Predicting purchasing behaviour by using B2B clickstream data

de Roij, C.M.

Award date:
2017

Technische Universiteit
Eindhoven
University of Technology

/ Department of
Industrial Engineering &
Innovation Sciences

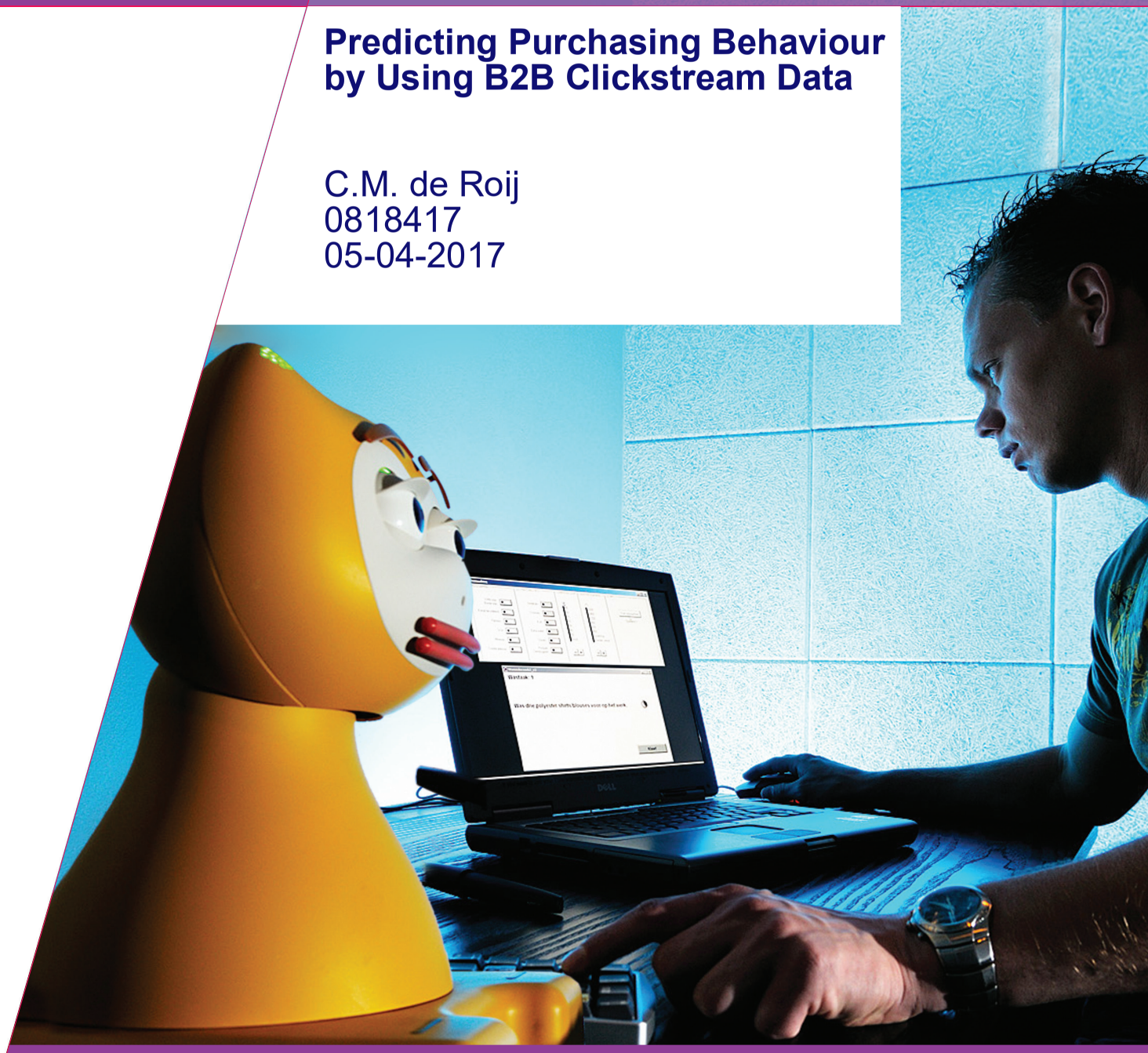
Master of Science
Human-Technology Interaction

Technische Universiteit
Eindhoven
University of Technology

/ Department of
Industrial Engineering &
Innovation Sciences

Predicting Purchasing Behaviour by Using B2B Clickstream Data

C.M. de Roij
0818417
05-04-2017



Master of Science **Human-Technology Interaction**

Where innovation starts

Eindhoven
05-04-2017

Predicting Purchasing Behaviour by Using B2B Clickstream Data

by Cas Michiel de Roij

Identity number 0818417

In partial fulfilment of the requirements for the degree of

Master of Science

in Innovation Sciences

Eindhoven University of Technology

Supervisors

prof. dr. C.C.P. Snijders

dr. ir. M.C. Willemsen

dr. G. Rooks

Abstract

This research uses B2B clickstream data to predict each customer's probability of a purchase based on observed (historical) behaviour. Logistic regression is used to model purchasing probability and show how user characteristics are related to a purchase. This research is also aimed at gaining insight in the dynamics of the purchasing process by examining when a meaningful prediction can be made about a customer's tendency to buy. These analyses are carried out by using 26 predictor variables from which 18 have been previously applied in B2C research. The predictor variables used in this research incorporate a variety of measurements from session based, session focus, visitor demographics and historical (purchasing) behaviour. The composition of these measures are aimed at capturing the diverse nature of visitor goals and distinguish different user types in order to predict the probability of a purchase. The results from these analyses show that B2B customers very frequently visited the website which resulted in a large amount of historical information. On average customers returned more than once a day and made their purchases at the end of a working day. Conversion rates proved to be a factor 3-7 higher than what is observed normal in B2C. The results illustrate the importance of historical information in predicting current purchasing probability, with roughly 50 % of the model's prediction power coming from historical predictors. Visitor typologies used in B2C that distinguish different user goals were observable in the results of the B2B model. The modelling on limited information showed high prediction power from very early in the session underlining the value of historical metrics. In total, this research shows predictive modelling on B2B generated data results in statistically significant and solid regression results. The findings from this research help to understand what determines online purchasing behaviour in B2B by examining the differences between B2B and B2C market structures.

1. Introduction

Increasing worldwide internet accessibility and its growing user base contributes to changing societal dynamics. The daily average time spend online has doubled over the last decade, partly caused by the increasing use of different (mobile) devices (OFCOM, 2015). The increase in internet use has also resulted in the rise of online marketplaces where goods can be purchased or sold. For businesses, these digital platforms provide advantages over conventional stores in terms of (global) accessibility, running cost and potential customer reach. For customers, online platforms are generally easy to use, enable self-service and price comparison from the comfort of home or work. From a research perspective, the rise of these platforms has resulted in a fruitful new source of information that can be used to analyse online behaviour. Researchers and business are interested in the way customers interact with e-commerce platforms to evaluate online behaviour and improve website usability and conversion rates. The digital nature of these sale channels make that behaviour of e-commerce customers cannot be physically observed and behavioural researchers are therefore bound to explore ways of dealing with this. One way of doing so is by using website clickstream data as a source for behavioural insights. Clickstream data contains user specific information that captures online behaviour on aggregated as well as disaggregated levels. This information can be used to deduct visitor goals and preferences as well as categorize different sort of visitors or analyse purchasing behaviour.

Clickstream data has previously been used to assess several aspects of website usage and online behaviour. This method of data collecting has proven to be useful for creating online visitor profiles and typology (Moe, 2003), and predicting purchasing behaviour (Moe & Fader, 2004; Poel & Buckinx, 2004; Verheijden, 2012). Research based on clickstream data has been mainly focussed on B2C (business to consumer) e-commerce channels and provides insights on the choice behaviour of online consumers. Their online purchasing behaviour can be explained (and partly predicted) by using metrics derived from sessions based actions. The way website visitors navigate says something about the purpose of their visit. Analysing clickstream data can reveal the underlying intentions of the visitors and therefore is of great value for researchers and businesses. In a way clickstream data captures online human behaviour and by doing so provides a relatively new source of information for behaviour scientists, marketing research and choice modelling.

Contributions to the field of online (purchasing) behaviour and decision making can roughly be dived by using three different scopes. The first relating to the research performed by behavioural researchers to gain an understanding into online choice and purchasing behaviour. Some influential examples of these contributions are the visitor typology analysis by (Moe, 2003; Venkatesh &

Agarwal, 2006; Montgomery, Li, Srinivasan, & Liechty, 2004), consumer behaviour trajectories and decision making (Senecal, Kalczynski, & Nantel, 2005; Lee, Ha, Han, Rha, & Kwon) and the modelling of purchasing behaviour (Moe & Fader, 2004; Poel & Buckinx, 2004; Verheijden, 2012). Second there is a body of knowledge that focuses on the prediction model performance rather than the output of the model. These contributions are technical and address the math and statistics behind choice behaviour and choice modelling. Thirdly a part of the existing literature focuses specifically on B2B e-commerce. Contributions in this niche are scarce with (Lord & Collins, 2002) describing a B2B visitor typology.

Recently the antecedents and dimensions of trust in B2B e-commerce were examined by Van Tilborg (2015). In his essay Van Tilborg (2015) raises the point that online B2B markets are inherently different from B2C since the nature of the relationship between the seller and buyer is more relationship driven. This complements the findings by (Parry, Rowley, Jones, & Kupiec-Teahan, 2012) who argue that buyer-seller trust is more important in B2B than in B2C caused by relative high value of products or services and a relatively high number of purchases from the same customer. Interestingly these contributions are one of the few that specifically addresses the differences between B2C and B2B e-commerce markets. In the early work by Moe (2003), Moe and Fader (2004) and Poel & Buckinx (2004) no clear distinction between B2C and B2B was made. What does stand out is the growing importance of B2B e-commerce with 9.3 % of total US business sales turnover coming from internet sales in 2015 (Forrester Research, 2015). In absolute numbers the volume of the global B2B e-commerce market is expected to become twice the size of the B2C market by 2020 (Frost & Sullivan, 2014), indicating that B2B online sales channels are rapidly growing. These notions provide the starting point for this research which broadens the current perspective on online purchasing behaviour by examining B2B clickstream data. The goal is to examine and explore B2B online purchasing behaviour and assess possible difference between B2C and B2B online purchasing behaviour. This results into the following research question:

RQ: *Can B2B clickstream data be used to infer and predict purchasing behaviour of users of an online marketplace?*

- *SQ 1:* *How do the predictors of B2B purchasing behaviour relate to existing findings in B2C and what are the differences?*
- SQ 2: *How does the performance of this B2B online purchasing model compare to existing B2C research in terms of accuracy and prediction power?*
- SQ 3: *How soon can a meaningful prediction be made about a customer's tendency to buy during his or her visit? How does this compare to B2C research?*

This study contributes to the field of online decision making by using B2B generated data to predict purchasing behaviour. This enriches the existing scientific field in multiple ways. First the use of B2B data in predicting purchasing behaviour has presumably been limited by the scare availability of inter-company generated data. Whilst it is likely companies have carried out online purchasing analysis indoors, supplying this data for research purposes has not been common. For this study, I teamed up with a large wholesaler which has resulted in a unique insight into their B2B e-commerce platform. This collaboration has made it possible to access large quantities of raw B2B clickstream data usable for the analysis of purchasing behaviour. Second the available volume data enables us to replicate the methods used in B2C research resulting in a level playing field comparison. Results will not be compared head-to-head; however, an evaluation of the direction and magnitude of comparable metrics is performed. Lastly this study provides new possibilities to assess whether (and how) differences between B2C and B2B markets can be observed, explained and measured.

2. Online purchasing behaviour in B2B markets

This review aims to provide an overview of the contributions related to online purchasing behaviour. The literature discussed here is predominantly empirical research and linked to online (purchasing) behaviour or the use of online generated data. Theoretical contributions on predictive modelling and segmentation of (online) choice behaviour that do not contain empirical analysis are not discussed in this review. In the section below we visit several scientific disciplines ranging from marketing research to applied statistics. In the past decade, a substantial portion of the scientific contributions has come from behaviour scientist. These contributions are generally aimed at covering the dynamics of online choice and purchasing behaviour. The relevance of understanding online choice behaviour is evident with the fast-increasing amount of internet users and online shoppers. Another source of clickstream literature contributions come from the domain associated with (online) marketing, where understanding customers' behaviour is vital in determining marketing strategy. Lastly a small part of the contributions come from the field of statistics and math. These contributions are mostly aimed at describing the technicalities of modelling and predicting online choice. Looking closely at the research field reveals there is an unequal divide between B2C and B2B contributions. The latter being underrepresented in terms of the number on contributions related to B2B, especially regarding B2B market share is growing rapidly. A handful of contributions, specifically dedicated to online purchasing or behaviour in B2B, have been published in the last 15 years. In this literature review I start by analysing the contributions in the B2C field followed by a review of the contributions in the B2B field. This is followed by a review of the differences between B2B and B2C and how these differences can be tested. From the literature, I propose a set of variables to predict purchasing behaviour in B2B.

2.1 Online purchasing behaviour in B2C

Previous research on online choice behaviour has resulted in several insights in how online visitors navigate through websites and how individual actions and behaviour can be used to explain purchasing behaviour. To start with: observed visitor behaviour can be used to compute a typology of visitors based on actions undertaken online. Research carried out by Moe (2003) indicates that online shoppers are driven by different motivations and that these motivational differences are observable by examining in-store navigational patterns. By using clustering techniques Moe found that online shoppers could be categorised as buying, browsing, searching or knowledge building shopping strategies. This typology differentiates between an immediate or future purchasing horizon and divides search strategy into directed or exploratory. This typology is aimed to match diverse types of user who show similar type of online behaviour. These customer groups all have their own

purchasing likelihoods resulting from different form of behaviour. Dissimilarities between the observed groups suggested that there are distinctive differences between the type of visitors and their underlying objectives (Moe, 2003). The differences in objectives are observable in the way visitors use the online shopping environment. This research by Moe (2003) was one of the first contributions that empirically studied the mechanisms of online behaviour and has served as a foundation for further research. We do see that other authors have different views on visitor typology. According to Venkatesh & Agarwal (2006) the visiting process can be seen as a dynamic process that cannot be captured in static profiling. A visiting session can result in multiple outcomes while the purpose and goal of the visit can change during the visit. In other words: a visitor's goal is not always fixed and can change under the influence of several factors. According to Venkatesj & Argerwal website usability is a key factor that influences visitor goals. Usability can therefore have a positive or negative effect on website use and conversion rates (Venkatesh & Agarwal, 2006). An example of this is seen when poor usability related to the checkout process causes customers to drop out. In this case, a customer visits the website with a purchasing intention and leaves without completing the transaction.

The notion that online behaviour can be captured in different metrics was used by Poel & Buckinx (2004) to predict online purchasing behaviour. Online visitor leave small pieces of information behind that can be used to model and predict future buying behaviour. Poel and Buckinx (2004) used this to construct various metrics in **four** categories to model purchasing probability. In their research a wide variety of predictive metrics was used which resulted in new insight on the constructs of choice behaviour. Let me borrow the categorization they used to group purchasing behaviour predictor variables: The **first** category Poel and Buckinx (2004) used were general session metrics. These general measures contain information on overall time spend on a website and the number of pages viewed. As pointed out by Lin et al. (2010) the metrics are similar in the way that they both measure a form of session stickiness. One could argue that measuring the total time spent and number of pages viewed is measuring the same thing twice: the amount of engagement a visitor displays towards the websites content (Panagiotelis, Smith, & Danaher, 2014). When relating these metrics to purchasing behaviour; previous research indicates that an increase in the amount of 'pageviews' and 'duration' are positively related to (the amount of) online purchases (Moe & Fader, 2004; Sismeiro & Bucklin, 2003; Lin, Hu, Olivia, Sheng, & Lee, 2010; Poel & Buckinx, 2004; Verheijden, 2012). With the number of unique pages viewed having greater significance in predicting purchases than total visit duration (Panagiotelis, Smith, & Danaher, 2014). The **second** category described by Poel and Buckinx (2004) are detailed clickstream metrics. These measures say something about how a visitor has interacted with a website and its content. In general, we see that detailed clickstream metrics are

on-site measurements that capture a specific element of website use. Poel and Buckinx (2004) divided detailed measurements into search, product or general related measurements and concluded that adding these measurements led to improved predictive performance. Other research indicates that measurements of product interest such as product (detail) page viewing, photo viewing or searching positively correlates with increasing purchasing probabilities (Verheijden, 2012) (Moe, Johnson, Fader, Bellman, & Lohse, 2004; Poel & Buckinx, 2004). In contrast to product interest metrics the use of search filters proved to be negatively correlated with purchasing probabilities (Verheijden, 2012), together with measurements for non-purchasing intentions such as the visiting of personal or company information pages (Verheijden, 2012; Poel & Buckinx, 2004). The **third** category consists of demographic variables which were found to be very useful when distinguishing buyers from non-buyers (Poel & Buckinx, 2004) or when making a customer segmentation (Sismeiro & Bucklin, 2003). Demographic variables such as gender, age, income and education are previously used for the construction of online user types (Padmanabhan, Zheng, & Kimbrough, 2001). The **fourth** category used by Poel and Buckinx (2004) describes variables related to historical (purchasing) behaviour. Metrics on historical purchasing behaviour are widely used in offline retailing as they contribute to a better understanding of future purchases (Schmittlein & Peterson, 1994). Displayed behaviour in previous visits can for example be used when modelling future profitability or overall customer lifetime value (Schmittlein & Peterson, 1994). This indicates that historic behaviour, such as the amount of store visits or overall spending, are relevant predictors when modelling future purchasing behaviour. What we see in general is that past actions provide guidance when modelling future (purchasing) behaviour. In modelling online purchasing behaviour, the total number of previous purchases is argued to be positively related to future purchasing probability (Poel & Buckinx, 2004; Verheijden, 2012; Moe & Fader, 2004). One could think of a regular customer that is familiar with a specific shop resulting in a lower purchasing threshold (Beatty & Ferrell, 1998), increasing the probability of future purchases. This resonates with the findings of Poel and Buckinx (2004) who found that an increase in the number of days since the last purchases is negatively correlated with purchasing probability. In other words, this is a different way of saying customers who buy frequent are more likely to make a future purchase compared to customers who have an irregular buying pattern.

2.2 Online purchasing behaviour in B2B and differences with B2C

Multiple scholars have addressed the workings of online purchasing behaviour in B2C. In examining the current body of knowledge, a clear distinction between B2C and B2B markets seems to be lacking. The research described previously solely uses B2C generated data and do not elaborate on the applicability of their findings in a B2B setting. There are some contributions that address specific

B2B behaviour, such as the works of Lord and Collins (2002) who explored different online search processes by organisational users. Their research was conducted with B2B generated data and with the use of factor analyses Lord and Collins (2002) distinguished 3 factors that represented different visitor goals. These goals (or website activities) where purchase related, quality or performance related or non-purchase related. These findings are equivalent to the visitor typology proposed by Moe (2003) as they differentiate between shopper's motivations and goals. Apart from the findings by Lord and Collins (2002) there seems to be no literature on modelling online purchasing behaviour which uses distinct B2B market data. Existing literature does not reveal why this might be the case however it could have something to do with the limited availability of appropriate data. Inter-company generated data usually contains a high degree of commercially sensitive information which companies like to keep for themselves.

Since the amount of B2B clickstream literature is limited, there is not a lot of empirical literature that can be reviewed. This is problematic, given several contributions on differences between online B2C and B2B markets underline the necessity to examine purchasing behaviour in B2B markets. To start with, Van Tilborg (2015) argues that buyer-seller trust is an essential part of online B2B trading. In B2B the buyer-seller relationship differs from B2C as it is more relationship driven. This relates to a difference in buying behaviour between B2C and B2B customers since buyer-seller relation stands central in B2B opposed to a product-central relation in B2C. The relationship in B2B is built on trust which is the most influential factor in the success of a B2B marketplace (Beige & Abdi, 2015). Trust is a complex and abstract concept and many scholars have devoted time and effort to define trust in (e)commerce. In this review, we will not dive into trust however use a common definition to explore B2B relations. The broad definition that (Gefen, 2000, p. 726) gives is: "*Trust,…, is the confidence a person has in his or her favourable expectation of what other people will do, based in many cases, on previous interactions".* What we can draw from these contributions is that trust is based on previous interaction and trust is one of the most influential factors for success in B2B. When trust is established the road for future transaction seems clear. We do know that B2B e-commerce sales repetition is relatively high with average purchases being of higher value compared to B2C purchases (Parry, Rowley, Jones, & Kupiec-Teahan, 2012). One of the effects of trust in B2B could therefore be a relatively high conversion rate, specifically compared to B2C where buyer-seller relation is assumed to be not as strong. Looking further into B2B, the customer bases in B2B seem to be more concentrated with a small number of customers generating a large part of the revenue (Tsiros, Ross, & Mittal, 2009). This further emphasises the importance of relationships in B2B since long-term commitments are generally the most profitable (Tilborg, 2015). This brings us back to the notion that

the relationship between buyer and seller might have a considerable influence on B2B buying behaviour.

When examining previous clickstream research, it should be noted that detailed clickstream measures are focussed on on-site action rather than on-site items and content. This makes results broadly interpretable since website specific design and content are not taken into account, contributing to the generalizability of clickstream research findings. As Panagiotelis, Smith and Danaher (2014) bring forth this focus does not account for differences in the behaviour of visitors caused by on-site content such as product characteristics or website design. What stands out is that any effect should be interpreted carefully and examined in the light of the examined data and research methodology. Nonetheless, the metrics discussed in the previous sections contribute to a better understanding of the mechanism influencing online choice behaviour and are therefore a logical starting point to start our analysis.

Research in B2C and B2B has shown purchasing behaviour can be predicted using different metrics. These metrics are derived from general session information, in-session information, visitor demographics and historical behaviour. The insights resulting from this mainly come from B2C related research where customers display distinctive user and buying patterns. In general, we encountered reasons to assume B2C and B2B purchasing behaviour might differ. Under the influence of long a term relation between buyer and seller, B2B visitors are presumed to visit more frequent. Research on B2C generated data by Verheijden (2012) indicated that historical variables only have a marginal effect in the prediction of purchasing behaviour. In other words, the power of Verheijden's (2012) prediction model did not significantly decrease when customer's historical purchasing information was left out. If B2B customer are indeed frequent visitors it is likely there is more historical information per customer compared to B2C, therefore enhancing the performance of historical variables.

2.3 Research Metrics

Research on B2C data provides a solid starting point to derive which metrics to use in a B2B context. The selection of predictor variables was carried out using a subset of my B2B dataset to test which variables could be deducted from the raw clickstream data. In doing so the set of predictor variables used by Verheijden (2012) and Poel & Buckinx (2004) were used as a starting point. The method of analyses in these contributions is equivalent to the one used in this research which contributes to comparability of different findings. In total, 26 variables were deducted from the dataset from which 18 have been used and tested in previous research. These variables are grouped in four categories. The categorization applied here is based on the works of Moe (2003) and Poel & Buckinx (2004). In research by Moe (2003) general and session focus measures were used to categorize visitors. The

distinction between general and session focus measures seems to be in place here since it provides a clear separation between general and page-tot-page related metrics. The third and fourth categories are session based metrics that can be deducted at customer level. Visit demographic tell us something about the nature of the visit and should not be mistaken with customer demographics. Metrics on historical behaviour provide information on previous website interaction and are deducted at customer level as well. Table 1 provides an overview of all the predictor variables, their reference and the direction of the effect between predictor and purchasing probability.

General session measures

General session metrics are high level measurements that indicate the total time spend on website as well as the number of pages that were viewed. The total number of pages viewed was used by Lin et al. (2010) and Verheijden (2012) and was found to be positively related to a purchase. The total duration of a visit was used by and array of authors and all speculate a positive relationship between increasing on-site duration and purchasing probability. Only few have tested on-site duration and the number of pages simultaneously, with Lin et.al. (2010) concluding that the number of pages viewed might be more indicative in predicting a purchase that the total session duration. The positive effect of general session measures on purchasing probability is expected to also exist in B2B. On-site interactions take time and trigger page views and it would therefore be logical to expect a positive relation with purchasing probability.

Session focus measures

The session focus metrics are derived from on-site actions and interaction with content. The metrics reflect the way a customer has interacted with the website during his or her visit. The number of used search filters indicates the total amount of unique search filters options that were used during a session. Research by Verheijden (2012) suggests that an increase in the number of filters has a negative impact on purchasing probability. As mentioned by Olbirch & Holsing (2011) the use of filters can be an indication for knowledge building behaviour. Similar to this non-purchasing intention is the visit of company and/or personal account related pages. The display of pages that are non-product related can be seen as an indication for a visit that is not purchase related and thus having a negative impact on purchasing probability. In contrast to this the viewing of pages that are content related has previously shown to be positively related to a purchase. As shown by Moe (2003) in her typology of store visits direct purchasing intention is shown in a high number of (single) product viewings. Captured by Verheijden (2012) is the ratio of product detail over product overview pages. With a high ratio indicating relatively low product interest (product detail pages are not viewed) and vice versa. Five metrics that were also introduced by Moe (2003) capture the content viewed during a session. Initially mend to categorize visitors the percent product, home, content, search and

assortment pages capture the focus of a visit. These metrics are incorporated in this research since the customers purchasing intention can be deducted from the distribution between these five variables. Lastly, I incorporated three additional metrics to this category that might to help further examine B2B purchasing behaviour. As the work of Moe & Fader (2004) indicates; measures of product interest contribute to a better understanding of the tendency to buy. I propose three metrics to capture this. First the number of (unique) product searches could be a measurement for search engine performance. As mentioned earlier an increase in search filters had a negative effect on purchasing behaviour. It would be in line with expectation to reveal a negative relationship between the number of unique searches and purchasing probability in this study. High search volumes can indicate knowledge building or browsing rather than the effect of a direct purchasing intention. This also holds for the second additional measurement that captures the total number of deleted filters. The third additional measure indicates the total number of product compares during a session. This metric is specifically aimed at capturing non-purchase intentions since a product specification comparison is expected to indicate a form of buyer uncertainty or knowledge building. The effects of the metrics mentioned above are expected to similarly exist in B2B. Here B2B website usage is anticipated to be comparable with B2C.

Visit demographics

The metrics on customer visit demographics are aimed to capture information from the visit itself rather than its content. Previously researched by Park & Chung (2009) and Verheijden (2012) the metric site transferred indicates if the website is being accessed by means of a search engine. As noted by Park & Chung (2009) it is believed that visitors that enter the website via a search engine display a form of exploratory search making them less like to make purchase. In both the study of Verheijden (2012) and Park & Chung (2009) site transferred visitors had a lower purchasing probability. We see a similar effect with the variable hurry, which indicates whether or not the average time per page is lower than that of previous sessions (for that particular customer). Investigated by Poel & Buckinx (2004) it turned out that visitor that were in a 'hurry' had a lower purchasing probability. Verheijden (2012) argues an online purchase requires a high level of involvement and deep consideration. Visits in which the final buying decision is made therefore take longer. Building on the findings of previous research I propose four additional metrics to further capture visit demographics. Two of these metrics account for a strongly deviating total time on-site. Here a shallow visit is described as a visit that has a total on-site time of less than 5 seconds. My argumentation here is that this captures a group of visitors that have no buying intention what-so-ever and accidently accessed the website. Furthermore, I capture the time of day on which a session was started. Since this research works with B2B data I suggest that sessions within business hours

have a higher purchasing likelihood than those outside these hours. Furthermore, I suggest a negative effect of the metric hurry in B2B based on the difference in browsing behaviour between B2C and B2B. A B2B customer is expected to make a better-informed purchasing decision and does so in relatively short session. The argument here is that longer session will typically reflect a customer that is searching for information while a short session indicates a (direct) purchase related action.

Historical behaviour

The metrics that can be placed in the category historical behaviour capture a variety of previously displayed behaviour. The metrics that are used in previous B2C research can be divided into historical visit and historical purchasing predictors. Here the visit frequency is a measure of a customer's recency that measures the total number of previous visits. The metric 'time since previous visit' accounts for the time since the last website visit measured in days. In research by Poel & Buckinx (2004) and Verheijden (2012) the number of previous visits turned out to be positively related to purchasing probability. Their research furthermore indicates a negative relationship between an increase in the number of days since the last visit and purchasing probability. This is possibly an indication that customer who frequently use a website are more likely to make a purchase compared to those to do not make frequent visits. Looking at the effects of historical purchasing predictors we see something quite similar with the amount of previous purchases being strongly positively related to purchasing probability. When applying these metrics in this research I foresee some important differences with previous B2C results. In his research Verheijden (2012) tested the impact of historical information by comparing models with and without previous knowledge. In this case previous knowledge stands for metrics that are deducted from previous website interactions, such as the total number of previous visits and previous purchases. Verheijden (2012) concluded that the exclusion of predictors deducted from previous behaviour did not significantly decrease the prediction power of his model. If B2B relations between buyer and seller is indeed more relationship driven I expect a more significant role for metrics that capture previous (purchasing) information and behaviour. If B2B relations are in fact more durable, concentrated and long-lived, historical variables are likely to account for a significant part of the models' prediction power.

Variable Category and Name	Description	Reference	Direction
General session measures			
Pageviews	Total number of pages viewed during a session	1,2,4,6	+
Duration	Total session duration in seconds	1,2,3,4,5,8	+
Session focus measures			
Number of product searches	Unique number of product searches		N.A.
Number of search filters	Total number of search filters used	2,8	-
Number of deleted filters	Total number of search filters deleted		N.A.
Number of product comparisons	Total number of product comparisons made		N.A.
Personal account pages visited	Dummy variable for visited personal pages	2,3,8	-
Company about pages visited	Dummy variable for visited about pages	2	-
Product detail pages visited	Dummy variable for visited detail pages	8	+
Product detail page ratio	Ratio of product detail pages over product overview pages	2,6	+
Percent product pages	Percentage of visited product related pages	6	+/-
Percent home pages	Percentage of visited home related pages	6	+/-
Percent content pages	Percentage of visited content related pages	6	+/-
Percent search pages	Percentage of visited search related pages	6	+/-
Percent assortment pages	Percentage of visited assortment related pages	6	+/-
Visit demographics			
Site transfer	Site accessed via search engine (dummy)	2,7	+/-
Shallow visit	On-site time is less than 5 seconds (dummy)		N.A.
Hurry	Average time per page is lower than average in previous sessions	2,3	-
Long session	In session time-out has been longer than 30 minutes		N.A.
Morning session	Session took place in the morning (dummy) 1-12 AM		N.A.
Afternoon session	Session took place in the afternoon (dummy) 1-6 PM		N.A.
Historical behaviour			
Visit frequency	Total number of previous visits	2,3,9	+
Time since previous visit	Number of days since previous visit	2,3,9	-
Purchasing history	Total number of previous purchases	2,3,9	+
Time since previous purchase	Number of days since the previous purchase	2,3,9	+/-
Purchasing ratio	Ratio of purchases over non-purchase visits		N.A.

Table 1. Overview of predictor variables used in this research, their reference and their observed effect on purchasing probability in other studies.

1. (Lin, Hu, Olivia, Sheng, & Lee, 2010)
2. (Verheijden, 2012)
3. (Poel & Buckinx, 2004)
4. (Panagiotelis, Smith, & Danaher, 2014)
5. (Padmanabhan, Zheng, & Kimbrough, 2001)
6. (Moe, 2003)
7. (Park & Chung, 2009)
8. (Olbrich & Holsing, 2011)
9. (Moe & Fader, 2004)

3. Data and Methodology

To examine purchasing behaviour in a B2B environment we consider actual clickstream data. The analysis in this research were established through collaboration with a large B2B wholesale distributor of technical products and supplies. This wholesaler is located in the Netherlands and mainly operates on the domestic market. Yearly revenue on this online channel is in excess of €660 million, providing a data set that is an order of magnitude larger than those used in the literature. To illustrate the size of this dataset we can conclude that it is over 400 times larger than the sets used in the contributions by Moe (2003) and Moe and Fader (2004), more than 100 times larger than the set used by Poel and Buckinx (2004) and more than 10 times larger than the data used in the research of Verheijden (2012) or Park and Chung (2009). The wholesaler in question provided full access to their online clickstream data and by doing so made available data on more than 3.2 million online sessions, generated between September and December 2015. This data was generated on an online sale platform where customers can search through a product catalogue that holds more than 2 million products. Products can be ordered directly from the website and are usually delivered within 24 hours. The distribution and delivery of products is managed in-house and products are distributed using company owned facilities. Being a B2B orientated seller, every online customer has to log-in before any products info can be searched or an order can be placed. A typical customer uses some form of product searching by using the on-site search engine. Products are categorized into 7 categories ranging from electronics, heating and sanitary supplies to consumer orientated products and safety equipment. Typical customers of this wholesaler are companies that operate in the construction industry, ranging from listed construction companies to self-employed construction workers. In general, order frequency ranges from several times a day for (very) frequent customer to an irregular order pattern for an infrequent customer, all of whom are captured in data that was provided. What makes this data unique is that customers can generally be identified very early in a session. This stems from the structure of the website where all product related content is only accessible after a user log-in. The data used in this research was captured using a JavaScript tracking application named Divolte. Similar to tools as Google Analytics this application collects clickstream data by the use of JavaScript tagging. The collected data is stored in a Hadoop environment to unsure fast data extraction and analysis. Table 2 provides some insight in the size of the dataset before pre-processing.

Total number of sessions	3,217,667
Total number of pageviews	46,692,080
Total number of customers[1] in dataset	36,175
Conversion rate	11.51 %

Table 2. Data descriptives from the B2B clickstream data used in this research.

B2B purchasing behaviour is captured in a session based, dichotomous variable which is '1' when a purchase was made and '0' when a purchase did not occur. In our case the outcome variable indicates a sale being completed and is triggered by the firing of the 'order confirmation' page. Logistic regression analysis enables us to calculate a purchasing probability for every session. Results from logistic regression analysis are relatively ease to interpreted. An array of authors has used similar analysis techniques (Poel & Buckinx, 2004; Moe & Fader, 2004; Olbrich & Holsing, 2011; Verheijden, 2012) which contributes to the comparability of our findings. In B2C literature we further see a widespread use of linear regression when one has an interval variables rather than a dichotomous outcome. For example, as Park and Chung (2009) have done when predicted online spending.

Before any analyses can be performed the raw data must be processed so that website interactions can be aggregated and captured in clickstream metrics. This transforms the data from a single pageview per row to aggregated (session) level information per row, significantly reducing data size while maintaining the necessary level of detail. The main challenge in handling this dataset was working with the size of the raw data. Computation power was limited since the data processing was carried out on a single workstation. The size of dataset required a stepwise approach and consisted of 5 steps and around 500 lines[2] of code. The first two steps of data processing were aimed at transforming and spitting the data into workable chunks. Each chunk of data was around one Gigabyte in size and contained over 30.000 sessions. Step 3 and 4 consisted of running some general computations on these chunks so a selection of variables could be dropped, reducing overall file size. At the end of step four the data was cleaned and separated into 100 lose day files. These files were merged in step 5 to compute the historical variables in the total dataset. The large amount of sessions and number of unique customers indicate that the B2B website in question has a large user-base. Part of this user-base is a group of so called 'hard core never buyers'[3] who use the wholesalers'

[1] A customer is defined as a unique user and is identified by a customer id and user id. In the dataset multiple users are able to use the same customer number. In our analysis customer are identified as by their user identification.

[2] Appendix 2 provides an overview of the computations carried out in these five steps

[3] A phrase adapted from Moe & Fader (2004)

website like a catalogue and have no intention of buying. Therefore, I exclude these customers from my analysis since there is no indication these 'hard core never buyers' have made or will make a purchase. The size of this group is substantial and about 1/3 of the total number of session. The number of average pageviews in this group is lower than that of buying customers. The amount of time spend on a single pages is relatively high and around 50 % higher than the time per page of the buying group. Interestingly the non-buying group visits more frequently than the buying group with time between visits being lower than 24 hours. These group characteristics resemble the characteristics of the 'knowledge building' cluster defined by Moe (2003).

The exclusion of 'hard core never buyers' reduces the amount of analysable sessions. The final analysis was performed on 2,099,732 sessions generated by 22,974 customers. The analysis itself was carried out using Stata 14. In STATA the 'logistic' command was used to perform the analysis. Significance levels were calculated using Stata's "cluster" option to take into account that observations are nested within users. Stata's built-in command that accounts for these effects somewhat more appropriately ("xtlogit" or "mixed") put a too heavy burden on my computer. An initial run of the xtlogit command with all relevant variables included suggested that the amount of variance at the user level conditional on the inclusion of these variables was small, which is why I chose to use the standard logistic regression command with clustered standard errors instead. Several predictors were transformed to a logarithmic scale to account for highly skewed distributions. Several predictors were included as dummy variables to try and further improve model performance.

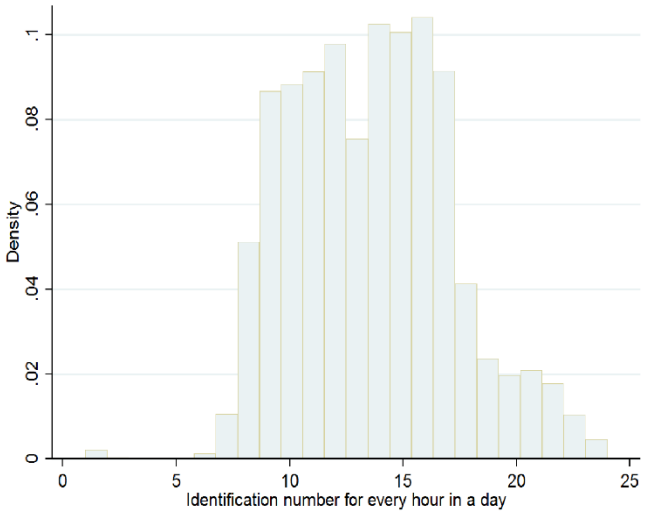
3.1 Composing a prediction model

The final prediction model was composed in several steps. The goal was to accomplish solid model performance while maintaining the significance of the predictor variables. The first step was therefore conducting several logarithmic transformations which resulted in all predictors being significant. To accomplish this the fact that the logistic regression model did not converge was ignored[4]. Accounting for the convergence failure was a challenge and the use of several STATA options[5] to account for this did not have the desired effect. Eventually the solution was found in transforming the variable 'number of searches' to a logarithmic scale. In addition to this (logistic) regression assumptions were considered and checked before selecting a final prediction model. Overall the composing of the predicting model proved to be a time intensive process.

[4] Convergence issues were ignored by setting a maximum number of iterations of 8.
[5] 'Gradient', 'difficult' and 'technique' functions were tried to cope with the issue of the model failing to converge.

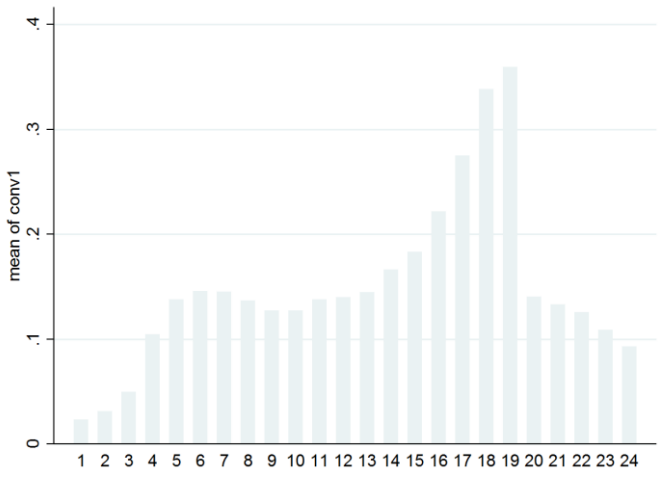
4. Results

The main goal of this thesis is to examine purchasing behaviour in a B2B setting. In the previous sections we saw how the current body of knowledge has developed in the last couple of decades. In predicting B2B purchasing behaviour 26 predictor variables were selected to gain inside in the mechanisms that influence online choice behaviour. To answer the research question two analyses were performed using this set of variables. The first analyses containing the full dataset, the second was aimed at modelling purchasing behaviour using sections of the dataset. To address the main research question, sub question number one and number two the first analyses will be used. The second model is used to shed light on the third sub question.



Before the results of these analyses are discussed let us zoom in on some general findings. The data shows most session occur during working hours, illustrated by figure 1. The graph in this figure shows most activity is between 8 AM and 7 PM. The website in question offers next-day delivery if an order is placed before 7PM which might explain the drop of the number of session after 7 PM. The distribution of the number of sessions shows a working day pattern.

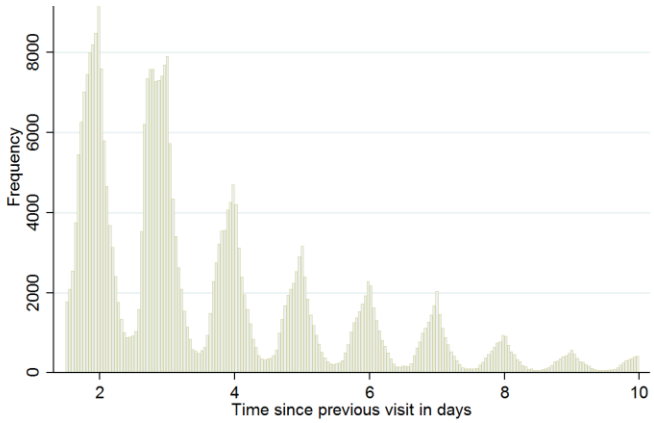
Figure 1 The distribution of the average number of sessions by every hour of the day



Overall conversion rate is high with 17.5 % of all visits leading to a purchase. In B2C literature we see that conversion rate roughly range from 8 % (Poel & Buckinx, 2004) to 1.75 % in the study by Verheijden (2012) and average B2C e-commerce conversion rates are generally between 3-5 % (Burstein, 2015). The absolute number of conversions is highest at the end of the working day. Interestingly the percentage of sessions that result in a purchase increases during the day.

Figure 2 The distribution of the average amount of conversions by every hour of the day

The conversion rates peaks between 6 and 7 PM and reaches more than 30 %. During the rest of the day the average conversion rate sits below 20 %.



A high number of customer seems to be specifically directed at purchasing when visiting the B2B website. On average, customers return regularly and visit the website daily. Figure 3 shows these visits even occur on roughly the same time every day. When we plot the time since previous visit we see a negative exponential relation between the number of visits and the time between these visits.

Figure 3 The distribution of the average time since previous visits by number of days

On average a customer made 122 visits in the 3.5 months. This average number is somewhat misleading since about 5 % of the total customer base makes more than 5 visits a day, strongly influencing the average number of visits. Because every customer must log-in we can establish such accurate visit frequency numbers. This stands in contrast with B2C where customers are commonly tracked by cookies rather than customer credentials which makes it difficult to determine if a customer has visited before. Having established this let us now focus on the effects and direction of the predictor variables.

4.1 Predicting purchasing behaviour in B2B - Interpretation of predictor effects

The effects of the 26 predictor variables are displayed in table 3. The effects themselves are shown as the regression coefficient and odds ratio. To facilitate the process of evaluating the predictors direction and magnitude a selection of variables was split into a dichotomous and interval variable. This helps to isolate different effects of the same predictor. For example: the number of search filters was split into a variable indicating if search filters were used (dichotomous) and a variable that contains the number of searches (interval). This isolates the effect of general search filter usage from the effect of an increase in the amount of search filters. This technique is used to evaluate the effects of the variables: number of product searches, number of search filters and number of deleted filters. The result from this logistic regression can be used for multiple purposes as in: the evaluation of model prediction power, classification performance or identification of a ‘likely buyer’ group. The evaluation of the model results is restricted to a direct comparison with results from previous research on predictor direction and magnitude.

General clickstream measures

The two general clickstream measures, [**Pageviews**] and [**Duration**], turn out to be both positively related to purchasing probability. As one might expect an increase in the number of pages viewed and overall time spent increases the chance of a purchase which is consistent with earlier research. We do see that the effect of the variable duration is larger than that of the variable pageviews in this B2B model. Interestingly this is not reflected by the marginal contribution of an increase in both variables in relation to the probability of a purchase. In other words, the effect of the amount of viewed pages is very small in this B2B model while the marginal effect of an increase in the number of pages viewed does significantly increase the change of a purchase. Figures 4 and 5 are indicative of the differences between these two variables. Similar to Lin et. al. (2010) the number of pages viewed is more indicative in predicting purchasing behaviour than the duration of a visit.

Session focus measures

An increase in the amount of [**Number of product searches**] turned out to be positive related to a purchase. This positive effect is relatively large which indicates that search engine use can be a predictor of purchasing behaviour. To gain a better understanding of the use of the search engine and its effect on purchasing probability the number of product searches was split up into a dichotomous and interval variable. Here the dichotomous variable indicates if the search engine is used or not. The effect of this dichotomous variable turned out strongly positive indicating that the use of the on-site search engine will generally result in higher purchasing likelihood. This does not come as a surprise since the search function in most cases be used in product selection or comparison. When looking at the effects of [**Number of search filter**] and [**Number of deleted filters**] we see a negative correlation between the amount of (deleted) filters and purchasing probability. Resonating with the findings of Verheijden (2012) the use of search filters is argued to be an indication of knowledge building behaviour (Olbrich & Holsing, 2011). When splitting up[4] the effects of these predictors a more nuanced view becomes visible. In session where search filters where used (or deleted) the customer was more likely to buy products compared to the session where no filter options where used. It is the increase in the amount of used (or deleted) filters that results in a negative effect on purchasing probability, not the use of the search option themselves. This is an important distinction to make since search option will likely be aimed at enhancing product findability. The effect of the [**Number of product compares**] is slightly negative and is presumed to be indicative of a form of buyer uncertainty. Similar to these results are the effects of [**Personal account pages visited**] and [**Company about pages visited**] which are both negative related to purchasing probability.

Logistic regression results

Variable group and name	Coefficient	Odds-ratio	Z-Value
General session measures			
Pageviews	0.0064	1.0064	22.11**
Duration[6]	0.3888	1.4752	98.56**
Session focus measures			
Number of product searches[3]	0.2701	1.3109	30.10**
Number of search filters[3]	-0.0796	0.9235	-12.71**
Number of deleted filters[3]	-0.1334	0.8751	-11.41**
Number of product comparisons	-0.0205	0.9797	-6.70**
Personal account pages visited	-0.1751	0.8394	-11.77**
Company about pages visited	-0.2790	0.7565	-11.42**
Product detail pages visited	0.6371	1.9809	47.03**
Product detail page ratio	-0.0761	0.9267	-4.20**
Percent product pages	-0.0512	0.9501	-79.72**
Percent home pages	-0.0655	0.9366	-116.07**
Percent content pages	-0.0884	0.9154	-24.81**
Percent search pages	-0.0703	0.9322	-88.56**
Percent assortment pages	-0.0916	0.9125	-89.99**
Visit demographics			
Site transfer	-0.0653	0.9367	-4.61**
Shallow visit	-1.0678	0.3437	-15.43**
Hurry	0.6150	1.8498	71.74**
Long session	-0.4099	0.6637	-29.77**
Morning session	-0.1334	0.8750	-9.49**
Afternoon session	0.4369	1.5480	35.53**
Historical behaviour			
Visit frequency[3]	-0.4895	0.6129	-33.72**
Time since previous visit	0.0254	1.0258	21.85**
Purchasing history[3]	0.4896	2.0711	42.00**
Time since previous purchase	-0.0126	0.9875	-16.64**
Purchasing ratio	0.0442	1.0452	86.68**
Constant	-4.1169	0.0163	-89.99**

McFadden R2	0.3956
Observation (sessions)	2,099,732
Nr. Clusters	22,974
Wald Chi2	84,015

Classification results (cut-off 0.45)	
Sensitivity	53.75 %
Specificity	93.68 %
Positive predicted values	64.40 %
Negative predicted values	80.44 %

Table 3 Logistic regression coefficients, odds ratio and significance of the B2B purchasing prediction model

[6] Logarithmic transformation applied. *Significance level <0.05 / ** Significance level <0.01

The negative effects of these non-purchase related predictors are all strong and are therefore argued to be indicative for identifying visitors that have no intention of buying. In contrast to these negative effects is the strong positive effect of [**Product detail pages visited**]. This is no wonder since the product detail page is part of the most common conversion path. Customers that do not consult product detail pages are presumed to have non-purchase related intentions as there was no direct need to review product details. Interestingly we see a negative effect of the [**Product detail page ratio**] on purchasing behaviour. What we have captured here is a group of customers that view a large amount of product detail pages without buying. This is an interesting finding regarding the positive effect of [**Product detail pages visited**]. A high amount of product detail pages over product overview pages turns out to negatively influence purchasing probability. We also see a negative effect of all the content percentage related predictors [**Percent product pages**], [**Percent home pages**], [**Percent content pages**], [**Percent search pages**] and [**Percent assortment pages**]. First introduced by Moe (2003), these five variables have a large contribution on the overall prediction power of the model. Running the B2B model without these predictors resulted in a McFadden R^2 drop of 8 %. The negative effects are presumed to be caused by the fact that a trade-off between percentages will always be present. A high percentage of any of these five predictors represents a highly-focussed session where customers have a very specific goal. Although the effects of these predictors might not be very informative, the combination of percentages does strongly contribute to the model's performance.

Visit demographics

The effect of the variable [**Site transfer**] turned out to be negatively related to purchasing probability. When a visitor entered the website via a search engine the probability of a purchase decreases. This is in line with findings from Park & Chung (2009) and Verheijden (2012) who have encountered similar effects. Interestingly we see a surprising effect in the magnitude and direction of [**Hurry**]. This variable is indicative of a viewing time per page that is lower than that of previous session (on customer level). The probability of a purchase increased in session where this was the case. This effect stand in contrast with findings from Poel & Buckinx (2004) and Verheijden (2012) who both encountered a negative effect of this predictor. A (direct) purchasing intention in B2B seems to take less time than other forms of browsing or website interaction. A strong negative effect is present in the relation between [**Long session**], [**shallow**] and purchasing probability. Sessions where the maximum time out time was longer than 30 minutes proved to be negatively related to a purchase. Session where the total on-site time was less than 5 seconds showed a similar effect. These predictors help identify a group of customers with low purchasing intentions, hence the relatively high or very low time on site. In the latter case a purchase is very rare since the order

process usually takes longer than 5 seconds. The time of day also seems to matter when predicting a purchase, as displayed by the effects of [**Morning session**] and [**Afternoon session**]. These two categorical dummy variables interact with each other and their effects must be interpreted relative to category that is missing. In this case that category is: sessions that were conducted during evening and night hours. We see that morning sessions have a slight negative purchasing probability compared to those conducted at night. Interestingly we see a very strong positive effect in afternoon sessions, indicating that the probability of a conversion is highest in the afternoon, as was illustrated in figure 2. This reflects the importance of accounting for the effect of office hours in B2B purchasing behaviour.

Historical behaviour

As mentioned before the predictors of historical behaviour proved to be very influential on overall model performance. To start with a negative relationship between [**Visit frequency**] and purchasing probability was encountered. This means a relatively high visiting frequency corresponds to a decrease in purchasing probability. Previous B2C research by Poel & Buckinx (2004),Moe & Fader (2004) and Verheijden (2012) showed a positive effects which is not reflected in this B2B analysis. The slightly positive effect of [**Time since previous visit**] also stands in contrast to their findings. As the time since previous visit increases purchasing probability increases as well. To help us understand what is happening here I ran a correlation analysis which revealed a slight negative correlation between the total number of visits and conversion rate. The number of previous visits also turned out to be positively related to the average number of session search queries. This indicates that customers who (very) frequently use the website make fewer purchases while the amount of search queries per session are higher on average. In our case (very) frequent visit behaviour would be more indicative of search oriented behaviour rather than purchase directed in B2C. As displayed by the graph in figure 6, the probability of a conversion steadily increases when the time since previous visit becomes larger. We do see a strong positive effect of [**Purchasing history**] on purchasing probability. In other words, an increasing amount of previous purchases is indicative of an increased likelihood of a purchase. This is similar to findings from Verheijden (2012), Poel & Buckinx (2004) and Moe & Fader (2004) who all measured similar effects. Customers who have purchased more in the past are more likely to purchase in the current session. This is also reflected in the positive effect of [**Purchasing ratio**] on purchasing probability. This relative measures accounts for the difference between the amounts of previous (purchasing) sessions per customer by turning the previous amount of purchasing into a ratio. Lastly an increase in the [**Time since previous purchase**] resulted in a lower purchasing probability.

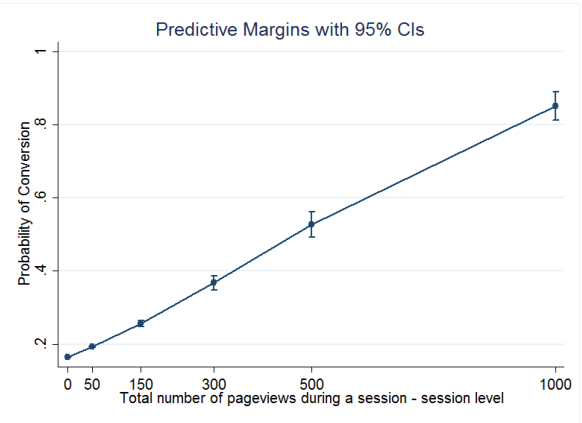


Figure 4 Marginal effect of variable 'pageviews' displayed over the number of pageviews

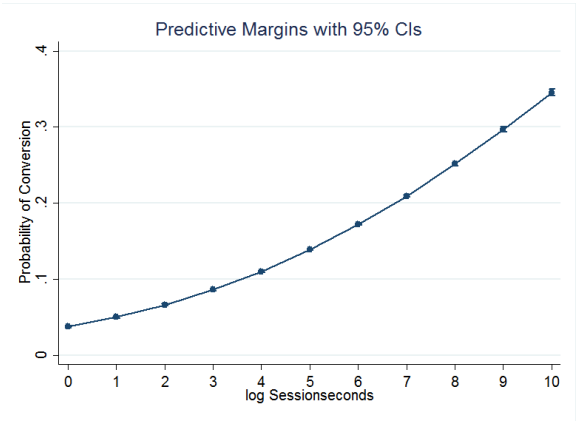


Figure 5 Marginal effect of variable 'duration' displayed over the number of sessionseconds

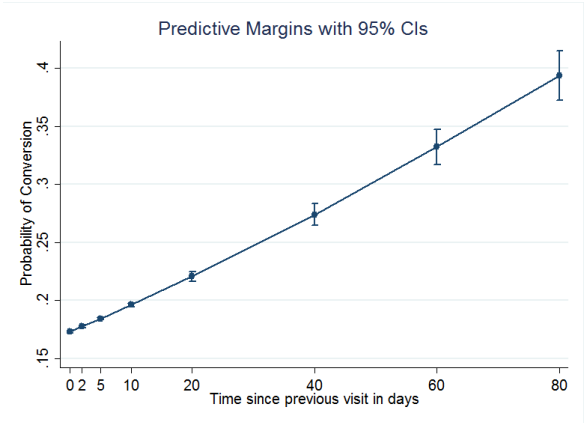


Figure 6 Marginal effect of variable 'time since previous visit' displayed over the time since previous visit in days

4.2 Comparing findings to previous research in B2C

We have already seen some differences between previous B2C research results and the outcomes of this research. The main differences relate to the nature of the relationship between buyer and seller. Historical behaviour turned out to have mayor influence on overall model performance. Most interestingly the relation between visit frequency and purchasing probability turned out to be negative[7]. This implies that (very) frequent users use this B2B platform for something different than buying alone. This arguably relates back to the relationship between buyer and seller where searched and bought goods will in most cases be used for third party assignments or projects. This creates a different form of search motivation since product search and product purchases will rarely be used for personal use. What we see in this analysis is behaviour that is displayed by well-informed, technical professionals who are oriented at (either) searching or buying. This distinction can be seen

[7] For an overview of the differences in predictor direction between B2B and B2C metric see Appendix 1.

in the difference between long, search based, sessions with below average conversion compared to short session with above average conversion. Visits are therefore distinctively goal directed and aimed at retrieving some form of information or making a purchase. This goal directed behaviour is reflected in the relative high website conversion rate of 17.5%. In a B2C setting websites are doing well when conversion rates reach 3-5 % (Burstein, 2015). Also, the high amount of sessions per customers reflect a form of professional usage where customers interact with the B2B website on daily bases. A B2B customer will therefore be familiar with the websites inner workings and its content. On average B2B customers conduced 120 sessions within 3 months. This results in roughly one session per day showing that B2B customers very frequently visited the webshop. This requires a website dynamic which is aimed at facilitating the needs of returning customers. In B2C customers will typically have far less website interactions and this forces B2C e-commerce platforms to aim at facilitating short-term purchases as best as possible. In this research, we saw some clear differences between the direction and magnitude of B2B and B2C metrics. Historical behaviour proved to be very important when predicting a purchase and this is no surprise given that customers typically have so many website interactions.

4.3 Model performance – Accuracy and prediction power compared to B2C

The logistic regression model performance was solid with 39.5 % (McFadden R^2) of the variance in the outcome variable explained while all predictors variables turned out to have a significant effect. The classification also scored well with a sensitivity of 54% while maintaining a specificity of 94%. During an average visit 16 pages were viewed and 2.4 searches were made. Most transactions occurred between 5 and 6 PM and about 2/3 of all transactions were completed between 12 and 6 PM. Looking into the regression results reveals that the directions of the effects of most predictors are relatively similar to previous B2C findings. Historical variables turned out to be very important in the overall prediction power of the model accounting for more than 2/3 of model prediction power. This stands in direct contrast with the findings of Verheijden (2012) who concluded that the effect of previous knowledge (historical variables) was limited in modelling B2C buying behaviour. In research carried out by (Poel & Buckinx, 2004) historical purchasing variables were labelled to be similarly important compared to session and customer related metrics. Results from this study indicate that the browsing behaviour of B2B customers is reasonably similar to B2C when it comes to website interactions. The main difference between B2C and B2B interactions is the weight of previous (historical) interactions on current behaviour. And that B2B customer might have more dichotomous way of interacting with the website, in the sense of searching or buying.

4.4 Predicting purchasing behaviour in B2B – Model prediction over time

The modelling of purchasing behaviour can be very informative when aiming to reveal the drivers of purchasing behaviour. However, the modelling of purchasing behaviour can also be done with limited information of the current session. This provides some guideline on how relevant my findings are when one would aim at predicting the outcome of a session that is still underway. In other words, how quickly can we say something useful about the probability of a purchase while a session is progressing? To answer this question the dataset was split into 9 separate groups using the variable "pageviews". The data was split using this metrics since every pageview provides a (additional) piece on information on session outcome. The more of these pieces are available the better they can be combined to form an estimation of the session outcome. These nine groups correspond to various levels of session maturation and are relatively similar to the ones used by Verheijden (2012). In his research Verheijden (2012) used a technique where the complete dataset was split into equal group based on the total number of session pageviews. Sessions with a maximum of 4 pageviews were places in group one, session with 5-6 pageviews in group two, 7-8 pageviews in group three and so on. In this research, the groups are cumulative meaning all data is included until a certain amount of pageviews. For example, group one includes all data with session being cut-off at 10 pageviews, group two includes all data with session being cut-off at 20 pageviews and so on. The goal here is to determine model performance on limited amount of data.

The results of this analysis are in table 7 and show that model performance is already relatively high at the beginning of a visit. We see a R2 value of roughly 16 % when only data from the first 10 pageviews was used. This indicates that around 40% of the total predicting power[8] of the model can be deducted from actions within the first 10 pageviews of a visit. This does not come as a surprise since customers are asked to log-in before any product related content can be viewed. When customers' log-in historical purchasing behaviour becomes available and this likely results in decent model performance from the get-go. As we saw earlier total model performance is strongly dependent on historical behaviour metrics, and this is underlined by the performance of this limited information model. In general, we see that the direction and magnitude of historical behaviour metrics changes only slightly over time. Furthermore, there are several interesting changes of predictor direction and magnitude when more data becomes available. To start with the product detail page ratio has a strong positive effect when the amount of data is limited. This effect weakens and eventually reverses when more data becomes available.

[8] R2 Max = 39%

Logistic regression results on sections of the data

Variable group and name	Odds-ratio's								
Seconds -->	10	20	40	80	150	300	500	1H	All
General session measures									
Pageviews	0.95[N]	1.03[N]	1.07*	1.08*	1.06*	1.04*	1.03*	1.01*	1.01*
Duration[9]	1.51*	1.44*	1.50*	1.59*	1.63*	1.62*	1.59*	1.49*	1.48*
Session focus measures									
Number of product searches[8]	1.38*	1.37*	1.34*	1.34*	1.33*	1.31*	1.27*	1.10*	1.32*
Number of search filters[4]	0.93*	0.92*	0.86*	0.84*	0.78*	0.81*	0.81*	0.85*	0.93*
Number of deleted filters[4]	1.20[N]	0.90[N]	0.87[N]	0.85*	0.82*	0.91*	0.92*	0.86*	0.88*
Number of product comparison	0.86*	0.82*	0.80*	0.83*	0.89*	0.92*	0.94*	0.97*	0.98*
Personal account pages visited	0.52*	0.44*	0.41*	0.44*	0.50*	0.62*	0.71*	1.02[N]	0.84*
Company about pages visited	0.59*	0.61*	0.58*	0.58*	0.59*	0.63*	0.66*	0.76*	0.76*
Product detail pages visited	0.66*	0.58*	0.59*	0.69*	0.80*	0.95[N]	1.05[N]	1.57*	1.99*
Product detail page ratio	2.38*	2.36*	2.22*	1.94*	1.70*	1.48[N]	1.36*	1.09*	0.93*
Percent product pages	1.01*	1.01*	1.01*	1.00*	0.99*	0.99*	0.98*	0.97*	0.95*
Percent home pages	1.01*	1.01*	1.00*	1.00*	0.99*	0.99*	0.98*	0.96*	0.94*
Percent content pages	0.99*	0.99*	0.99*	0.98*	0.97*	0.96*	0.97*	0.93*	0.92*
Percent search pages	1.02*	1.01*	1.00*	0.99*	0.99*	0.98*	0.96*	0.96*	0.93*
Percent assortment pages	0.99[N]	0.98[+]	0.98*	0.97*	0.96*	0.95*	0.95*	0.91*	0.91*
Visit demographics									
Site transfer	0.75*	0.77*	0.78*	0.77*	0.78*	0.78*	0.79*	0.88*	0.94*
Shallow visit	1.04[N]	1.17[N]	1.48*	1.94*	2.24*	2.31*	2.16*	0.90[N]	0.34*
Hurry	1.10*	0.96[N]	0.89[N]	0.91*	0.99*	1.13*	1.28*	1.85*	1.85*
Long session	1.92*	2.02*	2.09*	2.20*	2.27*	2.34*	2.41*	1.74*	0.66*
Morning session	0.75*	0.78*	0.80*	0.85*	0.88*	0.93*	0.95*	0.92*	0.88*
Afternoon session	1.34*	1.40*	1.43*	1.47*	1.50*	1.56*	1.58*	1.54*	1.55*
Historical behaviour									
Visit frequency[4]	1.03*	1.05*	1.07*	1.09*	1.11*	1.11*	1.12*	1.01[N]	0.61*
Time since previous visit	1.03*	1.03*	1.04*	1.04*	1.04*	1.04*	1.04*	1.05*	1.03*
Purchasing history[4]	1.09*	1.08*	1.07*	1.07*	1.07*	1.07*	1.08*	1.08[N]	2.07*
Time since previous purchase	1.03*	0.94*	0.97*	0.97*	0.97*	0.97*	0.97*	0.97*	0.99*
Purchasing ratio	1.06*	1.06*	1.06*	1.06*	1.06*	1.06*	1.06*	1.07*	1.05*
Constant	0.01*	0.01*	0.00*	0.00*	0.00*	0.00*	0.00*	0.00*	0.02*
McFadden R2	0.197	0.239	0.265	0.286	0.304	0.323	0.337	0.376	0.395
Observation (sessions)	2,6 m	2,5 m	2,5 m	2,5 m	2,5 m	2,5 m	2,5 m	2,5 m	2,1 m
Nr. Clusters	17 k	20 k	21 k	22 k	23 k	23 k	23 k	23 k	23 k
Wald Chi2	127 k	181 k	237 k	227 k	226 k	184 k	173 k	104 k	84 k
Classification results (cut-off 0.5)									
Sensitivity	19.3%	24.3%	27.3%	29.8%	31.9%	34.2%	35.7%	39.7%	53.6 %
Specificity	97.6%	97.4%	97.1%	96.9%	96.7%	96.5%	96.4%	95.6%	93.7 %
Positive predicted values	60.8%	61.4%	61.9%	62.5%	62.9%	63.3%	63.5%	63.5%	64.4 %
Negative predicted values	88.1%	88.4%	88.5%	88.8%	89.0%	89.3%	89.5%	89.9%	80.4 %

Table 4 Logistic regression results on sections of the data split up using the variable 'Sessionseconds'

[9] Logarithmic transformation applied.

[N] = Significance level >0.1 / [+] = Significance level <0.1 / * = Significance level <0.05

This says something about the viewing of product detail pages very early in a session. When this is the case purchasing probability strongly increases. We see a similar effect in the variable 'long session' which is an indication of a session having more than 30 minutes between pageviews. The effect of this predictor is positive when only limited information is available. This might be a result from the total visit duration increasing over a relative low amount of pageviews, increasing the probability of a purchase. Lastly, we see a positive effect of visit frequency throughout all models with exception of the model where all data was included. There is something happening in the inclusion of this final piece of information that reverses the effect of visit frequency.

4.5 Regression model validation

The logistic regression models were checked on multiple dimensions to determine the validity of the outcome. The large amount of observations (sessions) in the dataset provided a solid base for the modelling of online purchasing behaviour. All predictors turned out significant which is partly due to the large amount of observations. Model validation was firstly checked by a running a correlation analysis on all the model predictors as can be seen in table 5. This revealed several correlated predictors, with 7 correlations being above 0.5. Most of these correlations showed a positive relation between general clickstream measures and session focus measures. This is no surprise since most detailed clickstream measures increase when session duration increases and vice versa. A high correlation excised between the variables that indicated if the session was conducted in the morning or afternoon. These variables interact with each other since every session will fall in the category morning, afternoon or night. To check if this proved a problem for the model validity the Variance Inflation Factors (VIF) were checked as well. Model fit was tested using the Hosmer & Lemeshow goodness of fit test. Running this test on random samples of the data provided solid results.

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
General session measures																										
1	Pageviews	1																								
2	Duration	0.53	1																							
Session focus measures																										
3	Number of product searches	0.69	0.46	1																						
4	Number of search filters	0.62	0.44	0.52	1																					
5	Number of deleted filters	0.34	0.17	0.17	0.38	1																				
6	Number of product comparisons	0.37	0.2	0.25	0.27	0.13	1																			
7	Personal account pages visited	0.14	0.13	0.01	-0.02	0	0.01	1																		
8	Company about pages visited	0.08	0.07	0.02	0.02	0.02	0.02	0.03	1																	
9	Product detail pages visited	0.27	0.4	0.23	0.26	0.08	0.11	-0.06	-0.02	1																
10	Product detail page ratio	-0.17	-0.17	-0.28	-0.42	-0.17	-0.06	0.16	0	0.23	1															
11	Percent product pages	-0.02	-0.01	-0.02	-0.09	-0.05	0.07	-0.14	-0.05	0.46	0.4	1														
12	Percent home pages	-0.18	-0.21	-0.18	-0.21	-0.07	-0.08	-0.05	-0.03	-0.23	0.12	-0.23	1													
13	Percent content pages	-0.02	-0.01	-0.03	-0.03	-0.01	-0.01	-0.01	0.54	-0.06	0.03	-0.04	-0.02	1												
14	Percent search pages	0.17	0.19	0.32	0.37	0.13	0.06	-0.22	-0.04	0.06	-0.59	-0.14	-0.33	-0.05	1											
Visit demographics																										
15	Percent assortment pages	0	0	-0.03	-0.02	0	-0.01	-0.01	0.02	-0.04	-0.02	-0.05	-0.04	0	-0.04	1										
16	Site transfer	0.05	0.05	0	0	0.02	-0.01	-0.01	0.05	-0.01	-0.04	-0.06	0.03	0.03	-0.02	0.04	1									
17	Shallow visit	-0.17	-0.57	-0.16	-0.17	-0.05	-0.06	-0.08	-0.02	-0.26	0.14	0.09	0.08	0.01	-0.09	0	-0.06	1								
18	Hurry	-0.07	-0.5	-0.1	-0.04	0	-0.02	-0.04	-0.03	-0.11	-0.01	-0.03	0.06	0	-0.01	0	-0.01	0.1	1							
19	Long session	0.11	0.35	0.06	0.07	0.05	0.04	0.06	0.03	0.04	-0.01	-0.02	-0.02	0.01	0	0	-0.01	-0.06	-0.31	1						
20	Morning session	-0.03	0.01	-0.02	-0.03	-0.01	-0.01	0.02	0	-0.04	0.04	0	0.05	0	-0.03	0	-0.01	0.02	-0.04	0.08	1					
21	Afternoon session	-0.01	-0.01	0.01	0.01	-0.01	-0.01	-0.03	-0.02	0.03	-0.03	0.02	-0.04	-0.01	0.04	0	-0.02	-0.01	0.03	-0.05	-0.82	1				
Historical behaviour																										
22	Visit frequency	-0.07	-0.07	0.01	0.03	-0.05	0.01	0.01	-0.03	0.02	0.07	0.12	-0.04	-0.01	0.05	-0.03	-0.13	0.1	0.04	-0.03	0.02	0.03	1			
23	Time since previous visit	0.03	0.04	-0.01	0.01	0.03	-0.01	-0.01	0.02	-0.02	-0.04	-0.07	0.03	0.01	-0.02	0.03	0.12	-0.06	-0.03	0	0.04	-0.06	-0.28	1		
24	Purchasing history	0.01	0.01	0.01	0	-0.03	0.01	0.14	-0.02	0.02	0.04	-0.03	-0.02	-0.01	-0.05	-0.02	-0.06	-0.02	0.02	-0.02	-0.02	0.01	0.5	-0.14	1	
25	Time since previous purchase	-0.02	-0.02	0.01	0.04	0.01	0	-0.12	0	0	-0.01	0.07	0	0	0.08	0.01	0	0.04	0	-0.01	0.03	0	0.14	0.16	-0.56	1
26	Purchasing ratio	0.09	0.09	0.03	0.01	0.01	0.01	0.14	0	0.03	-0.01	-0.1	-0.03	-0.01	-0.08	-0.01	0.04	-0.09	-0.01	0	-0.04	0.01	-0.11	0.07	0.58	-0.45

Table 5 Correlation matrix of all predictor variables. Correlations that are highlighted are above 0.5.

5. Discussion and conclusion

This research was conducted to examine purchasing behaviour in an online B2B environment. Online clickstream data was used to model and predict B2B purchasing behaviour. The analyses were aimed at answering the main research question regarding if and how B2B data can be used for predicting purchases and where possible differences with B2C results might arise. The results from these analyses show online B2B purchasing behaviour can be modelled using similar techniques that have been previously applied in B2C research. The model that was constructed performed well with an R^2 value around 0.4. In general, we saw that there was a large group of frequently returning customers in the dataset that visited the website more than once a day. Most visits occurred during working hours and peaked at the end of the working day. In the late afternoon conversion rate was commonly above 30 %, meaning 1/3 of all visits resulted in a purchase. Average visit frequency turned out to be larger than one visit per customer a day. This influenced the prediction model performance in the sense that historical variables proved to be very important and contributed to roughly 50% of overall model performance.

5.1 Relating B2B model results to existing findings from B2C

The direction and magnitude of predictors was different from B2C on several occasions. The result showed that a section of the variables in B2B had a different relation to purchasing behaviour compared to B2C. Apart from the importance of the historical variables we encountered that predictors 'visit frequency' and 'time since previous visit' had different effects on purchasing probability than B2C research indicated. An increase in visit frequency resulted in lower purchasing probability which is likely caused by a group of customers that has no-purchasing intention. This customer group entered a relatively large amount of search queries which is assumed to be illustrative for their search oriented behaviour. In their case, the website presumable functions as a search database for product specifications or product pricing rather than a platform to purchase items.

A visit to a dedicated B2B website is not likely to be random and therefore B2B customers are assumed to be highly goal directed. The difference with B2C is that B2B customers are less likely to visit the website by change and therefore do not have to be persuaded into buying something. A B2B customer is possibly less subjective to techniques aimed to retain or trigger a purchase. The differences between B2C and B2B might therefore be traced back to the difference in relationship between buyer and seller rather than differences between the two markets. B2B customers will in general be rather 'fixed' compared to B2C customers and unable to switch between sellers at low cost. The reason historical variables turned out to be so valuable could therefore be attributed to the

fact that many visits from the same group of users results in a lot of historical information and is not necessarily related to specific B2B behaviour. The fact that we measured this effect is caused by the high number of returning customers which is a non-exclusive aspect of B2B markets. We might encounter the same sort of behaviour in B2C markets where the share of returning customers is high. I argue therefore that the nature of the buyer-seller relationship influences the effect of purchasing predictors more than market orientation alone.

5.2 Evaluating overall model performance compared to B2C

The B2B purchasing model constructed in this research is not inferior to B2C prediction models in terms of prediction power or classification ability. Prediction power is similar to, or higher than, models constructed by Verheijden (2012) or Poel & Buckinx (2004). Improving the prediction power of the B2B model could be desirable if the goal is to make the best purchasing prediction possible. Having a large customer base that frequently visits the website might help here since there is a lot of information available from almost every customer. This is different from B2C markets where the average number of visits per customer is relatively low which results in having less historical information to work with. This makes the modelling of purchasing probability more dependent on what is happening during the session rather than was has happened previously.

5.3 Making a meaningful prediction about a customer's tendency to buy during the visit

The importance of historical variables was confirmed by the 'predicting purchasing behaviour over time' models. Historical information was generally known early in a session which resulted in the model performing relatively well 10-20 seconds into the visit. This creates opportunities for modelling the probability of a purchase while the session is underway as the purchasing history of a becomes known when the customer can be identified. The results of our analyses show that the direction and influence of the majority of predictor variables does not change much during the visit implying the goal of a visitor might be relatively fixed. Being able to classify the visitors creates possibilities to alter the website to meet specific customer's needs. We see such techniques being applied by large (global) e-commerce websites that use data analysis to dynamically alter the structure of the website to boost conversion rates.

5.4 Implications and recommendations

This research has shown B2B purchasing behaviour can be predicted using online clickstream data. Overall model performance was comparable to B2C research as were the classification results. We have seen that a high visit frequency influences the importance of historical variables in the model. Computing these historical variables requires computing power compared to only using information that can be deducted from current session. The necessary computing power especially increases when longer periods of historical data are processed. In the case that the amount of returning

customers is relatively large I would argue that investing in the capacity to handle historical information would be beneficial. This stands in contrast with the recommendation by Verheijden (2012) who argued previous visiting behaviour could be omitted without significantly influencing model performance. This might be the case when there is almost no historical information available, however considering the results of this research this evaluation should rather be based on the amount of returning customers.

The results from this research have shown different user goals and behaviour patterns exist within the clickstream data. The company who supplied the data can benefit from this insight by analysing and categorizing different customer's groups in order to customize marketing campaigns, increase (after) sales and enhance website usability. A good place to start is distinguishing buyers from non-buyers since there is a relatively large group of customers who are 'hard core never buyers'. This group consists of roughly 1/3 of all users and showed knowledge building behaviour that did not result into a purchase. The clickstream data contains information about search queries and used filters that could help optimising search results or filter options. The company could examine which search queries do commonly not result into a purchase to determine where search and filter optimisation could start. In general, companies can use the result from this research to help determine which kind of user metrics are influential when determining purchasing probability. If visit and purchasing frequency is high, historical variables might contain a fruitful source of information.

5.5 Future research

Future research efforts could be directed at investigation the effect of historical variables in a B2C setting that has a large returning customer base. The relative importance of historical variables could be tested to examine if a similar effect could be measured in B2C. Another research opportunity could be improved this B2B model by using machine learning techniques. When applying such techniques, the prediction power of the model could likely be improved as well as the ability to find specific user patterns. This could be desirable if the purpose of the research would be to predict purchases as best as possible or categorize different user groups. A downside to using these techniques is that it becomes difficult to differentiate between the effects of induvial predictors. Future research can also give more insight into the effects of different market structures on the modelling of purchasing behaviour. This type of analysis could reveal if and how customers are influenced by the type of website they are interacting with. To examine the formation of user patterns a longitudinal study on multiple e-commerce platforms could provide means for comparing the effects of different market structures on purchasing behaviour.

Bibliography

Beatty, S., & Ferrell, M. (1998). Impulse buying: Modeling its Precursors. *Journal of Retailing, 74*(2), 169-191. doi:10.1016/S0022-4359(99)80092-X

Beige, S., & Abdi, F. (2015). On the critical success factors for B2B e-marketplace. *Decision Science Letters, 4*(1), 77-86.

Bigne, E., Ruiz, C., & Sanz, S. (2005). The impact of inter shopping paterns and demograhpics on consumer mobile buying behaviour. *Journal of Electronic Commerce Research*, 193-209.

Brucks, M. (1985). The effects of product class knowledge on information search behaviour. *Journal van Consumer Research*, 1-16.

Bucklin, R., Lattin, J., Ansari, A., Gupta, S., Bell, D., Coupey, E., . . . Steckel, J. (2002). Choice and the Internet: From Clickstream to Research Stream. *Marketing Letters*, 245-258.

Burstein, D. (2015, 09 15). *Ecommerce Research Chart: Industry benchmark conversion rates for 25 retail categories*. Retrieved from Marketing Sherpa: https://www.marketingsherpa.com/article/chart/conversion-rates-retail-categories

Elia, E., Boeck, H., Lefebvre, L.-A., & Lefebvre, E. (2005). Exploring B-to-B e-commerce adoption trajectories in manufactoring SME's. *Technovation 25*, 1443-1456.

Forrester Research. (2015, 04 02). *Andy Hoar's Blog.* Retrieved from Marketing & Strategy > eBusiness & Channel Strategy Professionals: http://blogs.forrester.com/andy_hoar/15-04-02-us_b2b_ecommerce_to_reach_11_trillion_by_2020

Forrester Research. (2015). *The Forrester Wave™: B2B Commerce Suites .* Cambridge: Forrester Research Inc. Retrieved from http://www.sap.com/bin/sapcom/en_us/downloadasset.2015-06-jun-27-07.the-forrester-wave-b2b-commerce-suites-q2-2015-pdf.bypassReg.html

Frost & Sullivan. (2014, 12 31). *Future of B2B Online Retailing.* Retrieved from Visionary Innovation Research - B2B e-sales: http://www.frost.com/c/5048246/sublib/display-report.do?searchQuery=MA4E&ctxixpLink=FcmCtx1&ctxixpLabel=FcmCtx2&id=MA4E-01-00-0000&bdata=aHR0cHM6Ly93d3cuZnJvc3QuY29tL3NyY2gvY3Jvc3MtY29tbXVuaXR5LXNlYXJjaC5kbz9zZWFyY2hUeXBlPWFkciZxdWVyeVRleHQ9TUE0RSZ4PTAme

Gefen, D. (2000). E-commerce: the role of familiarity and trust. *Omega - The International Journal of Management Science, 28*(6), 725-737.

IDC. (2012). *The digital universe in 2020: big data, bigger digital shadows, and biggest growth in the far east.* Framingham: International Data Corporation. Retrieved 03 04, 2016, from http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf

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

Lee, M., Ha, T., Han, J., Rha, J.-H., & Kwon, T. (June 28–July 01, 2015). Online Footsteps to Purchase: Exploring Consumers Behaviors on Online Shopping Sites. *WebSci'15, .* Oxford: ACM New York. doi:10.1145/2786451.2786456

Lin, L., Hu, J., Olivia, R., Sheng, L., & Lee, J. (2010). Is Stikiness Profitable for Online Retailers? *53*, pp. 132-136. ACM New York. doi:10.1145/1666420.1666454

Lord, K., & Collins, A. (2002). Supplier web-page design and organizational buyer preferences. *Journal of business & industrial marketing, 17*(2/3), 139-150. doi:10.1108/08858620210419772

Moe, W. (2003). Buying, Seraching, or Browsing: Differentiating Between Online Shoppers Using In-Store Navigational Clickstream. *Journal of Consumer Psychology, 13*(1), 29-39. doi:10.1207/S15327663JCP13-1&2_03

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

Moe, W., Johnson, E., Fader, P., Bellman, S., & Lohse, G. (2004). On the Depth and Dynamics of Online Search Behavior. *Management Science, 50*(3), 299-308. doi:10.1287/mnsc.1040.0194

Montgomery, A., Li, S., Srinivasan, K., & Liechty, J. (2004). Modeling Online Browsing and Path Analysis Using Clickstream Data. *Marketing Science, 23*(4), 579-595. doi:10.1287/mksc.1040.0073

OFCOM. (2015). *Adults' media use and attitudes.* London: Office of communications. Retrieved 02 16, 201, from http://stakeholders.ofcom.org.uk/binaries/research/media-literacy/media-lit-10years/2015_Adults_media_use_and_attitudes_report.pdf

Olbrich, R., & Holsing, C. (2011). Modeling Consumer Purchasing Behaviour in Social Shopping Communities with Clickstream Data. *The Service Industry Journal, 16*(2), 1451-1463. doi:10.2753/JEC1086-4415160202

O'Reilly, T. (2007). What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. *Communications & Strategies*, 17-39.

Padmanabhan, B., Zheng, Z., & Kimbrough, S. (2001). Personalization From Incomplete Data: What You Don't Know Can Hurt. *KDD '01 Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 154-163). ACM New York. doi:10.1145/502512.502535

Panagiotelis, A., Smith, M., & Danaher, P. (2014). From Amazon to Apple: Modeling Online Retail Sales, Purchase Incidence, and Visit Behavior. *Journal of Business & Economic Statistics, 32*(1), 14-29. doi:10.1080/07350015.2013.835729

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

Parry, S., Rowley, J., Jones, R., & Kupiec-Teahan, B. (2012). Customer-perceived value in business-to-business relationships: A study of software customers. *Journal of Marketing Management, 28*(7-8), 887-991. doi:10.1080/0267257X.2012.698637

Poel, D. v., & Buckinx, W. (2004). Predicting online-purchasing behavior. *European Journal of Operational Research, 166*(2), 557-575. doi:10.1016/j.ejor.2004.04.022

Practical Ecommerce. (2015, 04 09). *B2B Management*. Retrieved from B2B Ecommerce Growing; Becoming More Like B2C: http://www.practicalecommerce.com/articles/85970-B2B-Ecommerce-Growing-Becoming-More-Like-B2C

Reinartz, W., & Kumar, V. (2003). The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration. *Journal of Marketing*, 77-99.

Schmittlein, D., & Peterson, R. (1994). Customer base analysis: An industrial purchase process application. *Marketing Science, 13*(1), 1986-1998. doi:10.1287/mksc.13.1.41

Senecal, S., Kalczynski, P., & Nantel, J. (2005). Consumers' decision-making process and their online shopping behaviour: A clickstream analysis. *Journal of Business Research, 58*, 1599-1608. doi:10.1016/j.jbusres.2004.06.003

Sismeiro, C., & Bucklin, R. (2003). A Model of Web Site Browsing Bahaviour Estimated on Clickstream Data. *Journal of Marketing Research, 40*(3), 249-267.

Statistica.com. (2015, 11 04). *Key Figures of E-commerce*. Retrieved from Digital buyer penetration worldwide 2011-2018: http://www.statista.com/statistics/261676/digital-buyer-penetration-worldwide/

Tilborg, R. V. (2015). *A literature review of the antecedents and dimensions of trust in online B2B.* Enschede: University of Twente. Retrieved 01 16, 2016, from http://essay.utwente.nl/68517/1/VanTilborg_BA_MB.pdf

Tsiros, M., Ross, W., & Mittal, V. (2009). How Commitment Influences the Termination of B2B Exchange Relationships. *Journal of Service Research, 11*(3), 263-276. doi:10.1177/1094670508328982

Venkatesh, V., & Agarwal, R. (2006). Turning visitors into customers: A usability-centric perspective on pruchase behavior in electronic channels. *Management science, 53*(2), 367-382. doi:10.1287/mnsc.1050.0442

Verheijden, R. (2012). *Predicting purchasing behavior throughout the clickstream.* Eindhoven: Eindhoven University of Technology - Master Thesis.

Appendix 1. Predictor direction overview

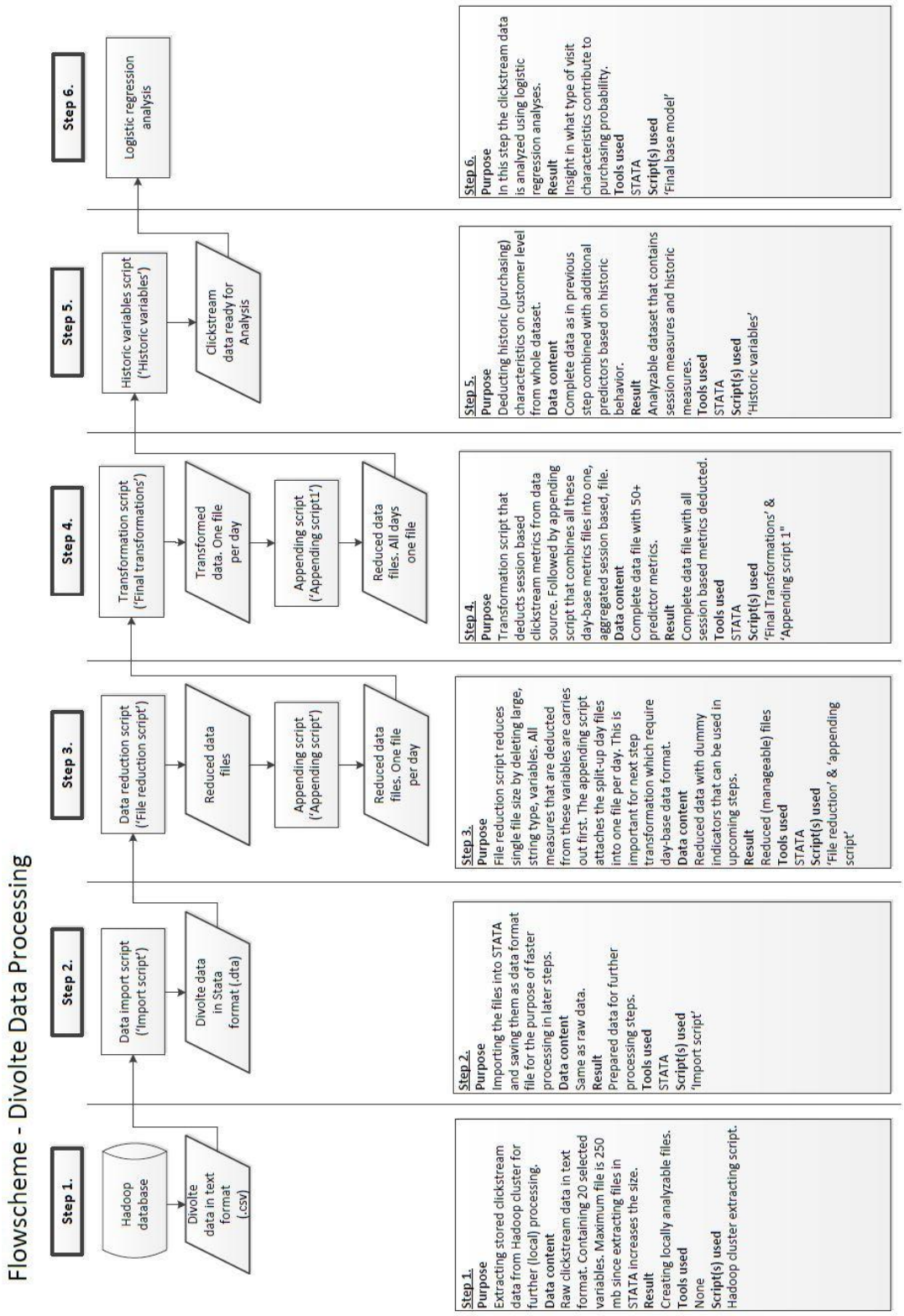
Variable Category and Name	Reference	Predictor direction Previous Research (B2C)	Predictor direction Results (B2B)
General session measures			
Pageviews	1,2,4,6	+	+
Duration	1,2,3,4,5,8	+	+
Session focus measures			
Number of product searches		N.A.	- [10]
Number of search filters	2,8	-	- [9]
Number of deleted filters		N.A.	- [9]
Number of product compares		N.A.	- [9]
Personal account pages visited	2,3,8	-	-
Company about pages visited	2	-	-
Product detail pages visited	8	+	+
Product detail page ratio	2,6	+	-
Percent product pages	6	N.A.	-
Percent home pages	6	N.A.	-
Percent content pages	6	N.A.	-
Percent search pages	6	N.A.	-
Percent assortment pages	6	N.A.	-
Visit demographics			
Site transfer	2,7	+/-	-
Shallow visit		N.A.	-
Hurry	2,3	-	+
Long session		N.A.	-
Morning session		N.A.	+
Afternoon session		N.A.	+
Historical behaviour			
Visit frequency	2,3,9	+	-
Time since previous visit	2,3,9	-	+
Purchasing history	2,3,9	+	+
Time since previous purchase	2,3,9	+/-	-
Purchasing ratio		N.A.	+

1. (Lin, Hu, Olivia, Sheng, & Lee, 2010)
2. (Verheijden, 2012)
3. (Poel & Buckinx, 2004)
4. (Panagiotelis, Smith, & Danaher, 2014)
5. (Padmanabhan, Zheng, & Kimbrough, 2001)
6. (Moe, 2003)
7. (Park & Chung, 2009)
8. (Olbrich & Holsing, 2011)
9. (Moe & Fader, 2004)

Note: The directions of the predictors in this table represent the direction of the relation between a predictor and purchasing probability.

[10] The effects of these predictors are positive when they are included as dummy variables. In other words, it is not the use of on-site search engines or product compares that reduce purchasing probability; it is the increase in the amount of usage that negatively relates to purchasing probability.

Appendix 2. Data preparation flowsheme



Appendix 3. Data transformation scripts

Script *** File reduction script ***

```
* Generaing a first in session identyfier
gen     firstinses  =               (firstinsession=="true")
drop    firstinsession

* ERP user identyfier
gen     erp  = regexm               (user, "erp")
recode  erp .=0
egen    erp_user     = max(         erp), by(sessionid)

* Generating a conversion measurement variable
generate conv  =        regexm (location, "cartconfirmation") if regexm(referer, "cartpreview") &
                        erp_user == 0 & eventtype == "pageTrack" & pagetype == "winkelwagen"
recode   conv .=        0

* Product detail
sort     sessionid
gen      detail =       regexm (location,"prd") & regexm(eventtype,"pageTrack")

* Use of product compare function (+nr. of compares)
generate compare =      regexm (location, "productcompare") & regexm(eventtype,"pageTrack")

* Site transferred
Generate site_transfer   = 1 if   regexm(referer, "bing") | regexm(referer, "google") |regexm(referer,
                                  "yahoo")

* Nr. of unique product searches during a session
bys sessionid :   egen nr_unique_searches = nvals(searchphrase)
```

Script *** Final transformations script ***

```
// 0. General stuff thats needs to be done before transformations
*-------------------------------------------------------------------------------

* Making a new N var on day basis
drop     n
gen      n   = _n

* Date and time vars
gen              date2     =   date          (date, "YMD")
format           date2 %tdDD/NN/CCYY
drop             date
rename           date2   date
gen              time2     =   clock         (time, "hms")
format           time2 %tcHH:MM:SS
```

```
drop            time
rename          time2   time
```

<u>* Generating a unique number for each session ID</u>

```
sort sessionid   time
egen uni_sesid = group(sessionid)
```

<u>*Identify own_employee searches</u>

```
recode  customer_number .       =       0
gen     own_employ              =       1 if customer_number >= 9000000
recode  own_employ      .       =       0
egen    own_employ1             =       max(own_employ), by(sessionid)
drop    own_employ
rename  own_employ1 own_employ
```

<u>* Identifying unique visits</u>

```
sort        uni_sesid  time
bys         uni_sesid (time)                :       gen unique = 1 if _n==1
replace unique                              =       0 if unique==.
tab         unique, mis
```

<u>* Sorting sessions by time</u>

```
gsort           -unique   time n
gen             n1                      =     _n
replace         n1                      = .   if unique == 0

bysort   uni_sesid (time n):        replace n1=n1[1] if missing(n1)
```

<u>*Generating customer ID</u>

```
gen     custid                  =       substr(user, strpos(user,"_")+1, .) if strpos(user,"_")>0
order   custid,   after(user)
```

<u>*Filling blank spots in page_view type rowes</u>

```
bys sessionid (custid): replace custid          =       custid[_N]
bys sessionid: replace customer_number          =       customer_number[_N]
bys sessionid: replace client_group             =       client_group[_N]
bys sessionid: replace user                     =       user[_N]
bys sessionid: replace vk_number                =       vk_number[_N]
```

<u>* Identifing long sessions (with time out longer than 30 minutes)</u>

```
sort            sessionid time
bysort          sessionid (time)                        :       gen time_between_activity     =
                                                                time - time[_n-1]
format          time_between_activity       %tcHH:MM:SS
recode          time_between_activity   .               =       0
gen             time_between_activity_sec               =        time_between_activity / 1000
gen             long_ses                                =       1 if time_between_activity_sec >=
                                                                1800
by              sessionid:          egen long_session   =       sum(long_ses)
replace         long_session                            =       1 if long_session >= 1
drop             long_ses
```

* Time measure for analysis when only part of the data is used

```
// Running visit duration in seconds
by sessionid    :       gen visit_duration_time = sum(time_between_activity_sec)

// Runing visit duration in pageviews
by  sessionid   :       gen page_view  =  regexm(eventtype,"pageView")
by  sessionid   :       gen visit_duration_pages = sum(page_view)
```

// 1. Depedent variable

```
*-------------------------------------------------------------------------------*
```

* Number of conversions deducted from ordernumbers in whole session

```
gen             conversion      =       1 if substr(ordernumber,1,1) == "5"
egen            convnr          =       sum(conversion), by(n1)
egen            convnr1         =       max(convnr), by(n1)
gen             conv1           =       0
replace         conv1           =       1 if convnr1 >= 1
drop            convnr
rename          convnr1 convnr
```

// 2. Sessions stickiness

```
*-------------------------------------------------------------------------------
```

* Counting the amount of pageviews per session

```
// Pageview dummy already defined above
sort            sessionid  time
by              sessionid       :       gen nr_pages   =  sum(page_view)
egen            nr_pages1       =       max(nr_pages), by(sessionid)
drop            nr_pages
rename          nr_pages1 pageviews
```

* Generating session duration

```
egen            sessiontimemax          =       max(time), by(uni_sesid)
egen            sessiontimemin          =       min(time), by(uni_sesid)
format          sessiontimemax %tcHH:MM:SS
format          sessiontimemin %tcHH:MM:SS

gen             duration                =       sessiontimemax-sessiontimemin
format          duration   %tcHH:MM:SS
drop            sessiontimemin sessiontimemax
```

* Duration into seconds

```
gen             sessionseconds                  =       duration/1000
```

* Indication per hour

```
gen             time_seconds = time/1000
gen             hour            =       1  if inrange(time_seconds, 0, 3600)     & unique == 1
replace         hour            =       2  if inrange(time_seconds, 3600, 7200)   & unique == 1
replace         hour            =       3  if inrange(time_seconds, 7200, 10800)  & unique == 1
```

```
etc. until hour 24
drop     time_seconds

* Splitting day in sections (dummy)
gen              morning          =        1 if inrange(hour, 1, 11)
recode           morning .        =        0
gen              afternoon        =        1 if inrange(hour, 12, 18)
recode           afternoon .      =        0
gen              evening          =        1 if inrange(hour, 18, 24)
recode           evening .        =        0

// 3. Website loyalty
*-------------------------------------------------------------------------------

* Shallow visits
sort      sessionid
by        sessionid: gen shallow   =        1 if sessionseconds <=5
recode    shallow .                =        0

* No log-in visits
sort      sessionid
by        sessionid: egen no_login         =        max(customer_number)
by        sessionid: gen no_login1         =        1 if no_login == 0
recode no_login1 .                         =        0

drop no_login
rename no_login1 no_login


// 4. Historical purchase behaviour
*-------------------------------------------------------------------------------
// See historical purchasing behaviour script

// 5. Focussed search
*-------------------------------------------------------------------------------

* Site transferred
recode           site_transfer   .        =        0
sort             sessionid time
by sessionid:    egen  site_transfer1     =        sum(site_transfer)
replace          site_transfer1           =        1 if site_transfer1 >=1
drop             site_transfer
rename           site_transfer1    site_transfer

* Absolute number of filters in a sessions
sort             sessionid time
gen              filter             =        regexm(action,"Filter")
by               sessionid          :        gen total_filter =  sum(filter)
egen             total_filter1      =        max(total_filter), by(sessionid)
drop             total_filter
rename           total_filter1 total_filter
drop             filter
```

* Absolute number of deleted filters

```
gen         filter_deleted          =       regexm(label,"Verwijder filter")
egen        deleted_count           =       sum(filter_deleted), by (sessionid)
egen        filter_deleted1         =       max(deleted_count), by(sessionid)
drop        deleted_count filter_deleted
rename      filter_deleted1 filters_deleted
```

* Detail to overview pages

```
// nr. total search related (overview) pages / nr. detail pages
egen        detail_count            =       sum(detail), by (sessionid)
egen        detail_count1           =       max(detail_count), by(sessionid)
drop        detail detail_count
rename      detail_count1 nr_detail_pages

gen         overview                =       regexm(eventtype,"pageView") &
                                            regexm(pagetype,"zoeken")
egen        overview_count          =       sum(overview), by (sessionid)
egen        overview_count1         =       max(overview_count), by(sessionid)
drop        overview overview_count
rename      overview_count1 nr_overview_pages

gen         detail_page_ratio       =       nr_detail_pages / nr_overview_pages
replace     detail_page_ratio       =       1 if detail_page_ratio >=1
```

//6. Product interest

```
*-------------------------------------------------------------------------------
```

* Product detail

```
//nr. detail pages already defined above
gen         detail_pages            =       nr_detail_pages
replace     detail_pages            =       1 if detail_pages >=1
```

* Use of product compare function (+nr. of compares)

```
egen        compare_count           =       sum(compare), by (sessionid)
egen        compare_count1          =       max(compare_count), by(sessionid)
drop        compare_count
rename      compare_count1 nr_prd_compares
drop        compare
gen         compare                 =       nr_prd_compares
replace     compare                 =       1 if compare >=1
```

* Nr. of unique product searches during a session

```
egen        nr_searches_count1      =       max(nr_unique_searches),  by(sessionid)
drop        nr_unique_searches
rename      nr_searches_count1      nr_unique_searches
recode      nr_unique_searches .    =       0
```

//7. Non purchase intentions

*----------

* Personal account pages viewed

```
sort            sessionid
generate        account             =       regexm(pagetype, "invoices")
generate        account1            =       regexm(pagetype, "orders")
egen            account_count       =       sum(account),  by (sessionid)
egen            account_count1      =       sum(account1), by (sessionid)
gen             account_count2      =       account_count + account_count1
egen            account_count3      =       max(account_count2), by(sessionid)
drop            account_count account_count1 account_count2
rename          account_count3 nr_account_pages
drop            account account1
gen             account_pages       =       nr_account_pages
replace a       ccount_pages        =       1 if account_pages >=1
```

* About pages viewed

```
sort            sessionid
gen             about               =       regexm(pagetype,"content")
egen            about_count         =       sum(about), by (sessionid)
egen            about_count1        =       max(about_count), by(sessionid)
drop            about about_count
rename          about_count1 nr_about_pages
gen             about_pages         =       nr_about_pages
replace         about_pages         =       1 if about_pages >=1
```

* Hurry

```
// how much time is sent per pageview, preperation for histroical variable script
gen             seconds_per_page    =       sessionseconds / pageviews
format          seconds_per_page %9.0f
```

//8. Additional stuff

*----------

* Sort for interpretation

```
sort sessionid time
```

* Creating dummies for pagetypes

```
gen     home            =       regexm(pagetype, "home") & regexm(eventtype,"pageView")
gen     content         =       regexm(pagetype, "content") & regexm(eventtype,"pageView")
gen     zoeken          =       regexm(pagetype, "zoeken") & regexm(eventtype,"pageView")
gen     assortment      =       regexm(pagetype, "assortment") & regexm(eventtype,"pageView")
gen     product         =       regexm(pagetype, "product") & regexm(eventtype,"pageView")

gen             click               =       1
bysort          n1:gen sumclicks    =       sum(click)
egen            sumclicks1          =       max(sumclicks), by(n1)
drop            sumclicks
rename          sumclicks1 sumclicks
```

```
egen    sumhome         =       sum(home),      by(n1)
egen    sumcontent      =       sum(content),   by(n1)
egen    sumzoeken       =       sum(zoeken),    by(n1)
egen    sumassortment   =       sum(assortment), by(n1)
egen    sumproduct      =       sum(product),   by(n1)

gen     perchome        =       (sumhome      / sumclicks) * 100
gen     perccontent     =       (sumcontent   / sumclicks) * 100
gen     perczoeken      =       (sumzoeken    / sumclicks) * 100
gen     percassortment  =       (sumassortment / sumclicks) * 100
gen     percproduct     =       (sumproduct   / sumclicks) * 100

* Averge time per page
gen     averagetime     =       sessionseconds / sumclicks

* Put all the stuff in readable order
sort n1 time
```

Script * Historical variables *****

```
*Visit frequency
by              customer_number custid          :       gen visit_frequency = _n // 02-10-2016
replace         visit_frequency                 =       0 if customer_number == 0

*Time since previous visit
Gen             tspv            =       0
bys             customer_number custid (timestamp)   :       replace tspv = ((timestamp –
                                                              timestamp[_n-1]) / 1000) / 3600
gen             tspv1           =       tspv /24
drop            tspv
rename          tspv1 tspv
format          tspv %10.0f
recode          tspv .          =       0

*Time since previous purchase
Gen             tspp            =       0
order           customer_number custid conv1 timestamp
bys             customer_number custid (conv1 timestamp)     :       replace tspp = ((timestamp –
                                                                     timestamp[_n-1]) / 1000) /
                                                                     3600 if conv1 == 1
bys             customer_number custid (timestamp): gen conv_sum            =sum(conv1)
bys             customer_number custid conv_sum(timestamp): replace tspp    = sum((timestamp –
                                                                             timestamp[_n-1]) /
                                                                             1000) / 3600 if
                                                                             conv1 == 0
replace         tspp    =       0 if customer_number == 0
replace         tspp    =       0 if tspp <= 0
gen             tspp1   =       tspp /24
drop tspp conv_sum
rename tspp1 tspp
format tspp %10.0f
```

```
//recode tspp .=0

*Defening recent purchases
gen      recent_purchase        =        0
replace recent_purchase         =        1 if tspp <=7

*Nr. of total (previous) purchases
Bys      customer_number custid (timestamp)   :        gen total_nr_purchases = sum(conv1)

*Ratio of purchases over total nr. visits
gen              purchasing_ratio        =        total_nr_purchases / visit_frequency
recode           purchasing_ratio .      =        0
format           purchasing_ratio %3.2f
gen              purchasing_ratio1       =        purchasing_ratio * 100
drop             purchasing_ratio
rename           purchasing_ratio1 purchasing_ratio

*Been before today
bys              customer_number custid date (time)     :        gen been_before_today = _n-1 if
                                                                 !missing(custid)
recode           been_before_today .=0

*Hurry
egen             mean_seconds_per_page = mean(seconds_per_page), by (customer_number custid)
format           mean_seconds_per_page %9.0f
gen              hurry     =      1 if seconds_per_page <= mean_seconds_per_page
recode           hurry .   =      0

*Rename some vars for clearness
rename perchome                  perc_home
rename perccontent               perc_content
rename perczoeken                perc_zoeken
rename percassortment            perc_assortment
rename percproduct               perc_product
rename averagetime               average_time

*Changing some formats
format perc_home                  %3.1f
format perc_content               %3.1f
format perc_zoeken                %3.1f
format perc_assortment            %3.1f
format perc_product               %3.1f
format purchasing_ratio           %3.1f
format detail_page_ratio          %3.1f
format average_time               %10.1f
```