Bus Inf Syst Eng 61(4):413–431 (2019)
https://doi.org/10.1007/s12599-018-0528-2

RESEARCH PAPER

The Price of Privacy

An Evaluation of the Economic Value of Collecting Clickstream Data

Annika Baumann · Johannes Haupt · Fabian Gebert · Stefan Lessmann

Received: 4 April 2017 / Accepted: 20 December 2017 / Published online: 21 February 2018
© Springer Fachmedien Wiesbaden GmbH, part of Springer Nature 2018

Abstract The analysis of clickstream data facilitates the understanding and prediction of customer behavior in e-commerce. Companies can leverage such data to increase revenue. For customers and website users, on the other hand, the collection of behavioral data entails privacy invasion. The objective of the paper is to shed light on the trade-off between privacy and the business value of customer information. To that end, the authors review approaches to convert clickstream data into behavioral traits, which we call clickstream features, and propose a categorization of these features according to the potential threat they pose to user privacy. The authors then examine the extent to which different categories of clickstream features facilitate predictions of online user shopping patterns and approximate the marginal utility of using more privacy adverse information in behavioral prediction models. Thus, the paper links the literature on user privacy to that on e-commerce analytics and takes a step toward an economic analysis of privacy costs and benefits. In particular, the results of empirical experimentation with large real-world e-commerce data suggest that the inclusion of short-term customer behavior based on session-related information leads to large gains in predictive accuracy and business performance, while storing and aggregating usage behavior over longer horizons has comparably less value.

Keywords Predictive analytics · e-Commerce · Privacy · Behavioral targeting · Clickstream data

Accepted after two revisions by Prof. Dr. Suhl.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s12599-018-0528-2) contains supplementary material, which is available to authorized users.

A. Baumann (✉) · J. Haupt · S. Lessmann
Chair of Information Systems, Humboldt University of Berlin, Spandauer Straße 1, 10178 Berlin, Germany
e-mail: annika.baumann@hu-berlin.de

J. Haupt
e-mail: johannes.haupt@hu-berlin.de

S. Lessmann
e-mail: stefan.lessmann@hu-berlin.de

F. Gebert
Akanoo GmbH, Mittelweg 121, 20148 Hamburg, Germany
e-mail: fabian@akanoo.com

1 Introduction

The e-commerce sector is steadily growing and estimated to have reached $1.915 trillion of sales turnover worldwide in 2016 (eMarketer 2016). With customers increasing spending, web usage mining has been established as a common practice by e-shops to offer website visitors an enhanced user experience and to better understand customer behavior (Cooley et al. 1997). The underlying data are collected in the form of clickstreams, which might include information such as the pages visited and the time spent on each page (Senécal et al. 2005). Clickstream data is seen as one of the top value adding data sources by businesses (Statista 2016a) with applications in online marketing, customer analysis, or website development. Within online marketing, clickstream mining has been readily adopted by business and academia to understand the behavior of website visitors. Use cases of individual-level clickstream data include customer targeting (e.g., Pai et al. 2014), understanding navigational preferences (e.g., Montgomery et al. 2004), and predicting customer conversion (e.g., Buckinx and Van den Poel 2005). But since no good comes without harm, the collection of user data

www.manaraa.com

always brings with it the possible hazard of privacy related issues, which pose ethical and economic risks to both customers and companies.

The informational privacy of website visitors is of concern for e-shops because the success of converting the visitors into customers (or lack thereof) depends, amongst other things, on the potential risks of the transaction as perceived by the visitor (Metzger 2004). From a user perspective, perceived risks of privacy exist in the form of third-party access to personal information, misuse of exposed information, unconsented secondary use of provided information and unintended mining or mapping of individual behavior (Dinev et al. 2013). The perceived risk of e-commerce transactions can be mitigated and user decisions positively influenced by increasing trust in the website through comprehensive privacy protection (Kim et al. 2008; Nofer et al. 2014).

One way to improve perceived privacy is to avoid use of user data unless it has been provided willingly by the customer (Liu et al. 2005). Clickstream data on the other hand is collected without action or consent by the website visitor. Consequently, privacy concerns do not only include existing customers who need to provide their sensitive personal information to complete the buying process, but also prospective customers who are anonymous and have not actively provided consent for the use of their data. In addition, online advertising companies such as DoubleClick collect user data in form of clickstream across the users' whole browsing history, combining several data sources and therefore intervening with their privacy in order to offer them the most fitting advertisements based on their aggregated website visits and search engine requests (Akrivopoulou and Stylianou 2009, p. 125).

In general, privacy preserving data collection and analysis has been in the center of attention of big data research (e.g., Agrawal and Srikant 2000). However, as we detail in Sect. 3, there is still a lack of research focusing on the collection of clickstream data and the prediction of customer behavior under the restriction to simultaneously maintain a certain level of privacy. We argue that the collection of customer data is a strategic business decision and needs to be evaluated according to its marginal gain in relation to incurred risks and costs by managers and customers alike. Since the amount and type of data collected and stored is in the control of the e-shop and clickstream data is dispensable for the direct operational sales processes, the strategic question is what level of privacy in data collection is suitable to maximize sales performance under minimum risk exposure.

To answer this question, we review approaches to convert raw clickstream data into behavioral traits, which we call clickstream features, and identify groups of clickstream features based on their relevance for privacy issues. We then examine the economic value of clickstream features from different privacy categories through the lens of predictive modeling. In particular, we consider an e-commerce context and assume a company to gather clickstream data with the intention to predict customer behavior. Accurate behavior predictions can, for example, inform the company's marketing strategy and, more generally, aid in achieving growth targets. Drawing upon the literature on cost-sensitive learning, we link the economic value of clickstream data to the accuracy of a behavior prediction model. This allows us to quantify the marginal profit gain associated with employing a set of clickstream features and the opportunity costs of refraining from using these features, respectively.

So far, existing research considering the privacy aspect of clickstream data collection has focused on whether several data sources (Padmanabhan et al. 2006) or a larger amount of data comprising a longer observation period (Stange and Funk 2015) yield advantages in predictive accuracy. We contribute to existing literature by focusing on what kind of clickstream features need to be included in a predictive model to obtain sufficiently accurate conversion predictions based on empirical evidence for two e-shops. Additionally, we provide an economic analysis of the privacy-accuracy trade-off to inform managerial decision-making. For example, we show that the inclusion of short-term clickstream data derived from session-related information leads to large gains in targeting accuracy, while long-term-based clickstream features over several sessions facilitates only a marginal gain in accuracy and value for the observed shops.

The remainder of the paper is organized as follows: Sect. 2 discusses the background and motivation of our work. Section 3 reviews related literature. Section 4 explains our methodology. Empirical results are presented and discussed in Sect. 5. Section 6 concludes the paper, states limitations, and gives an outlook for future research.

2 Background and Motivation

In this section, we will discuss the concept of clickstream data collection in more detail and highlight its relevance for privacy-related aspects.

In general, the collection of user data on the Internet occurs in two distinct ways. Internet users may provide information actively and consciously, e.g. by creating a user account or by conducting a transaction where process completion requires the provision of personal information. They also pass information passively as a byproduct of visiting webpages in that every visit – or 'click' – leaves a digital footprint that is stored in web server logfiles and, in conjunction with subsequent page visits, provides what is

www.manaraa.com

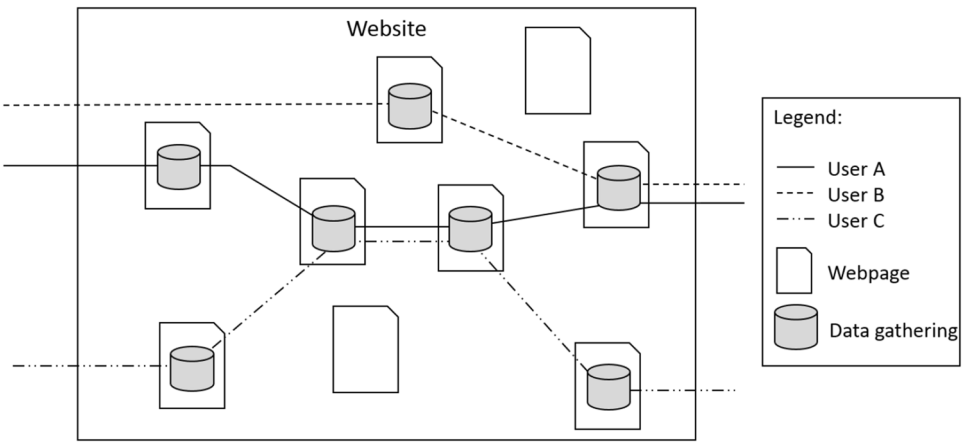

Fig. 1 Three examples of the clickstream data collection process

called clickstream data (Skok 2000). Clickstream can be defined as a "record [which contains information about] the Internet service provider, the type of computer and software used, the website linked from, the amount of time spent perusing each page, and exactly what parts of the website were explored and for how long" (Solove 2001). Therefore, clickstream might not only include information about the path a user has taken through the website but also details about the interaction with the website in form of click-, scroll-, tab switch and basket events. Additionally, user agent data transferred with the clickstream such as the access device, browser information and screen resolution can be derived from it.

Figure 1 shows an example of three sessions of users visiting a website and how clickstream is collected in this process. A session is comprised of a series of webpages visited by a user that is terminated when no interaction takes place for a specified duration. In this example, each user takes a different path through the website. At each single traversal through a webpage data is gathered in the form of clickstream. For example, when user A first visits the website the overall visit count is set to one and the number of webpages visited is updated at each page traversal, i.e. the webpage count is set to one when visiting the first webpage, set to two when the second webpage is visited and so on until the user leaves the website. Informational bits can also be continued based upon historical clickstream data. For example, once user A re-visits the website the visit count is then updated to two.

Personal data is protected under the aspect of informational privacy, which is defined as "the individual interest in avoiding disclosure of personal matters" (Lin 2002, p. 1094). Informational privacy has become especially relevant in the new area of the Internet and information technology where the collection and processing of data became beyond measure. In general, the collection and use of Internet user data is regulated in different ways depending on the country. For example, regulation in the US is sparse, while the European Union requires websites to obtain the user's permission regarding cookie placement and informing them whether data is collected and how it is used (Baumer et al. 2004). Legal restrictions define data security standards for certain types of data according to the sensitivity of the information, e.g. anonymous, personally identifiable information (PII), or medical data.[1] Data is considered as personally identifiable when a connection between the data and an individual is possible with reasonable effort. Such PII might be for example an e-mail address, a name, telephone number or other identifiers such as a social security number (Lin 2002). Clickstream is not classified as PII but still poses privacy threats such as potential de-anonymization, secondary use of data, unknown extent of data collection and the possibility to combine non-PII clickstream data with personal data (Sipior et al. 2011; Pollach 2011).

In this regard, many Internet users are not aware of the information they transmit while browsing and what kind of data is collected by whom (Hoofnagle et al. 2012). They are left with the feeling as if they "lost all control over how personal information is collected and used by companies" (Turow et al. 2009). Users who are not registered or logged into a site can be considered as anonymous by choice. Nevertheless, their clickstream data is collected and used to track their behavior when visiting a website. From a shop owners' perspective, motivations to do so include developing user profiles, for example to inform marketing actions. Given that anonymous visitors have not agreed to the collection and use of their clickstream data, they may

[1] For example, see the *Health Insurance Portability and Accountability Act of 1996* or the California *Online Privacy Protection Act of 2003* for the US or the *General Data Protection Regulation* for EU regulation.

www.manaraa.com

hold a certain "expectation of privacy in clickstream data" (Skok 2000).

Clickstream data can constitute a severe threat with respect to website visitors' privacy. For example, clickstream data has been shown to facilitate the de-anonymization and access to personal information of users through revealing URLs (Libert 2015; Greis 2016). Since clickstream data contains the URLs of the webpages a user has visited, it is possible to track what is of interest for a specific user. Here, strongly sensitive information such as personal preferences or healthcare information can be revealed when this specific information is part of a URL (Libert 2015). Furthermore, URLs which contain account access information such as an e-mail address, being classified as PII, could be also revealed through non-hidden URLs (Greis 2016). In the e-commerce setting the URLs of e-shops might reveal in what kind of sensible products a user might be interested in and could possibly combined with personal information among log-in.

In addition, the long-term observation of behavioral user patterns can be used to de-anonymize users by matching recurring visit and page interaction patterns, collected in the past or on other websites, to an anonymous visitor (Yang 2010). Here, behavioral patterns such as the specific journey the user takes on the website, how long she stays on specific pages and where click and scroll events take place might be an indication who is visiting the website through the match of reoccurring patterns. In this case a user can even be de-anonymized when cookie deletion takes place since no identifier in the form of an ID is necessary. However, this approach is only applicable when a lot of data is available (Yang 2010). Another method, which can constitute a threat to online user privacy is browser fingerprinting. Research shows a high success rate of browser (re-)identification on the basis of user agent information (e.g., Eckersley 2010; Nikiforakis et al. 2014). Here, specific information about which browser a website visitor uses in combination with the underlying version is often so unique that single users can be identified based upon the user agent information collected altogether with the clickstream data. Since no long-term observation and no revealing URLs are necessary, this can be seen as the most obtrusive approach.

These cases illustrate how the collection of clickstream data may impede user privacy. More specifically, they show how raw clickstream data can be converted into features that characterize and potentially predict user behavior, which can be considered an invasion of user privacy in itself. In combination with increasing privacy awareness by consumers, data privacy statements have also become a part of trust-related marketing communications for companies (Bansal et al. 2015). Consequently, management has an incentive to reflect the degree to which they collect and store sensitive customer data.

3 Related Research

Using clickstream data as a means to predict a specific object of interest has been widely adopted in the literature. Possible prediction targets include the likelihood of customer churn (Moertini and Ibrahim 2015), user personalization approaches (e.g., Pai et al. 2014), or the prediction of purchase behavior and conversion (e.g., Buckinx and Van den Poel 2005). We provide an overview on relevant literature in the field of conversion prediction from two perspectives which are the features used for prediction and in what regard the privacy aspect in relation with clickstream data has been considered by literature so far.

3.1 Conversion Prediction and Clickstream Features

This section will give a detailed overview of features extracted from raw clickstreams to predict conversion as a basis for our own set of clickstream in Sect. 4.1. We focus on previous work related to conversion modeling because purchase prediction is one of the most common fields in prior literature. Furthermore, since conversion (e.g., a purchase) occurs on a single website, clickstream data collection and privacy are under direct control of the site owner; as opposed to online advertisement, where data collection routinely involves third party providers such as ad networks (e.g., Stange and Funk 2014).

Table 1 provides an overview on related literature focusing on the features used for predictive modeling. We group those features into classes depending on whether they belong to clickstream data or additional information. Clickstream features are further sub-grouped into the more fine-grained categories Page, Time, Monetary, Page Interaction and User Agent. Furthermore, we highlight those papers which have a focus on one of the main topics of our paper which is whether they cover a privacy and/or a profit analysis.

Existing research in predictive modeling made use of a number of clickstream features which we group into five categories. The first three, *Page*, *Time* and *Monetary*, are based on the well-known concept of recency, frequency and monetary value analysis (Zhang et al. 2015). *Page* combines data related to the path a website visitor traverses and how often specific pages or page categories have been visited. *Time* contains information about the time spent on each page or aggregated page categories. *Monetary* collects outcomes of historical and present purchase behavior. The monetary value of the purchase can be taken from

www.manaraa.com

Table 1 Overview of focus, feature categories and time horizons used in research for conversion prediction (alphabetically ordered)

References	Privacy focus	Profit/ business value	Feature horizon		Feature category					
					Clickstream					Additional information
			Current session	Across session	Page	Time	Monetary	Page interaction	User agent	Demographics
This paper	x	x		x	x	x	x	x	x	
Banerjee and Ghosh (2001)			x		x	x				
Chan et al. (2014)				x	x			x		x
Iwanaga et al. (2016)				x	x	x				
Jiang et al. (2012)			x		x	x				
Lee et al. (2010)		x		x	x	x				
Lu et al. (2005)			x		x					
Moe (2003)			x		x	x				
Moe and Fader (2004)				x		x	x			
Moe et al. (2002)		Lift	x		x	x				
Padmanabhan et al. (2006)	x	Lift		x	x	x	x			x
Park and Park (2015)			x		x					
Pitman and Zanker (2010)			x		x			x		
Sarwar et al. (2015)			x		x	x	x			
Sato and Asahi (2012)				x	x		x			
Senécal et al. (2014)			x		x	x		x		
Sismeiro and Bucklin (2004)		x		x	x	x		x		
Stange and Funk (2015)	x			x	x	x	x	x		
Suh et al. (2004)		Lift	x		x	x				
Buckinx and Van den Poel (2005)				x	x	x	x			x
Vroomen et al. (2005)		x		x		x	x	x		x
Wu et al. (2005)			x		x					
Zhao et al. (2016)				x	x	x	x			
Zheng et al. (2003)		Lift		x	x	x	x			x

clickstream data as a form of basket value. Aggregated amounts of basket values in the clickstream data sum up to historical purchase information of website visitors. The fourth category, which we call *page interaction*, includes variables related to basket actions (e.g., an item is placed in the basket during a session of a user), click, scroll and tab switch events. The last clickstream feature category, *user agent*, consists of information related to the access device, browser, screen resolution and IP-resolved location. The additional category *demographics*, which does not belong to clickstream data, but is included in our survey to derive a comprehensive picture of the feature categories used in literature. Demographics contains, for example, data related to gender, income and education of a user. We use these five categories to classify prior work in conversion modeling in terms of the employed data (Table 1).

Furthermore, with respect to the temporal reference of the clickstream features, we note that varying time horizons of clickstream features have been in the focus of research, which is also depicted in Table 1. The data used can be solely based on the current user session or alternatively can contain information across sessions, capturing historical information in terms of earlier website visits and purchases.

As shown in Table 1, nearly all prior studies consider features belonging to the page category, whereas most of them additionally include time-related information. Monetary-related features, features capturing direct

www.manaraa.com

interactions with the website and demographics are still used fairly often. However, user agent is a category which has been used only to be able to combine aggregated demographic data with available clickstream data (Chan et al. 2014) but not as specific feature category. However, we will consider this feature category for our prediction task.

From a feature horizon perspective, several studies also consider the use of a broader time horizon of information by collecting data over a longer period to add historical session and purchase information (e.g., Buckinx and Van den Poel 2005; Sismeiro and Bucklin 2004). In general, the literature is almost equally divided into studies that use only information with respect to a current website visit and studies that, in addition, use historical data related to earlier website visits and purchases. The broad use of feature categories and varying time horizons supports the relevance to consider privacy-related aspects since the broader the more information used and the longer the time horizon considered, the more privacy severe an approach might be.

3.2 Conversion Prediction and Privacy

We will next take a detailed look at all papers considered and examine those which are closely related to our work in that they raise privacy issues in combination with assessing the linked business value.

The second column *Privacy Focus* in Table 1 depicts whether prior work on clickstream-based conversion modeling makes reference to user privacy. Of all papers considered, only two papers raise privacy issues at all. Moreover, while some studies examine the link between the accuracy of a behavior prediction model and its business performance (see column *Profit/Business Value*), the potential trade-off between performance and privacy has eluded research. In appraising this result, it is important to note that some studies assess predictive accuracy using the lift measure. Although they do not investigate the business performance of their models (e.g., in terms of costs and revenues), it is possible to relate lift, under certain assumptions, to profitability (Masand and Piatetsky-Shapiro 1996). To acknowledge that at least an implicit link to business performance exists in corresponding work, we highlight usage of the lift measure in Table 1.

This section will present the two studies which are closest related to this work in that they raise the issue of privacy, which are Padmanabhan et al. (2006) and Stange and Funk (2015). The former consider the privacy aspect of user data in terms of the trade-off of using a single data source compared to using data collected across several websites. Cross-site data provides a more comprehensive picture of user behavior. However, such data is normally only available via acquisition from third-party vendors. To that end, the authors define features that are either user- or site-centered. While site-centered data only uses information from a single data source, i.e. one website at a time, and is therefore more privacy preserving, the user-centered approach captures the behavior of website visitor across all websites in their dataset. The authors investigate the prediction accuracy of their tree-based model with regard to three different dependent variables: conversion during the session, conversion during any consecutive session, and return website visit. Using the lift measure and predictive models based on all available data, the authors show that the user-centric approach always outperforms the privacy-friendly site-centric approach. However, since third-party data is expensive to obtain, it is often not a sensible option for e-commerce websites to have complete information for all visitors across websites. A fraction of 45% of user-centric data is necessary to build a model which is able to outperform a site-centric model trained on all available data. Therefore, including only a small extent of privacy adverse information on browsing behavior across several websites might reduce prediction accuracy compared to using comparatively privacy friendly single site data only.

Stange and Funk (2015) examine how sample size affects the predictive accuracy of a clickstream prediction model. This relates to privacy in that gathering larger samples requires companies to collect data over longer horizons, and thus act in a relatively more privacy adverse manner. Using 1-month data of two online retailers, they find that including only 1% of all available clickstream data is already sufficient to predict the likelihood of conversion with satisfying accuracy. Despite looking at privacy from the perspective of the amount of data needed, their dataset still contains several privacy-harming features such as the link between advertisement and website interaction of a single user.

While existing papers with a relation to the privacy aspect of clickstream data collection focus either on the amount of data needed (Stange and Funk 2015) or whether collecting data across multiple websites exhibits benefits (Padmanabhan et al. 2006), we focus on understanding what kind of data from a single data source is sufficient to obtain accurate conversion predictions. Therefore, our research contributes to the existing literature in three ways. First, we define privacy categories for site-centric clickstream data. Second, we investigate the incremental benefits of successively including more privacy adverse data into a predictive model to understand the informational gain of the identified categories. Third, since the collection and usage of clickstream data is a business decision, we consider a specific use case to analyze the monetary value of the different privacy-relevant feature subsets.

المنارة للاستشارات
www.manaraa.com

Table 2 Description of our defined settings with varying privacy horizons on the dimensions of clickstream feature category and time horizon

Privacy Relevance	Setting	Description	Clickstream Feature Category					Feature Horizon	
			Page	Time	Monetary	Page Interaction	User Agent	Current Session	Across Session
Lower ↑ Site-centered	Session Content	...uses only information of the current session of a user related to page visited and time spent on page.	x	x	x			x	
	Session Behavior	...considers interactions with the website with respect to basket, click, scroll and tab switch events.	x	x	x	x		x	
	Cross Session	...contains information spanning a longer time horizon over all current sessions of the observation period.	x	x	x	x			x
Higher ↓ User-centered	Identifiable	...contains user agent related information such as IP-resolved location, screen resolution, access device and software.	x	x	x	x	x		x

A detailed overview of the features contained in each setting can be found in the appendix

4 Methodology

This chapter discusses the construction of the different clickstream feature sets based on the features' risk to data privacy and clarifies our predictive modeling methodology.

4.1 Definition of Privacy Settings and Feature Extraction

To grasp the connection between clickstream data, its utility for website owners, and threat for user privacy, we categorize clickstream features into groups according to the severity with which they might invade privacy. The categorization is based on the time horizon and user-centricity of the data during site visits. In general, privacy risk increases with the amount and dimensions of data, which, in turn, increase with the time horizon over which a visitor is monitored. For example, gathering clickstream data for a specific visitor over one session is less severe than monitoring this visitor's behavior for multiple sessions. Surveys show that around 69% of Internet users do not delete their cookies at least on a monthly basis, making it easy to re-identify the majority of revisiting website users (Statista 2016b; comScore 2007). Hence, the horizon of clickstream data gathering is one determinant of the severity of its privacy impact.

In addition to the time dimension, privacy implications vary with the type of data being collected. This is especially relevant since we will focus on website visitors who visit a shopping website anonymously (i.e., without registration). In general, the more data is available, the more holistic the picture of a visitor and the more conclusions about future behavior can be derived from the data (Bennett et al. 2012). Therefore, the richness of data collection and information extraction is a second factor that we consider in our clickstream feature categorization.

In particular, we consider website-centric data as less privacy intrusive than user-centric data. The former is related to information such as the webpages a user has visited, whereas we define user-centric data as information related to user agent and page engagement in the form of basket actions, click- and scroll events. Drawing upon the two determinants of potential privacy issues, monitoring time horizon and data richness, we propose four categories of clickstream features, which we label *SessionContent*, *SessionBehavior*, *CrossSession*, and *Identifiable*. Table 2 summarizes those feature sets where we provide an indicator of privacy relevance on a high-level basis and a description of each feature set. Furthermore, we adopted the classification approach of Sect. 3 and summarize the kind of information contained at the specific privacy level, i.e. features of the current and all less critical levels of privacy.

www.manaraa.com

Table 3 Summary of the datasets of both shops used in the empirical study

Summary	Shop 1		Shop 2	
	Tablet	Computer	Tablet	Computer
Users	2055	9247	1463	6087
Sessions	10,947	51,349	11,171	47,087
Views	120,845	585,570	182,726	631,277
Purchases	744 (6.80%)	4112 (8.01%)	538 (4.82%)	1968 (4.18%)

Table 4 AUC values for predictive models build on the feature sets separately (left) and the incrementally increasing feature set (right)

Feature set	AUC			
	Sets separately		Incremental extension	
	Shop 1	Shop 2	Shop 1	Shop 2
SessionContent	0.797	0.759	0.797	0.759
SessionInteraction	0.781	0.758	0.801	0.765
CrossSession	0.760	0.763	0.832	0.801
Identifiable	0.535	0.528	0.834	0.803

Table 5 Basket abandonment rates (in %) for each step in the purchase process (four steps in case of shop 1, three steps in case of shop 2)

	Purchase step				
	0	1	2	3	4
Shop 1 (%)	89	61	38	24	13
Shop 2 (%)	86	58	39	25	–

We argue that privacy benefits stem from smaller observation periods and more site-centric features. In the *SessionContent* setting, we only include page, path, category and basic time (i.e., time on a webpage and session duration) and monetary (i.e. monetary amount in basket) related information of a session. These features are based on information directly available through the browser requests and consequent page views. In other words, features belonging solely to the current session are our information baseline, which contains the least privacy invasive information.

The *SessionBehavior* setting is related to click and scroll events, basket actions and the time spent in total on different page categories (e.g. on product and shopping basket pages). The setting is defined as more privacy intrusive since the interaction of the client on the page is observed in addition to the page visit itself. Interaction with an e-commerce webpage hints at an interest for a specific product, for example via basket actions or if several click/scroll events on a specific page signal a strong interest in the information displayed on that page. Research has shown that mouse movements map to a certain extent the gaze movement of a website visitor therefore hinting at the relevance of a specific page (Guo and Agichtein 2010a; Rodden et al. 2008). Furthermore, website interactions in form of click and scroll events have been shown to provide a stronger signal with respect to a purchase intention compared to only using content-based information (Guo and Agichtein 2010b). In addition, website interaction information can be used for re-identification purposes (O'Connell and Walker 2014). Nevertheless, privacy risks are reduced by the time frame, which is still restricted to one session.

The *CrossSession* setting contains features related to all preceding site visits within a two-month period, implying that the monitoring horizon is larger compared to previous settings. Tracking a user over multiple sessions implies that data needs to be stored over the full time period, increasing both the risk of misuse and the amount of data at risk, thus leading to a stronger privacy impact. By storing session information and connecting it via a user identifier, e.g. through cookies or browser fingerprinting, long-term user profiles can be constructed. The observation of a longer time span of user behavioral patterns of website visits and

Table 6 Campaign revenue matrix for the coupon campaign setting

	Actual decision	
	Purchase planned	No purchase planned
Prediction		
Purchase/no coupon	r	0
No purchase/issue coupon	r − c	p × (r − c)

www.manaraa.com

Table 7 Derived cost matrix for the coupon campaign setting

	Actual decision	
	Purchase	No purchase
Prediction		
Purchase/no coupon	0	$-p \times (r - c)$
No purchase/issue coupon	$-c$	0

interactions give a stronger indication of the intention of the users' website visit (Bennett et al. 2012). Additionally, this enables user profiling and in the end to match the behavioral patterns over time with the identity of the website visitor (Yang 2010). Here, the sequences of and time spent on webpages visited by a user is used to detect re-occurring patterns in user navigational behavior so that the yet unknown web session can be assigned with a certain probability to a specific user. This approach is even more intrusive since it can refrain from traditional tracking techniques in the form of cookies, where the website visitor can control and hinder the tracking attempts through the regular deletion of cookies from the system. Instead, via the large-scale observation of user behavior over time, sessions can be matched to particular website visitors without the definitive need of technological identifiers. Features in this category include, for example, aggregates such as the overall number of page views of a single user, her mean time spend on a page, or differences in interaction patters of the current visit compared to previous visits (e.g., time on page compared to this user's mean time on page).

Defined as our most privacy intrusive setting, the category *Identifiable* contains user agent information such as IP-resolved location and details related to browser, access device and screen resolution. Clickstream data in itself can be collected anonymously and restricted to the tracking of a user within one session, e.g. by the deletion of cookies. To facilitate behavioral pattern matching for user identification, a certain observation horizon is necessary to obtain reasonable results (Yang 2010). However, browser fingerprinting can be used to recognize and track online users by the setup of their system (e.g., Eckersley 2010; Boda et al. 2012), since it is transmitted automatically with a page

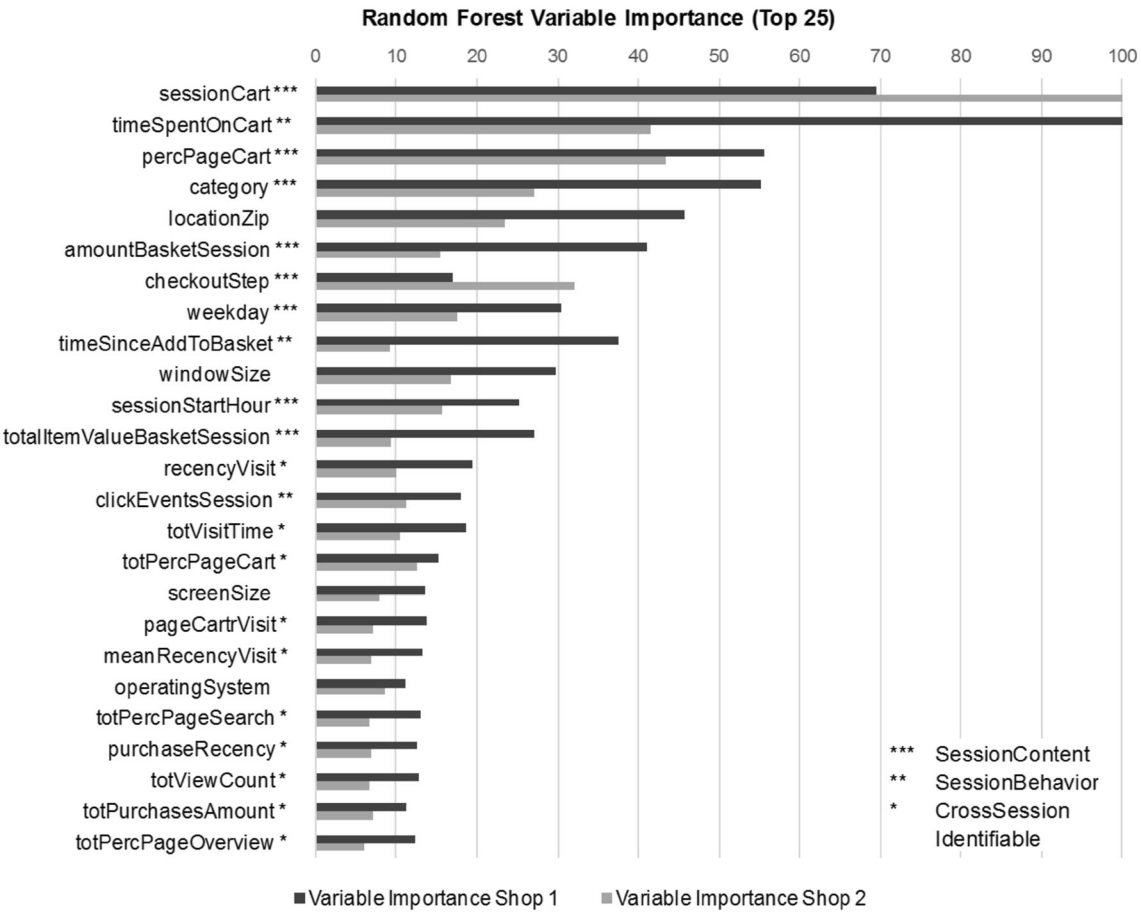

Fig. 2 Random forest variable importance for the 25 most important variables ordered according to their average relevance for both shops

www.manaraa.com

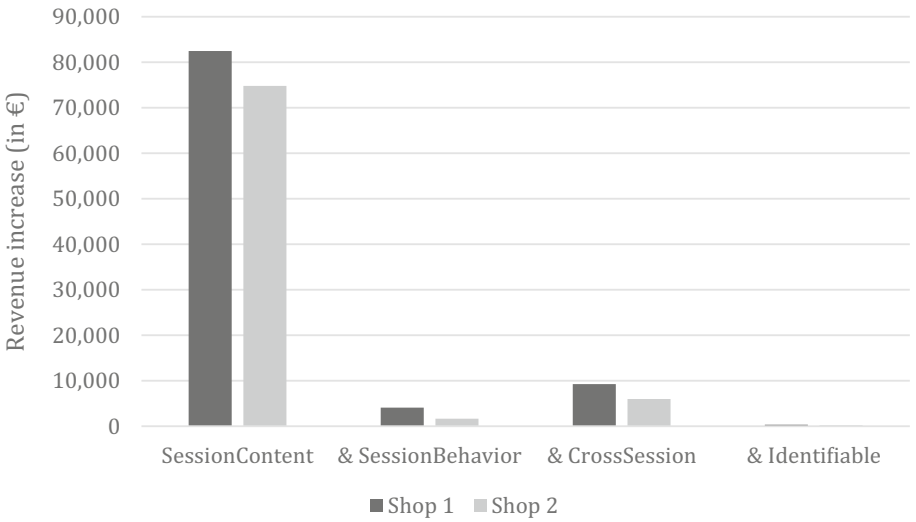

Fig. 3 Revenue increase as compared to campaign without customer targeting (all/no coupons), averaged over coupon success rates from 1 to 5%

request. Features of this setting include data on the customer's access infrastructure, e.g. device type, device brand, operating system, and browser, location and update recency of the user's system. They also include the IP address and the inferred location of the user. Since this data can be used to locate and identify the user in context, we classify these features as potentially personally identifiable information and most privacy concerning.

To facilitate the prediction of user behavior, raw clickstream data is converted into clickstream features. In total, we derive a set of 84 clickstream features. The full list of features used can be found in the appendix.

4.2 Predictive Modeling

This chapter describes our predictive modeling approach in terms of algorithms used and the specific set-up to clarify how our derived clickstream feature sets influence predictive accuracy. Not gathering any clickstream data might be most desirable from a privacy perspective. However, this clearly conflicts with website owners' business goals and their interest to gather data for user behavior prediction. To clarify the trade-off between respecting user privacy and collecting informative data, we approximate the value of clickstream data in a predictive modeling context.

We train prediction models to estimate the purchase probability for each visitor in the current session after each click given the features described above. This method is known as "clipping at every click" (Van der Meer et al. 2000; Senécal et al. 2014). More specifically, we aim at predicting whether an e-shop visitor will make a purchase in her current session (e.g. Padmanabhan et al. 2001), where one user session comprises multiple page views. If a purchase is made at any point within a session, we define the target variable to be positive for all page views in that session; and as negative otherwise. We then estimate a statistical model that classifies each new observation, i.e. each page view of a user on the website, into one of the two categories "user will purchase during this session" and "user will not purchase during this session". We predict purchases at the page view level since marketing stimuli like e-coupons can be offered at any point in a session and therefore require page-level granularity.

In formal terms, we face a binary classification problem with groups purchase/no purchase. Several machine learning algorithms are available to estimate classification models (e.g., Lessmann and Voß 2010) and the analysis of model performance and variable importance is dependent on the model choice. For the purpose of this study, we select the random forest algorithm due to its prevalence in practice and good track record in many applications (Kuhn and Johnson 2013) and because we observe it to perform best in terms of overall prediction error when compared to other models on our data. We determine model performance by pretests comparing the predictions of random forest, C5.0 and gradient boosting including parameter tuning on the full data.

Random forest is an ensemble algorithm. It combines hundreds of decision trees build on subsamples of the observation and feature space to ensure diversity among individual trees. Each tree is a sequence of binary splits of the data that maximize class purity in leaf nodes. For each observation, the random forest model estimates the probability of it to belong to class *purchase* by the ratio of trees that predict this class.

To build the model and test it on unseen observations, we split our data into a training and test set consisting of data from August and September 2015, respectively. We

www.manaraa.com

Table 8 Simulation results for asymmetric cost in terms of total campaign revenue (left) and relative gain compared to the next less sensitive data subset (right)

Conversion rate	Total campaign revenue (in 1000€)					Net gain to less sensitive set (in 1000€)				
	1%	2%	3%	4%	5%	1%	2%	3%	4%	5%
Shop 1										
Naïve: all/no coupons	2612	2612	2612	2694	2780	–	–	–	–	–
SessionContent	2629	2679	2736	2804	2874	17.2	66.6	124.5	109.8	94.3
& SessionBehavior	2629	2683	2743	2808	2879	0.2	4.5	6.8	4.3	4.8
& CrossSession	2633	2690	2753	2821	2892	3.6	7.1	9.6	12.9	13.1
& Identifiable	2633	2690	2753	2821	2893	0.1	− 0.3	0.3	0.7	0.8
Shop 2										
Naïve: all/no coupons	2665	2665	2720	2820	2921	–	–	–	–	–
SessionContent	2668	2744	2829	2917	3007	2.9	78.8	108.9	96.8	86.6
& SessionBehavior	2667	2746	2831	2920	3010	− 0.8	1.9	1.7	2.4	3.1
& CrossSession	2680	2752	2835	2923	3014	13.2	6.3	3.8	3.0	3.6
& Identifiable	2679	2751	2835	2924	3015	− 1.0	− 1.2	0.4	1.3	1.4

estimate the model from the training data and use the test data to assess predictive accuracy. Prior to that, we perform fivefold cross-validation on the training data to identify suitable values for the algorithm parameters (i.e., the number of decision trees in the random forest and the number of randomly selected variables per tree split), as is standard practice in predictive modeling (e.g., Kuhn and Johnson 2013).

Using the above approach, we examine how the accuracy of purchase prediction models varies with the clickstream features they embody. Therefore, model selection, training, and testing are conducted separately for each of the clickstream feature subsets. More specifically, we consider two experimental settings. First, the models are trained on each feature subset in isolation. Results of this setting provide an estimate of the predictive value contained in the underlying features. Second, we train models on an incrementally increasing set of features, where more privacy adverse features are added in each step. For example, we start with training a model using only the least sensitive features of the *SessionContent* setting. Next, we add the features of the *SessionBehavior* setting and train a second model. We continue the incremental addition of feature subsets in the order of privacy adverseness until all feature subsets are considered. Taking both the incremental expansion of feature sets and the test of feature sets in isolation into account, we assess random forest models based on a total of seven distinct feature sets (i.e. four times each individual subset plus three times incrementally developed subsets), each with a different number of features and degree of sensitivity with regard to customer privacy.

This modeling approach allows us to quantify the marginal value of adding features of a presumably more comprehensive, but also more privacy adverse category. However, we also acknowledge a limitation of our approach, namely that it disregards any additional value that clickstream features and gathering the corresponding raw data, respectively, provide to the website (i.e., shop) owner beyond facilitating predictive modeling. Our justification for concentrating on predictive modeling is twofold. First, analysis of the user journey and product-centered browsing behavior can largely be performed on aggregated clickstream data. Thus, gathering clickstream data at the individual user level is likely dispensable for strategic site management tasks. Second, the individual purchase and personal information used in most business intelligence applications is provided willingly by users upon registration or purchase. Unlike the anonymous visitors who we focus on in this paper, registered customers have explicitly agreed to further data collection.

5 Empirical Results

Adopting the methodology discussed above, we analyze the performance of the classifier for the defined subsets of features in two ways. First, we employ statistical performance measures to create a comparable benchmark of the general predictive power of each feature set. To complement the accuracy assessment, we approximate the economic value of different models. While being specific to one application context (e.g., specific cost and revenue consideration) and thus less general, we consider the economic analysis to add useful insight from a managerial perspective.

www.manaraa.com

5.1 Data and Data Preparation

As a basis of our study, we obtain clickstream data from two large European online retailers of two e-shops selling apparel and shoes, respectively. Both shops are comparable in size as measured by the number of users, sessions and views within the two-month observation period as described in Table 3. The clickstreams include desktop and mobile users who accessed the shop websites in a two-month period from August to September 2015. As part of data preparation, we exclude the first three page views within each session, since we assume constant coupon success probability throughout the profit analysis (see Sect. 5.4). An empirical analysis of the data shows very high exit rates for these views suggesting that a large number of visitors does not enter the page with a strong intention to interact. In case of both shops, around 55% of website visitors leave the website after only three webpage views. High exit rates further suggest a strong dependency between the current view and coupon success probability, which would require explicit modeling of redemption rates for the profit simulation beyond the scope of this paper. From business perspective, coupon marketing on the first three page views amplifies the number of played coupons at a very low redemption rate. This is generally not in line with company expectations due to concerns about customer price expectations regarding the availability of coupons and the brand image.

Based upon the empirical analysis of the distribution of the length of sessions, we further deleted sessions with more than 500 page views each under the assumption that such sessions come from bots (Banerjee and Ghosh 2001). However, this affected only one session in both datasets. Furthermore, since we include clickstream features that span several user sessions, we select customers with at least four visits during the two-month period. This is to ensure that all experimental settings have access to the same observations. Table 3 summarizes the resulting data, which includes 120,554 unique user sessions from 18,852 customers with 1,520,418 page views in total. The overall conversion rate is about 6.11% (7362 sessions).

5.2 Analysis of Predictive Performance

In this section, we analyze the predictive value of each feature set. On the basis of the clickstream features corresponding to an observation (i.e., a page view), the random forest model estimates the probability that a purchase will be made in the corresponding session. Comparing this probability estimate to a cutoff, one obtains a discrete classification of observation into the two groups purchase/no purchase. It is common practice to assess the accuracy of classification models using receiver operating characteristic (ROC) analysis.

Without any further assumptions about the application setting, we report the value of the feature sets measured by the area-under-the-ROC-curves (AUC) of each model build on the respective features, where an AUC of 0.5 and 1 indicate the performance of random assignment and perfect separation, respectively (Kuhn and Johnson 2013). Comparing the feature sets based on model performance has the advantage to capture any redundancy or interaction effects between features of different subsets. Table 4 (right) shows the marginal improvement of clickstream features via a stepwise extension adding a set of more privacy adverse features at each step. For both shops, we observe that extending the *SessionContent* feature set by including page interaction features does not substantially improve the predictive model. Starting with an AUC of 0.797 for shop 1 and 0.759 for shop 2 in the *SessionContent* setting, the inclusion of behavioral information only accounts for an improvement of 0.004 (shop 1) and 0.006 (shop 2), respectively. We tentatively conclude that there seems to be little if any (predictive) value in combining the two sets of clickstream features in this behavior prediction model. In contrast, extending the feature set by *CrossSession* features increases performance by about 0.03 AUC points. Given the baseline level of AUC equal to 0.80 and 0.76 for shop 1 and shop 2, respectively, an increase of 0.03 may signal a sizeable improvement of model performance in economic terms. Finally, the features concerning user and system information appear to not add any predictive information beyond that already embodied in the random forest model of the previous step. This follows from the, once again, very small performance increase of about 0.002 AUC points (both shops). At the same time, this data is most likely to reveal a user's demographic information or identity, thus bearing the highest potential risk to user privacy.

The AUC performance for the feature sets separately presented in Table 4 (left), indicates to what extend more privacy adverse features contain information already captured by less invasive variables. In line with marketing intuition, characteristics representing information originating from a single session and on-page behavior within a session are the strong predictors of customer conversion (Chaffey 2015). It is worth noting that *SessionContent* information, identified as the least privacy adverse type of customer information, matches and outperforms both more privacy intrusive feature sets and long-term profile data on this metric. Indeed, the aggregation of session and behavior data in the form of cross-session features, which can be used to create a customer profile, is substantially less informative on its own for shop 1, while being slightly higher for shop 2. The weak predictive power of the

www.manaraa.com

Identifiable set in the incremental setting is mirrored when considering the set on its own. Apparently, the data on the user agent header with regard to information on a user's location and system does not show substantial predictive power as indicated by an AUC value of just above 0.5. We will come back to these findings in Sect. 5.3 when discussing the importance of individual variables.

In summary, the analysis of predictive accuracy using the AUC provides evidence for our datasets that simple features based on page information within the current session are sufficient to achieve a performance level close to using the full set of clickstream features considered in the study. On the other hand, a notable performance increase over the baseline setting using only the least privacy adverse features has been observed for our data when adding *CrossSession* features.

5.3 Importance of Individual Variables

Before we go on to analyze the effect of the observed performance gains in terms of monetary profit, we analyze the importance of the variables to the model individually in order to identify the main predictors in each of the feature subsets. This allows us to further differentiate the marginal benefit of collecting very specific data with the potential to (1) reduce the number of sensitive features to be collected by focusing on a (small) subset of important predictors and (2) develop less sensitive proxy data for important predictors. At the same time, this section extends research to identify important predictors of customer purchase behavior. This analysis is limited by the dependence of the importance score on the random forest algorithm used to build the model and evaluate the feature importance.

The random forest algorithm provides a measure of relative variable importance, which captures the degree to which corrupting a variable decreases the predictive performance of the classification model (Breiman 2001). To determine the importance rank of a variable, random forest calculates the classification accuracy of each individual decision tree using the observations not employed for growing this tree, adulterates the variable by adding random noise, re-assesses the component trees' accuracy, and averages the difference in accuracy before and after variable corruption across all trees in the forest. The larger the decrease in accuracy, the more important the variable.

Figure 2 shows the variable importance ranks for the 25 most predictive variables in the best-performing random forest, which is based on the full feature set. The variables are ordered according to their average relevance for both shops, i.e. the averaged importance values for each variable of both shops. Importance estimates are normalized in the range of 100–0 indicating maximal and minimal importance, respectively. We use stars to identify the membership of a variable to one of our four feature sets.

Overall, variable importance develops consistently across shops. Notable differences can be observed for features that capture information concerning basket or checkout interaction. The corresponding clickstream features belong to the *SessionContent* and *SessionBehavior* set, which the previous analysis has shown to encompass similar information. In other words, features can substitute each other, so that variation in the importance ranking across shops is plausible. Pearson correlation scores for the discussed features support this interpretation and are included in the appendix. Correlation between variables may also impact the importance scores of the correlated variables by mitigating (acerbating) the accuracy decrease resulting from permutation in case of positively (negatively) correlated variables (Gregorutti et al. 2017). For the results of Fig. 2, the correlation patterns (see Fig. A in the Appendix; available online via springerlink.com) suggest that random forest importance scores might underestimate the actual relevance of the top three features due to positive correlation, whereas importance scores of features describing the operating system, the purchase recency and the time since adding a product to the basket, which are negatively correlated with some other features, might have been overestimated. However, in view of a relatively large forest size of 700 trees, such effects, it they exists, are likely to be small and should not distort overall tendency in feature importance.

Figure 2 provides three main insights. First, the *SessionContent* features account for six of the ten most important features. Within these, features describing the time of visit and an interaction with the shopping cart are most predictive. This result comes with the caveat that basket-related features are informative only after interaction with the basket has taken place, although it is important to note that this interaction includes viewing or removing products from the basket during search phase. The fact that the most important features of the *SessionContent* setting and the *SessionBehavior* setting both convey information related to shopping cart interactions also explains why the inclusion of *SessionBehavior* features does not significantly improve prediction performance (see Sect. 5.2).

Second, variable importance is highly skewed overall and within each subset. This suggests that it might be possible to reduce the number of features and thus the amount of data being collected about shop visitors without sacrificing predictive accuracy. The *CrossSession* features are an exception, which make up the body of important features at a rather low relative importance, but have been found to significantly increase predictive performance when they are included as a set. This suggest that there is

www.manaraa.com

no small subset of *CrossSession* features that could be singled out to provide the performance gain while less privacy-sensitive features are collected.

Third, we find information on user location and device size to be important, although they have not increased predictive performance in Sect. 5.2. This suggests that the predictive information contained in these features is also embodied in clickstream features from other sets.

An important qualification to these results is that the information captured in *SessionContent* and *SessionBehavior* is created over the course of the session and is therefore only available at later stages. This is different for features of *Identifiable* which are readily available at the very beginning of a session and also for features of *CrossSession* in case of returning visitors. This restriction is particularly relevant for applications where marketing contact is fixed to a specific page view before or at which the prediction must take place. From a data perspective, providing an incentive to reduce basket abandonment is potentially profitable even at late stages of the purchase process. Table 5 shows the ratio of customers in our dataset that do not complete their purchase after each page in the purchase process. Here, in case of shop 1 four steps are necessary for purchase completion whereas in case of the shop 2 only three steps are required. When finalizing the purchase process, customers undergo steps such as reviewing basket contents, entering their shipping information and the final confirmation of their purchase. Even at this point and at the last purchase step, abandonment rates are still as high as 13 and 25% for shop 1 and 2, respectively. High basket abandonment rates can be caused by a sudden change in customer intention potentially acerbated by unintuitive website design or (lack of) shipping and payment options.

5.4 Economic Value of Customer Data

Statistical measures of predictive performance and variable importance avoid assumptions on the application setting and thus represent a universal indicator of predictive power. This advantage is also a downside. In particular, an interpretation of the AUC or a variable importance score might not capture the characteristics of a specific application context. Moreover, the performance indicators used in management practice typically comprise measures of economic values. In this sense, managers might find it difficult to appreciate a difference in terms of the AUC and make decisions on the ground of such information. More specifically, the results of Sect. 5.2 indicate that features belonging to *SessionInteraction* and *Identifiable* setting are irrelevant for prediction. This suggest that there is no need to gather corresponding data. Likewise, including *CrossSession* features has been found to improve accuracy, which implies that the e-shop should continue to collect user information across multiple sessions. However, the consequences of these decisions remain abstract when examined in the dimension of AUC differences. A cost–benefit-analysis, although being less general, provides useful additional information for managerial decision making. We therefore simulate a specific business scenario, namely coupon targeting, in order to analyze the monetary value associated with the use of different clickstream feature subsets. This achieves two goals. First, it provides a realistic reference value regarding the business value of sensitive customer information, and second, it outlines the process that is required to express the question of data collection in monetary terms and to make informed business decisions.

To pursue these goals, we consider the marketing context associated with the data and assume that the e-shop strives to increase sales by means of couponing. e-coupons are dynamically incorporated into a webpage and thus each user's session and have gained substantial popularity to stimulate purchases in e-commerce (e.g., Khajehzadeh et al. 2014). When a coupon is offered to a visitor who is not inclined to buy, there is a probability p that she will purchase, which we assume to be constant over users. If a purchase takes place the e-shop receives expected revenue of r reduced by the cost of the marketing incentive c, where generally $c < r$ by design. However, when a coupon is offered to a customer who would buy naturally, the company faces an opportunity cost equal to the coupon value c. Assuming no other strategic restrictions on coupon offerings apply, it is optimal to offer a coupon to all those and only those customers, who do not plan to purchase naturally, as identified by the classification model. The cost–benefit matrix for the setting considered here (Table 6) has, to the best of our knowledge, not been described in previous literature, but differs from the standard coupon targeting setting only in so far as the cost associated with the coupon is realized only when a purchase takes place.

We can express the net revenue matrix (Table 6) in the form of a decision-equivalent cost matrix (Table 7), where the costs on the diagonal are normalized to be equal to zero without an effect on the optimal probability threshold (Margineantu 2001; Elkan 2001). This cost matrix better expresses the optimization problem faced by the decision model. The model aims at distinguishing purchasers from non-purchasers under the constraints that (1) issuing a coupon to a purchaser unnecessarily reduces sales profit by the coupon value c and (2) not targeting a non-purchaser foregoes a chance to convince the customer, which is associated with an opportunity cost of the expected sales value.

The performance of a classifier in monetary terms then depends on its ability to distinguish accurately between

www.manaraa.com

actual buyers and non-buyers (i.e., classification accuracy), the ratio between r and c, and the success probability of the coupon, p (i.e., conversion rate). The dependence on parameters such as c, r, and p explains why an economic evaluation of a predictive model is less general than an evaluation based on the AUC. We compute total sales revenue by multiplying the number of customers in each class with the respective revenue for the class.

In the following, we set basket revenue r to the average basket value observed in our data, which is 54.37€ and 49.45€ for shop 1 and shop 2, respectively. For c, we select a 10% reduction on the basket value approximately matching the 5€ coupon value employed by the respective shops in their campaigns, which we consider consistent with general marketing practice. The average face value of online coupons in the non-food area has been shown to be around 2€ in 2016 (KantarMedia 2016) indicating that our approach is more pessimistic in terms that the wrong classification of a non-buyer as buyer yields to a more severe punishment, i.e. a higher financial loss.

In addition to discount values, coupon conversion rates (i.e. how many customers accept a coupon and complete a purchase) are likely to depend on industry, online shop and product category characteristics as well as other criteria. To the best of our knowledge, prior literature does not offer insights concerning average coupon conversion probabilities across these categories. Likewise, publicly available information on this matter is limited, which is intuitive considering that corresponding information is sensitive. Some evidence is available for China where most successful websites achieve conversion rates up to 6% (Statista 2017). However, this data comes from 2011 and does not distinguish between coupon types, values, industries, etc.

In the interest of generality and to ensure robustness of results, we therefore consider several coupon conversion rates p between 1 and 5% and simulate the business value of a coupon targeting model for these settings. The choice of the conversion rate interval centers around the global conversion average for online shoppers of 2.5% (Statista 2016c). The interval is also consistent with (Statista 2017). Note that savings associated with the correct identification of a buyer stay constant over coupon conversion rates, while the cost of misclassification increases with coupon success probability.

From the conversion rate shown in the dataset description (Table 3) and the cost ratio given in Table 7, it is clear that the prediction problem is imbalanced, i.e. that the non-purchase class is more common than the purchase class, and cost-sensitive, i.e. that the misclassification of purchasers as non-purchasers is more costly than vice versa. To account for both issues, we apply a post-processing method for each feature set and choose the revenue-optimal probability threshold empirically on the training data (Sheng and Ling 2006). To obtain discrete class assignments from the random forest classifier, which produces purchase probabilities, we compare probabilistic predictions to a threshold and classify users as buyers if the random forest predicts a purchase probability above the threshold; and non-buyers otherwise. By setting a higher (lower) threshold, less (more) users are classified as purchasers and receive a coupon, thus adjusting for the class distribution and cost setting. We select the revenue-maximal threshold for each feature set and coupon effectiveness by calculating the revenue on the training data for a range of thresholds in [0; 1]. This way, we identify the threshold that leads to the highest revenue for each model and use this threshold when applying the model to classify the users in the test set.

Given these assumptions, Fig. 3 shows the net revenue generated over 247,325 and 251,786 customers in the test set for shop 1 and 2, respectively, by employing a customer targeting model based on each of the feature sets averaged over the range of coupon success rates. We consider as benchmark the revenue of a no-model solution, i.e. a hypothetical campaign where either no or all customers receive a coupon, whichever is more profitable given the respective coupon success rate. We calculate the revenue gain of the decision model by subtracting the revenue of the benchmark from the total model revenue. A substantial average increase in revenue of 82,482€ and 74,792€ for shop 1 and shop 2, respectively, is generated by the predictive model employing *SessionContent* features. Additional gains achieved by the inclusion of *SessionBehavior* and *CrossSession* features are comparatively smaller at below 5000€ and 10,000€, respectively. The overall revenue of the campaigns and the net revenue gain of each feature set compared to the next less sensitive set for each coupon success rate, which is the basis for Fig. 3, are reported in Table 8.[2]

The results provide two main insights. First, substantial cost savings can be achieved by better coupon targeting using the least privacy invasive feature set. Compared to a hypothetical benchmark campaign, where either no coupons are handed out or all customers receive a coupon, the savings amount to between 65,000€ and 125,000€ per month for all but the 1% coupon success rate scenario, which are realized by targeting only customers with a low conversion probability or excluding expected buyers from the coupon campaign, respectively. Even when coupons are assumed to be least effective at p = 1%, the most basic

[2] The calculations are based on the actual number of correctly and incorrectly classified customers across the 50 (2 shops × 5 feature sets × 5 conversion rate) settings. Interested readers find results at this level of detail in the Appendix.

www.manaraa.com

SessionContent model creates savings of 24,000€ and 3000€ for shop 1 and 2, respectively. The high gains indicate that the collection of session data and the development of a predictive model are highly profitable in this example.

Second, making use of more sensitive customer data does not lead to a linear increase in campaign results. The marginal gain from *SessionBehavior* features lies between 1000€ and 7000€ with the average at approximately 4500€ and 2000€ for shop 1 and 2, respectively. The addition of *CrossSession* features entails an observed revenue gain of between 3000€ and 13,000€ for shop 1 and 2, respectively. In four cases, the addition of features results in a small observed revenue loss, particularly for the *Identifiable* features. As apparent from the AUC (see Table 4), the extended model does not in fact provide worse predictions. However, the model and its predictions are so similar to the less invasive feature set that slight variation in the empirically tuned probability threshold cause slightly better or worse performance on the test set. The generally insubstantial difference in performance of the *Identifiable* set to the *CrossSession* set undermines the naïve credo that more data is always better.

In summary, while there are monetary gains achievable through the employment of more sensible features for coupon targeting for this data and application, the largest part of the realized gains is achieved with comparably privacy-friendly data. In all but the 1% redemption rate scenarios, the most basic *SessionContent* data allows the realization of at least 90% of the highest feasible savings, not adjusted for the costs of data collection, storage and protection, and additional risks that come with the handling of more sensitive data. The simulation in Table 8 thus facilitates the conclusion that the collection and application of a sensitive set of customer features beyond non-behavioral session data is unnecessary to achieve large returns on investment.

At the same time, the absolute gain provided by the collection of *SessionBehavior* and *CrossSession* information may be judged to be substantial from a business perspective. Expressing the net revenue difference in terms of the maximum gains obtainable by making use of the full set of features, the simulation suggests that a company in the assumed setting foregoes an average of 15% (shop 1) and 7% (shop 2) of potential revenue by refraining from collecting information more sensitive than *SessionContent* features. These numbers exclude the special case of a conversion rate of 1% for shop 2, where 80% of revenue is associated with *CrossSession* features. Especially for a repeated campaign setting, the expected gain from *CrossSession* features would have to be weighed against privacy considerations. More clearly, the set of *Identifiable* features, which we classify as most sensitive with regard to customer privacy, show no substantial advantage in predictive power and revenue gain in this simulation.

6 Conclusion

We investigate the marginal gain of employing clickstream data for purchase prediction of website visitors in relation to the risks to data privacy associated with data collection. The goals of our study are three-fold. First, we define four categories of clickstream information based on the threat to data privacy, namely *SessionContent*, *SessionBehavior*, *CrossSession* and *Identifiable* information in order of increasing risk. We use this framework to classify the features extracted from large clickstream datasets from two online retailers. Second, based upon this data we empirically analyze the marginal gain in predictive accuracy for the prediction of purchase behavior associated with using more sensitive customer information. This encompasses an evaluation of the importance of each feature and the performance of privacy-based feature sets both individually and aggregated. Third, we simulate a specific marketing application as undertaken by these retailers to estimate the monetary value of targeted marketing actions associated with refraining from using privacy adverse types of clickstream data.

Using a random forest model, we show that for the considered datasets the most privacy preserving *SessionContent* setting delivers competitive results in terms of customer behavior prediction. These results are improved by combining the data with *CrossSession* information about past site visits, whereas the collection of on-page behavior during the session represented by *SessionBehavior* and *Identifiable* user information do not significantly improve prediction performance.

In order to estimate the business value of extending the collected data, we simulate a coupon marketing campaign through which the e-commerce shops increase conversion rates by offering coupons to website visitors. The random forest model is used to optimize campaign targeting by identifying users that will purchase without the marketing incentive. The simulation confirms that *SessionContent* or *CrossSession* information provide a sizeable economic benefit for the considered e-commerce shops. In this setting, we estimate the opportunity costs of not collecting behavioral data and aggregating clickstream data over time at about 15% (shop 1) and 7% (shop 2) in terms of the maximum revenue obtainable by making use of the full set of features. These results imply some variation between shops and some space for e-commerce businesses to decide whether the costs and risks associated with data collection and storage are worth the marginal gain.

المنارة للاستشارات
www.manaraa.com

With respect to individual variables, we attribute the good performance of privacy-preserving *SessionContent* features to information about the page category, value of the current basket, and the time of the page view. Overall, more than half of the 15 most important variables are classified in the *SessionContent* setting, while the second most important setting is *SessionBehavior* ranking as the second-best privacy preserving setting.

Our study also exhibits limitations that could be addressed in future work. First, there is some potential to extend the information sources considered in the feature set. We focus on site-centric data and disregard user-centered data collected over a range of websites by third-party entities, since this kind of data is costly to acquire for e-commerce shops. Extending the feature set by cross-site information would further increase the potential for behavior prediction, while aggravating the potential for personal identification of users and the misuse of their data. Future research could also extend the *Identifiable* setting by more involved data collection methods to extract information by cross-referencing IP addresses or retrieving installed plug-ins, language settings supplied and similar information provided by the browser. Likewise, focusing on the trade-off between privacy and profitability, we analyze empirical results across groups of variables with different privacy implications. Given the large number of variables, an analysis of privacy implications at the level of an individual variable seems impractical. However, such analysis would be useful from a business perspective to provide insights concerning the predictive and economic value of individual variables and inform shop owners which data to gather. For example, a comprehensive analysis of the partial dependence plots for the random forest model could provide further insights into the specific non-linear effects of each variable on the model prediction.

Second, we report model performance and variable importance at any view during the session. While our analysis of basket abandonment rates shows potential for marketing activities even at late stages of the purchase process, applications that are restricted to data collected until an early point during the session will likely observe a higher relevance of information that is unrelated to the current session. The optimal point in time to play a coupon and, somewhat related, the most effective type of coupon to be used, e.g. percentage-deduction or free-shipping, are interesting in themselves, but must be left for future analysis.

Third, we look at the monetary value of privacy preserving clickstream prediction in isolation and disregard any additional value of the collected data. While sales data and aggregated clickstream data are expected to be sufficient for standard marketing analyses, there clearly is potential for a more comprehensive value analysis. In particular, live testing in a real-world setting would be a promising approach to validate the monetary costs of restricting data usage determined in the simulation.

References

Agrawal R, Srikant R (2000) Privacy-preserving data mining. ACM SIGMOD Record 29:439–450. https://doi.org/10.1145/335191.335438

Akrivopoulou C, Stylianou A (2009) Navigating in Internet: privacy and the socioeconomic and legal implications of electronic intrusion. IGI Global, Hershey

Banerjee A, Ghosh J (2001) Clickstream clustering using weighted longest common subsequences. In: Proceedings of the web mining workshop at the 1st SIAM conference on data mining

Bansal G, Zahedi F, Gefen D (2015) The role of privacy assurance mechanisms in building trust and the moderating role of privacy concern. Eur J Inf Syst 24:624–644. https://doi.org/10.1057/ejis.2014.41

Baumer D, Earp J, Poindexter J (2004) Internet privacy law: a comparison between the United States and the European Union. Comput Secur 23:400–412. https://doi.org/10.1016/j.cose.2003.11.001

Bennett PN, White RW, Chu W, Dumais ST, Bailey P, Borisyuk F, Cui X (2012) Modeling the impact of short-and long-term behavior on search personalization. In: Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval. ACM, pp 185–194

Boda K, Földes Á, Gulyás G, Imre S (2012). User tracking on the web via cross-browser fingerprinting. In: Information security technology for applications, pp 31–46

Breiman L (2001) Random forests. Mach Learn 45:5–32

Buckinx W, Van den Poel D (2005) Predicting online-purchasing behaviour. Eur J Oper Res 166:557–575. https://doi.org/10.1016/j.ejor.2004.04.022

Chaffey D (2015) Digital business and e-commerce management, 6th edn. Pearson, London

Chan T, Joseph I, Macasaet C, Kang D, Hardy RM, Ruiz C, Porras R, Baron B, Qazi K, Hannon P, Honda T (2014) Predictive models for determining if and when to display online lead forms. In: Proceedings of the twenty-eighth AAAI conference on artificial intelligence (AAAI), pp 2882–2889

comScore (2007) Cookie-based counting overstates size of web site audiences. In: comScore, Inc. http://www.comscore.com/chi/Insights/Press-Releases/2007/04/comScore-Cookie-Deletion-Report. Accessed 22 Dec 2016

Cooley R, Mobasher B, Srivastava J (1997) Web mining: information and pattern discovery on the world wide web. In: Proceedings of the ninth IEEE international conference on tools with artificial intelligence. IEEE, pp 558–567

Dinev T, Xu H, Smith JH, Hart P (2013) Information privacy and correlates: an empirical attempt to bridge and distinguish privacy-related concepts. Eur J Inf Syst 22:295–316

Eckersley P (2010) How unique is your web browser? In: International symposium on privacy enhancing technologies symposium. Springer, Heidelberg, pp 1–18

Elkan C (2001) The foundations of cost-sensitive learning. Int Jt Conf Artif Intell 17:973–978

eMarketer (2016) Worldwide retail e-commerce sales will reach $1.915 trillion this year. In: Emarketer.com. https://www.emarketer.com/Article/Worldwide-Retail-Ecommerce-Sales-

www.manaraa.com

Will-Reach-1915-Trillion-This-Year/1014369. Accessed 22 Dec 2016
Gregorutti B, Michel B, Saint-Pierre P (2017) Correlation and variable importance in random forests. Stat Comput 27:659–678. https://doi.org/10.1007/s11222-016-9646-1
Greis F (2016) Browser-Addons: Browserverläufe von Millionen deutschen Nutzern verkauft. In: Golem.de. http://www.golem.de/news/browser-addons-browserverlaeufe-von-millionen-deutschen-nutzern-verkauft-1611-124171.html. Accessed 22 Dec 2016
Guo Q, Agichtein E (2010a) Towards predicting web searcher gaze position from mouse movements. In: Proceedings on extended abstracts on human factors in computing systems (CHI), pp 3601–3606
Guo Q, Agichtein E (2010b) Ready to buy or just browsing? Detecting web searcher goals from interaction data. In: Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval. ACM, pp 130–137
Hoofnagle C, Urban J, Li S (2012) Privacy and modern advertising: most US internet users want 'do not track' to stop collection of data about their online activities. In: Amsterdam privacy conference
Iwanaga J, Nishimura N, Sukegawa N, Takano Y (2016) Estimating product-choice probabilities from recency and frequency of page views. Knowl Based Syst 99:157–167. https://doi.org/10.1016/j.knosys.2016.02.006
Jiang Q, Tan CH, Wei KK (2012) Cross-website navigation behavior and purchase commitment: a pluralistic field research. In: Proceedings of the Pacific Asia conference on information systems (PACIS)
KantarMedia (2016) CPG digital coupon circulation grows by 23.4% in 1H16, reaching 3.7 billion. In: Kantarmedia.com. http://www.kantarmedia.com/us/newsroom/press-releases/cpg-digital-coupon-circulation-grows-by-23-4-in-1h16. Accessed 1 March 2017
Khajehzadeh S, Oppewal H, Tojib D (2014) Consumer responses to mobile coupons: the roles of shopping motivation and regulatory fit. J Bus Res 67:2447–2455. https://doi.org/10.1016/j.jbusres.2014.02.012
Kim DJ, Ferrin DL, Rao HR (2008) A trust-based consumer decision-making model in electronic commerce: the role of trust, perceived risk, and their antecedents. Decis Support Syst 44:544–564. https://doi.org/10.1016/j.dss.2007.07.001
Kuhn M, Johnson K (2013) Applied predictive modeling. Springer, Heidelberg
Lee M, Ferguson ME, Garrow LA, Post D (2010) The impact of leisure travelers' online search and purchase behaviors on promotion effectiveness. Working paper, Georgia Institute of Technology
Lessmann S, Voß S (2010) Customer-centric decision support: a benchmarking study of novel versus established classification models. Bus Inf Syst Eng 2:79–93. https://doi.org/10.1007/s12599-010-0094-8
Libert T (2015) Privacy implications of health information seeking on the web. Commun ACM 58:68–77
Lin E (2002) Prioritizing privacy: a constitutional response to the Internet. Berkeley Technol Law J 17:1085–1154
Liu C, Marchewka J, Lu J, Yu C (2005) Beyond concern: a privacy–trust–behavioral intention model of electronic commerce. Inf Manag 42:127–142. https://doi.org/10.1016/j.im.2004.01.002
Lu L, Dunham M, Meng Y (2005) Mining significant usage patterns from clickstream data. In: Advances in web mining and web usage analysis. Springer, Heidelberg, pp 1–17
Margineantu DD (2001) Methods for cost-sensitive learning. Doctoral dissertation, Department of Computer Science, Oregon State University
Masand B., Piatetsky-Shapiro G (1996) A comparison of approaches for maximizing business payoff of prediction models. In: Proceedings of the 2nd international conference on knowledge discovery and data mining, Portland, OR, USA. AAAI Press Menlo Park, pp 195–201
Metzger M (2004) Privacy, trust, and disclosure: exploring barriers to electronic commerce. J Comput Med Commun. https://doi.org/10.1111/j.1083-6101.2004.tb00292.x
Moe W (2003) Buying, searching, or browsing: differentiating between online shoppers using in-store navigational clickstream. J Consum Psychol 13:29–39. https://doi.org/10.1207/153276603768344762
Moe W, Fader P (2004) Capturing evolving visit behavior in clickstream data. J Interact Mark 18:5–19. https://doi.org/10.1002/dir.10074
Moe WW, Chipman H, George EI, McCulloch RE (2002) A Bayesian treed model of online purchasing behavior using in-store navigational clickstream. Revising for 2nd review at Journal of Marketing Research
Moertini VS, Ibrahim N (2015) Efficient techniques for predicting suppliers churn tendency in e-commerce based on website access data. J Theoret Appl Inf Technol 74(3):300–309
Montgomery A, Li S, Srinivasan K, Liechty J (2004) Modeling online browsing and path analysis using clickstream data. Mark Sci 23:579–595. https://doi.org/10.1287/mksc.1040.0073
Nikiforakis N, Kapravelos A, Joosen W, Kruegel C, Piessens F, Vigna G (2014) On the workings and current practices of web-based device fingerprinting. IEEE Secur Priv 12:28–36
Nofer M, Hinz O, Muntermann J, Roßnagel H (2014) The economic impact of privacy violations and security breaches: a laboratory experiment. Bus Inf Syst Eng 6:339–348. https://doi.org/10.1007/s12599-014-0351-3
O'Connell BM, Walker KR (2014) User-browser interaction-based fraud detection system. In: USPTO Patent Full-Text and Image Database. http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PALL&p=1&u=%2Fnetahtml%2FPTO%2Fsrchnum.htm&r=1&f=G&l=50&s1=8,650,080.PN.&OS=PN/8,650,080&RS=PN/8,650,080. Accessed 22 Dec 2016
Padmanabhan B, Zheng Z, Kimbrough SO (2001) Personalization from incomplete data: what you don't know can hurt. In: Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining, San Francisco, California
Padmanabhan B, Zheng Z, Kimbrough SO (2006) An empirical analysis of the value of complete information for eCRM models. MIS Q 30(2):247–267
Pai D, Sharang A, Yadagiri MM, Agrawal S (2014) Modelling visit similarity using click-stream data: a supervised approach. In: Web information systems engineering (WISE). Springer, Heidelberg, pp 135–145
Park CH, Park YH (2015) Investigating purchase conversion by uncovering online visit patterns. SSRN Electron J. https://doi.org/10.2139/ssrn.1685469
Pitman A, Zanker M (2010). Insights from applying sequential pattern mining to e-commerce click stream data. In: IEEE international conference on data mining workshops (ICDMW). IEEE, pp 967–975
Pollach I (2011) Online privacy as a corporate social responsibility: an empirical study. Bus Ethics Europ Rev 20:88–102
Rodden K, Fu X, Aula A, Spiro I (2008) Eye-mouse coordination patterns on web search results pages. In: Proceedings of extended abstracts on human factors in computing systems (CHI'08)
Sarwar SM, Hasan M, Ignatov DI (2015) Two-stage cascaded classifier for purchase prediction. arXiv preprint arXiv:1508.03856

www.manaraa.com

Sato S, Asahi Y (2012) A daily-level purchasing model at an e-commerce site. Int J Electric Comput Eng (IJECE). https://doi.org/10.11591/ijece.v2i6.1816

Senécal S, Kalczynski P, Nantel J (2005) Consumers' decision-making process and their online shopping behavior: a click-stream analysis. J Bus Res 58:1599–1608. https://doi.org/10.1016/j.jbusres.2004.06.003

Senécal S, Kalczynski P, Fredette M (2014) Dynamic identification of anonymous consumers' visit goals using clickstream. Int J Electron Bus 11:220. https://doi.org/10.1504/ijeb.2014.063036

Sheng VS, Ling CX (2006) Thresholding for making classifiers cost-sensitive. In: Proceedings of the 21st national conference on artificial intelligence. AAAI Press, Boston, MA, USA

Sipior JC, Ward BT, Mendoza RA (2011) Online privacy concerns associated with cookies, flash cookies, and web beacons. J Internet Commer 10:1–16

Sismeiro C, Bucklin R (2004) Modeling purchase behavior at an e-commerce web site: a task-completion approach. J Mark Res 41:306–323. https://doi.org/10.1509/jmkr.41.3.306.35985

Skok G (2000) Establishing a legitimate expectation of privacy in clickstream data. Michigan Telecommun Technol Law Rev 6:61–88

Solove DJ (2001) Privacy and power: computer databases and metaphors for information privacy. Stanf Law Rev 53:1393–1462

Stange M, Funk B (2014) Real-time-advertising. Bus Inf Syst Eng 6(5):305–308. https://doi.org/10.1007/s12599-014-0346-0

Stange M, Funk B (2015) How much tracking is necessary? The learning curve in Bayesian user journey analysis. In: Proceedings of the 23rd European conference on information systems

Statista (2016a) Executive survey: big data sets that add the most value 2012. In: Statista. https://www.statista.com/statistics/249054/executive-survey-on-big-data-sets-that-add-the-most-company-value/. Accessed 22 Dec 2016

Statista (2016b) Löschen oder Unterdrücken von Cookies bei deutschen Internetnutzern bis 2015| Umfrage. In: Statista. https://de.statista.com/statistik/daten/studie/168870/umfrage/nutzung-von-programmen-die-cookies-loeschen/. Accessed 22 Dec 2016

Statista (2016c) Global online shopping conversion rate 2016. Statistic. In: Statista. https://www.statista.com/statistics/439576/online-shopper-conversion-rate-worldwide/. Accessed 12 Jan 2017

Statista (2017) The ten coupon websites with the highest conversion rate in China in June 2011. Statistic. In: Statista. https://www.statista.com/statistics/278752/coupon-websites-by-conversion-rate-in-china/. Accessed 08 Nov 2017

Suh E, Lim S, Hwang H, Kim S (2004) A prediction model for the purchase probability of anonymous customers to support real time web marketing: a case study. Expert Syst Appl 27(2):245–255. https://doi.org/10.1016/j.eswa.2004.01.008

Turow J, King J, Hoofnagle C, Bleakley A, Hennessy M (2009) Americans reject tailored advertising and three activities that enable it. SSRN Electron J. https://doi.org/10.2139/ssrn.1478214

Van der Meer D, Dutta K, Datta A, Ramamritham K, Navanthe SB (2000) Enabling scalable online personalization on the web. In: Proceedings of the 2nd ACM conference on electronic commerce. ACM, pp 185–196

Vroomen B, Donkers B, Verhoef P, Franses P (2005) Selecting profitable customers for complex services on the Internet. J Serv Res 8(1):37–47. https://doi.org/10.1177/1094670505276681

Wu F, Chiu IH, Lin JR (2005) Prediction of the intention of purchase of the user surfing on the web using hidden Markov model. In: Proceedings of international conference on services systems and services management (ICSSSM'05). IEEE, pp 387–390

Yang Y (2010) Web user behavioral profiling for user identification. Decis Support Syst 49(3):261–271. https://doi.org/10.1016/j.dss.2010.03.001

Zhang Y, Bradlow E, Small D (2015) Predicting customer value using clumpiness: from RFM to RFMC. Mark Sci 34(2):195–208. https://doi.org/10.1287/mksc.2014.0873

Zhao Y, Yao L, Zhang Y (2016) Purchase prediction using Tmall-specific features. Concurr Comput Pract Exp 28(14):3879–3894. https://doi.org/10.1002/cpe.3720

Zheng Z, Padmanabhan B, Kimbrough S (2003) On the existence and significance of data preprocessing biases in web-usage mining. INFORMS J Comput 15:148–170. https://doi.org/10.1287/ijoc.15.2.148.14449

www.manaraa.com