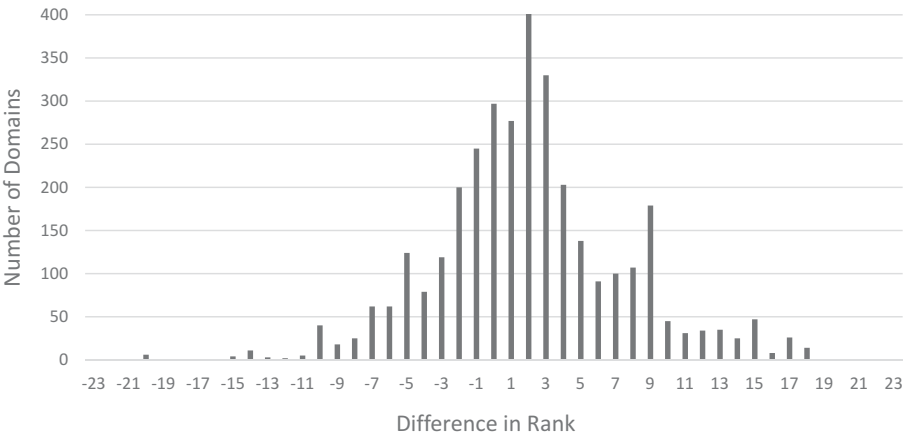


Figure 4. Distribution of Domains by Difference in Rank (5% train, 95% test)

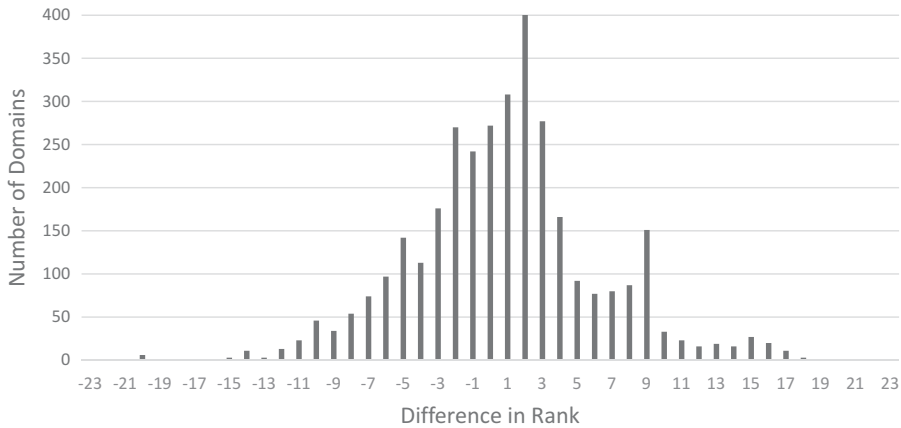


Figure 5. Distribution of Domains by Difference in Rank (2.5% train, 97.5% test)

rank declines to 1,828 while the number of domains with a negative difference in rank rises to 1,307. Also, the average difference in rank for the entire data set drops dramatically to 1.35, almost 34 percent less than the previous test. This shows the lower threshold of the model's performance. Given the tested training sets, 5 percent appears to be an appropriate minimum benchmark for operational performance. Table A10 of the online appendix summarizes the performance of each train/test ratio.

The improvement in performance of the model as the size of the training set increases is expected since the accuracy of the classifier most usually correlates with the training set size. In real-world conditions, only a small subset of the panel members has made purchases that can be logged and analyzed. We, however, cannot know a priori what percentage of panel users will have their purchases analyzed once our system is deployed. That is why we explore the entire spectrum of training/test set ratios and conclude that even with very small ratios (such as 5 percent) our model can obtain positive results.

Comparative Analysis of Our Model's Performance

Note that in our experimental data set each and every user has made purchases, so we can establish the gold standard of our model's performance based on the 100 percent training data: We aggregate the purchase behavior of all our users who visited a specific domain into that domain's profile. This would be the upper limit of our model's performance.

On the other hand, we can establish the baseline performance for each training/test split ratio, which would be based on the training set only, without the prediction aspect of our model. For the baseline, we aggregate purchase behaviors of only training set users, while simply ignoring the rest

of the users. We expect our model to perform within the window of baseline on the bottom and gold standard on the top. Next we refer to our model as “baseline plus predicted,” as it is based on aggregating known purchases of users from the training set together with predicted purchase behavior of users from the test set. By comparing our model’s performance with the baseline and with gold standard, we evaluate the predictive aspect of our model.

In addition to the ADR measure that we just used, we report on a standard measure called mean reciprocal rank (MRR): the reciprocal rank of a domain’s subject category 1 over the rank of the subject category in the domain’s PCD. For example, the reciprocal rank of the domain *bandcamp.com* is 0.25 because its subject category Music was ranked fourth in the domain’s PCD. The MRR of a data set is calculated by averaging the reciprocal rank of the subject category of all 3,408 domains in the data set. Informally speaking, ADR is an absolute measure of our model’s performance as it captures the deltas in the ranks, while MRR is the relative measure, taken as a fraction of the best result possible (rank 1 of the desired category). ADR ranges $[-N, N]$, where N is the number of categories, while MRR ranges between 0 and 1.

The following chart shows the performance of the baseline and baseline plus predicted models (using ADR), relative to each other and relative to the gold standard. All five training/test ratios are plotted on the chart. We compare the performance of GBDT classifier with the RF classifier that is considered a strong runner-up. We average all the results over five random restarts and show the standard error of the mean as an error bar for each training/test ratio (although some of the error bars are too small to be visible).

As can be seen in [Figure 6](#), both GBDT and RF are staying within the borders of the baseline from the bottom and gold standard from the top, which is very much predictable. While the performance of both classifiers is statistically indistinguishable at the 2.5 percent mark, starting from the 5 percent mark GBDT performs significantly better, and by reaching the 40 percent mark RF stops improving over the baseline. GBDT, in contrast, stays somewhat in the middle between the baseline and the gold standard over all the training/test set ratios, while the gap between the baseline and the gold standard reduces.

The [Figure 7](#) graph is the repetition of [Figure 6](#), while ADR is replaced with MRR. Here, too, GBDT behaves very well as predicted, staying in between the baseline and the gold standard. It is closer to the baseline at the 2.5 percent mark but manages to retain its middle position following this. RF, on the other hand, dips below the baseline at the 40 percent mark. This means that it is more beneficial to use 40 percent of users with available purchases to build domain profiles, over inferring the purchase behavior of the other 60 percent and aggregating it together with the available purchases. While we can only speculate about the cause of the RF degradation, we can say that RF is not considered as strong as GBDT, so this outcome is not surprising.

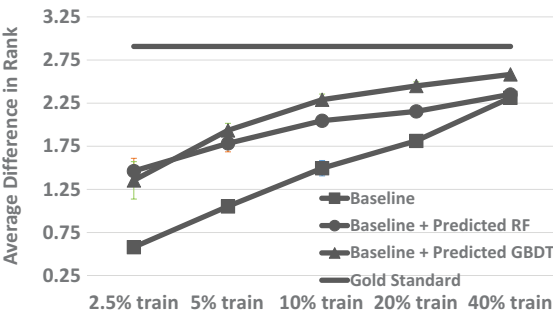


Figure 6. Performance of the GBDT classifier as compared to the RF classifier, measured in ADR

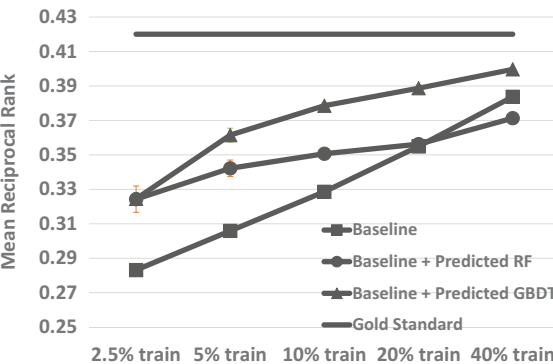


Figure 7. Performance of the GBDT classifier as compared to the RF classifier, measured in MRR

Since the GBDT classifier performed well over all versions of the training/ test set ratios, as measured by both ADR and MRR, while approaching the gold standard for high ratios, we can conclude about the good overall performance of the prediction element of our model, showing its efficacy in predicting the purchase categories of website visitors, and therefore our model’s general ability to profile website audiences.

Conclusion

Internet advertising and the related field of online ad targeting are constantly evolving. The industry, made up of many players including advertisers, ad networks, and advertising technology companies, is a large and vibrant realm. A lot of time, money, and human resources are spent on developing new data collection and targeting methods that increase the efficiency and effectiveness of this industry’s efforts. Behavioral tracking and retargeting have been shown in literature and practice to be successful models for

delivering relevant ads to consumers. However, a model proposed in this paper that explicitly looks at purchase behavior takes this one step further and removes the noise of behavioral indicators that are irrelevant, nonindicative, or entirely erroneous when it comes to the actual bottom line. Relying on concrete outcomes innovates by narrowing down to what is crucial for online advertisers and their clientele—the actual purchases of products. In this paper, we proposed a novel panel-based purchase-focused ad targeting paradigm and developed a machine learning model for it. Our model provides a solution that is both accurate and forward-thinking, and therefore must be highly sought after by the major players in the online advertising industry.

In the meantime, as concerns over third-party tracking and violations of consumer's privacy increase, these major advertising players need to seek ways to balance accurate targeting while respecting individual's privacy boundaries. Since these two imperatives are inherently opposed, an ad targeting model that uses real purchase data taken from a panel of online users and is used to deduce the audience segments of any given Website zooms out from the individual onto a broader and less personal whole. This has positive implications for stakeholders in the online industry. Individuals who participate in the panel can easily be incentivized to participate (as used to be done with traditional forms of media audience research such as TV rating systems) and are less inclined to feel morally infringed upon when they know that their individual browsing and purchasing behavior is responsibly used in aggregate for the purpose of estimation. In addition, the eeriness and discomfort of individuals who experience the current forms of tracking and retargeting is replaced by the more palatable and useful targeting that is driven by aggregated estimates and purchase-based audience profiling.

While being broader, our methodology is actually more relevant and beneficial to consumers—and by extension to advertisers—than existing approaches. By targeting ads based on audiences and audience segments, individuals may find themselves grouped into segments and product categories in which they may have not been associated with prior (i.e., if advertisers were targeting them based only on their individual browsing behavior). For example, a runner visiting a site that is considered, based on its audience's purchases, attractive to both runners and healthy eaters may find suggestions for food-related products that may not necessarily have been presented to her based on her typical browsing behavior.

The methodology presented here has been shown to be highly effective in determining the profile of a Website's audience. The study developed a predictive model that was trained on large quantities of clickstream data and then applied to much larger quantities of unlabeled data while profiling thousands of Websites, for some of which little to none training data were available. There are also limitations to this research. The purchase data used in this work were limited to the purchases made over the course of a year. It is recommended in further research that this be expanded to a larger time span. The more that purchase data are included in the model training, the more accurate the model can be. In addition, the diversity of the data can be

increased as well. Rather than just sampling data from Amazon.com, purchase data can theoretically be gleaned from multiple sources. It is worth investigating the URL structure of other e-commerce platforms to see if purchases can be ascertained from clickstream data of users who made purchases on those sites.

It is believed that if this work is expanded upon and further developed by vested businesses, the effectiveness and applicability of this model can be made even more far-reaching. The advertising, ad-tech, and online marketing industries, with their vast resources, access to large quantities of data, and highly motivated incentives to develop effective targeting models (which still respect user privacy), are welcomed and encouraged to take the foundation laid forth in this work and expand and improve upon it. As those industries are facing a regulatory threat of tightening the user privacy preservation policies, panel-based approaches appear to be a plausible substitute to modern behavioral targeting. Panels, in turn, provide tremendous flexibility for online advertisers to analyze sensitive data that would have never been available to them through any other channel. Since many Web users are willing to trade their online privacy for a variety of rewards, while some other users are becoming extremely guarded about their data privacy, using data of the former category to attract the latter category might be the only way for the online advertisement industry to overcome a potential regulatory shake-up and thrive into the future.

Acknowledgments

We thank SimilarWeb's team for giving us the opportunity to use their data in our research. We thank Turi's team for answering numerous technical questions. Ron Bekkerman thanks his wife, Anna, for her constant support.

NOTES

1. Note that those users who are not interested in participating in the panel will be able to browse the Web freely without being tracked, so their privacy will not be violated. It is expected, though, that many users will find benefits in trading their privacy off for various rewards—and those will be the users who will decide to participate in the panel. Currently, millions of people participate in online panels, and their number is likely to grow over time.
2. In the preliminary phase of this research, we discovered that it was possible to see in the Amazon URL structure the actual products that were purchased by visitors to Amazon.com. See the Methodology chapter for more details.
3. Relative to other solutions in the industry, ComScore is known to have a more thorough but comparatively small panel of users.
4. For a good review on implementation issues of the large-scale machine learning methodology, see [3].
5. While the cutoff parameter of three purchases was chosen rather arbitrarily, it is consistent with the data-mining and machine-learning practice of cutting off the long tail of power-law distributions to distill the predictive signal from surrounding noise (see, e.g., [4]).
6. <https://turi.com/>

7. The top 31 categories (of 71 total) account for 96.1 percent of all the purchases made in our purchase data set.
8. <http://www.similarweb.com/category>
9. The splitting is done uniformly at random using the packaged `random_split()` function in GraphLab Create.
10. ADR is the most intuitive measure of the model's performance: If it is positive, the model is helpful, whereas if it is negative, the model is harmful.

Supplemental File

Supplemental data for this article can be accessed on the publisher's website.

REFERENCES

1. Baumann, A.; Haupt, J.; Gebert, F.; and Lessmann, S. Changing perspectives: Using graph metrics to predict purchase probabilities. *Expert Systems with Applications*, 94 (March 2017), 137–148.
2. Beales, H. The value of behavioral targeting. *Network Advertising Initiative*, 1 (January 2010). Available at https://www.networkadvertising.org/pdfs/Beales_NAI_Study.pdf
3. Bekkerman, R.; Bilenko, M.; and Langford, J. (eds.). *Scaling Up Machine Learning: Parallel and Distributed Approaches*. New York: Cambridge University Press, 2011.
4. Bekkerman, R.; El-Yaniv, R.; Tishby, N.; and Winter, Y. Distributional word clusters vs. words for text categorization. *Journal of Machine Learning Research*, 3 (March 2003), 1183–1208.
5. Bilenko, M.; and Richardson, M. Predictive client-side profiles for personalized advertising. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM, 2011, pp. 413–421.
6. Bucklin, R.E.; and Sismeiro, C. Click here for Internet insight: Advances in clickstream data analysis in marketing. *Journal of Interactive Marketing*, 23, 1 (February 2009), 35–48.
7. Chapelle, O.; Shivaswamy, P.; Vadrevu, S.; Weinberger, K.; Zhang, Y.; and Tseng, B. Multi-task learning for boosting with application to web search ranking. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM, 2010, pp. 1189–1198.
8. Chatterjee, P.; Hoffman, D.L.; and Novak, T.P. Modeling the clickstream: Implications for web-based advertising efforts. *Marketing Science*, 22, 4 (November 2003), 520–541.
9. Chen, T.; and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international Conference on Knowledge Discovery and Data Mining*. New York: ACM, 2016, pp. 785–794.
10. Chen, Y.; Pavlov, D.; and Canny, J. F. Large-scale behavioral targeting. *Proceedings of the 15th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining*, ACM, 2009, pp. 209–218.

11. De Bock, K.; and Van den Poel, D. Predicting website audience demographics for web advertising targeting using multi-website clickstream data. *Fundamenta Informaticae*, 98, 1 (January 2010), 49–70.
12. Drozdenko, R. G.; and Drake, P. D. *Optimal database marketing: Strategy, development, and data mining*. Thousand Oaks: Sage, 2002.
13. Fredrikson, M. and Livshits, B. Repriv: Re-imagining content personalization and in-browser privacy. In *2011 IEEE Symposium on Security and Privacy*. Piscataway, NJ: IEEE, 2009, pp. 131–146.
14. Goss, J. “We know who you are and we know where you live”: The instrumental rationality of geodemographic systems. *Economic Geography*, 71, 2 (1995), 171–198.
15. Guha, S.; Cheng, B.; and Francis, P. Privad: Practical privacy in online advertising. *USENIX Conference on Networked Systems Design and Implementation*, 2011, 169–182. Available at https://www.usenix.org/legacy/events/nsdi11/tech/nsdi11_proceedings.pdf#page=179.
16. Hansen, T.; Jensen, J.M.; and Solgaard, H.S. Predicting online grocery buying intention: a comparison of the theory of reasoned action and the theory of planned behavior. *International Journal of Information Management*, 24, 6 (2004), 539–550.
17. Jones, J.; and Saad, L. *USA Today/Gallup poll*. Available at http://www.gallup.com/file/poll/145334/Internet_Ads_Dec_21_2010.pdf. Accessed on December 9, 2017.
18. King, R.C.; Schilhavy, R.A.; Chowa, C.; and Chin, W.W. Do customers identify with our website? The effects of website identification on repeat purchase intention. *International Journal of Electronic Commerce*, 20, 3 (2016), 319–354.
19. Lambrecht, A.; and Tucker, C. When does retargeting work? Information specificity in online advertising. *Journal of Marketing Research*, 50, 5 (2013), 561–576.
20. Liu, H.; Pardoe, D.; Liu, K.; Thakur, M.; Cao, F.; and Li, C. Audience expansion for online social network advertising. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM, 2016, pp. 165–174.
21. Lu, B.; Fan, W.; and Zhou, M. Social presence, trust, and social commerce purchase intention: An empirical research. *Computers in Human Behavior*, 56 (March 2016), 225–237.
22. McDonald, A.; and Cranor, L. F., *Beliefs and behaviors: Internet users’ understanding of behavioral advertising*. Available at <https://ssrn.com/abstract=1989092>. Accessed on December 9, 2017.
23. Mitchell, S. Birds of a feather. *American Demographics*, 17, 2 (February 1995), 40.
24. Moe, W.W.; and Fader, P.S. Capturing evolving visit behavior in clickstream data. *Journal of Interactive Marketing*, 18, 1 (2004), 5–19.
25. Morrison, D.G. Purchase intentions and purchase behavior. *The Journal of Marketing* (April 1979), 65–74.
26. Nishimura, N.; Sukegawa, N.; Takano, Y.; and Iwanaga, J. A latent-class model for estimating product-choice probabilities from clickstream data. *Information Sciences*. in press.

27. Nottorf, F. Modeling the clickstream across multiple online advertising channels using a binary logit with Bayesian mixture of normals. *Electronic Commerce Research and Applications*, 13, 1 (February 2014), 45–55.
28. Olbrich, R.; and Holsing, C. Modeling consumer purchasing behavior in social shopping communities with clickstream data. *International Journal of Electronic Commerce*, 16, 2 (December 2011), 15–40.
29. Purcell, K.; Brenner, J.; and Rainie, L. Search engine use 2012. Available at http://www.pewinternet.org/files/old-media/Files/Reports/2012/PIP_Search_Engine_Use_2012.pdf. Accessed on December 9, 2017.
30. Sen, A.; Dacin, P.A.; and Pattichis, C. Current trends in web data analysis. *Communications of the ACM*, 49, 11 (November 2006), 85–91.
31. Shen, J.; Geyik, S.C.; and Dasdan, A. Effective audience extension in online advertising. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM, 2015, pp. 2099–2108.
32. Shi, S.W.; Xia, M.; and Huang, Y. From Minnows to Whales: An empirical study of purchase behavior in freemium social games. *International Journal of Electronic Commerce*, 20, 2 (December 2015), 177–207.
33. Sismeiro, C.; and Bucklin, R. E. Modeling purchase behavior at an e-commerce web site: A task-completion approach. *Journal of Marketing Research*, 41, 3 (August 2004), 306–323.
34. Su, Q.; and Chen, L. A method for discovering clusters of e-commerce interest patterns using clickstream data. *Electronic Commerce Research and Applications*, 14, 1 (February 2015), 1–13.
35. Toubiana, V.; Narayanan, A.; Boneh, D.; Nissenbaum, H.; and Barocas, S. Adnestic: Privacy preserving targeted advertising. *Proceedings Network and Distributed System Symposium*, (2010).
36. TRUSTe & Harris Interactive. Privacy and online behavioral advertising. Available at <https://www.eff.org/files/truste-2011-consumer-behavioral-advertising-survey-results.pdf>. Accessed on December 9, 2017.
37. Tsai, C.Y.; and Chiu, C.C. A purchase-based market segmentation methodology. *Expert Systems with Applications*, 27, 2 (August 2004), 265–276.
38. Turow, J.; King, J.; Hoofnagle, C.J.; Bleakley, A.; and Hennessy, M. Americans reject tailored advertising and three activities that enable it. Available at <http://ssrn.com/abstract=1478214>. Accessed on December 9, 2017.
39. Van den Poel, D.; and Buckinx, W. Predicting online-purchasing behaviour. *European Journal of Operational Research*, 166, 2 (October 2005), 557–575.
40. Van der Heijden, H.; Verhagen, T.; and Creemers, M. Understanding online purchase intentions: Contributions from technology and trust perspectives. *European Journal of Information Systems*, 12, 1 (March 2003), 41–48.
41. Venkatesh, V.; and Agarwal, R. Turning visitors into customers: A usability-centric perspective on purchase behavior in electronic channels. *Management Science*, 52, 3 (March 2006), 367–382.
42. Wedel, M.; and Kamakura, W. A. *Market Segmentation: Conceptual and Methodological Foundations*. Boston: Springer, 2000.
43. Wu, X.; Yan, J.; Liu, N.; Yan, S.; Chen, Y.; and Chen, Z. Probabilistic latent semantic user segmentation for behavioral targeted advertising. In

Proceedings of the SIGKDD International Workshop on Data Mining and Audience Intelligence for Advertising. New York: ACM, 2009, pp. 10–17.

SAAR KAGAN (saarkagan@gmail.com) is an independent consultant, helping companies with systems analysis, solution architecture, and product management. Between 2013 and 2016 he was a Senior Product Manager at SimilarWeb, a leader in online traffic measurement and analytics. Previously, he worked for MyHeritage, Nielsen, and Accenture. He holds a B.A. in Sociology from UCLA and an M.A. in Data Science from the University of Haifa, Israel.

RON BEKKERMAN (ron.bekkerman@gmail.com; corresponding author) is an Assistant Professor of Data Science at the University of Haifa, Israel. His research interests are in applying Big Data methodologies to a variety of social and natural sciences such as Law, History, Linguistics, Psychology, Ecology, Epidemiology, and others. Previously, he was CTO of CandorMap, Chief Data Officer of Carmel Ventures, Senior Research Scientist at LinkedIn, and Research Scientist at HP Labs. He holds a Ph.D. in Machine Learning from the University of Massachusetts, Amherst. Dr. Bekkerman has coauthored more than 20 papers published in leading venues, such as *PLOS ONE*, *Journal of Machine Learning Research*, *CVPR*, *SIGKDD*, *ICML*, *WWW*, *SIGIR*, *IJCAI*, and *EMNLP*.