Predicting purchasing behavior throughout the clickstream

Verheijden, R.M.C.

Award date:
2012

Technische Universiteit
Eindhoven
University of Technology

Eindhoven
9 November 2012

Predicting purchasing behavior throughout the clickstream

by Ruud Verheijden

identity number 0609445

in partial fulfilment of the requirements for the degree of

Master of Science
in Innovation Sciences

Eindhoven University of Technology

Supervisors:

prof. dr. Chris Snijders
dr. ir. Martijn Willemsen

Abstract

This research aims to predict anonymous visitors' purchasing behavior on an e-commerce website using clickstream data. Clickstream data provides a detailed record of visitors' actions on a website and has been found useful in predicting online behavior (R. E. Bucklin & Sismeiro, 2009). This research proposes a set of 17 variables used for predicting online purchasing behavior categorized to one of six categories: (1) session stickiness, (2) website loyalty, (3) historical purchase behavior, (4) focused search behavior, (5) product interest and (6) non-purchasing intentions. Subsequently a model to predict online purchasing behavior based on data of anonymous users of an e-commerce website is developed. More specifically, to gain more insight into predicting purchasing behavior throughout the clickstream a series of three analyses is conducted. Results indicate that the use of disaggregate data in predicting online purchasing behavior is favored over aggregated data, as used in previous research. Furthermore, it was found that using clickstream data of previous visits to predict purchasing behavior only marginally improves the model's predictive capacity. This provides three main advantages: (1) it requires less computation power, which can be quite costly on high traffic websites, (2) there is no need to store previous visit data, which save large amounts of data storage costs, and (3) there is no need to use cookies to track returning visitors, which eases the process of meeting up with privacy regulations. Furthermore, the effects of individual predictor variables throughout the clickstream process were investigated and were found to change significantly. Finally, the predictive performance throughout the clickstream was investigated and was found to increase steadily as more data becomes available. All in all, these results provide important new insights into predicting purchasing behavior based on clickstream data during a user's visit session.

Introduction

The ever-increasing popularity of the Internet coincides with its importance as an e-commerce platform as a result of many firms started offering their products via the online marketplace. E-commerce has seen a vast growth over the past decade and is expected to continue increasing its importance as a retail channel, both in terms of absolute value as well as relative to the total retail market (Thuiswinkel.org, 2012; U.S. Department of Commerce, 2012). The online marketplace clearly provides advantages for firms in contrast to physical stores in terms of potential global visibility, large market area and relatively low investments. However, exploiting an e-commerce website also comes with downsides such as anonymity of its visitors.

In contrast to physical retail stores, webshops are facing relatively anonymous potential customers due to the lack of direct observability of customers' rationale and intentions, for example via salesmen. Over the years, technological developments in web technology have provided new ways to overcome this problem by allowing more insight into website visitors' behavior via so-called *clickstream* data (R. Bucklin et al., 2002; R. E. Bucklin & Sismeiro, 2009). Given the complexity of the online environment and anonymity of its users a thorough understanding and accurate prediction of people's online behavior, for example via clickstream data, has become important for websites' success (R. E. Bucklin & Sismeiro, 2009).

Clickstream data typically contains information about visitor's actions on a website, such as pages viewed, buttons clicked and items purchased, and provides a rich source of information to increase understanding of people's online behavior. Given that it provides data of real visitors without artificial interruption it provides a rich source of information for empirical modelers seeking a better understanding of online behavior (R. Bucklin et al., 2002; R. E. Bucklin & Sismeiro, 2009). Many researchers already have acknowledged the importance of clickstream data and covered a wide range of opportunities it provides, including improved insight into website usage, online advertising and shopping behavior (Andersen et al., 2000; R. E. Bucklin & Sismeiro, 2009). The latter having attracted growing interest of researchers and online businesses in the e-commerce domain.

Like in offline retail, also online retailers' primary point of concern is understanding purchase behavior and acting accordingly, especially in an environment where conversion of visitors into paying customers rarely exceeds 5% (Ayanso & Yoogalingam, 2009; Lin, Hu, Sheng, & Lee, 2010). Several researchers, like Moe & Fader (2004a, 2004b), Van den Poel & Buckinx (2005), Park & Chung (2009) and Olbrich & Holsing (2011), have picked up on this by modeling online purchasing behavior based on clickstream data in order to find factors underlying consumers' online purchasing behavior. These researches used clickstream data of visitors' actions and mainly focused on modeling purchasing behavior using aggregated clickstream data of a specific completed visit, resulting in a model to predict purchase behavior during the next visit.

However, obvious implementations of these kinds of purchase prediction models, such as website personalization based on purchasing likelihood, require purchase behavior predictions before the actual purchase has been made or before the visitor has left the website. Other examples of possible applications are displaying personalized advertisements or providing individual promotions for visitors inclining to purchase. Both of which require purchasing behavior prediction well in advance to the end of the visit.

Furthermore, due to fierce competition between e-commerce websites and online visitors' fickle navigational behavior (Park & Chung, 2009) understanding and predicting visitors' purchasing behavior during the current visit is of great importance. Besides, the very low cost of online information acquisition, for example of additional product information, and the overload in terms of product supply (Fesenmair, Werthner, & Wöber, 2006) even further increases the necessity of acting during the current visit.

Therefore, it is important to be able to predict purchasing behavior early during the visiting process, in contrast to previous research that predicted purchase behavior at the end of the visit. To my knowledge, none of the existing researches regarding clickstream based purchase prediction modeling have investigated predicting purchasing behavior throughout the clickstream. So given that predicting online purchasing behavior throughout the visitors' clickstream is a relatively unknown area this research will act as an important first step to gain more insight into this topic.

Although consumer behavior in traditional offline stores has a large scientific base, it is in many ways substantially different from online consumer behavior (R. Bucklin et al., 2002). Given that clickstream data research is still in its infancy, much research in the field of online consumer behavior still needs to be done (R. E. Bucklin & Sismeiro, 2009; Olbrich & Holsing, 2011). Therefore, this research tries to add to the current knowledge base to gain a better understanding of people's online purchasing behavior. More specifically, I try to elaborate on the prediction of purchasing behavior throughout the clickstream process. This might lead to scientific advance in understanding possible changes in factors underlying purchasing behavior throughout the clickstream. Furthermore, this research will provide a basis for improving applicability of online purchase prediction models in the commercial domain since it provides insight into predicting purchasing behavior in advance of the end of the visit.

In this study, I develop a model to predict online purchasing behavior based on data of anonymous users of an e-commerce website. The proposed model consists of a set of new and previously tested variables, for which I propose a new categorization to provide order and consistency to the wide array of possible variables. To gain more insight into modeling purchasing behavior through the clickstream I conduct a series of three analyses. Firstly, I start by investigating the differences between visit level aggregated clickstream data, as used in previous research, and pageview level disaggregated clickstream data, which incorporates more

detailed information needed to model purchasing behavior throughout the clickstream. Differences between an aggregate and disaggregate data structure are also visually explained in Image 2.

Secondly, to gain further insight into differences throughout the clickstream with respect to repeating visits a comparison is made between a model including knowledge about previous visits and a model prediction merely based on the current visit session. Thirdly, I investigate changes in model prediction throughout the clickstream by splitting up data into multiple pieces based on either the number of pageviews, time on the site or different stages in the purchasing process. Using this splitting, changes in the models' predictive performance and effects of its variables are investigated.

Literature review

As earlier mentioned, understanding and predicting consumer behavior and acting accordingly is of primary importance to the success of e-commerce websites (Sismeiro & Bucklin, 2004). Especially given that the online marketplace is a complex and rather anonymous environment in which competition is high and costs of switching are very low (Fesenmair et al., 2006; Park & Chung, 2009). Given that the costs and effort to visit and re-visit an e-commerce website is very low, it is likely that many visitors visit the website without any motivation to purchase. This contrasts with physical offline stores were it is less likely that a shopper will put time and effort in making a trip to the store without any purchase intentions (Ayanso & Yoogalingam, 2009; Moe & Fader, 2004b).

As a result, online shoppers might cover a wide range of browsing and purchasing behavior when visiting a website (Moe & Fader, 2004b). Janiszewski (1998) provided more insight into browsing behavior by dichotomizing search behavior as either goal-directed search or exploratory search based on the amount of time spent viewing a certain piece of information. Moe (2003) elaborated on this and developed a typology of website visits varying in their underlying objectives and purchasing propensity. She classified visits based on navigational patterns into one of four types of browsing strategies: directed buying, search/deliberation, hedonic browsing or knowledge building.

Visitors following a *directed buying* navigational pattern intend to make a purchase and poses substantial information before making the purchase decision. They tend to follow a focused and goal directed search pattern since their search is nearing and making a purchasing decision is nearby. *Search/deliberation* also follows a goal directed search pattern but visitors following this pattern have planned their purchase in the near future. However, they have not yet decided which product to buy in a specific category (Moe, 2003).

In contrast to directed buying and search/deliberation, visitors following a hedonic browsing and knowledge building search pattern show significantly less purchasing propensity. *Hedonic browsing* visitors tend to show exploratory search patterns, are more stimulus driven and have not yet decided in what product category to buy. Finally, visitors following a *knowledge building* pattern also show high levels of exploratory browsing but are not considering any specific purchase (Moe, 2003).

Montgomery, Li, & Srinivasan (2004) followed a similar classification and concluded that navigation behavior can be categorized as either browsing or deliberation. Their classification shows resemblance to that of Moe (2003) and Janiszewski (1998) in the sense that *browsing* is similar to exploratory search behavior with little purchasing intentions and *deliberation* is similar to focused search and is purchase oriented (Montgomery et al., 2004). Furthermore, Montgomery et al. (2004) found that visitors might easily switch between different types navigational modes during a single session.

Since visitors might change their browsing behavior or purchasing intentions during a session it is difficult to understand and predict their purchasing behavior solely on visit level classification. To gain a more detailed understanding of visitors' online purchasing behavior Moe & Fader (2004b) developed a predictive model to accommodate for different types of behavior. They were one of the first to use clickstream data to predict purchasing behavior and developed a model predicting purchase conversion behavior based on the history of visits and purchases. Their dynamic individual-level probability model described purchasing behavior as a function of visit effects and purchasing threshold. *Visit effects* refers to the differences in underlying objectives and purposes in the purchasing process that visits can have, based on the earlier described typology of visit types by Moe (2003). The *purchasing threshold* refers to the psychological resistance to buying online, which is based on previous experience with the purchasing process of a specific website (Moe & Fader, 2004b).

Furthermore, Moe & Fader (2004a) investigated the effect of visitors' return to a website and their evolving visit behavior. They found that people visiting a website more often have a greater propensity to buy. Limitations of Moe & Fader (2004a, 2004b) are that they only investigated the visit itself and although they acknowledged the notion that within visit activities can have a large influence on the propensity to buy, it was not part of their analyses.

Subsequent research by Van den Poel & Buckinx (2005) did investigate within session effects on purchasing behavior. They empirically tested the effect of different types of predictors on forecasting purchasing behavior. Using variable selection techniques on a large set of variables they proposed four different categories of variables: (1) general visit-level clickstream behavior, (2) more detailed within visit clickstream behavior, (3) customer demographics and (4) historical purchasing behavior. By combining both between visit variables, like the number of previous visits, and within visit variables, like the type of page visited, Van den Poel & Buckinx

(2005) came up with a set op nine variables significantly improving predictive performance over previous studies. Including recency of the last visit, number of pageviews on product pages and the total number of purchases among others. Furthermore, they acknowledged Moe & Fader's suspicion by showing that within session detailed clickstream behavior is very important for predicting purchasing behavior (Van den Poel & Buckinx, 2005).

In contrast to earlier researches where clickstream data of webshops selling physical consumers goods, such as books (Moe & Fader, 2004b), CD's (Moe & Fader, 2004a) or wine (Van den Poel & Buckinx, 2005), was used, Park & Chung (2009) extended the online purchase prediction-modeling concept to the area of travel websites. Their research aimed at predicting e-traveller's purchasing behavior on travel websites. They explained purchasing behavior as a function of search motivation and on-site involvement. *Search motivation*, which is closely related to the concept of focused search as defined by Janiszewski (1998) and Moe (2003), is referred to as the way in which visitors enter the website. Park & Chung (2009) argue that visitors who directly enter a website, i.e. by typing in the URL, search in a more goal-directed and focused way. While visitors entering the website via a referring website, e.g. via a search engine, follow a more exploratory search pattern.

In line with Moe (2003), also Park & Chung (2009) found that visitors following a focused search pattern have a higher propensity to buy. Furthermore, they defined *on-site involvement* in terms of the duration on the website and number of pageviews. They found that a high duration on the website and a low number of pageviews results in a higher likelihood to purchase (Park & Chung, 2009). Lin et al. (2010) supported the notion of on-site involvement but named it *session stickiness*, defined as the amount of time a visitor spends on the website during a single visit.

Similar to Park & Chung (2009), also Lin et al. (2010) found that the duration of a visit on a website positively influences the purchasing likelihood. However, they also found the number of pageviews on a website to positively influence purchasing propensity, which contrasts with Park & Chung (2009). This might be explained by the type of website on which the model was tested and its available products. Park & Chung (2009) tested their model on visitors searching for travel related purchases, like airplane tickets and hotel reservation, of which the purchasing process generally require high involvement and includes reading customer reviews. Lin et al. (2010) verified their hypothesis on websites selling a wide variety of consumer product, like books, cd's, drugs and clothing, which require less involvement and might explain the contradicting findings.

Finally, Olbrich & Holsing (2011) empirically modeled purchasing behavior using clickstream data on social shopping communities. Social shopping communities offer user-generated collages of fashion product and are linked to e-commerce websites selling the actual products. These websites' product catalogs are designed similarly to the average product catalog on a modern webshop in the sense that they consist of an overview list of products, with name and

photo, selected based a set of filters, like price or category. However, instead of purchasing a product on the website a conversion is defined as a "click-out" to a related webshop the specific product. Olbrich & Holsing (2011) found that, in line with (Moe, 2003), visitors' focused search behavior results in a higher likelihood to convert. Furthermore, they concluded that the use of product filters, which is related to exploratory search behavior, results in a lower probability of visitor conversion. The above-mentioned predictors' effects on purchasing behavior can also be found in Table 1.

All of the above-mentioned studies mainly looked at understanding and predicting purchasing behavior based on clickstream data of complete visits. As a result, their models predicted purchasing behavior at the end of the visit, which can be acted upon in next visits. However, as explained earlier, given people's online fickle navigational behavior and websites' low visiting and re-visiting costs, the chances of people returning to the website are low. So acting based on people's predicted purchasing behavior well in advance of the end of the visit is necessary. Therefore, this research aims at gaining more insight into predicting online purchasing behavior throughout the clickstream.

To be able to investigate changes in purchasing behavior prediction during a visiting session detailed clickstream data of each action conducted by a visitor needs to be gathered. Previous researches aggregated all data of visitors' actions to visit level data resulting in one single row in the dataset for each visit. However, the aggregation process results in a loss of information as individual variables' values are aggregated into one single value. Since I am interested in predicting purchasing behavior throughout the clickstream an aggregated dataset is not sufficient. Therefore, in this study a disaggregated dataset is used in which each row represents a visitor's action. The differences in data structures are visually represented in Image 2.

Since this study uses a different data structure as compared to previous studies, a first analysis will be conducted showing differences between the two structures when modeling purchasing behavior. Subsequent, the effects of clickstream data of possible previous visits are investigated. Many of the previously mentioned researches have incorporated clickstream data of both the current and previous visiting sessions. However, none of the existing researches have questioned the effect and usefulness of using clickstream data of prior visits. Given that linking visitors to previous visits' clickstream data is a difficult process and requires large amounts of data storage it is of interest to investigate its importance. Therefore, the effect of limiting the clickstream dataset to predicting purchasing behavior by only incorporating data of the current visit is investigated.

Finally, changes in the model's predictive performance and effects of individual variables on purchasing behavior are investigated. To be able to investigate the effects throughout the clickstream the dataset is split up into multiple time wise successive pieces. Subsequently, analyses will be run to predict purchasing behavior based on each of the individual pieces to be

able to investigate the effects of predicting purchasing behavior throughout the clickstream. The dataset splitting will be conducted in three different ways to ensure solid results that are not dependent on a single splitting method. Splitting will be done based on (1) the number of pageviews, (2) the time spent on the site and (3) different stages in the purchasing process.

All of the three analyses conducted in the study have, to my current knowledge, never been done before in the field on clickstream based purchase behavior prediction research. To be able to conduct these three analyses a set of variables needs to be determined to use for predicting online purchasing behavior. Therefore, the next chapter will introduce a set of variables used for conducting all three analyses.

Model variables

To test the effect and predictive performance of factors predicting online purchasing behavior throughout the clickstream, I propose a set of variables covering a wide variety of aspects. The list consists of a total of 17 variables covering predictors related to both the current and previous visits as well as page type related variables, search behavior and on-page events. To be consistent and comparable to prior studies many important predictor variables from earlier studies are included. Furthermore, several new variables are added as they are expected to be of importance in predicting purchasing behavior.

Van den Poel & Buckinx (2005) already proposed a categorization of variables to provide meaning and order to the list of variables. They categorized a total of nine variables into four categories, namely (1) General clickstream data, (2) Detail clickstream data, (3) Customer demographics and (4) Historical purchase behavior. However, since this study focuses on anonymous users the third category is not applicable to this research. Furthermore, I think that the first two categories can be categorized more precisely since their naming is quite universal. Therefore, in this study I will propose a new categorization by applying more detailed and meaningful category naming to the set of variables used in this study.

The categorization was carefully chosen based on intuition and logic since finding underlying factors using statistical analyses on the dataset used in this study, for example using factor analysis, turned out result in meaningless categories. This might be explained by the fact that variables can apply to different types of behavior. For example a person visiting many pages might do so because the required information is difficult to find resulting in a high number of pageviews, however it might also mean that the person is interested in buying a product and is therefore reading additional information. Given that these different types of purchasing behavior make it difficult to find underlying factors, the categorization of variables based on intuitions seems justified.

As a result, the individual variables are grouped into six different categories, namely: session stickiness, website stickiness, historical purchasing behavior, focused search, product interest and non-purchasing intentions. The paragraphs below describes each of the categories and the, both new and earlier tested, variables, which are also presented in Table 1 together with their expected effect on the likelihood to purchase.

Session stickiness

According to Lin et al. (2010) session stickiness is defined as the amount of time spent on a website during a visiting session. They introduced the term to online purchasing behavior research and empirically investigated the effect of stickiness on visitors' conversion behavior. The authors argue that session stickiness can be either measured by the number of pages viewed (*Pageviews*) during a single visit or the amount of time spent during the session (*Duration*). Their results indicated that both session time and pageviews are positive predictor variables for purchasing likelihood (Lin et al., 2010).

Though not directly referred to as session stickiness also others have acknowledged the importance of the number of pageviews and time spent during the session as positive indicators of purchasing likelihood. Park & Chung (2009) tested the effect of both variables on conversion behavior and found significant results. Furthermore, Olbrich & Holsing (2011) and Van den Poel & Buckinx (2005) also found that the time spent during a session positively influences purchasing propensity.

Website loyalty

Similar to the concept of website stickiness, as mentioned by Lin et al. (2010), website loyalty also relates to the amount of interaction with a website. However, website loyalty refers to interaction with the website in multiple sessions within a specific time period. Though website loyalty can be defined in many ways, I propose a set of two variables for explaining website loyalty based on variables repeatedly used in prior research. Firstly, the number of past visits to a website (*VisitFrequency*) has been investigated by Moe & Fader (2004a, 2004b) and Van den Poel & Buckinx (2005). They showed that frequent visitors to a website have a higher likelihood to purchase as compared to infrequent users.

Secondly, to capture more of the visitors' loyalty to a website the number of days since its last visit (*VisitRecency*) can be added. The same three researches that also included VisitFrequency also investigated to effect of VisitRecency on purchasing propensity. They found that a longer time between the current and last visit results in a higher likelihood to purchase (Moe & Fader, 2004a, 2004b; Van den Poel & Buckinx, 2005).

Historical purchase behavior

Van den Poel & Buckinx (2005) introduced the category of historical purchase behavior in his research given that it had been proved to be useful in predicting purchasing behavior in the offline world. Therefore, the two variables describing historical purchasing behavior found after best subset selection processes by Van den Poel & Buckinx (2005) are also tested in this research. The first variable describes the number of previous purchases made at the website (*TotalPurchases*), which was also found to be positively influencing purchasing behavior by Moe & Fader (2004b) as it lowers the purchasing threshold (Beatty & Ferrell, 1998).

The second variable describes the number of days since last purchase (*PurchaseRecency*). Contradicting to the total number of purchases, PurchaseRecency was found to be a negative predictor for purchasing propensity by Van den Poel & Buckinx (2005). This might be explained by the idea that people will not buy again shortly after a purchase but merely visit the website again to find additional information regarding their purchase, like shipping or billing information.

Focused search

As explained earlier, both Moe (2003) and Park & Chung (2009) embraced Janiszewski's (1998) dichotomized classification of search behavior into goal-directed focused search and exploratory search. Moe (2003) classified online visitors in one of four categories and showed that visitor's with a high degree of focused search also showed the highest conversion rates. According to her research, visitors with a focused search pattern search within a single product category and visit multiple products in that category. Resulting in a high number of product detail page visits compared to a relative low number product overview pages (*ProdOvervRatio*). As also explained later on, product detail pages refer to a dedicated page displaying all information about a single product, whereas product overview pages refer to pages consisting of a, catalog like, array of small product photo's with a link to the corresponding product detail page.

Though tested in an offline setting, Janiszewski (1998) argued that paging through a catalog resembles an exploratory search behavior, which is very similar to paging through product overview pages on websites. To describe this behavior in terms of focused search, I propose a new variable, which is described as the percentage of pageviews on the first product overview page relative to all pageviews on product overview pages (*PercOvervFirst*). So the more people are following a focused search pattern the higher the value of PercOvervFirst, since they will not follow an exploratory search pattern of paging through product overview pages.

Park & Chung (2009) also acknowledge the importance of focused search in prediction purchasing behavior. However, they defined exploratory search in terms of the way in which the website was accessed during the current visit. Their research suggested that visitors entering the website via a transferring website, like a search engine or referrer website, resemble

exploratory search patterns and are therefore less likely to purchase (Park & Chung, 2009). To be consistent with Park & Chung (2009) I included the SiteTransferred variable, which negatively predicts focused search behavior.

Furthermore, since many modern websites offer functionality to filter product selections on product overview pages, it is worthwhile investigating this behavior. Olbrich & Holsing (2011) already tested the effects of using filters on social shopping websites, which were found to negatively influence conversion. Though not mentioned directly by Olbrich & Holsing (2011), I argue that the number of unique filters used on product overview pages (*Filters*) resembles exploratory search since the use of multiple filters might relate to knowledge building behavior (Olbrich & Holsing, 2011). Given that Filters resembles exploratory search behavior it is expected to negatively influence purchasing behavior.

Product interest

It can be expected based on logic that people who show interest in certain products are more likely to purchase those specific products. Moe (2003) acknowledged the importance of the visitor's interest in certain products, as visitors with direct purchasing intentions tend to show high levels of repeat product viewings, which involves deep deliberation (Moe, 2003). Based on Moe's classification, it was found that the maximum number of times a single product was viewed (*MaxRepProduct*) correlated with a high conversion rate. Therefore, I argue to include the variable to test its effects in a purchasing prediction model.

Furthermore, Bellman, Lohse, & Johnson (1999) also acknowledged the importance of product interest by claiming that looking at product information is an influential predictor for predicting online purchasing behavior. Since looking at product information is difficult to measure based on clickstream data, I propose a set of two new variables approximating the degree of looking at product information. Firstly, I propose to use the average number of times a product photo is enlarged on a product detail page (*AvgPhotoZoom*). A high average of photo zooming indicates that a visitor is interested in certain products since it provides additional information that requires additional effort to gather. Secondly, I propose to use the average number of times a different product size is selected (*AvgSizeSwitch*), which provides additional product information, like product availability.

Another way visitors might show their interest in a specific product is by adding it to the shopping cart (*AddToCart*). Since many people use the shopping cart rather as a wish list it is not necessarily a precursor of purchase. Furthermore, AddToCart has been added to the model because it has never been tested before in online purchase predicting research and its effect throughout the clickstream might be interesting to explorer.

Table 1: Predictor variables included in the model and their expected effect on purchasing behavior

Variable	Description	Effect
Session stickiness		
Pageviews	Number of pages viewed during the session	+ [1, 3]
Duration	Time spent during the session (excluding last pageview)	+ [1, 2, 3, 6]
Website loyalty		
VisitFrequency	Number of past visits	+ [4, 5, 6]
VisitRecency	Number of days since last visit	+ [4, 5, 6]
Historical purchase behavior		
TotalPurchases	Number of purchases during previous sessions	+ [5, 6]
PurchaseRecency	Number of days since last purchase	– [6]
Focused search		
PercOvervFirst	Percentage of pageviews on first page relative to all overview pageviews	+ [7]
ProdOvervRatio	Ratio of product to overview pageviews	+ [8]
SiteTransferred	Visitor accessed the website via either a search engine, referrer website or email newsletter (dummy variable)	– [3]
Filters	Number of unique filters used	– [2]
Product interest		
AvgPhotoZoom	Average number of times product photo enlarged per product	+
AvgSizeSwitch	Average number of times product size switched per product	+
MaxRepProduct	Maximum number of times a single product was viewed	+ [8]
AddToCart	Product(s) added to shopping cart (dummy variable)	+
Non-purchasing intentions		
Hurry	Average time per page lower than the average in previous sessions	– [6]
PersonalPages	Number of pageviews on pages related the visitor's personal account (like account registration, newsletter subscription, RMA requests and shipment status)	– [6]
AboutPages	Number of pageviews on information about the company and its physical stores (like vacancies, franchising information and physical store opening hours)	–

Sources: [1] Lin et al. (2010), [2] Olbrich & Holsing (2011), [3] Park & Chung (2009), [4] Moe & Fader (2004a), [5] Moe & Fader (2004b), [6] Van den Poel & Buckinx (2005), [7] Janiszewski (1998), [8] Moe (2003)

Non-purchasing intentions

Though most visitors will visit a webshop with a certain degree of planning to purchase a product, not all visits necessarily are intended for shopping. Given that many webshops are related to a physical offline counterpart and also provide additional information about the company, it might be the case that people merely visit the website to gather information about the company or physical store. Furthermore, also returning visits after a recent purchase might only be intended to gather for example shipment information. As a result these kinds of visits will not be likely to convert to a purchase.

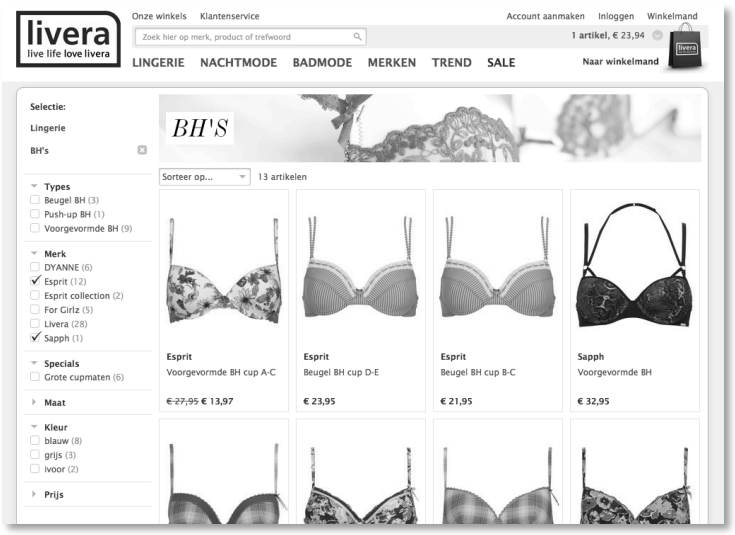
Therefore, two variables are added to the model, which are expected to negatively influence purchasing behavior given that they try to capture behavior of visitors not having any intentions to purchase. Firstly, a variable describing the number of pageviews on pages related to information about the company, vacancies or physical stores (*AboutPages*) is added. Secondly, a variable capturing the number of pageviews on personal account pages, like newsletter subscription, RMA requests and shipment information (*PersonalPages*), is added.

Furthermore, making an online purchase is assumed to require high involvement into the decision-making process as well as deep consideration of making the purchase. As a result, the visit in which the final purchase decision is made is expected to take longer than possible earlier visits. Therefore, visits in which people spend less time as compared to earlier visits are expected to show higher degrees of non-purchasing intentions. To capture this behavior the "Hurry" variable introduced by Van den Poel & Buckinx (2005) is used which describes a dummy variable indicating whether or not the average time per page in the current session is lower than the average of previous sessions (*Hurry*).

Data & methodology

For this research I used disaggregate anonymous clickstream data from the website of a major retailer selling lingerie, nightwear and bathing fashion in The Netherlands. The website sells a selected assortment of products available in over 120 physical stores following the franchise formula. The website is structured, similar to many modern webshops, using several categories each consisting of multiple product overview pages, customizable via a set of filters. Overview pages consist of an array of product photos and links to product detail pages consisting of product information, like high-resolution photos, price, description, available colors and sizes. See Image 1 for a screenshot of a product overview page.

Image 1: Screenshot of a product overview page of the Dutch website the clickstream data was collected from.



The clickstream data was collected using the popular open-source Piwik website analytics software. Piwik gathers visitor's clickstream data via Javascript code, which triggers a server-side script to register the pageview of an event in a MySQL database. Due to its large user and developer community Piwik provides a solid and accurate Javascript based visitor tracking system. Although Javascript can be disabled by the user, thereby excluding him from data collection, previous research suggests that 92% to 98% of the online users have Javascript enabled (W3Techs, 2012; WebAIM, 2010; Zakas, 2010).

The Piwik Javascript tracker has three main advantages over the more standard server logs used in earlier research, as was used for example by Van den Poel & Buckinx (2005). Firstly, Javascript tracking works regardless of browser caching since it is triggered every time the page is displayed, resulting in more accurate clickstream data. Secondly, it allows to track on-page visitor behavior, like tracking of clicks that do not trigger a page reload. Thirdly, Javascript allows returning visitors to be identified using cookies and a browser configuration hash, which is more accurate than IP-address identification via server logs.

Clickstream data on all publically accessible pages, including a set of on-page events, shopping basket contents and purchases, were registered using Piwik's default asynchronous Javascript visitors-tracker. On-page events include a handful of actions that do not trigger page reload, like enlarging the product photo and pressing the "add to cart"-button, which as far as I can tell have never been tested in purchase predicting research before.

The data was collected during a time period of about four months, dating from May 5th 2012 until September 10th 2012. This time period is considered long enough to reliably track repeated visit behavior since research by Google (2011) indicated that people shopping for clothing online on average take 27 days before making a purchase. Besides, within 90 days almost all

purchasers have finished their shopping journey and have made their purchase (Google, 2011). During the four month time period roughly 460,000 visits with a total of 4.8 million pageviews and 680,000 on-page events were registered, which can be considered a high data volume compared to most previous studies (Bucklin & Sismeiro, 2009; Olbrich & Holsing, 2011). Table 2 displays an overview of some summary statistics of the final dataset after preprocessing as described later on.

Table 2: Descriptives of the website of a major lingerie retailer in The Netherlands from May 5th 2012 until September 10th 2012 after preprocessing.

Statistic	Frequency
Number of visitors	363,389
Number of visits	461,278
Number of purchases	8,044
Number of pageviews	4,873,972
Number of on-page events	686,241
Conversion rate	1.74%

Preprocessing

Before obtaining a meaningful dataset the Piwik relational MySQL database had to be manually queried to get the required data. The database size and complexity of the data transformations, to obtain a statistically easily analyzable single-table file, resulted in a rather complex and time-consuming process of developing and executing the data transformations. Since this research focuses on predicting purchasing behavior throughout the clickstream, the dataset was transformed to a format in which each row consisted of one pageview or on-page event. This contrasts with earlier research using datasets aggregated at the session level, in which possibly valuable information is lost in the aggregation process.

To clarify the differences between disaggregated and aggregated data Image 2 shows an example of both types of data structure based on the same dataset. Based on the aggregated dataset one would argue that the two visits are identical. However, when looking at the disaggregated data one could extrapolate more detailed information from the visit. For example, instead of the time spent on the website, disaggregate data provides more detailed information in the form of time spent on each individual page. Furthermore, based on the disaggregate data one could notice that the first visitor added products to the cart early in the visiting process and purchased when the ratio of product to overview pages was high, while the second visitor showed the opposite effect. Therefore, disaggregated data provides possibilities to subtract more detailed information.

Image 2: Data structures of both disaggregated and aggregated data of the same dataset including two fictive visits.

Disaggregate dataset

VisitID	ServerTime	TimeOnPage	Purchase	AddToCart	ProdOvervRatio
1	08-14-2012 15:47:29	25	0	1	20%
1	08-14-2012 15:47:54	9	0	1	20%
1	08-14-2012 15:48:03	29	0	0	40%
1	08-14-2012 15:48:32	42	0	0	40%
1	08-14-2012 15:49:14	80	0	0	80%
1	08-14-2012 15:50:34	0	1	0	80%
2	08-14-2012 15:47:29	25	0	0	80%
2	08-14-2012 15:47:54	9	0	0	80%
2	08-14-2012 15:48:03	29	0	0	40%
2	08-14-2012 15:48:32	42	0	1	40%
2	08-14-2012 15:49:14	80	0	1	20%
2	08-14-2012 15:50:34	0	1	0	20%

Aggregate dataset

VisitID	ServerTime	TimeOnSite	Purchase	AddToCart	ProdOvervRatio
1	08-14-2012 15:47:29	195	1	2	40%
2	08-14-2012 15:47:29	195	1	2	40%

Variables to uniquely identify each visit and visitor were taken from Piwik's visitor identification process, based on browser cookies combined with a semi-unique browser configuration hash, and used to identify returning visits. Similar to Van den Poel & Buckinx (2005), a time interruption of more than 30 minutes between two pageviews is considered a new visit.

The dataset was cleaned from visits by the employees maintaining the website and visits from multiple search engine crawlers and one monitoring robot. Furthermore, consistent with Van den Poel & Buckinx (2005) and Olbrich & Holsing (2011) single page visits and sessions with a viewing time below 1 second are removed. These cases are removed because they do not resemble a real browsing session and could be the result of system errors or user mistakes. Besides, outliers consisting of visits with more than 100 pageviews, 1 hour of viewing time or more than 40 returning visits were removed from the data. Finally, all pageviews and on-page events in a session after a purchase has occurred are removed from the data since I am interested in predicting in advance of the purchase. After data preprocessing the original dataset that consisted of 547,366 visits with a total of 5,877,949 pageviews and events is trimmed down to 461,278 visits with 5,560,214 pageviews and events.

Data analysis

Given that the proposed model tries to predict whether or not a visitor will purchase, an obvious approach is to use logistic regression. Despite the fact that Moe & Fader (2004b) found that logistic regression might be outperformed by other types of models, several other studies, including Van den Poel & Buckinx (2005) and Olbrich & Holsing (2011), did use logistic regression. Likewise, I also argue to use logistic regression given that it is conceptually simple and consistent with important preceding research. To test for external validity of the predicted models the dataset was split up using 80% as a training set for the model and 20% as a testset.

Analyses were conducted with STATA 12 using the *logistic* command including clustering of sessions to correct z-values since multiple cases belong to the same observation. In fact, the *xtlogit* command would provide a better model representation due to its correction for both z-values and coefficients. However, due to its more complex calculations in combination with the

huge size of the dataset this resulted in an unusable long computation time given the available computation power.

Several variable transformations have been applied to the set of variables in Table 1 in order to make them suitable for logistic regression analysis. Firstly, the number of Pageviews and Duration were transformed respectively using a square root and a natural logarithm transformation to ensure linearity with the log odds, which is one of the assumptions of logistic regression. Furthermore, given its highly positively skewed distributions AddToCart and VisitFrequency were transformed into respectively a dummy variable and a scale of 1, 2, 3, 4 or 5 and above previous visits.

The dependent variable in all logistic regressions is the dummy variable whether or not a purchase has occurred during the current visit. This variable is true for all pageviews and on-page events during that specific visit. Although this adds artificial knowledge of future purchasing behavior to the dataset, I argue that this is justifiable given that I am only interested in predicting whether or not a purchase occurs in the current session.

Results

Since I am interested in investigating modeling through the clickstream I conduct a series of three analyses to gain more insight into this topic. Firstly, I will compare the use of aggregated data against the use of disaggregated clickstream data in predicting online purchasing behavior. As explained earlier, all previous online purchasing prediction researches have used clickstream data aggregated at session level, which basically ignores information throughout the session. However, to be able to predict purchasing behavior throughout the clickstream, more detailed pageview level disaggregated clickstream data is needed. To investigate the differences between both data structures I compare both a model based on aggregate data with a model based on disaggregate data from the same dataset.

Secondly, a comparison will be made between a “Previous knowledge”-model including knowledge about previous visits of a specific visitor and a “Current visit”-model predicting merely based on clickstream data of the current visit. This might provide more insight into the effectiveness of having additional data of a returning visitor in predicting purchasing behavior. To my knowledge this has never been tested before. Based on the list of variables stated in Table 1 it can be noticed that five variables include data from previous visits, namely VisitFrequency, VisitRecency, TotalPurchases, PurchaseRecency and Hurry. The “Previous knowledge”-model will incorporate all variables stated in Table 1, while the “Current visit”-model will lack the five previously mentioned variables, resulting in a model containing only current session data. To strengthen the results, the same comparison, between the “Previous knowledge”-model and the “Current visit”-model, will also be conducted using clickstream data only of returning visitors. In this analysis the effect of previous knowledge about the returning visitor is expected to be

stronger since only returning visitors are included, resulting in a possibly clearer distinction between to the two models.

Thirdly, I will investigate changes in model prediction over time as more clickstream data of a visitor becomes available. This is possible because the preprocessed dataset consists of the running sum of all included variables with each row representing a pageview or on-page event. By splitting up each visitors' visit into multiple time wise successive pieces and separately analyzing these pieces using data of all visitors, changes throughout the clickstream in the sense of model's performance and effects of individual variables can be analyzed. Using more than one method the results are expected to be more solid as the splitting does not depend on one specific method. Time wise successive splitting is conducted based on three different splitting variables, namely (1) the number of pageviews, (2) time on the site or (3) different stages in the purchasing process. For the first two splitting variables the groups are composed to contain an approximately equal number of observations.

Furthermore, the third splitting method was based upon research by Sismeiro & Bucklin (2004) who argued that the completion of a purchase consists of multiple sequential Nominal User Tasks. Despite the fact that their idea was based on a different kind of e-commerce website, the sequential stages splitting method can also be applied to other kinds of webshops. Therefore, I propose a split based on entering the checkout process for the first time, which is basically one of the few stages the visitor has to go through to complete the purchase. As a result, three different types of splitting methods are used to investigate effects of predicting purchase based on partial data.

Aggregated versus disaggregated clickstream data

Logistic regressions on the same dataset using both data aggregated at visit level and disaggregated data at pageview level is displayed in Table 3. The results in Table 3 show that there are large differences in outcomes of the two analyses. The visit level clickstream data analysis shows that the variables PurchaseRecency, PercOvervFirst, SiteTransferred and AvgPhotoZoom do not appear to be significant, although they are significant predictors when modeling using disaggregate data.

When interpreting results from logistic regression it is useful to look at the odds ratio since it provides a more intuitive result as compared to the coefficient. The odds ratio relates to the percentage of increase in purchasing likelihood for one additional unit of that specific variable (Van den Poel & Buckinx, 2005). For example, an odd ratio of 1.06 for VisitFrequency can be interpreted as a 6% increase in likelihood to purchase for every additional visit.

Table 3: Comparison between two logistic regressions on pageview level (disaggregate) clickstream data and visit level (aggregate) clickstream data.

	Pageview level (disaggregate data)			Visit level (aggregate data)		
Variable	Coefficient	Z-value	Odds-ratio	Coefficient	Z-value	Odds-ratio
Pageviews[1]	0.10	8.45***	1.11	0.25	21.47***	1.28
Duration[2]	0.04	3.73***	1.04	0.65	35.76***	1.92
VisitFrequency	0.06	4.07***	1.06	0.20	14.19***	1.23
VisitRecency	-0.01	-4.75***	0.99	-0.01	-5.81***	0.99
TotalPurchases	0.33	3.55***	1.40	0.31	2.96***	1.37
PurchaseRecency	0.01	2.19**	1.01	0.00	0.74	1.00
PercOvervFirst	0.41	9.36***	1.50	0.13	1.81	1.14
ProdOvervRatio	0.03	2.58***	1.03	0.09	4.07***	1.09
SiteTransferred	-0.10	-3.17***	0.90	0.01	0.35	1.01
Filters	-0.07	-3.49***	0.93	-0.20	-9.71***	0.82
AvgPhotoZoom	0.14	4.98***	1.14	0.07	1.69	1.07
AvgSizeSwitch	0.47	13.28***	1.60	0.67	10.12***	1.95
MaxRepProduct	0.04	2.78***	1.04	-0.08	-6.97***	0.92
AddToCart	2.48	91.56***	11.99	4.36	71.54***	78.20
Hurry	0.23	5.10***	1.26	-0.30	-4.52***	0.74
PersonalPages	0.08	6.71***	1.09	0.05	5.84***	1.05
AboutPages	-0.31	-4.24***	0.73	-0.78	-10.98***	0.46
Constant	-4.47	-101.82***	0.01	-11.16	-92.69***	0.00
McFadden R^2	0.2463			0.5240		
Observations	4,456,388			369,448		
Sessions	369,766			369,448		
Wald chi2	31,548.77			14,674.31		
Sensitivity[1]	3.25% (3.57%)			11.33% (4.79%)		
Specificity[1]	99.74% (99.76%)			99.76% (99.69%)		
Pos. predictive value[1]	45.90% (49.85%)			45.44% (50.53%)		
Neg. predictive value[1]	93.74% (93.99%)			98.44% (94.06%)		

[1]: Square root transformation applied, [2]: Natural logarithm transformation applied, [3]: Percentages between brackets result from classification on the testset to test the external validity.
Significance level: ***: < 0.01, **: < 0.05, *: < 0.1

When looking at the odds ratio large differences can be noticed between the two models. The effects of Duration and AvgSizeSwitch seem to be much larger when modeling using visit level data. On the other hand, the effects of PercOvervFirst and AboutPages seem to indicate the opposite. However, the most important differences between the disaggregate data and aggregate data seem to be differences variables' direction of their effect. The SiteTransferred variable was predicted to positively influence purchasing behavior when modeling using visit level data, consistent with the original research of Park & Chung (2009). However, when modeling using pageview level data, which allows for more detailed information and thereby providing a better representation of real visit behavior, results show that SiteTransferred negatively influences the purchasing likelihood.

Also the effect of MaxRepProd seems to have an opposing effect direction when comparing the two models. Furthermore, the Hurry variable modeled using visit level data is consistent with research of Van den Poel & Buckinx (2005), while pageview level modeling show that the effect of Hurry positively influences purchasing behavior. Differences between the pageview level model and the visit level model might be explained by the loss of information due to aggregation

of data in the case of the visit level model. Given this consistency with previous research, it might be that, at least in the case of the Hurry variable, previous researches might have found results incorrectly representing reality due to information lost as a result of data aggregation.

When comparing the predictive performance measurements of the models, I mainly focus on the McFadden R^2, which is an indicator of model fit and values between 0.2 and 0.4 are indicating a good model fit (Louviere, Hensher, & Swait, 2000), sensitivity and positive predictive value. *Sensitivity* is defined as the percentage of correctly positively classified purchases given the set of all real purchases. The *positive predictive value* can be defined as the percentage of correctly positively classified purchases given the set of all cases classified as purchases.

The fact that only 1.7% all visits in the data actually converted to a purchase makes the interpretation of the classification results difficult. Although dependent on the exact implementation, I argue that sensitivity and the positive predictive value are the most important predictors of the performance of these kinds of models since the desired behavior of the model is predicting as much purchases correctly as possible. Given that this research mainly focuses on the scientific value of understanding predicting purchase behavior, I am largely interested in the relative differences between two models rather than absolute predictive performance values.

When looking at the McFadden R^2 a clear difference can be noticed. However, I argue that the pseudo R2 resulting from the visit level model might be due to an error given its extremely high model fit. Furthermore, classification numbers are difficult to compare between the two models since the visit level model classifies purchasing behavior for the session as a whole while the pageview level model classifies individual pageviews within a single visit. Given that the first few pageviews are always less likely to convert, a lower degree of sensitivity can be expected.

Previous knowledge versus current session based modeling

The results of the logistic regression based on models with knowledge about previous session versus models based on only current session data are shown in Table 4. The model with knowledge of previous visits consists of all variables as described in Table 1, while the model only based on current visit data lacks the variables VisitFrequency, VisitRecency, TotalPurchases, PurchaseRecency and Hurry. For easy comparison purposes the left column of Table 4 shows the original pageview level model as displayed in Table 3. Furthermore, both types of models are also modeled using data of only returning visitors.

Table 4: Comparison between logistic regressions on a model with previous knowledge (identical to pageview level model in Table 3) and a model based on current session data only. Both types are modeled using both all data and data of only returning visitors.

	All data included				Only returning visitors data included			
	Prev. know. model		Current visit model		Prev. know. model		Current visit model	
Variable	Coef.	Z-value	Coef.	Z-value	Coef.	Z-value	Coef.	Z-value
Pageviews[1]	0.10	8.45***	0.11	8.58***	0.07	3.75***	0.07	3.33***
Duration[2]	0.04	3.73***	0.04	3.33***	0.02	1.40	0.03	2.06**
VisitFrequency	0.06	4.07***			-0.10	-4.99***		
VisitRecency	-0.01	-4.75***			-0.02	-7.59***		
TotalPurchases	0.33	3.55***			0.34	3.81***		
PurchaseRecency	0.01	2.19**			0.01	3.71***		
PercOvervFirst	0.41	9.36***	0.41	9.44***	0.39	5.84***	0.37	5.62***
ProdOvervRatio	0.03	2.58***	0.03	2.65***	0.03	1.42	0.03	1.41
SiteTransferred	-0.10	-3.17***	-0.13	-4.27***	-0.20	-3.74***	-0.22	-4.28***
Filters	-0.07	-3.49***	-0.07	-3.65***	-0.11	-3.23***	-0.11	-3.24***
AvgPhotoZoom	0.14	4.98***	0.12	4.49***	0.09	1.89*	0.08	1.82*
AvgSizeSwitch	0.47	13.28***	0.46	13.10***	0.51	8.25***	0.52	8.33***
MaxRepProduct	0.04	2.78***	0.04	2.75***	0.02	1.19	0.02	0.97
AddToCart	2.48	91.56***	2.50	91.81***	2.27	50.85***	2.30	51.91***
Hurry	0.23	5.10***			-0.05	-1.10		
PersonalPages	0.08	6.71***	0.09	6.96***	0.06	3.64***	0.08	4.90***
AboutPages	-0.31	-4.24***	-0.32	-4.35***	-0.63	-4.52***	-0.59	-4.31***
Constant	-4.47	-101.82***	-4.38	-102.72***	-3.34	-44.37***	-3.70	-58.86***
McFadden R^2	0.2463		0.2430		0.2247		0.2135	
Observations	4,456,388		4,456,388		1,120,260		1,120,260	
Sessions	369,766		369,766		86,120		8,746.70	
Wald chi2	31,548.77		31,742.40		0.0000		0.0000	
Sensitivity[1]	3.25% (3.57%)		2.09% (2.18%)		5.54% (9.70%)		2.43% (2.58%)	
Specificity[1]	99.74% (99.76%)		99.82% (99.82%)		99.43% (99.20%)		99.67% (99.75%)	
Pos. predictive value[1]	45.90% (49.85%)		44.61% (44.67%)		49.94% (44.46%)		43.37% (40.77%)	
Neg. predictive value[1]	93.74% (93.99%)		93.67% (93.91%)		91.05% (94.32%)		90.80% (93.93%)	

[1]: Square root transformation applied, [2]: Natural logarithm transformation applied, [3]: Percentages between brackets result from classification on the testset to test the external validity.
Significance level: ***: < 0.01, **: < 0.05, *: < 0.1

Results displayed in Table 4 indicate that when looking at the models based on all data the current visit model performance is almost as good as the model incorporating previous visit knowledge. The slight decrease in McFadden R^2 from 0.2463 to 0.2430 is hardly noticeable and the decrease in sensitivity and positive predictive value are also very small. When looking at the models based only on previous visit behavior the differences between the models should become more clear-cut. The differences in McFadden R^2 are once again marginally as they change from 0.2247 to 0.2135. Furthermore, sensitivity and positive predictive value lower respectively from 5.54% to 2.43% and from 49.94% to 43.37%. Interpretation of whether these differences are perceived small or large heavily depends on the final implementation of the model. However, in cases where this degree of loss in sensitivity and positive predictive value is considered acceptable, the model based on only current visit data provides a clear advantage since it incorporates less variables likely making it easier to implement.

Model prediction over time

Finally, the results of the set of logistic regression analyses regarding the purchasing prediction over time are presented. The dataset is split up based on either the number of pageviews, time on site or the moment checkout process was entered the first time. Splitting based on pageviews and time is split up into eight groups with approximately equal sizes. Since in total 18 logistic regressions were conducted for this analysis the full set of results are displayed in Appendix A. Given that large amount of output data only a selection of six variables and three model characteristics will be displayed below. To provide a clear insight into variable change and the change in model performance characteristic are plotted in the graphs visible in Image 3.

Based on the series of graphs displayed in Image 3 it can be noticed that Duration, TotalPurchases, PercOvervFirst and AboutPages show a clear increase in coefficient value as more clickstream data becomes available. For the first three, this can be interpreted as the variable becoming more influential in predicting purchasing behavior later in the clickstream process. The time spent on the website is found to be an increasingly important predictor as more clickstream data is gathered. Furthermore, the number of purchases made in previous sessions plays a less important role early during the visit but increases as time progresses. The percentage of the number of pageviews on the first overview page shows a similar pattern, which might be explained by the fact that the variable probably starts at 100% when the first overview page is hit after which it will lower as more overview pages are viewed.

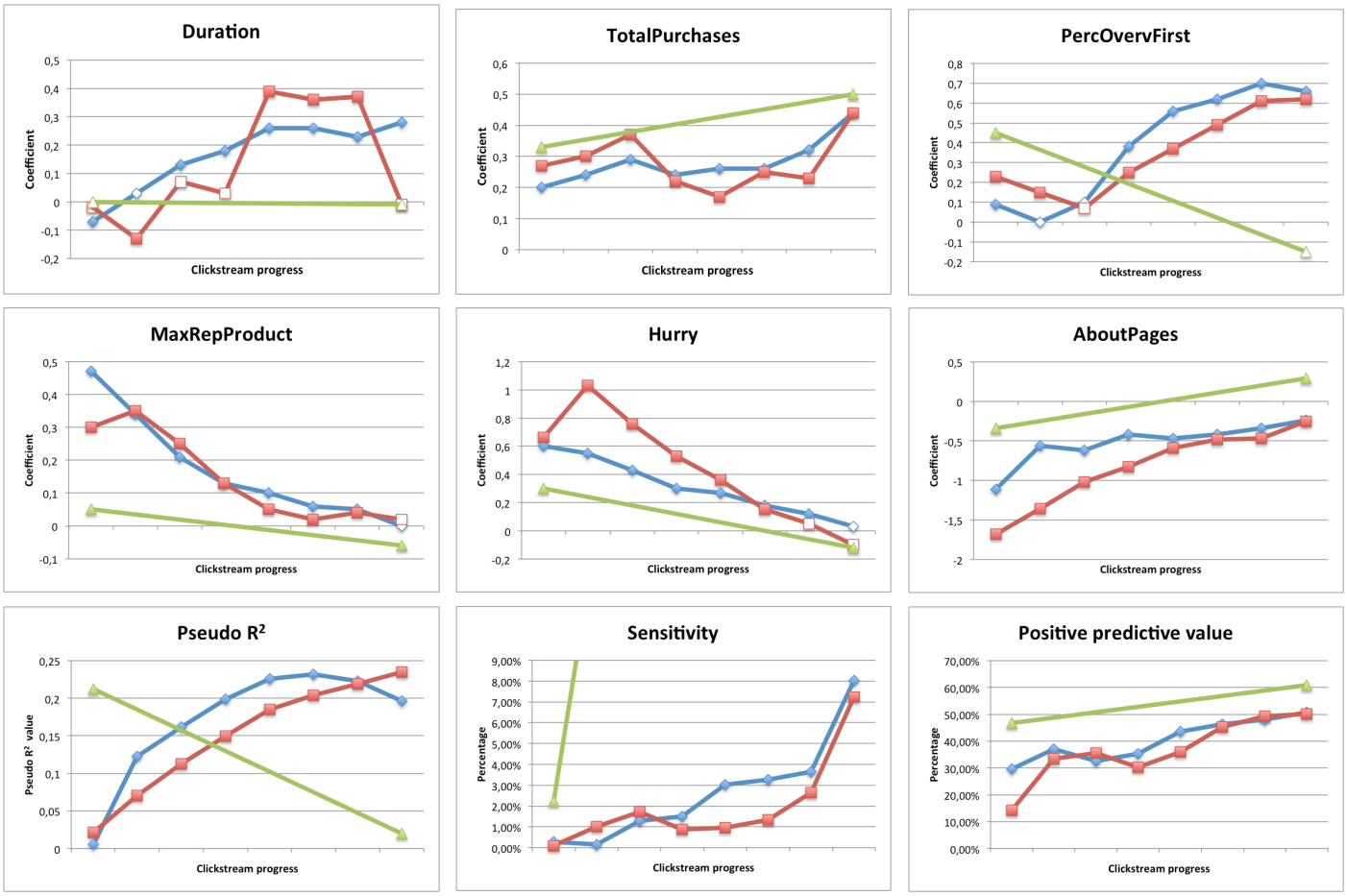
In the case of the AboutPages variable the variable becomes a less negative predictor for purchasing behavior as more clickstream data becomes available. This might be explained by the idea that some visitors might briefly visit the website to gather company related information, for example about opening hours of related physical stores. Furthermore, MaxRepProduct and Hurry show the opposite since their effect on predicting purchase propensity decreases throughout the clickstream process. The number of times a product is repeatedly viewed seems to be of less importance later in the clickstream process, which might be the result of many visitors following a exploratory search pattern and thereby revisiting product pages regardless of the purchasing behavior.

It can be clearly noticed that with all six variables the coefficient changes rather independently of the chosen method for splitting up the dataset, which supports the reliability of the found results. The green line following a deviant slope for both the Duration and PercOvervFirst can be explained by the fact that their coefficient in the “after purchase”-stage was not significant in both cases.

The bottom three graphs in Image 3 display performance characteristics for the different models. Based on these graphs it can be seen that the predictive performance of the model increases as more clickstream data becomes available. Especially the increase in sensitive is very strong, although the extremely high sensitivity (91,77%) for the “after checkout”-stage is likely

to be caused by an error. Based the three plotted performance measures it can be concluded that the model is able to provide more accurate purchasing behavior predictions as more clickstream data becomes available.

Image 3: Series of graphs plotting the found coefficients (see Appendix A) throughout the clickstream. Blue line: splitting based on pageviews, Red line: splitting based on time on site, Green line: splitting based on before and after purchase. White markers show insignificant coefficients.



Conclusion & discussion

Due to the fierce competition between webshops, people's fickle online navigational behavior and the low costs of switching to another site the chance a visitor will revisit the website is low. Hence, it is important to predict online purchasing behavior, in contrast to earlier research, already during the current visit. Therefore, this study aimed at gaining more insight into predicting online purchasing behavior throughout the clickstream based on a series of three analyses.

However, to be able to conduct the series of analyses, I proposed a new categorization to a set of 17 selected, both new and previously tested, predictor variables. The categorization consists of six categories, namely (1) session stickiness, (2) website loyalty, (3) historical purchase behavior, (4) focused search behavior, (5) product interest and (6) non-purchasing intentions.

The aim of this categorization is to provide more insight into the underlying factors influencing online purchasing behavior and to provide a starting point for future research. Furthermore, with the introduction of the new variables AvgPhotoZoom, AvgSizeSwitch and AddToCart, I argue that adding on-page events could increase online purchasing prediction performance. Since on-page events have not been tested before in this type of research I try to encourage future research to investigate on this since many modern websites include on-page actions that cannot be tracked by merely looking at pageview based clickstream data.

This research, however, mainly focused on a series of three analyses to gain more insight into predicting purchasing behavior throughout the clickstream. First, a comparison was made between the use of disaggregate pageview level data and aggregated visit level data for modeling purchasing behavior. In contrast to previous research, I argue that the use of disaggregate pageview level data better represents real purchasing behavior since the use of aggregated data ignores changes throughout the clickstream. This claim was supported by the results of the logistic regression models using both data structures of the same dataset. Results indicated that the direction of the effects on multiple variables differed between the two models.

For example, the effect of Hurry, which was found to negatively influence purchasing likelihood by Van den Poel & Buckinx (2005) consistent with out visit level data, was found to be a positive predictor when using disaggregated data. The opposite turned out to be true for the SiteTransferred variable, which was found to be a positive predictor according to (Park & Chung, 2009). Given that disaggregated data does better represent real purchasing behavior as compared to aggregated data, differences between the models might be explained by the possible loss of information. As a result, I argue that the use of disaggregate data better represents purchasing behavior and should be used in future research.

Secondly, the influence of previous visit clickstream data on the predictive performance of the model was tested by comparing a “Previous knowledge”-model, including knowledge of previous visits, with a “Current visit”-model, including only current visit data. Results indicate that including previous knowledge only marginally increases model performance. Although highly dependent of the exact implementation, I argue that in many cases the additional costs of tracking previous visit data does not weight against the marginally added predictive capacity of the model. Omitting previous visit behavior has three main advantages: (1) it requires less computation power, which can be quite costly on high traffic websites, (2) there is no need to store previous visit data, which save large amounts of data storage costs, and (3) there is no need to use cookies to track returning visitors, which eases the process meeting up with privacy regulations. These advantages can possibly provide large financial and operational benefits for companies implementing purchasing prediction models.

Thirdly, the changes in effect of different variables and the model's predictive performance throughout the clickstream were investigated. By splitting the dataset into multiple time wise successive pieces and modeling purchasing behavior based on each of the individual pieces of data differences throughout the clickstream were investigated. Results showed that a total of 6 out of 17 variable significantly changed in their influence on predicting purchasing behavior throughout the clickstream. Both MaxRepProduct and Hurry had a decreasing influence when more clickstream data became available, while the effect of Duration, TotalPurchases, PercOvervFirst and AboutPages clearly increased later on in the clickstream process.

For example, the number of purchases in previous sessions plays a less important role for predicting purchasing behavior early during the visit. However, later on in the visit the number of previous purchases plays a significantly more important role. Based on these results it can be concluded that the influence of individual predictor variables significantly changes throughout the clickstream process, which favors the use of disaggregate clickstream data for modeling purchasing behavior.

Furthermore, the predictive performance of the model over time was investigated. Results indicate that the predictive capacity of the model steadily increases over time. I argue that the classification sensitivity and the positive predictive value are important predictors in this kind of model since the desired behavior of the model, in most cases, is to predict as much purchases correctly as possible. Since their performance together with the McFadden R^2 value increases over time, I conclude that the overall model performance increases throughout the clickstream process. This is a potentially important conclusion for people implementing these kinds of purchasing behavior models since the results provide valuable insight into changes in predictive performance over time. For example, when personalizing a website based the predicted purchasing likelihood one should keep in mind that personalization will be less accurate early during the visit and will become more accurate as more clickstream data becomes available. Likewise, results of this study regarding changes in individual predictor variables' effects could be interpreted similarly.

Although the use of purchasing prediction models is useful in many cases, there always remains a trade-off between optimizing the website for all visitors and personalizing the website based on models like this. However, this heavily depends on the exact implementation and important to remember is that predicting purchasing behavior strongly depends on the type of website and the intentions of customers visiting the website. Furthermore, it is important to notice that there is a theoretical limit to the predictive performance of predicting purchasing behavior merely on clickstream data since not all consumer behavior, like a person's shopping mood, can be measured using clickstream data.

Future research

Given that the field of clickstream data research is still in its infancy much research still needs to be done. This study was the first to investigate predicting purchasing behavior throughout the clickstream and much more research in this area is needed. Future research might look into the effects on model performance of interaction effects between variables in the model. Due to time constraints and available computation power interaction effects were not included the models used in this study. However, Park & Chung (2009) showed that adding interaction effects in their model significantly improved the model's performance.

Furthermore, future research could investigate the optimal cut-off value for classifying visits as either a buyer or non-buyer. In this research I held on to the generally accepted cut-off value of .5, however, lower values might also be justifiable since obtaining a very high likelihood to purchase is very difficult based merely on clickstream data. Although not part of the analysis discussed in this article, I did conduct a series of brief analyses to find a cut-off value optimizing the predictive performance of the model in terms of four performance measures, namely sensitivity, specificity, positive predictive value and negative predictive value. Based on the dataset and disaggregate model used in the first of three analysis in this study I found an optimal cut-off value of 0.3 in terms of predictive performance of the model. However, additional research is needed to scientifically support this finding.

References

Andersen, J., Giversen, A., Jensen, A. H., Larsen, R. S., Pedersen, T. B., & Skyt, J. (2000). Analyzing clickstreams using subsessions. *Proceedings of the 3rd ACM international workshop on Data warehousing and OLAP - DOLAP '00* (pp. 25–32). New York, New York, USA: ACM Press. doi:10.1145/355068.355312

Ayanso, A., & Yoogalingam, R. (2009). Profiling Retail Web Site Functionalities and Conversion Rates: A Cluster Analysis. *International Journal of Electronic Commerce*, *14*(1), 79–114. doi:10.2753/JEC1086-4415140103

Beatty, S. E., & Ferrell, M. E. (1998). Impulse Buying : Modeling Its Precursors. *Journal of Retailing*, *74*(2), 169–191.

Bellman, S., Lohse, G. L., & Johnson, E. J. (1999). Predictors of Online Buying Behavior. *Communications of the ACM*, *42*(12), 32–38.

Bucklin, R. E., & Sismeiro, C. (2009). Click Here for Internet Insight: Advances in Clickstream Data Analysis in Marketing. *Journal of Interactive Marketing*, *23*(1), 35–48. doi:10.1016/j.intmar.2008.10.004

Bucklin, R., Lattin, J., Ansari, A., Gupta, S., Bell, D., Coupey, E., Little, J., et al. (2002). Choice and the internet: from clickstream to research stream. *Marketing Letters*, *13*(3), 245–258.

Fesenmair, D., Werthner, H., & Wöber, K. (2006). *Destination Recommendation Systems: Behavioral Foundations and Applications* (p. 11). London: CABI.

Google. (2011). *Beyond last click: Understanding your consumers' online path to purchase* (pp. 1–20).

Janiszewski, C. (1998). The Influence of Display Characteristics on Visual Exploratory Search Behavior. *Journal of Consumer Research,*, *25*(3), 290–301.

Lin, L., Hu, P. J.-H., Sheng, O. R. L., & Lee, J. (2010). Is Stickiness Profitable for Electronic Retailers? *Communications of the ACM*, *53*(3), 132–136. doi:10.1145/1666420.1666454

Louviere, J. J., Hensher, D. A., & Swait, J. D. (2000). *Stated Choice Methods: Analysis and Applications*. Cambridge: Cambridge University Press.

Moe, W. W. (2003). Buying , Searching , or Browsing : Differentiating Between Online Shoppers Using In-Store Navigational Clickstream. *Journal of Consumer Psychology*, *13*(1&2), 29–39.

Moe, W. W., & Fader, P. S. (2004a). Capturing evolving visit behavior in clickstream data. *Journal of Interactive Marketing*, *18*(1), 5–19. doi:10.1002/dir.10074

Moe, W. W., & Fader, P. S. (2004b). Dynamic Conversion Behavior at E-Commerce Sites. *Management Science*, *50*(3), 326–335. doi:10.1287/mnsc.1040.0153

Montgomery, A., Li, S., & Srinivasan, K. (2004). Modeling Online Browsing and Path Analysis Using Clickstream Data. *Marketing Science*, *23*(4), 579–595. Retrieved from http://www.jstor.org/stable/10.2307/30036690

Olbrich, R., & Holsing, C. (2011). Modeling Consumer Purchasing Behavior in Social Shopping Communities with Clickstream Data. *International Journal of Electronic Commerce*, *16*(2), 15–40. doi:10.2753/JEC1086-4415160202

Park, J., & Chung, H. (2009). Consumers' travel website transferring behaviour: analysis using clickstream data-time, frequency, and spending. *The Service Industries Journal*, *29*(10), 1451–1463. doi:10.1080/02642060903026254

Sismeiro, C., & Bucklin, R. E. (2004). Modeling Purchase Behavior at an E-Commerce Web Site : A Task Completion Approach. *Journal of Marketing Research*, *41*(3), 306–323.

Thuiswinkel.org. (2012). *Thuiswinkel Markt Monitor 2011-2*.

U.S. Department of Commerce. (2012). *Quarterly retail e-commerce sales 1st quarter 2012* (pp. 1–5). Washington.

Van den Poel, D., & Buckinx, W. (2005). Predicting online-purchasing behaviour. *European Journal of Operational Research*, *166*, 557–575. doi:10.1016/j.ejor.2004.04.022

W3Techs. (2012). Usage of JavaScript for websites. Retrieved September 24, 2012, from http://w3techs.com/technologies/details/cp-javascript/all/all

WebAIM. (2010). Screen Reader User Survey #3 Results. Retrieved September 24, 2012, from http://webaim.org/projects/screenreadersurvey3/#javascript

Zakas, N. (2010). How many users have JavaScript disabled? Retrieved September 24, 2012, from http://developer.yahoo.com/blogs/ydn/posts/2010/10/how-many-users-have-javascript-disabled/

Appendix A

Table A1: Comparison of logistic regression analyses on the dataset split up in equally groups (except from the first group) based on pageviews.

	Pageviews							
	1 - 4	5 - 6	7 - 8	9 - 11	12 - 15	16 - 21	22 - 31	32 - 100
Variable	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.
Pageviews[1]	0.14***	-0.16***	-0.10	-0.09**	-0.06	0.02	0.08***	0.07**
Duration[2]	-0.07***	0.03	0.13***	0.18***	0.26***	0.26***	0.23***	0.28***
VisitFrequency	0.12***	0.07***	0.08***	0.07***	0.05***	0.04**	0.02	-0.02
VisitRecency	-0.02***	-0.02***	-0.01***	-0.01***	-0.01***	-0.01***	-0.01***	-0.00
TotalPurchases	0.20***	0.24***	0.29***	0.24***	0.26***	0.26**	0.32***	0.44**
PurchaseRecency	0.01***	0.01**	0.01**	0.01**	0.01**	0.01**	0.01*	0.01
PercOvervFirst	0.09***	0.00	0.10	0.38***	0.56***	0.62***	0.70***	0.66***
ProdOvervRatio	0.10***	0.15***	0.11***	0.09***	0.08***	0.08***	0.02	-0.03
SiteTransferred	-0.15***	-0.23***	-0.21***	-0.18***	-0.13***	-0.08**	-0.05	0.02
Filters	0.03	0.01	-0.02	-0.05**	-0.07***	-0.08***	-0.11***	-0.08**
AvgPhotoZoom	0.16***	0.10***	0.11***	0.13***	0.10***	0.10**	0.11**	0.09
AvgSizeSwitch	0.43***	0.45***	0.36***	0.39***	0.52***	0.44***	0.53***	0.72***
MaxRepProduct	0.47***	0.34***	0.21***	0.13***	0.10***	0.06***	0.05***	-0.00
AddToCart	1.85***	2.12***	2.30***	2.41***	2.41***	2.46***	2.49***	2.70***
Hurry	0.60***	0.55***	0.43***	0.30***	0.27***	0.18***	0.12*	0.03
PersonalPages	0.03	0.16***	0.17***	0.19***	0.17***	0.15***	0.10***	0.04***
AboutPages	-1.11***	-0.56***	-0.62***	-0.42***	-0.47***	-0.42***	-0.34***	-0.24**
Constant	-4.17***	-3.70***	-4.30***	-4.72***	-5.33***	-5.63***	-5.73***	-6.02***
Pseudo R^2	0.0548	0.1226	0.1616	0.1987	0.2256	0.2318	0.2226	0.1967
Observations	1,442,603	534,685	427,133	489,666	451,890	417,164	352,560	340,687
Sessions	369,662	245,433	194,827	156,316	112,652	73,847	41,934	18,724
Wald chi[2]	11475.17	10239.95	11228.50	12239.30	10446.68	7658.93	4824.04	1655.36
Sensitivity[1]	0.28% (14.38%)	1.55% (4.07%)	1.28% (4.31%)	1.50% (3.85%)	3.02% (4.57%)	3.27% (4.76%)	3.64% (4.89%)	8.01% (4.38%)
Specificity[1]	99.98% (98.44%)	99.91% (99.59%)	99.89% (99.67%)	99.85% (99.72%)	99.72% (99.70%)	99.62% (99.71%)	99.40% (99.70%)	97.65% (99.67%)
Pos. predictive value[1]	29.58% (37.91%)	37.06% (39.65%)	32.70% (46.37%)	35.38% (47.95%)	43.65% (49.86%)	46.26% (51.96%)	47.93% (51.87)	50.85% (46.48%)
Neg. predictive value[1]	97.76% (94.56%)	96.61% (94.01%)	95.95% (94.03%)	94.98% (94.01%)	93.51% (94.05%)	91.18% (94.06%)	87.08% (94.07%)	77.75% (94.03%)

[1]: Square root transformation applied, [2]: Natural logarithm transformation applied, [3]: Percentages between brackets result from classification on the testset to test the external validity.
Significance level: ***: < 0.01, **: < 0.05, *: < 0.1

Table A2: Comparison of logistic regression analyses on the dataset split up in equally groups (except from the first group) based on time on the site in seconds.

	Time on site (seconds)							
	1 - 12	13 - 40	41 -75	76 - 120	121 - 185	186 - 300	301 - 550	551 - 3600
Variable	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.
Pageviews[1]	-0.20	0.14**	-0.19	-0.17***	-0.14***	-0.05**	0.01	0.12***
Duration[2]	-0.02	-0.13***	0.07	0.03	0.39***	0.36***	0.37***	-0.01
VisitFrequency	0.16***	-0.04**	0.01	0.07***	0.07***	0.08***	0.07***	0.00
VisitRecency	-0.02***	-0.03***	-0.02***	-0.01***	-0.01***	-0.01***	-0.01***	-0.00
TotalPurchases	0.27***	0.30***	0.37***	0.22**	0.17*	0.25**	0.23**	0.44***
PurchaseRecency	0.01***	0.01***	0.01***	0.01***	0.01	0.01**	0.01***	0.01
PercOvervFirst	0.23***	0.15***	0.07	0.25***	0.37***	0.49***	0.61***	0.62***
ProdOvervRatio	-0.24**	0.22***	0.14***	0.06**	0.05**	0.02	0.01	0.01
SiteTransferred	-0.16***	-0.20***	-0.22***	-0.20***	-0.15***	-0.12***	-0.04	-0.01
Filters	0.04	0.07**	0.08***	-0.01	-0.05*	-0.10***	-0.11***	-0.08***
AvgPhotoZoom	-0.11	-0.02	0.09***	0.12***	0.12***	0.09***	0.10**	0.19***
AvgSizeSwitch	0.96***	0.55***	0.52***	0.37***	0.42***	0.35***	0.41***	0.73***
MaxRepProduct	0.30***	0.35***	0.25***	0.13***	0.05**	0.02*	0.04**	0.02
AddToCart	1.25***	2.10***	2.29***	2.42***	2.47***	2.46***	2.47***	2.57***
Hurry	0.66***	1.03***	0.76***	0.53***	0.36***	0.15**	0.05	-0.10
PersonalPages	-0.83***	-0.22***	0.00	0.14***	0.17***	0.18***	0.14	0.06***
AboutPages	-1.68***	-1.36***	-1.02	-0.83***	-0.59***	-0.48***	-0.47	-0.25***
Constant	-3.88***	-3.99***	-3.94	-3.77***	-5.56***	-5.72***	-6.11	-4.35***
Pseudo R^2	0.0213	0.0700	0.1129	0.1497	0.1849	0.2043	0.2192	0.2351
Observations	562,830	547,529	568,660	552,061	543,121	570,864	554,704	556,619
Sessions	369,278	280,378	235,766	190,371	145,961	106,840	68,477	45,888
Wald chi[2]	1810.96	3545.52	5935.34	7892.48	8295.17	7529.24	5784.40	3648.96
Sensitivity[1]	0.01% (0.43%)	1.00% (10.72%)	1.72% (2.68%)	0.88% (1.16%)	0.97% (3.31%)	1.35% (4.27%)	2.64% (5.57%)	7.23% (4.43%)
Specificity[1]	100.00% (99.93%)	99.95% (98.89%)	99.91% (99.74%)	99.93% (99.90%)	99.92% (99.75%)	99.88% (99.72%)	99.66% (99.65%)	98.38% (99.61%)
Pos. predictive value[1]	14.29% (28.43%)	33.42% (39.03%)	35.66% (40.58%)	30.32% (44.87%)	35.91% (47.13%)	45.26% (50.63%)	49.28% (51.61%)	50.23% (42.62%)
Neg. predictive value[1]	98.12% (93.82%)	97.58% (94.37%)	97.17% (93.94%)	96.55% (93.86%)	95.36% (93.98%)	93.32% (94.03%)	89.19% (94.10%)	82.46% (94.03%)

[1]: Square root transformation applied, [2]: Natural logarithm transformation applied, [3]: Percentages between brackets result from classification on the testset to test the external validity.
Significance level: ***: < 0.01, **: < 0.05, *: < 0.1

Table A3: Comparison of logistic regression analyses on the dataset split up based on the first occurrence of visiting the checkout page.

	Steps	
	Before checkout	After checkout
Variable	Coef.	Coef.
Pageviews[1]	0.10***	-0.02
Duration[2]	-0.00	-0.01
VisitFrequency	0.04***	-0.05*
VisitRecency	-0.01***	0.00
TotalPurchases	0.33***	0.50**
PurchaseRecency	0.01**	0.00
PercOvervFirst	0.45***	-0.15
ProdOvervRatio	0.03**	0.02
SiteTransferred	-0.09***	0.03
Filters	-0.05**	0.07*
AvgPhotoZoom	0.17***	0.31***
AvgSizeSwitch	0.48***	0.34**
MaxRepProduct	0.05***	-0.06***
AddToCart	2.37***	0.34***
Hurry	0.30***	-0.12
PersonalPages	0.11***	-0.12***
AboutPages	-0.34***	-0.30*
Constant	-4.38***	0.29
Pseudo R^2	0.2123	0.0195
Observations	4,368,140	88,248
Sessions	369,654	9,798
Wald chi[2]	23210.46	92.65
Sensitivity[1]	2.24% (2.55%)	91.77% (89.81%)
Specificity[1]	99.85% (99.86)	17.33% (12.22%)
Pos. predictive value[1]	46.73% (53.83%)	60.94% (6.34%)
Neg. predictive value[1]	94.71% (93.94%)	59.98% (94.77%)

[1]: Square root transformation applied, [2]: Natural logarithm transformation applied, [3]: Percentages between brackets result from classification on the testset to test the external validity.
Significance level: ***: < 0.01, **: < 0.05, *: < 0.1