Multi-Channel Advertising Effectiveness

Modeling User Behavior and Approaches for Decision Support in Real-Time Advertising

Von der Fakultät Wirtschaftswissenschaften
der Leuphana Universität Lüneburg zur Erlangung des Grades
Doktor der Wirtschafts- und Sozialwissenschaften

- Dr. rer. pol. -

genehmigte Dissertation von Martin Stange
geboren am 17.12.1984 in Hamburg

Eingereicht am:	19. September 2016
Mündliche Verteidigung (Disputation) am:	20. Januar 2017
Erstbetreuer und Erstgutachter:	Prof. Dr. Burkhardt Funk
Zweitgutachterin:	Prof. Dr. Nadia Abou Nabout
Drittgutachter:	Prof. Dr. Peter Niemeyer

Die einzelnen Beiträge des kumulativen Dissertationsvorhabens sind wie folgt veröffentlicht:

Stange, M. and B. Funk (2014). "Real-Time Advertising". In: Business & Information Systems Engineering 56.5, pp. 305–308.

Stange, M. and B. Funk (2015). "How Much Tracking Is Necessary – The Learning Curve in Bayesian User Journey Analysis". In: Proceedings of the 23rd European Conference on Information Systems, Muenster, Germany.

Stange, M. (2015). "The Impact of TV Ads on the Individual User's Purchasing Behavior". In: Proceedings of the 36th International Conference on Information Systems, Fort Worth, United States.

Stange, M. and B. Funk (2016). "How Big Does Big Data Need To Be?" In: Enterprise Big Data Engineering, Analytics, and Management. Ed. by M. Atzmueller, S. Oussena, and T. Roth-Berghofer. IGI Global. Chap. 1, pp. 1–15.

Stange, M. and B. Funk (2016). "Predicting Online User Behavior Based on Real-Time Advertising Data". In: Proceedings of the 24th European Conference on Information Systems, Istanbul, Turkey.

Veröffentlichungsjahr: 2017

Danksagung

Den folgenden Personen gilt aufgrund ihrer großen Unterstützung auf dem Weg zu meiner Dissertation mein besonderer Dank: Herrn Professor Dr. Burkhardt Funk, Frau Professor Dr. Nadia Abou Nabout, Herrn Micha Edlich, meiner Frau Tanja, meinen Eltern und meiner Oma.

Abstract

Over the last two decades, online advertising has become one of the most important dimension of corporate communications. Nowadays, companies promote their products and services through multiple online marketing channels, for example newsletters, display and video advertising, and most notably, search engine advertising. In this context, statistical models developed in research and practice can be used to measure the effectiveness of advertising activities. The results from these kinds of analyses can, for instance, be used to attribute marketing success (sales, registration, etc.) to individual advertising activities and, thus, to support budget planning for future advertising campaigns. In recent years, a new form of advertising on the Internet has emerged: real-time advertising. Among others, it allows companies to identify potential customers and target them with respect to their interests. In this way, real-time advertising can increase advertising effectiveness and it could, at the same time, improve user experience. With the emerge of this new form of advertising, statistical models have become even more important because they are now being increasingly used to predict online user behavior.

The articles included in this dissertation analyze user-level clickstream data generated during multi-channel advertising campaigns (including TV advertising) and during real-time auctions. The goal of the analyses conducted here is to better understand advertising effects and to support decision-making in this context. Most of the analyses are based on Bayesian models. These models allow for a very flexible structure, which enables researchers to model, for instance, heterogeneity across different types of users or non-linear parameters such as users' reaction times and exponential decay of advertising effects. In addition, these models allow for the inclusion of prior knowledge of parameter distributions, and, therefore, they are well suited for iterative analyses based on clickstream data.

Bayesian models can be evaluated in different ways. Instead of only relying on statistical metrics, the articles included in this dissertation aim to estimate the economic value of these models based on their predictive performance. Although this measure can only approximate their true economic value, this approach can be used to compare and evaluate different models and to illustrate the impact of predictive analyses for companies in the context of big data.

This dissertation contributes to both information systems research and marketing research and has many managerial implications. First, a process is developed to determine optimal sample sizes representing the best balance between computational costs and predictive accuracy in e-commerce in particular and big data contexts in general. In practice, this process can be used to reduce infrastructure and computational costs. Second, the articles included here describe

models that can be used to measure the impact of television ads on users' online shopping behavior. The models can provide insights concerning the effectiveness of individual television ads, the interactions between different advertising channels and the difference in user behavior of TV-induced customers and their non-TV-induced counterparts. Thereby, the models could support decision-making with respect to future advertising campaigns and targeting. Third, the articles describe several possibilities to extend and improve decision support systems currently used in e-commerce and marketing. These improvements enable practitioners to predict users' interests for arbitrary products and services by using corresponding websites as dependent variables. This approach can be used to improve the effectiveness of real-time advertising campaigns, especially those intended to raise brand awareness among customers.

In addition to these contributions, the articles describe possibilities for future research projects at the intersection of information systems and marketing. These kinds of projects could aim to develop methods to take advantage of new possibilities resulting from technological progress, to increase profits from advertising campaigns and selling ad inventories, to provide deeper insights concerning the effectiveness of multi-channel advertising campaigns, and to improve targeting of individual consumers by considering their interests and privacy concerns.

Contents

List of Figures

List of Tables

1 Introduction

In the last 25 years, the Internet has transformed the lives of billions of people. It has lead to innovations such as search engines, online shops, or social networks and will most likely generate further innovations and new business models. Nowadays, a majority of Internet users expects news pages, social networks, and search engines, among other services, to be free of charge (Ritzer and Jurgenson, 2010). For this reason, the majority of free services on the Internet is ad-funded. User interactions in online advertising generate data that is in focus of this dissertation.

Advertisers can use this data to analyze the effectiveness of their advertising activities (Montgomery et al., 2004). In multi-channel advertising, where different online and offline channels are used to advertise products or services, the results of this kind of analysis can be used to support decision-making. This includes, for instance, the allocation of future budgets across individual advertising channels, such as search engine advertising, e-mail advertising, or display advertising (Anderl et al., 2016). With the emerge of real-time advertising[1] as a new form of trading ad slots on the Internet, statistical models have become even more important because they can now be used to predict individual users' interests and purchase intentions and thereby to account for heterogeneous user reactions to advertisements (Bleier and Eisenbeiss, 2015a; Funk and Nabout, 2016).

The articles included in this cumulative dissertation present possible ways to address the challenges combined with these kinds of analyses and present ways to make use of the massive amount of data that is generated during multi-channel advertising campaigns in general and in real-time advertising in particular. They contribute to information systems (IS) and marketing research by developing and evaluating methods that can be used to reduce the amount of data in e-commerce contexts, to measure the effect of TV advertising on online behavior, and to predict user behavior based on real-time advertising data.

The introduction of this dissertation is structured as follows: First, the purpose and the goals of this dissertation are described. Second, a brief overview of related fields of research is provided and their connection to IS is explained. Third, the methods used in the articles are described. Fourth, the main contributions are summarized and the strengths and weaknesses of each article are evaluated. Finally, the introduction concludes by discussing possibilities for further research.

[1] Real-time advertising is also known as programmatic advertising, real-time bidding, or data driven display advertising.

1.1 Purpose and Goals

This section describes three challenges that can be addressed using the methods and tools developed in the articles included in this cumulative dissertation.

1.1.1 Reducing the Amount of Data in E-Commerce Contexts

Due to the increasing use of ad blockers (Statista, 2016b) and the resulting losses in revenue, there is a need for a more accurate match between ad content and user. Statistical analyses can be used to achieve this goal by, for instance, predicting users' target groups or conversion probabilities at the level of the individual user. Such analyses rely on data that is readily available in online advertising, particularly in real-time advertising. Processing and storing all of the available data in real-time advertising in particular and in online advertising in general is, on the other hand, associated with significant costs. In addition to considerable investments in infrastructure, these costs also include significant computational effort, particularly when Bayesian models are used in conjunction with MCMC methods (Wilkinson, 2005). The trade-off between the benefits associated with predictive analyses and the infrastructural costs raises the question of how much data is required to make useful predictions concerning individual users (Crone and Finlay, 2012), which is the first challenge addressed in the articles included in this dissertation. The goal is to develop a method to determine a minimum amount of data required to make accurate predictions about future user behavior. This kind of method can also be beneficial for companies because it would allow them to limit data collection and thereby to respond to the growing concerns by users with regard to extensive tracking activities (Acquisti et al., 2015).

1.1.2 Insights Concerning Multi-Channel Advertising Effects

Particularly in conjunction with multi-channel advertising data, the purpose of statistical models is not limited to prediction (Agarwal and Dhar, 2014). In this context, data generated by user interactions with different advertising media can be used to analyze how individual advertising activities affect user decisions (Naik and Peters, 2009). These kind of analyses make use of statistical models and can be utilized to evaluate the relative contribution of different advertising activities to marketing success, which is important for future allocations of the advertising budget across channels (Anderl et al., 2016). Since it is relatively difficult to measure user contacts with offline advertising (Gal-Or et al., 2006; Kitts et al., 2010) such as radio or television ads, these kinds of models are usually limited to online channels. However, offline advertising is still the major method to promote products and services for many companies (Statista, 2016a). Therefore, effective ways to consider the impact of offline ads on online user behavior could help advertisers to improve their advertising campaigns across offline and online channels (Joo et al., 2014), which is the second challenge addressed in this dissertation. The goal is to close this gap by considering TV advertising campaigns at a user individual level to provide insights concerning offline-online advertising effects and to improve the cross-channel advertising effectiveness.

1.1.3 Prediction of User Behavior

These days, real-time advertising is the most rapidly growing form of advertising on the Internet with a share of approximately 31% in the US in 2015 (Statista, 2016c). This number shows the importance for advertisers and website owners (i.e., the publishers) to engage in real-time advertising. From the perspective of publishers, this kind of advertising allows for both a reduction of transaction costs and an increase in revenues due to higher utilization (Balseiro et al., 2014). For advertisers, the advantage of this new form of advertising lies in the possibility to show ads only to users who are interested in the advertised products (Pandey et al., 2011), and, thereby, to optimize profits from display advertising campaigns (Wu et al., 2015). The increasing use of ad blockers (Statista, 2016b), however, suggests that real-time advertising still lacks effective methods and tools to target users with appropriate ads (Goldfarb and Tucker, 2011b). In particular, methods are needed to target users early in the sales funnel when only little information on their interests is available. Addressing this challenge is the third major goal of this dissertation.

1.2 Theoretical Background

This section briefly describes the theoretical basis of the articles included here and discusses how it relates to the academic discourse of IS.

1.2.1 Modeling User Behavior on the Internet

Designing and evaluating tools to provide the right kind and amount of information to users such as managers or consumers at the right time has a long tradition in IS (Agarwal and Dhar, 2014). In this sense, using data to provide consumers with information that fits their individual requirements can be regarded as a distinct dimension of IS research, which can provide novel insights concerning unresolved marketing problems. The direction in IS research proposed by Agarwal and Dhar (2014) is pursued in the articles included in this dissertation.

The models presented here are used to understand and predict user behavior on the Internet to, for instance, provide them with information on products or services of their interest (Montgomery et al., 2004). The models are based on the work of Chatterjee et al. (2003) who developed a user journey[2] model consisting

[2]In this dissertation, the terms "user journey" and "customer journey" are used interchangeably and describe the sequence of contacts (or touch points) of a given user (or customer) with ads or websites.

of variables that represent long-term and short-term advertising effects[3]. Their model is based on clickstream data generated in the context of display advertising campaigns. They conducted, among others, a hierarchical logistic regression to account for heterogeneity across users. In addition, the model accounts for wear-in and wear-out effects by including quadratic terms of several covariates. Their results allowed for extracting the impact of individual advertising activities on users' click probabilities.

These kinds of analyses can explain the effectiveness of different advertising activities in terms of clicks. However, their economic usefulness is limited because click probability is not directly correlated with conversion probability (Lee et al., 2012; Manchanda et al., 2006; Pandey et al., 2011). Therefore, recent studies often aim to explain conversion rates (Chen and Berkhin, 2011; Zhang et al., 2014). The model developed by Xu et al. (2014), for instance, can be used to predict conversion probabilities by considering decaying advertising effects over time. In real-time advertising, such models can be used to optimize bids for ad slots (Adikari and Dutta, 2015; Lee et al., 2013), to select the right advertising material for the right user at the right time (Perlich et al., 2012), and to properly distribute advertisers' budgets over time (Lu et al., 2015; Yuan et al., 2013). Using real-time advertising for branding campaigns, where conversions and clicks are not of primary interest, is more difficult because methods that can be used to determine users' potential interests in certain products can hardly be found.

In general, the effects of different advertising activities are interdependent (Anderl et al., 2016; Batra and Keller, 2016; Chatterjee, 2010; Kireyev et al., 2015; Piercy, 2012; Yang and Ghose, 2010). For this reason, advertising channels should not be analyzed separately, but have to be seen in the context of other channels. In other words, analyses need to consider cross-channel effects. These effects have, for instance, been observed between the search engine advertising channel and the organic search channel (Yang and Ghose, 2010) or between display ads and organic search conversions (Kireyev et al., 2015). Cross-channel effects are, however, not limited to online advertising (Dinner et al., 2014; Duan and Zhang, 2014; Joo et al., 2014; Yang and Ghose, 2010). Liaukonyte et al. (2015) showed that offline data of TV ads can be used to explain customers' online search and shopping behavior. Therefore, it might, for instance, make sense to combine a TV advertising campaign with a complementary search engine advertising campaign instead of treating these advertising activities separately (Joo et al., 2014).

[3]In this dissertation, the term "effects" (or "impact") is often used to describe user reactions to ads. It must not be misunderstood as a causal influence of ads on user behavior, since the analyses presented here cannot be used to identify whether ads make users behave in a certain way or just attract a certain type of users who are characterized by different patterns of behavior than other users.

1.2.2 Predictive Analytics and Value of Data

An important principle of IS research is the generation of (economic) value through data. In this context, statistical methods and tools are more likely to be examined with respect to their potential to generate business value from (big) data, than from a statistical or algorithmic perspective (Agarwal and Dhar, 2014). Predictive analytics, an exemplary direction in IS research, follows this principle by using historical data to make predictions that are likely to be of economic value (Shmueli and Koppius, 2011). For instance, companies can use predictive analytics to reduce warehousing or maintenance costs, or to prevent churn (Waller and Fawcett, 2013). From a scientific perspective, predictive modeling aims, for instance, to uncover new causal mechanisms, to generate new hypotheses, or to develop new operationalizations of constructs (Shmueli and Koppius, 2011).

Statistical models proposed in the IS and marketing literature are often evaluated by reporting measures that indicate how much variance in the data can be explained by the model (Chatterjee et al., 2003; Li and Kannan, 2014; Liaukonyte et al., 2015). These methods are suitable, for example, to compare different versions of nested models (Gelman et al., 2014, p. 178). However, when sample sizes become very large, the practical relevance of evaluations that only rely on the significance of model parameters, for instance, decreases (Kruschke, 2010; Lin et al., 2013). By contrast, metrics concerning the predictive performance of models are practically meaningful because they can be associated with economic value (Drummond and Holte, 2000; Shmueli and Koppius, 2011). The evaluation method used in this dissertation is based on assigning costs to false predictions and profits to true predictions. Thereby, the economic value of the application of a statistical model can be determined (Domingos, 1999; Elkan, 2001; Nottorf and Funk, 2013).

Due to the potential benefits of predictive analytics, the collection of as much information as possible is often assumed to be very beneficial for many companies (Beath et al., 2012). However, handling the massive amount of data that companies are confronted with can lead to high computational costs. As discussed above, it is often not clear how much data is required to make accurate predictions (Crone and Finlay, 2012). Learning curve sampling has been used to determine the amount of data required to appropriately support decision-making. This method is based on the observation that an increase in sample size reduces uncertainty in the parameter estimates of a learned model (Meek et al., 2002). Based on this method, the articles presented in Chapter 3 and 4 present a framework to determine the required amount of data in the context of real-time advertising.

1.3 Methods

The articles presented here make use of primarily three different methodological tools that are discussed in this section. When it comes to modeling, the articles consistently make use of user-level data that is transformed into user journeys (Section 1.3.1). These models are created based on a Bayesian approach and estimated using MCMC methods (Section 1.3.2). An excursus on how to create an MCMC sampler to estimate Bayesian models is described in Section 1.3.3. In this dissertation, the evaluation of these models is mostly based on predictions and their valuation from the advertisers' perspective (Section 1.3.4).

1.3.1 Advertising Effectiveness Models Based on Clickstream Data

This dissertation develops and evaluates different variations of the model for measuring advertising effects introduced by Chatterjee et al. (2003). It is based on user-level data and has primarily two advantages over other models used for this purpose: First, it is a very flexible model that can easily be extended by additional variables and hierarchy levels. Second, since it is based on a logistic regression, it can be used to calculate probabilities for certain user behaviors (click, registration, conversion). These two aspects make the model well suited to be used in the context of real-time advertising and multi-channel attribution modeling (Shao and Li, 2011).

In its original form, the model proposed by Chatterjee et al. (2003) was used to estimate the users' probability to click on banner ads. Within the scope of this dissertation, the model is extended to predict conversions and product interests. This change of purpose is feasible, because clicks on banner ads do not directly correspond to economical benefit. Conversions, instead, are directly associated with companies' revenues and, therefore, can make analyses economically more meaningful (Manchanda et al., 2006). Other models proposed in the literature to analyze consumer behavior on the Internet are either not flexible enough (e.g., Anderl et al., 2016) or cannot be used to calculate probabilities for future user behavior on an individual basis (Joo et al., 2014; Liaukonyte et al., 2015) and have, therefore, not been used for analysis in this dissertation.

The model proposed by Chatterjee et al. (2003) is based on clickstream data that refers to data generated by individual users while surfing the web (Hui et al., 2009). All data sets used within this dissertation are generated in this manner. They have been recorded by advertising agencies and provided for research purposes. These companies recorded every user action on different online shops including user identifications, timestamps, and advertising channels consumers have used to open the respective website. Recorded actions include clicks on internal links, user registrations and purchases. These data sets have been used to test the process to find the minimum sample size in predictive analytics (Chapters 3 and 4) and to estimate the effect of TV ads on online behavior of users

(Chapters 5 and 6). One data set involves bid requests from a time period of four days received from a big ad exchange. Each bid request includes a user identification, a time stamp and the URL of the website from where it was triggered. The data set has been used to build a model for predicting user interests (Chapter 7).

To analyze this kind of clickstream data, it can be transformed into user journeys. In contrast to the raw data, the user journeys used here contain information concerning previous user activity (Braun and Moe, 2013; Chatterjee et al., 2003; Poel and Buckinx, 2005). These include, for instance, the time between two sessions, the time since the last conversion, and the number of previous contacts with certain advertising channels. These variables allow, for example, to estimate the long-term effects of different advertising activities.

1.3.2 Bayesian Methods

All articles except the one presented in Chapter 2 use statistical methods to analyze the data. More specifically, the methods are used to understand the impact of advertising exposure on decision making at the level of the individual user. The statistical models developed here involve more than one hierarchical level, regularization, or non-linear parameters. Frequentist methods such as maximum likelihood estimation are often not suitable to estimate these kinds of models. Instead, the use of Bayesian models in combination with Markov Chain Monte Carlo (MCMC) methods allows for nearly unbounded complexity including, for instance, heterogeneity across unit-level parameters and non-normal parameter distributions (Allenby and Rossi, 2006; Kruschke, 2010). With the availability of affordable computational power, these methods have become increasingly popular in marketing research (Bucklin and Sismeiro, 2009; Chatterjee et al., 2003; Manchanda et al., 2006; Rossi et al., 2006; Xu et al., 2014). In the context of user journey analyses, Bayesian methods are feasible because they can be used, for instance, to estimate the decay rates of advertising effects (Section 1.3.3), non-linear transformations of user journey variables (Section 5.4.1), or cluster-level parameters (Section 5.4.3).

Apart from the possibility to estimate models of high complexity, Bayesian models have another advantage over frequentist methods: They enable researchers to include prior beliefs about reality and to update this belief as new data is considered for analysis (Kruschke, 2010; Gelman et al., 2014, p. 34 ff.). The model presented in Section 7.3.2, for instance, follows this principle by iteratively updating knowledge concerning the correlation of the users' clickstreams and interests as new data is available. Depending on the degree of prior certainty, prior belief either has a weak (small degree of certainty) or strong (high degree of certainty) influence on the results of Bayesian parameter estimation (i.e., posterior distributions), which, thus, does not only rely on data, but also on prior knowledge obtained from previous analyses or the literature (Gelman et al., 2014, p. 35). In this sense, Bayesian data analysis enables cumulative scientific progress (Kruschke, 2010).

Bayesian estimations result in parameter distributions (Kruschke, 2010). For simple models, these posterior distributions can be calculated analytically or be approximated using, for instance, variational Bayesian methods. More complex models require MCMC methods that allow for a high flexibility in model creation (Gelman et al., 2014, p. 259 ff.). These methods are based on random number generators that are used to draw samples from posterior parameter distributions. Since these methods are frequently used in the articles presented in this dissertation, this introduction deserves more detailed information on Bayesian model creation and estimation. For this purpose, the next section describes a non-standard Bayesian model that can be used to estimate decaying advertising effects and shows how to estimate it using MCMC sampling.

1.3.3 Exemplary Creation and Estimation of a Bayesian Model

Although a variety of software packages to estimate Bayesian models exists, it is often required to implement specific MCMC samplers, for instance, to parallelize involved sampling steps or to implement non-linear extensions that are not covered by available software packages. For this reason, this section describes the creation and estimation of a logistic model that allows for determining exponentially decaying advertising effects. The knowledge of channel-specific decay parameters may be useful for attributing advertising success to individual online channels and for planning multi-channel marketing campaigns.

The model presented here extends the model proposed by Chatterjee et al. (2003) by using decay parameters to transform the design matrix X. It is described with Equation 1.1:

$$p(conv)_{jk} = \frac{1}{1 + \exp(-\tilde{X}_{jk}\beta)} \tag{1.1}$$

In this equation, $p(conv)_{jk}$ represents the probability that user j converts in his/her k^{th} session. The design matrix X can be transformed according to Equation 1.2, where X_{jk} represents the k^{th} observation of user j at time t_k ($k > 1$). The parameter α_l represents the unknown decay constant of the advertising effect of channel l.

$$\tilde{X}_{jkl} = X_{jkl} + \sum_{m=1}^{k-1} \exp\left\{-(t_k - t_m)\alpha_l\right\} X_{jml} \tag{1.2}$$

Using this equation, Table 1.1 shows the exemplary transformation of X_{jkl} (second and third column) into $\tilde{X}_{jkl}$ (fourth and fifth column).

To create a sampling algorithm that can be used to estimate the model parameters α and β, first, a general procedure on how to create a Gibbs sampler is described.

t	X_{jk1}	X_{jk2}	$\tilde{X}_{jk1}$	$\tilde{X}_{jk2}$
t_1	1	0	1	0
t_2	0	1	$\exp\{-(t_2-t_1)\alpha_1\}$	1
t_3	1	0	$1+\exp\{-(t_3-t_1)\alpha_1\}$	$\exp\{-(t_3-t_2)\alpha_2\}$
t_4	1	0	$1+\exp\{-(t_4-t_1)\alpha_1\}+\exp\{-(t_4-t_3)\alpha_1\}$	$\exp\{-(t_4-t_2)\alpha_2\}$

TABLE 1.1: Transformation of covariates for user j at time t_k.

General Procedure

First, the probabilistic graphical model that relates to the (research) question is created. An appropriate method to draw these models is plate notation. The graphical model includes acyclic dependencies between parameters and the multiplicities of the parameters' distributions. Although it is a useful tool to get a common understanding of the model, the graphical model is not a prerequisite to create the Gibbs sampler (Murphy, 2012, p. 321 ff.). Second, the joint posterior density of the parameters is derived. It consists of the product of likelihood and prior distributions, according to Bayes rule $p(M|D) = p(D|M)p(M)/p(D)$. Considering multiplicity, which is represented by plates in the graphical model, is crucial in this step (Gelman et al., 2014, p. 288). Third, the marginal distributions for each parameter θ are calculated, ignoring normalizing terms. All factors not including parameter θ can be removed from the equation because they are constant with respect to parameter θ. Fourth, the conditional parameter distributions from the previous step have to be transformed into a common density (e.g., normal or gamma distribution) in order to sample from them directly, if possible (Gelman et al., 2014, pp. 289-290). If no closed form can be found, the parameters can be estimated using a Metropolis step. Fifth, in each sampling iteration, samples are drawn from the conditional distributions calculated in the previous step and used for the subsequent sampling iteration. Higher order parameters are sampled only once per iteration whereas group level parameters can be sampled in parallel due to their conditional independence (Gelman et al., 2014, p. 289).

Creating the Probabilistic Graphical Model

To set up a Gibbs sampler that allows for sampling from the empirical distribution of α and β, the approach proposed by Albert and Chib (1993) is followed and the graphical model is created in accordance with the first step described in the general procedure (Figure 1.1). The parameters λ_i and z_i are introduced by Albert and Chib (1993). These variables are used in the Gibbs sampler to draw samples from a mixture of truncated normal distributions to approximate the logistic distribution. The graphical model shows the dependencies of the parameters and their multiplicities. The parameter N refers to the total number of touch points. The index i replaces the index jk, i.e., touch point k of user j, for convenience.

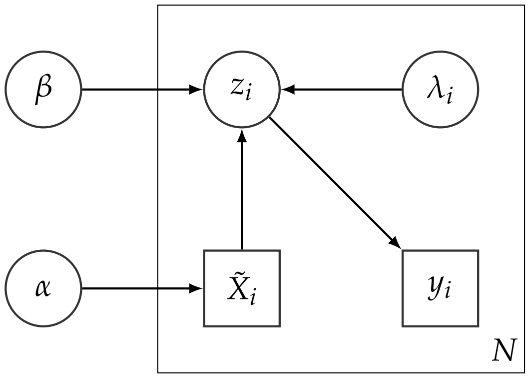

FIGURE 1.1: Probabilistic graphical model of a logistic regression with transformation of covariates using parameter α.

Deriving the Joint Posterior Distribution and Defining Prior Distributions

The joint posterior distribution of the model $p(\alpha, \beta, \lambda_i, z_i, \tilde{X}_i, y_i)$ can be derived from the graphical model presented in Figure 1.1 by applying the chain rule for conditional distributions.

$$p(\alpha, \beta, \lambda_i, z_i, \tilde{X}_i, y_i) = p(\alpha)p(\beta)\prod_i p(\lambda_i)p(z_i|\beta, \lambda_i, \tilde{X}_i, y_i)p(\tilde{X}_i|\alpha) \tag{1.3}$$

Based on Equation 1.3, the marginal posterior densities of the involved parameters can be calculated. Since it is not possible to sample from the logistic probability density function directly, Albert and Chib (1993) approximate it by introducing a latent variable z_i which is distributed following a mixture of truncated normal distributions following Equation 1.4:

$$p(z_i|\beta, \lambda_i, \tilde{X}_i, y_i) = \begin{cases} N_{0,\infty}(\tilde{X}_i\beta, \lambda_i^{-1}) & \text{if } y_i = 1 \\ N_{-\infty,0}(\tilde{X}_i\beta, \lambda_i^{-1}) & \text{else} \end{cases} \tag{1.4}$$

To approximate the logistic distribution, the precision parameters λ_i are sampled from a Gamma distribution with $\lambda_i \sim \Gamma(\nu/2, \nu/2)$ with $\nu \approx 8$ (Albert and Chib, 1993; Gelman et al., 2014, pp. 293-294). The likelihood of the latent variable z_i (neglecting the truncation) is given in Equation 1.5. The truncation can be done by inverse sampling, which is not going to be described in further detail here.

$$z_i \;\propto\; \lambda_i^{\frac{1}{2}} \exp\left\{-\frac{\lambda_i}{2}(\tilde{X}_i\beta - z_i)^2\right\} \tag{1.5}$$

The prior distributions of λ_i and β are given in Equations 1.6 and 1.7.

$$p(\beta) \;\propto\; \exp(-\frac{1}{2}\beta^T\Omega\beta) \tag{1.6}$$

$$p(\lambda_i) \;\propto\; \lambda_i^{\frac{\nu}{2}-1}\exp(-\frac{\nu}{2}\lambda_i) \tag{1.7}$$

In Equation 1.6, the parameter Ω represents the prior precision matrix for the parameters β.

Calculating Marginal Posterior Distributions

Given the equations described above, the marginal posterior distributions can be calculated. According to Equation 1.3, the parameters β are distributed following

$$p(\beta|\cdot) \propto p(\beta) \prod_i p(z_i|\beta, \lambda_i, \tilde{X}_i, y_i). \tag{1.8}$$

Inserting priors and likelihood from Equations 1.5, 1.6, and 1.7 and eliminating index i by applying matrix notation yields

$$p(\beta|\cdot) \propto \exp\left\{-\frac{1}{2}\beta^T\Omega\beta - \frac{1}{2}[\tilde{X}\beta - z]^T\Lambda[\tilde{X}\beta - z]\right\}. \tag{1.9}$$

To sample from the marginal distribution, Equation 1.9 needs to be transformed to a closed-form distribution. This is done by completing the squares in the exponent of Equation 1.9. The exponent, denoted ω for convenience, is given by

$$\omega = -\frac{1}{2}\left\{\beta^T\Omega\beta + \beta^T(\tilde{X}^T\Lambda\tilde{X})\beta - 2\tilde{X}\beta\Lambda z + z^T\Lambda z\right\}. \tag{1.10}$$

All factors of $p(\beta|\cdot)$ that do not include β can be integrated out, as they are constant with respect to β:

$$\omega = -\frac{1}{2}\left\{\beta^T(\tilde{X}^T\Lambda\tilde{X} + \Omega)\beta - 2\tilde{X}\beta\Lambda z\right\} \tag{1.11}$$

This exponent can be transformed into the exponent of a multivariate normal distribution of the form $-1/2(\mu-\beta)^T\Theta(\mu-\beta) = -1/2(\mu^T\Theta\mu - 2\mu^T\Theta\beta + \beta^T\Theta\beta)$ with precision matrix Θ and mean μ. By comparing the summands, the precision matrix and means of the multivariate normal distribution ($\beta \sim \text{Normal}(\mu, \Theta)$) can be derived as follows:

$$\Theta = \tilde{X}^T\Lambda\tilde{X} + \Omega \tag{1.12}$$

$$\mu = \Theta^{-1}\tilde{X}\Lambda z \tag{1.13}$$

In analogy to the process described for β, the precision parameters λ_i can be sampled as described with Equation 1.14:

$$
\begin{aligned}
p(\lambda_i|\cdot) &\propto p(\lambda_i)p(z_i|\beta,\lambda_i,\tilde{X}_i,y_i) \\
&\propto \lambda_i^{\frac{\nu}{2}-1}\,\lambda^{\frac{1}{2}}\,\exp(-\frac{\nu}{2}\lambda_i)\,\exp\left\{-\frac{1}{2}\lambda_i(\tilde{X}_i\beta-z_i)^2\right\} \\
&\propto \lambda_i^{\frac{\nu}{2}+\frac{1}{2}-1}\exp\left\{-\frac{1}{2}(\nu\lambda_i+\lambda_i(\tilde{X}_i\beta-z_i)^2)\right\} \\
&\propto \lambda_i^{\frac{\nu}{2}+\frac{1}{2}-1}\exp\left\{-(\frac{\nu}{2}+\frac{1}{2}(\tilde{X}_i\beta-z_i)^2)\lambda_i\right\}
\end{aligned}
\tag{1.14}
$$

Equation 1.14 has the form of a Γ distribution with shape $\frac{1}{2}(\nu+1)$ and scale $2(\nu+(\tilde{X}_i\beta-z_i)^2)^{-1}$. Consequently, the parameters λ_i can be sampled directly from this distribution in each iteration of the Gibbs sampler (Albert and Chib, 1993).

Metropolis Step to Sample from Decay Constants

To sample from the decay constants α used to transform X into $\tilde{X}$, a Metropolis step is needed because there is no closed form of the related marginal distribution (Gelman et al., 2014, p. 281 f.). The Metropolis step consists of three parts: First, a new decay constant α_l^* is sampled from a proper jumping distribution (Gelman et al., 2014, p. 278). For example, the following jumping distribution could be assumed to sample α_l^*:

$$\alpha_l^* \sim \text{Normal}(\alpha_l,\sigma^2) \tag{1.15}$$

In this equation, α_l^* represents the proposed value for the decay parameter of referral channel l, while α_l refers to the decay parameter from the previous iteration. Second, the design matrix X needs to be transformed using Equation 1.2 and α_l^*. This step needs to be done for each user and referral channel l and, therefore, it is well suited to be parallelized. Third, the ratio of probabilities for α_l^* and α_l is calculated. Since the jumping distribution is symmetric, i.e., $p(\alpha_l|\alpha_l^*)=p(\alpha_l^*|\alpha_l)$, this ratio does not depend on the jumping distribution (Gelman et al., 2014, p. 278 f.):

$$r_l=\frac{\exp\left\{-[\tilde{X}^*\beta-z]^T\Lambda[\tilde{X}^*\beta-z]\right\}}{\exp\left\{-[\tilde{X}\beta-z]^T\Lambda[\tilde{X}\beta-z]\right\}} \tag{1.16}$$

In this equation, $\tilde{X}^*$ represents the transformed design matrix where covariate l is exchanged by transforming it using α_l^*. All other covariates are transformed using α from the previous sampling iteration. The new value α_l^* is accepted with probability p_l^{acc} following:

$$p_l^{acc}=\min\{1,r_l\} \tag{1.17}$$

For proper mixing of chains, the parameter σ^2 of the jumping distribution (Equation 1.15) can be tuned prior to the actual analysis to achieve a desired acceptance rate (Gelman et al., 2014, p. 279 ff.).

Due to the approximation of the logistic distribution with a mixture of truncated normals, the values of β have to be divided by 0.634 after the estimation has finished (Albert and Chib, 1993). The sampling algorithm described here has been implemented and tested using simulated data.

Implementation

Drawing posterior samples from a Gibbs sampler is relatively easy, since most statistical software packages support drawing from simple distributions such as the ones used here. However, these implementations might not necessarily be the most efficient ones to obtain posterior distributions (Gelman et al., 2014, p. 280, 296). In addition to more efficient methods to draw samples from posterior distributions, for example Hamiltonian Monte Carlo estimation (Gelman et al., 2014, p. 300 ff.), MCMC algorithms can often be parallelized to speed up computation times (Da Silva, 2010; Wilkinson, 2005). A simple form of parallelization is deploying multiple instances of a Gibbs sampler on different CPU kernels which results in multiple MCMC chains. This kind of parallelization is often implemented by common software packages used to estimate Bayesian models (Gelman et al., 2014, p. 307; Plummer, 2003). Parallelizing different sampling steps within one iteration, however, is only possible for steps that involve independently drawn distributions or independent transformations of variables. In the Gibbs sampler described here, these steps include the transformation of each user journey using α, i.e., $\tilde{X}_i = f(X_i, t_i, \alpha)$, or sampling the individual λ_i. These steps could, for instance, be parallelized on GPU cores (Da Silva, 2010). Due to the increasing interest in Bayesian methods, further research that aims to develop efficient implementations that take advantage of widely available multicore hardware (such as GPUs) would be useful in this context.

1.3.4 Valuation of Data

In the literature, several ways to assess and compare statistical models have been proposed (Gelman et al., 2014, p. 165 ff.). In this dissertation, the predictive performance of models is the most important measure for evaluation because predictions can be used in the context of decision support systems, particularly in real-time advertising. Higher predictive accuracies enable advertisers to make better decisions. Therefore, the model evaluation conducted here includes the area under the receiver operator characteristic (AUC) or misclassification errors. These metrics, however, cannot be used to assess the utility of predictions, i.e., the actual economic value that is associated with the data or, to be more precise, the model that was trained using this data. This is particularly true for data sets with highly imbalanced numbers of occurrences of different outcomes in multinomial classifications and different costs associated with false predictions (Domingos, 1999).

For this reason, the articles included in this dissertation evaluate models based on cost/benefit considerations in the following way: First, a $k \times k$ cost matrix $C = (c_{ij})$ is defined that includes the marginal profits and costs of correct and incorrect classifications. The entries $c_{i \neq j}$ represent the marginal costs for falsely labeling an instance of class j with label i, while the diagonal entries $c_{i=j}$ represent the marginal profits (i.e., negative costs) for correctly labeling an instance of class j with label i. Second, for classifications based on probabilities (as in the case of logistic regressions), k threshold probabilities p_i^T, representing the minimum probability for labeling an unknown instance with label i, need to be determined so that they maximize the utility $U(p^T) = -\sum_{ij} c_{ij} \cdot a_{ij}(p_i^T)$. The variable $a_{ij}(p_i^T)$ represents an entry of the $k \times k$ confusion matrix that is generated by testing the classifier on a holdout set. Each entry $a_{ij}(p_i^T)$ holds the number of instances of class j classified with i (using p_i^T). The optimal values of the threshold probabilities can, for instance, be determined by simulated annealing (Murphy, 2012, p. 869 f.).

This kind of model evaluation is a helpful method that can be used in practice to chose between different classifiers and to assess the profitability of applying a classifier (Domingos, 1999; Drummond and Holte, 2000). As demonstrated in Chapter 7, the evaluation could be conducted using cross-validation to increase its robustness. In practice, it may, however, be difficult to determine the costs for false predictions and the profits for true predictions, particularly in real-time advertising because of the varying costs for ad impressions (Chawla, 2005). Therefore, the approach presented here can only be used to approximate the costs and profits of predictive analytics in this context.

1.4 Discussion

This dissertation consists of six interrelated articles. This section summarizes their contributions and managerial implications and discusses how they are linked with each other. In addition, it evaluates each article's strengths and weaknesses. The articles follow a pragmatic approach of IS research that focuses, for instance, on developing new arguments, or uncovering new relationships and patterns of user behavior (Avison and Malaurent, 2014). This perspective allows researchers to address new challenges in the industry and to adopt new methods developed by other disciplines (Agarwal and Dhar, 2014). In contrast to more traditional directions in IS research, which is very much concerned with theory, this recent direction emphasizes the primacy of data and methods. The articles included here follow this direction by discussing several current challenges in online advertising and developing methods and tools to address them.

1.4.1 Contributions and Implications

As indicated by their ranking by VHB-Jourqual, the three conference papers presented in Chapter 3 (Stange and Funk, 2015), Chapter 5 (Stange, 2015), and Chapter 7 (Stange and Funk, 2016b) represent the most relevant contributions of this dissertation. In addition, the article presented in Chapter 2 has been published in Business & Information Systems Engineering. Although it is unranked according to VHB-Jourqual V3, the book chapter presented in Chapter 4 was reviewed in a double blind review process. The study presented as a working paper in Chapter 6 has been presented on the European Marketing Agency Conference 2016 in Oslo and on the Marketing Science Conference 2016 in Shanghai. The major contributions and implications and the methodical basis of the articles included in this dissertation are illustrated in Figure 1.2.

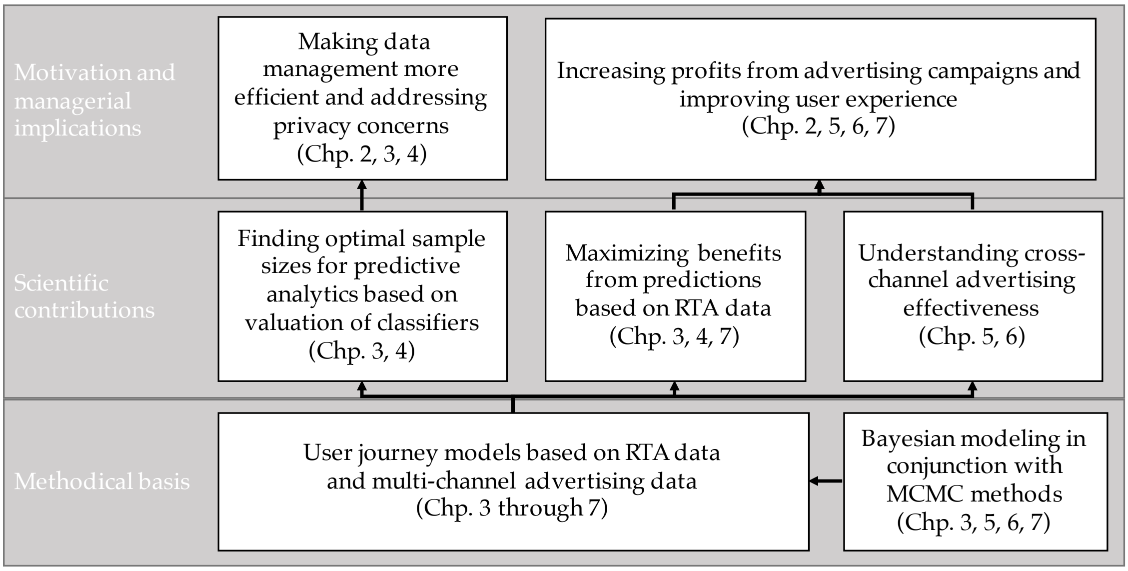

FIGURE 1.2: Methodical basis, contributions and implications, and connections between the articles.

This figure also implies the relationship among the articles: First, Chapter 2 identifies research gaps in the context of real-time advertising regarding data management (addressed in Chapters 3 and 4) and decision support (addressed in Chapters 5, and 6, and 7). Second, an extended version of the model developed in Chapter 3 is presented in Chapter 4. Third, Chapter 5 proposes a model to measure the impact of TV ads on the users' conversion probability and this model is extended in Chapter 6. Fourth, the model developed in Chapter 7 could serve as an extension of the models developed in Chapter 5 and 6. Table 1.2 summarizes the contributions and implications of the articles presented in the following chapters.

Table 1.2: Contributions, managerial implications, and ranking of journal by VHB Jourqual V3

Title	Scientific contribution	Managerial implications	Journal
Real-Time Advertising (Chp. 2)	The study (1) outlines research possibilities for IS scholars; it (2) presents a BPMN diagram to illustrate the process of real-time advertising and involved actors.	The article provides an overview of challenges for practitioners using real-time advertising for exposing users to ads.	BISE (ranking: B)
How Much Tracking Is Necessary? (Chp. 3)	The article (1) proposes a process to determine a minimum sample size in e-commerce contexts based on learning curve sampling; (2) six different Bayesian analyses are conducted that show the saturation of predictive accuracy in user journey analyses; (3) more complex models require more data to outperform simpler models.	The study shows (1) that advertisers only need to store a small fraction of the available data produced by user interaction; (2) the method can be used in practice to reduce infrastructure costs and computational costs generated by storing data in an unsystematic manner.	ECIS 2015 (ranking: B)
How Big Does Big Data Need to Be? (Chp. 4)	The chapter provides (1) an extension of Stange and Funk (2015) by considering the costs for including additional data records into the analysis; it (2) shows that benefits obtained from predictive analyses decrease as more and more samples are considered; it (3) generalizes the idea of Stange and Funk (2015).	The chapter provides a process model that (1) can be used to determine the optimal sample size in a predictive analytics application; it (2) encourages a critical discussion on data collection and storage in different industries.	Book chapter, IGI Global (no ranking)

Table 1.2: Scientific contributions, managerial implications, and ranking of journals according to VHB Jourqual V3

Title	Scientific contribution	Managerial implications	Journal
The Impact of TV Ads on the Individual User's Purchasing Behavior (Chp. 5)	The article (1) provides a Bayesian non-linear extension of the model proposed by Chatterjee et al. (2003); it (2) describes a model that can be used to estimate the time-dependent short-term TV effects on conversion probability; it (3) presents a hierarchical Bayesian model used to estimate cross-channel advertising effects.	Advertisers can use the method (1) to gain insights into offline-online advertising effects and (2) to improve predictions, e.g., in real-time advertising; the model (3) can support managers during planning of advertising campaigns in terms of the allocation of budgets to individual online and offline channels.	ICIS 2015 (ranking: A)
The Reduced Customer Revenue of TV-Induced Online Shoppers (Chp. 6)	The results suggest that customers who open a shop's website in response to TV ads exhibit (1) lower conversion probabilities, (2) smaller shopping baskets, and, consequently, are characterized by (3) lower customer revenue than other customers.	Advertisers (1) can use the modeling approach to analyze the short-term advertising effect of multiple TV ads during a campaign; they are (2) advised to consider reduced conversion probabilities and customer revenues, for instance, when calculating the return on investment of TV advertising campaigns.	Working paper (no ranking)
Predicting Online User Behavior Based on Real-Time Advertising Data (Chp. 7)	The article (1) provides a framework to extend and improve existing decision support systems employed in e-commerce and marketing; it (2) provides a method to measure the impact of bid request data based on the valuation of the analysis; it (3) encourages scholars to focus on this type of big data.	The method developed here (1) enables practitioners to make predictions about users' interests for arbitrary products or services by using corresponding websites as dependent variables; it (2) can be used in decision support systems to increase profits from (awareness-oriented) cross-channel advertising campaigns.	ECIS 2016 (ranking: B)

1.4.2 Evaluation of the Contributions

The purpose of this section is to reconsider the contributions, implications and limitations of the articles included in this dissertation.

Research Opportunities in the Context of RTA

The article presented in Chapter 2 (Stange and Funk, 2014) provides a short overview of possibilities for further research in the context of real-time advertising from the point of view of IS and, thereby, presents a contextual framework for this dissertation. It contains a business process diagram that provides a systematic overview of the real-time advertising process and involved actors. This article is not a systematic literature review, instead it proposes three distinct further directions of IS research. More specifically, it is argues that IS research could contribute to data management, decision support, and users' perception of this new type of advertising.

Learning Curve Sampling in User Journey Analyses

The method proposed in Chapter 3 (Stange and Funk, 2015) can be used to determine the minimum sample size required to make accurate predictions about future user behavior in the context of multi-channel advertising. The article provides strong evidence for the saturation of predictive accuracy based on two different data sets and three different Bayesian models. Results suggest that e-commerce companies only need to consider a very small fraction of their data to optimize profits from predictive analyses. There are only few theoretical studies dealing with this issue, probably because data costs are often of minor concern in studies dealing with predictions in the context of big data. In addition, companies' actual requirements to make data management more efficient has yet to be addressed by the literature. To identify further managerial implications, qualitative research to examine companies' perspectives on the trade-off between infrastructural costs and benefits from predictive analytics in greater detail would be required. However, as the article intends to be a starting point to strengthen practitioners' and researchers' awareness for possibilities to reduce computational costs in the context of big data analyses, these drawbacks become less significant. The article suggests that extensive tracking activities by companies, which many customers are increasingly concerned about (Acquisti et al., 2015), are not necessarily economically feasible. Thus, applying the proposed methods may also help to regain customer trust and satisfaction (Bleier and Eisenbeiss, 2015b; Malhotra et al., 2004).

The book chapter (Stange and Funk, 2016a) included in Chapter 4 extends the model proposed in Chapter 3 by considering the marginal costs for data storage and collection to determine the maximum utility for predictive analytics. In contrast to Stange and Funk (2015), the book chapter applies regularized maximum likelihood estimation and, thus, does not contain a random slope model. In contrast to the article presented in Chapter 3, the book chapter uses 10-fold cross-validation to calculate the area under the curve metric. Due to their similarity, the paper has the same weaknesses as the paper presented in Chapter 3. Although the process model developed here could be beneficial for arbitrary uses of predictive analytics, its applicability is only tested using e-commerce data. To provide evidence for its general applicability, further research that applies the process to predictive big data analytics outside of the e-commerce field would be needed.

The Impact of TV Ads on Online Customer Behavior

Stange (2015) presents a novel approach to measure the effect of television advertising on online customer behavior. In contrast to related work investigating the effect of TV ads on online behavior at an aggregated level, the model developed here allows for estimating this effect at the level of the individual user. For this reason, the model can be used to better understand customer behavior in e-commerce with respect to offline advertising. At the same time, the model can improve predictions, for instance, in the context of real-time advertising. While it offers a new method, the article also has three drawbacks: First, the structure of this article is not as clear as it could be. The first step of the analysis describes a non-linear extension of the user journey model proposed by Chatterjee et al. (2003). The increase in predictive accuracy of this extension was compared with one of the inclusion of the TV effects (Table 5.6) to test which approach is the more effective means to improve predictive accuracy. This comparison would certainly deserve further attention (maybe in a separate paper). The major goal of the article, however, is the analysis of TV ad effects on online behavior. This analysis would have deserved a more detailed discussion, for instance, with regard to the interpretation of the results obtained from the Bayesian model. Due to the first part of the analysis and the space limitations, this extended discussion was not possible in the original article. Second, inconsistencies in the data set used here were found during data preparation for the article presented in Chapter 6. These inconsistencies are discussed in greater detail in Section 5.8. While they do not limit the major methodological contributions of the article, these issues were investigated and the approach was tested again using a corrected sample. However, it was decided to include the unchanged version of the article in Chapter 5 to preserve its originality. Nevertheless, several footnotes were added in Chapter 5 to indicate drawbacks related to these inconsistencies. Third, during the investigation of these inconsistencies, strong correlations of effects that originate from the time of the day and effects that are related to TV ads (aired at a certain time of the day) were found. These correlations make the identification of TV advertising effects more difficult and, at the same time, explain why the inconsistencies in the data set were not noticed in the first place. This issue is discussed in greater detail in Section 5.8.1. In addition, an alternative modeling that is more robust with respect to these kinds of correlations is discussed in Section 5.8.2.

The study presented in Chapter 6 extends the method proposed in Chapter 5 to measure the difference between customers that open the online shop in response to TV ads and those that found their way to the online shop by other means with respect to conversion probabilities, shopping baskets, repeat purchase behavior, and a 90-day revenue per customer. The analysis is based on calculating the probability that a given visit is a direct response to a TV ad. The results suggest that customers who open the website in response to TV ads are less likely to convert and are characterized by smaller shopping baskets. Advertisers are advised to anticipate this difference, for instance, by granting discounts to customers who are likely to have watched a certain TV ad recently. In its current form as a working paper, the article could be improved by applying hierarchical models to account for heterogeneity across individual customers. In addition, its theoretical foundation could be extended. Due to the relatively short time period covered by the data (4 months), the study cannot reveal insights concerning long-term customer behavior. For instance, it would have been be very interesting to compare customer lifetime values of consumers who tend to respond to TV ads and those who do not. Thus, for future research, it is recommended to increase the time period of data collection.

Using RTA Data to Predict User Behavior

The article presented in Chapter 7 (Stange and Funk, 2016b) provides a method to predict user behavior based on real-time advertising data using an iterative Bayesian multinomial logistic regression with regularization. The article directly addresses challenges related to decision support that are outlined in Chapter 2. In addition to introducing a new method, this article contributes to the literature by discussing the practical implementation of the model as an extension to existing decision support systems. These systems could consider the results from the model as an additional piece of information to select ads and calculate bids. The iterative Bayesian model described here could be easily modified to meet companies' requirements and extended by hierarchical layers (e.g., using spatial, geographical, or demographic data). As the article involves a valuation of the analysis considering different benefit/cost ratios, it also addresses one of the research possibilities outlined in Stange and Funk (2015). The Bayesian analysis proposed here is, however, computationally demanding and, therefore, might be impractical in some real-life situations. The suggestion that it could be useful in cross-channel advertising scenarios might need additional discussions and further evidence. In addition, the study does not provide a comparison to other models used in real-time advertising or proposed by the literature. Instead, it has a rather exploratory character and is more practically motivated rather than theoretically. According to Agarwal and Dhar (2014), however, such studies are one means of IS research to uncover new patterns of knowledge and, thus, to help develop new theories.

1.5 Conclusion

The articles included in this dissertation propose methods that can be used to support decision-making in e-commerce contexts. More specifically, the articles address three important dimensions of this process: managing data, explaining advertising effects, and predicting user behavior.

The first important dimension addressed in this dissertation is management of big data in e-commerce. Users are increasingly aware of tracking activities and not willing to disclose private information on the Internet (Acquisti et al., 2015). This shift toward data privacy and regulation could potentially hurt advertisers and publishers (Goldfarb and Tucker, 2011a). The articles included here contribute to this debate by showing that only a small fraction of the data currently generated by user activity is needed to make accurate predictions about future user behavior. In light of the current situation and this finding, companies are well advised to reconsider the usefulness of extensive tracking activities in order to regain the trust of users, and, also to lower infrastructural costs. Further research involving different data sets could be helpful to advance this discussion.

Regarding the second dimension, explaining advertising effects, the models proposed in this dissertation can be used by e-commerce companies to attribute marketing success to individual advertising channels. To assess to what extent different online and offline advertising channels contribute to users' purchasing probability, the models can, however, be only a first step. To apply them in a commercial context, more research would be necessary to transform the estimated parameters into a managerially more meaningful form with due regard to possible saturation effects and scalability constraints of different advertising activities (Funk and Nabout, 2016).

Concerning the third dimension, predicting user behavior, the methods proposed here can serve as extensions of bidding agents employed by advertisers and their service providers in the context of real-time advertising. This rapidly growing form of advertising could be used to target individual users by predicting their specific interests. Applying the appropriate methods and tools for targeting, advertisers can increase their profits due to more effective campaigns. In light of general market dynamics and evolving business models, researchers should continue to improve models and decision support systems used in this context, for instance by including novel data from different sources such as devices used, previously visited websites, geographical information, information on offline advertising, or information on the viewability of ads.

In summary, the intersection of IS and marketing involves many opportunities for further research, some of which are discussed in greater detail in the articles included here. Apart from theoretical contributions, this research can provide practical methods and tools for advertisers and their service providers that have the potential to improve decision support systems in e-commerce, and, thereby to increase companies' profits and, at the same time, customer trust and satisfaction.

1.6 References

Acquisti, A., L. Brandimarte, and G. Loewenstein (2015). "Privacy and human behavior in the age of information". In: *Science* 347.6221, pp. 509–515.

Adikari, S. and K. Dutta (2015). "Auto pricing Strategy for Real time bidding in advertising: A dynamic programming approach". In: *Proceedings of the 36th International Conference on Information Systems, Fort Worth, United States*.

Agarwal, R. and V. Dhar (2014). "Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research". In: *Information Systems Research* 25.3, pp. 443–448.

Albert, J. H. and S. Chib (1993). "Bayesian analysis of binary and polychotomous response data". In: *Journal of the American Statistical Association* 88.422, pp. 669–679.

Allenby, G. M. and P. E. Rossi (2006). "Hierarchical Bayes models: A Practitioners Guide". In: *SSRN 655541*.

Anderl, E., I. Becker, F. von Wangenheim, and J. H. Schumann (2016). "Mapping the customer journey: Lessons learned from graph-based online attribution modeling". In: *International Journal of Research in Marketing (Forthcoming)*.

Avison, D. and J. Malaurent (2014). "Is theory king?: Questioning the theory fetish in information systems". In: *Journal of Information Technology* 29.4, pp. 327 –336.

Balseiro, S. R., J. Feldman, V. Mirrokni, and S. Muthukrishnan (2014). "Yield optimization of display advertising with ad exchange". In: *Management Science* 60.12, pp. 2886–2907.

Batra, R. and K. L. Keller (2016). "Integrating Marketing Communications: New Findings, New Lessons and New Ideas". In: *Journal of Marketing, forthcoming*.

Beath, C., I. Becerra-Fernandez, J. Ross, and J. Short (2012). "Finding value in the information explosion". In: *MIT Sloan Management Review* 53.4, pp. 18–20.

Bleier, A. and M. Eisenbeiss (2015a). "Personalized Online Advertising Effectiveness: The Interplay of What, When, and Where". In: *Marketing Science* 34.5, pp. 669–688.

Bleier, A. and M. Eisenbeiss (2015b). "The Importance of Trust for Personalized Online Advertising". In: *Journal of Retailing* 91.3, pp. 390–409.

Braun, M. and W. W. Moe (2013). "Online Display Advertising: Modeling the Effects of Multiple Creatives and Individual Impression Histories". In: *Marketing Science* 32.5, pp. 753–767.

Bucklin, R. E. and C. Sismeiro (2009). "Click here for Internet insight: Advances in clickstream data analysis in marketing". In: *Journal of Interactive Marketing* 23.1, pp. 35–48.

Chatterjee, P. (2010). "Multiple-channel and cross-channel shopping behavior: role of consumer shopping orientations". In: *Marketing Intelligence & Planning* 28, pp. 9–24.

Chatterjee, P., D. L. Hoffman, and T. P. Novak (2003). "Modeling the clickstream: Implications for web-based advertising efforts". In: *Marketing Science* 22.4, pp. 520–541.

Chawla, N. V. (2005). "Data Mining for Imbalanced Datasets: An Overview". In: *Data Mining and Knowledge Discovery Handbook*. Ed. by O. Maimon and L. Rokach. Springer US, pp. 853–867.

Chen, Y. and P. Berkhin (2011). "Real-time bidding algorithms for performance-based display ad allocation". In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1307–1315.

Crone, S. F. and S. Finlay (2012). "Instance sampling in credit scoring: An empirical study of sample size and balancing". In: *International Journal of Forecasting* 28.1, pp. 224–238.

Da Silva, A. F. (2010). "cudaBayesreg: Bayesian Computation in CUDA." In: *R Journal* 2.2, pp. 48–55.

Dinner, I., H. J. van Heerde, and S. A. Neslin (2014). "Driving Online and Offline Sales: The Cross-channel Effects of Traditional, Online Display, and Paid Search Advertising". In: *Journal of Marketing Research* 51.5, pp. 527–545.

Domingos, P. (1999). "MetaCost: A general method for making classifiers cost-sensitive". In: *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '99*, pp. 155–164.

Drummond, C. and R. C. Holte (2000). "Explicitly representing expected cost: An Alternative to ROC Representation". In: *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 198–207.

Duan, W. and J. Zhang (2014). "The Impact of Referral Channels in Online Customer Journey". In: *Proceedings of the 35th International Conference on Information Systems, Auckland, New Zealand*.

Elkan, C. (2001). "The foundations of cost-sensitive learning". In: *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pp. 973–978.

Funk, B. and N. A. Nabout (2016). "Cross-Channel Real-Time Response Analysis". In: *Programmatic Advertising: The Successful Transformation to Automated, Data-Driven Marketing in Real-Time*. Ed. by O. Busch. Springer International Publishing Switzerland, pp. 141–151.

Gal-Or, E., M. Gal-Or, J. H. May, and W. E. Spangler (2006). "Targeted Advertising Strategies on Television". In: *Management Science* 52.5, pp. 713–725.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2014). *Bayesian data analysis*. Vol. 3. Taylor & Francis.

Goldfarb, A. and C. E. Tucker (2011a). "Privacy Regulation and Online Advertising". In: *Management Science* 57.1, pp. 57–71.

Goldfarb, A. and C. E. Tucker (2011b). "Online Display Advertising: Targeting and Obtrusiveness". In: *Marketing Science* 30.3, pp. 389–404.

Hui, S. K., P. S. Fader, and E. T. Bradlow (2009). "Path Data in Marketing: An Integrative Framework and Prospectus for Model Building". In: *Marketing Science* 28.2, pp. 320–335.

Joo, M., K. C. Wilbur, B. Cowgill, and Y. Zhu (2014). "Television Advertising and Online Search". In: *Management Science* 60.1, pp. 56–73.

Kireyev, P., K. Pauwels, and S. Gupta (2015). "Do display ads influence search? Attribution and dynamics in online advertising". In: *International Journal of Research in Marketing, Available online 13 October 2015, In Press*.

Kitts, B., L. Wei, D. Au, A. Powter, and B. Burdick (2010). "Attribution of conversion events to multi-channel media". In: *Proceedings - IEEE International Conference on Data Mining, ICDM*, pp. 881–886.

Kruschke, J. K. (2010). "What to believe: Bayesian methods for data analysis." In: *Trends in cognitive sciences* 14.7, pp. 293–300.

Lee, K., B. Orten, A. Dasdan, and W. Li (2012). "Estimating conversion rate in display advertising from past erformance data". In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, Beijing, China*, pp. 768–776.

Lee, K. C., A. Jalali, and A. Dasdan (2013). "Real Time Bid Optimization with Smooth Budget Delivery in Online Advertising". In: *Proceedings of the Seventh International Workshop on Data Mining for Online Advertising, Chicago, IL, USA*, pp. 1–13.

Li, H. A. and P. K. Kannan (2014). "Attributing Conversions in a Multichannel Online Marketing Environment: An Empirical Model and a Field Experiment". In: *Journal of Marketing Research* 51.1, pp. 40–56.

Liaukonyte, J., T. Teixeira, and K. C. Wilbur (2015). "Television Advertising and Online Shopping". In: *Marketing Science* 34.3, pp. 311–330.

Lin, M., H. C. Lucas, and G. Shmueli (2013). "Too Big to Fail: Large Samples and the p-Value Problem". In: *Information Systems Research* 24.4, pp. 906–917.

Lu, S., Y. Zhu, and A. Dukes (2015). "Position Auctions with Budget-Constraints: Implications for Advertisers and Publishers". In: *Marketing Science* 34.6, pp. 897 –905.

Malhotra, N. K., S. S. Kim, and J. Agarwal (2004). "Internet Users' Information Privacy Concerns (IUIPC): The Construct, the Scale, and a Causal Model". In: *Information Systems Research* 15.4, pp. 336–355.

Manchanda, P., J.-P. Dubé, K.-Y. Goh, and P. K. Chintagunta (2006). "The effect of banner advertising on internet purchasing". In: *Journal of Marketing Research* 43.1, pp. 98–108.

Meek, C., B. Thiesson, and D. Heckerman (2002). "The Learning-Curve Sampling Method Applied to Model-Based Clustering". In: *Journal of Machine Learning Research* 2, pp. 397–418.

Montgomery, A. L., S. Li, K. Srinivasan, and J. C. Liechty (2004). "Modeling Online Browsing and Path Analysis Using Clickstream Data". In: *Marketing Science* 23.4, pp. 579–595.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. Cambridge, Massachusetts: The MIT Press.

Naik, P. and K. Peters (2009). "A Hierarchical Marketing Communications Model of Online and Offline Media Synergies". In: *Journal of Interactive Marketing* 23.4, pp. 288–299.

Nottorf, F. and B. Funk (2013). "The economic value of clickstream data from an advertiser's perspective". In: *Proceedings of the 21st European Conference of Information Systems, Utrecht, Belgium*.

Pandey, S., M. Aly, and A. Bagherjeiran (2011). "Learning to target: what works for behavioral targeting". In: *Proceedings of the 20th ACM international conference on Information and knowledge management, Glasgow, United Kingdom*, pp. 1805–1814.

Perlich, C., B. Dalessandro, R. Hook, O. Stitelman, T. Raeder, and F. Provost (2012). "Bid optimizing and inventory scoring in targeted online advertising". In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China*, pp. 804–812.

Piercy, N. (2012). "Positive and negative cross-channel shopping behaviour". In: *Marketing Intelligence & Planning* 30.1, pp. 83–104.

Plummer, M. (2003). "JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling". In: *Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna*.

Poel, D. van den and W. Buckinx (2005). "Predicting online purchasing behaviour". In: *European Journal of Operational Research* 166.2, pp. 557–575.

Ritzer, G. and N. Jurgenson (2010). "Production, Consumption, Prosumption: The nature of capitalism in the age of the digital prosumer". In: *Journal of Consumer Culture* 10.1, pp. 13–36.

Rossi, P. E., G. M. Allenby, and R. McCulloch (2006). *Bayesian statistics and marketing*. Wiley, pp. 136–142.

Shao, X. and L. Li (2011). "Data-driven multi-touch attribution models". In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 258–264.

Shmueli, G. and O. R. Koppius (2011). "Predictive analytics in information systems research". In: *MIS Quarterly* 35.3, pp. 553–572.

Stange, M. (2015). "The Impact of TV Ads on the Individual User's Purchasing Behavior". In: *Proceedings of the 36th International Conference on Information Systems, Fort Worth, United States*.

Stange, M. and B. Funk (2014). "Real-Time Advertising". In: *Business & Information Systems Engineering* 56.5, pp. 305–308.

Stange, M. and B. Funk (2015). "How Much Tracking Is Necessary - The Learning Curve in Bayesian User Journey Analysis". In: *Proceedings of the 23rd European Conference on Information Systems, Muenster, Germany*.

Stange, M. and B. Funk (2016a). "How Big Does Big Data Need To Be?" In: *Enterprise Big Data Engineering, Analytics, and Management*. Ed. by M. Atzmueller, S. Oussena, and T. Roth-Berghofer. IGI Global. Chap. 1, pp. 1–15.

Stange, M. and B. Funk (2016b). "Predicting Online User Behavior Based on Real-Time Advertising Data". In: *Proceedings of the 24th European Conference on Information Systems, Istanbul, Turkey*.

Statista (2016a). *Distribution of media and marketing spending worldwide in 2015, by medium/target; http://www.statista.com/statistics/268975/worldwide-advertising - expenditure-since-2008/ (retrieved on 24 May 2016)*.

Statista (2016b). *Number of Smartphone users worldwide from 2014 to 2019; http:// www.statista.com/statistics/435252/ adblock-users-worldwide/ (retrieved on 7 May 2016)*.

Statista (2016c). *Programmatic vs. non-progrmmatic share of digital advertising spending in the United States from 2015 to 2020; http://www.statista.com/statistics/271499/ forecast-for-the-market-volume-of-online-advertising-in-the-us/ (retrieved on 7 May 2016)*.

Waller, M. A. and S. E. Fawcett (2013). "Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management". In: *Journal of Business Logistics* 34.2, pp. 77–84.

Wilkinson, D. J. (2005). "Parallel Bayesian Computation". In: *Handbook of Parallel Computing and Statistics*. Chapman and Hall/CRC; 1 edition, pp. 477–509.

Wu, W. C.-H., M.-Y. Yeh, and M.-S. Chen (2015). "Predicting Winning Price in Real Time Bidding with Censored Data". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*. New York, USA, pp. 1305–1314.

Xu, L., J. A. Duan, and A. Whinston (2014). "Path to Purchase: A Mutually Exciting Point Process Model for Online Advertising and Conversion". In: *Management Science* 60.6, pp. 1392–1412.

Yang, S. and A. Ghose (2010). "Analyzing the Relationship Between Organic and Sponsored Search Advertising: Positive, Negative, or Zero Interdependence?" In: *Marketing Science* 29.4, pp. 602–623.

Yuan, S., J. Wang, and X. Zhao (2013). "Real-time bidding for online advertising: measurement and analysis". In: *Proceedings of the Seventh International Workshop on Data Mining for Online Advertising, Chicago, IL, USA*.

Zhang, W., S. Yuan, and J. Wang (2014). "Optimal real-time bidding for display advertising". In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining New York, NY, USA*.

2 Real-Time Advertising

Stange, M. and B. Funk (2014). "Real-Time Advertising". In: Business & Information Systems Engineering 56.5, pp. 305–308.

2.1 Data-Driven Online Advertising

Over the last two decades, online advertising has become one of the most important elements of corporate communications. Whereas static banner ads dominated initially, search advertising (Varian, 2007) now encompasses the largest part of global online advertising spending. In recent years, a new form of online advertising, real-time advertising[1] (RTA), has been increasingly used. RTA is based on auctions in which individual advertising spaces are sold within a few milliseconds after calling a website. Advertisers or their media agencies participate in these auctions. RTA was first established in the U.S. and is now represented in the German market, with a market share of approximately 10 % (BVDW, 2013). RTA will progressively replace the traditional forms of purchasing online advertising space, which nowadays is being sold in large quotas and at predetermined prices that are mediated by marketers and media agencies. Thus, from the perspective of publishers, RTA allows both a reduction of transaction costs and an increase in revenues due to higher utilization. Advertisers can target ads based on the product affinity of user groups, which allows for the optimization of advertising campaigns in a short period of time (Ghosh et al., 2009). In the context of RTA, information systems research can contribute to a number of research topics due to its interdisciplinary orientation. For example, research topics include the automated decision support within the auction process, management of the large quantities of data, perceptions of RTA by various actors, and development of sustainable digital business models. The present article describes RTA from the perspective of stakeholders and outlines selected research questions.

2.2 Actors and the RTA Process

In addition to Internet users, publishers and advertisers, additional technology service providers are involved in the RTA process (Figure 2.1). These service providers include supply-side platforms (SSPs) that are commissioned by the publishers to offer advertising space in the market places. These market places are known as ad exchanges. The advertising spaces are auctioned on behalf

[1] In practice, the terms "real-time bidding", "programmatic buying", and "data-driven display advertising" are used synonymously with RTA.

of the advertisers by demand-side platforms (DSPs). Other actors include data management platforms (DMPs), which offer individual user profiles and interest data to support decisions in the auctions that are handled by the DSPs.

The RTA process starts when a website that contains RTA ads is displayed in the browser (Figure 2.1, Browser). The returned HTML code causes the browser to send an HTTP request to the SSP; the request is similar to the classic banner display advertising. The SSP (Figure 2.1, SSP) recognizes the size and position of the advertising space and adds additional information to prepare for the auction (e.g., minimum price, allowed and excluded forms of advertising content, the environments characterization). While the browser waits for an answer, the SSP sends the data to the ad exchange to trigger an auction (Figure 2.1, Ad Exchange).

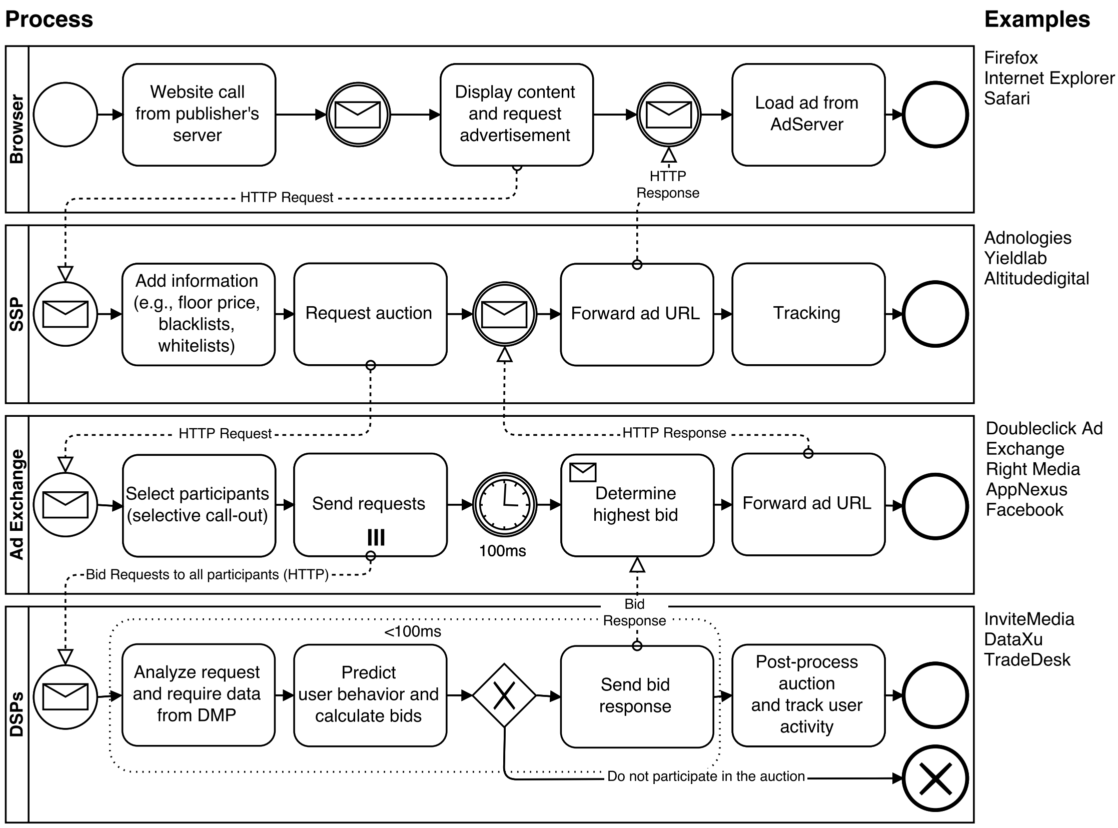

FIGURE 2.1: The business process model of RTA and examples for the involved actors.

The ad exchange does not forward the request to all of the connected DSPs. Rather, the participating DSPs are selected based on the information that is contained in the request and on findings from previous auctions. This method of selection reduces the data volume that needs to be transferred and processed by the different actors. Afterwards, the ad exchange forwards the information about the advertising space to the selected DSPs in the form of a bid request. The bid request is usually formatted in JavaScript Object Notation and includes information about the user, context, and advertising space (Figure 2.2).

User information	
Request ID	`id: "Mv\2005\n\345\177"`
Encoded IP address of the user	`ip: "\314j\310"`
Operating system and device	`user_agent: "Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US) AppleWebKit/534.13"`
User ID	`google_user_id: "CAESvb-4SLDjMqsY9"`
Time zone of the user	`timezone_offset: -300`
Cookie age	`cookie_age_seconds: 7685804`
Origin of the user	`region: "US-MA"; city: "Boston"`
Context information	
URL of the website	`url: http://www.example.com/`
Language of the website	`detected_language: "en"`
Detected website content and weights	`detected_vertical { id: 22 weight: 0.67789277 }`
Ad slot specification	
Ad slot ID	`id: 1;`
Dimensions	`width: 300; height: 250`
Unallowed content	`excluded_attribute: 7`
Floor price	`matching_ad_data { adgroup_id: 3254984134 minimum_cpm_micros:2000 }`
Visibility of the ad slot	`slot_visibility: BELOW_THE_FOLD`

FIGURE 2.2: Example of a bid request (Google, 2014).

Within approximately 20 milliseconds, the DSPs must decide on the advertiser and campaign for which to place a bid (Figure 2.1, DSPs). To make this decision, all of the available data about the current user can be considered. When a user is visiting the website for the first time, the available information is limited to the bid request. Later, the user can be detected by cookie matching. This allows the cookies of the SSP/ad exchange to be linked to the user IDs of the advertiser. The users activities (e.g., ad contacts, website visits, and purchasing activity) can be stored in a customer journey and used for automated decisions during the bidding process. Additionally, DSPs can use third-party data (e.g., socio-demographic characteristics, user interests) that are offered by DMPs to support the decisions within the auction process.

Regarding auction participation, the DSPs send a bid response within a pre-determined time interval. The response contains the bid, advertisement URL, target web page, and content information of the advertising material. The ad exchange selects the highest bidder and forwards the advertising media to the SSP, which then forwards the media to the waiting browser.

Because second-price auctions are used in RTA, the URL of the ad contains the paid price in an encrypted form that can be decrypted by the DSP when the banner is loaded from the ad server. Ad servers are commonly used for user activity tracking and provide extensive tracking capabilities. On the side of the DSPs, these data can be added to the customer journey and later used to support decisions in the auction process.

In practice, different variants of the illustrated RTA process can be found. For example, Facebook operates its own ad exchange, where only advertising spaces from Facebook are traded. Hence, the differentiation among publisher, SSP, and ad exchange is absent.

2.3 Research Topics and Contributions of IS Research

In the following paragraphs, specific IS research topics are presented. Because of the wide array of topics, a systematic assessment of RTA-related research topics is not within the scope of this article.

2.3.1 Decision Support

In RTA, all of the involved actors make numerous decisions. For example, actors must decide which areas of a website – and under what conditions and with which intermediaries – should be offered to a specific selection of advertisers (Balseiro et al., 2014). IS research can contribute to the dynamic selection of suppliers, i.e., the intermediaries and the technical infrastructure (Probst and Buhl, 2012).

Previous research primarily focused on the bidding process with its millions of individual auctions. In addition to budget constraints, campaign periods, and the maximum frequency of impressions per user and bid request, the customer journey is used to determine the price a DSP should bid. Statistical models (Nottorf, 2014) enable companies to predict user behavior (e.g., the probability of a purchase). Essentially, conditional probabilities in the following form are derived: "The impression of an RTA display ad at the present time for user X with profile attributes Y and customer journey Z increases the probability of a purchase within a given time interval by A%". With the help of these conditional probabilities, the economic assessment of the potential RTA ad and the calculation of the maximum bid are possible (Perlich et al., 2012). Today, it is common practice, however, that bids are based on the available budget rather than being determined by the user or context. This is especially true for branding campaigns, where the measurement of success is more difficult to calculate than in performance-oriented campaigns.

Research in this context offers several methodologically sophisticated questions that are highly relevant in practice. For example, how can decision-making consider user-specific and context-specific factors? How can heterogeneity in user behavior be modeled? What interactions between RTA and other marketing channels in the customer journey can be measured? What are the dynamics of user behavior, and how do these dynamics impact the predictive power of the models?

2.3.2 Data Management

RTA produces large amounts of data. For example, Googles DoubleClick Ad Exchange processes tens of thousands of bid requests per second and forwards them to the connected DSPs. For the DSPs, this process results in data volumes that exceed one TByte in a few days, which quickly leads to data volumes in the order of PBytes. If DSPs and advertisers keep these data for later analysis, they face high costs of data storage and processing. Various research questions can address the ratio of the cost of data collection, storage, and processing to the benefit of data analyses, and research can consider how this ratio is used in RTA in terms of the value of the data (Nottorf and Funk, 2013). This involves questions regarding the optimal amount of data needed per user and whether an aggregation of user leads to a reduction in the amount of data. In practical terms, there is the question of whether the handling of data is economically sensible in terms of decisions that need to be made within the auctions. For these decisions, there is the question of how to sample from the data prior to the analysis (i.e., model estimation). This is important because statistical methods can be computationally intensive (e.g., simulation-based approaches to model estimation); thus, in addition to the cost of data storage, the cost of the computing capacity must be considered when estimating the total costs of data management. The answer to these questions depends on many factors, such as the dynamics of user behavior, seasonal effects and competition.

In todays practices, systematic decisions about which data will be collected, stored, and processed are rare. Instead, all of the available data are collected but primarily remain unanalyzed; thus, RTA is a prime example of Big Data. Research in IS can estimate the economic value of the data and develop processes that can be used to manage the data. These contributions are especially important in the context of the increasing use of cloud services.

2.3.3 User Perceptions

Online advertising is an essential financing strategy of many websites, whereas user fees are not always enforceable because of users Internet experience and competition. RTA allows to analyze user behavior and target users individually. With regard to banner advertising in general and, thus, RTA in particular, Goldfarb and Tucker (2011) have shown that the personalization of advertising messages has a positive effect on sales. In the U.S., this study has also shown that privacy concerns of users are connected with obtrusive, personalized ads. This aspect of advertising has not been well studied, and there may be long-term or negative effects. Therefore, sensitive and transparent handling of user-specific profiles is advisable, particularly in Europe. The study of user behavior (e.g., cookie acceptance, use of ad blockers, acceptance of fingerprinting) and users willingness to disclose personal data represents an exciting IS research field (Carrascal et al., 2013). With regard to emerging legal restrictions, there is uncertainty about whether current business models of RTA actors will remain over time and about what changes are necessary and sustainable.

2.4 Implications

The auction principle of RTA will become more prominent in marketing in coming years. Existing business models will continue to evolve, and at the same time, new business models will be developed. Changes in data protection legislation in Europe and changes in user behavior, as well as the growing convergence of media channels, are expected to play an important role in this context. Furthermore, the quality of available advertising space is expected to continue to improve, which means that RTA can also be used for branding campaigns. DMPs will become increasingly important because automated decisions rely on high-quality data. In the era of online video stores and streaming services, the use of RTA as personalized advertising will likely increase in the classic online arena and even in radio and television. Several companies (e.g., http://wywy.com) already rely on this trend. The RTA process also shows certain parallels to the products in financial markets. However, whether other forms of selling media will evolve in RTA, such as in analogy of financial futures, depends on the development of appropriate business models (Veit et al., 2014).

In summary, RTA offers an exciting, interdisciplinary field for IS research due to its high degree of practical relevance and is characterized by a range of economic, methodological, technical, and social issues.

2.5 References

Balseiro, S. R., J. Feldman, V. Mirrokni, and S. Muthukrishnan (2014). "Yield optimization of display advertising with ad exchange". In: *Management Science* 60.12, pp. 2886–2907.

BVDW (2013). *Real Time Advertising Kompass 2013/2014*. Tech. rep. Düsseldorf: Bundesverband Digitale Wirtschaft e.V.

Carrascal, J. P., C. Riederer, V. Erramilli, M. Cherubini, and R. de Oliveira (2013). "Your browsing behavior for a big mac: Economics of personal information online". In: *Proceedings of the 22nd international conference on World Wide Web*, pp. 189–200.

Ghosh, A., P. McAfee, K. Papineni, and S. Vassilvitskii (2009). "Bidding for Representative Allocations for Display Advertising". In: *Lecture Notes in Computer Science* 5929, pp. 208–219.

Goldfarb, A. and C. E. Tucker (2011). "Online Display Advertising: Targeting and Obtrusiveness". In: *Marketing Science* 30.3, pp. 389–404.

Google (2014). *Real Time Bidding Protokoll*.

Nottorf, F. (2014). "Modeling the clickstream across multiple online advertising channels using a binary logit with bayesian mixture of normals". In: *Electronic Commerce Research and Applications* 13.1, pp. 45–55.

Nottorf, F. and B. Funk (2013). "The economic value of clickstream data from an advertiser's perspective". In: *Proceedings of the 21st European Conference of Information Systems, Utrecht, Belgium*.

Perlich, C., B. Dalessandro, R. Hook, O. Stitelman, T. Raeder, and F. Provost (2012). "Bid optimizing and inventory scoring in targeted online advertising". In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China*, pp. 804–812.

Probst, F. and H. U. Buhl (2012). "Lieferantenportfolio-Management für IT- Services unter Berücksichtigung von Diversifikationseffekten". In: *Wirtschaftsinformatik* 54.2, pp. 69–81.

Varian, H. R. (2007). "Position auctions". In: *International Journal of Industrial Organization* 25.6, pp. 1163–1178.

Veit, D., E. Clemons, A. Benlian, P. Buxmann, T. Hess, D. Kundisch, J. M. Leimeister, P. Loos, and M. Spann (2014). "Business Models". In: *Business & Information Systems Engineering* 6.1, pp. 45–53.

3 How Much Tracking Is Necessary?

The Learning Curve in Bayesian User Journey Analysis

Stange, M. and B. Funk (2015). "How Much Tracking Is Necessary – The Learning Curve in Bayesian User Journey Analysis". In: Proceedings of the 23rd European Conference on Information Systems, Muenster, Germany.

Abstract

Extracting value from big data is one of todays business challenges. In online marketing, for instance, advertisers use high volume clickstream data to increase the efficiency of their campaigns. To prevent collecting, storing, and processing of irrelevant data, it is crucial to determine how much data to analyze to achieve acceptable model performance. We propose a general procedure that employs the learning curve sampling method to determine the optimal sample size with respect to cost/benefit considerations. Applied in two case studies, we model the users' click behavior based on clickstream data and offline channel data. We observe saturation effects of the predictive accuracy when the sample size is increased and, thus, demonstrate that advertisers only have to analyze a very small subset of the full dataset to obtain an acceptable predictive accuracy and to optimize profits from advertising activities. In both case studies we observe that a random intercept logistic model outperforms a non-hierarchical model in terms of predictive accuracy. Given the high infrastructure costs and the users' growing awareness for tracking activities, our results have managerial implications for companies in the online marketing field.

3.1 Introduction

Online advertising produces large data sets. For instance, consider the amount of data that is produced in a real-time advertising (RTA) setting for a specific advertiser (Stange and Funk, 2014): On a publishers website, each touch point for each user generates a bid request to all potential advertisers. Assume 10,000 auctions per second on an ad exchange, such as AppNexus, and approximately 500 bytes per auction generates 400 Gbytes per day of data for an advertiser. Advertisers that collect and store these types of messages from several ad exchanges for future analyses rapidly acquire Tbytes or Pbytes of data, which are associated with significant costs. Considering these costs, advertisers should carefully assess payoffs from related analyses. Another challenge in online marketing is the velocity and variability of data due to variable consumer behavior, competitive dynamics and varying customer requirements. To provide better insight into the implication of these factors, we consider the following RTA setting: Demand-side platforms place bids on behalf of their customers (i.e., agencies and advertisers) as a response to incoming bid requests (Lee et al., 2013). Their bids must consider the customers campaign goals and budget and the success probability (i.e., click or conversion) of the individual user. This task requires the collection of information about the user, e.g., the user journey or demographic data. The data are used to predict a click or conversion probability based on classifiers (e.g., the logistic regression model). However, multiple external and internal factors enable users to change their behaviors over time (Bucklin and Sismeiro, 2003). In addition, campaign goals and budget constraints may change over time. Thus, a well-performing model may no longer be appropriate for predicting future user decisions. Frequent model updates based on new data are required, which is related to the cost of data collection, data storage and data preparation.

The models that have been proposed to describe user behavior under the influence of online advertising increasingly employ Bayesian data analysis and Markov Chain Monte Carlo (MCMC) estimation techniques (Bucklin and Sismeiro, 2009; Chatterjee et al., 2003; Nottorf and Funk, 2013). Although Bayesian data analysis supports high flexibility in model building, it is computationally demanding (Lee et al., 2012). The need for a speed up of these methods is demonstrated by researches who investigate opportunities to parallelize the underlying algorithms (Da Silva, 2010; Wilkinson, 2005). Despite existing cloud offerings, the computational power required to estimate these models requires significant costs (Deelman et al., 2008). Thus, there is a trade-off between the predictive accuracy of a model and the related computational cost of the parameter estimation.

In this paper, we propose a process to minimize computational costs by minimizing the amount of data required for the analysis. This process helps to determine the optimal sample size for a data analysis using the learning curve sampling method. The proposed process is a general process that can be applied to many types of data analysis in classification and regression problems. In this manner, we contribute not only to the field of online marketing and privacy on the Internet, but also provide a guideline for practitioners and researches in other areas of predictive analytics based on big data. Using two case studies, we apply this process to the user journey analysis of two German online retailers and demonstrate that the optimal sample size is far less than the total amount of available data. Thus, data collection, storing and processing efforts and costs can be significantly reduced.

The paper is structured as follows: First, we review related studies of the learning curve sampling method, model performance and clickstream data analysis. Second, we describe the general approach to determining the optimal sample size, which consists of four different steps. Last, we apply this approach to our empirical data sets, discuss our results and derive managerial implications.

3.2 Related Studies

Our study is based on two research topics: The first topic is the learning curve sampling method, which represents the observation that the predictive accuracy of a model increases as a function of the amount of processed data (Meek et al., 2002). The second topic is clickstream data analysis, which is frequently employed in online marketing research to model and predict user decisions on the Internet.

3.2.1 Learning Curve

Model performance as a function of sample size is a frequently discussed topic in publications of the medical or sociology fields (Brutti et al., 2009; Sahu and Smith, 2006; Santis, 2007). The costs associated with data collection in these fields are relatively high compared with the costs associated with data collection in the online marketing field. However, articles about the effect of sample size on the predictive power of models in the online advertising domain are not available. A general approach for obtaining an appropriate sample size is the learning curve sampling method. This method is driven by the observation that an increase in the sample size reduces the uncertainty in the parameter estimates of the learned model (Gu et al., 2001; Meek et al., 2002). Meek et al. (2002) formalized this approach by introducing a stopping criterion, which is based on the following two assumptions: First, the computational effort increases as a function of sample size, which is related to cost. Second, reduced uncertainty in the parameter estimates is related to benefit. Thus, by increasing the sample size and iteratively evaluating model performance, an optimum in the utility can be obtained. In this study, we employ this sampling method to obtain the optimal sample sizes in clickstream data analyses.

A common method for measuring the predictive accuracy is to integrate the receiver operator characteristic (ROC) curve to obtain the area under the curve (AUC; Bradley, 1997). The AUC represents the probability that a randomly chosen unknown object is correctly classified. In our case, we employ different logistic regression models, which we use to determine the posterior predictive densities of conversion probabilities for unknown users. We show that the AUC converges to a maximum value when the sample size is increased.

Numerous methods for measuring model quality exist. One of these methods focuses on the length of the highest density interval (HDI) of the estimated parameters (Joseph et al., 1995). This average length criterion (ALC) converges to a minimum value when the sample size is increased, as shown for simulated data (Wang and Gelfand, 2002). In our paper, we show that this case is also valid for clickstream data.

No published studies of online marketing employ the learning curve sampling method to determine the minimum sample size required to appropriately compute a model. The clickstream data literature neither provides an analytical comparison of the sample size and the predictive accuracy nor an indication how many user journeys they used and why. We use the learning curve sampling method to determine an optimal sample size that represents the best balance between computational costs and predictive accuracy and thereby identify the concrete number of needed user journeys. In addition, the paper contributes to the big data research field due to the general applicability of the proposed process described in Section 3.3.

Bayesian models and MCMC methods provide high flexibility in model building and estimation. The strength of these models is the ability to sample from a variety of distributions in combination with a hierarchical model structure. In the context of customer journey analysis they are feasible, because they allow it to determine variables such as decay rates from marketing activities and parameters for non-linear transformations of customer journey variables. On the other hand, they are computationally demanding. Some authors propose a different approach to minimize the computational cost related to these methods. They parallelize the computation on multiple processor cores (Da Silva, 2010; Henriksen et al., 2012; Jacob et al., 2011) to reduce computation times for a given amount of data. This method requires a deep understanding of the specific sampling algorithm and parallel programming languages, such as CUDA C. The process proposed in this paper, however, does not depend on a specific algorithm or method, but is generally applicable to arbitrary model structures.

3.2.2 Analysis of Clickstream Data

Clickstream data consists of data records produced by user interactions on the Internet. Each time a user is exposed to a display ad or searches for a brand-related keyword, an interaction is recorded that represents one entry in the clickstream data. Clickstream data are also referred to as user journey or customer journey data.

As part of the website usage mining discipline, investigations in clickstream data over the past ten years can be categorized into website usage and navigation, online shopping behavior and advertising on the Internet (Bucklin and Sismeiro, 2009). We focus on the latter. Chatterjee et al. (2003) developed a model to predict a users individual click proneness based on clickstream data. In their study, a random effects logit model was employed to predict a consumers response to banner advertisement. They concluded that a model that includes heterogeneity terms across sessions and users best describes the click behavior. Using this model, Nottorf and Funk (2013) show that advertisers can significantly reduce advertising costs if the advertisement is only exposed to users with the highest click probabilities. We use this approach to calculate the costs of the prediction. This finding has also been demonstrated by other authors using different methods, such as hypothesis tests (Klapdor, 2013) or higher-level Markov chains (Anderl et al., 2014).

Many authors point out the importance of cross channel marketing (e.g., Anderl et al., 2014; Klapdor, 2013). However, cross channel marketing is not limited to online channels. As shown by Joo et al. (2014), offline data from TV advertising spots can be used to predict users' online behavior. They find that the more brand-related TV spots are broadcasted, the more users search for these brands using search engines. Thus, the inclusion of offline data in user journey models, can result in significant improvements of the models' performance. However, modeling offline advertising effects is not as straight forward as modeling online advertising effects, since it is hardly possible to determine if a user in fact was exposed to the offline advertisement. In the first case study of this paper, we use TV spot data as an additional independent variable, which has not yet been done in published literature.

The studies about clickstream data often present descriptive statistics of the used samples. However, none of them provides a systematic comparison of different sample sizes and the resulting predictive accuracy. Thus, we contribute to this research field by providing the needed amount of user journeys.

3.3 Estimation of the Optimal Sample Size

We propose a four-step process to determine the optimal sample size. In this paper, we focus on the highlighted steps in Figure 3.1: (1) Select the initial sample size and sampling strategy, (2) determine the learning curve and predictive accuracy using the estimated parameters, (3) determine the cost of collecting, storing, and processing data, and (4) select the optimal sample size for repeated analysis.

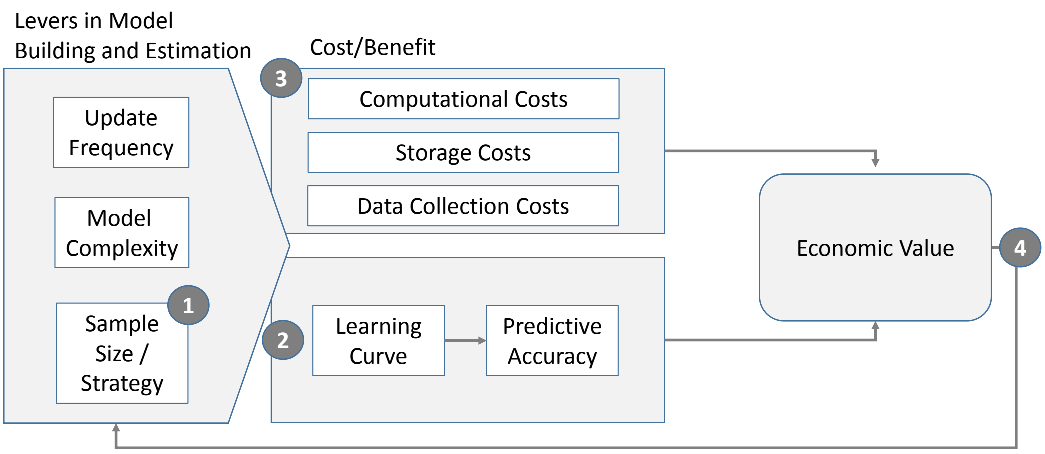

FIGURE 3.1: Levers in model building and estimation.

Beginning with a set of data, we select the sample size and sampling strategy. First, for a scenario of rare events, such as clicks or conversions in the online marketing context (Cho, 2003), the choice of sampling strategy is crucial for the estimation of the parameter values (King and Zeng, 2001). Stratified sampling is an appropriate method for estimating the parameters of the logit model (Falk et al., 2004). However, the total amount of collected data is higher for stratified sampling compared with a simple random sample. For example, in RTA, the amount of bid requests is greater than the amount of related clicks or conversions by orders of magnitude. However, there is no indication of a future click or conversion in the absence of a prediction mechanism. Thus, all bid requests should be stored and the stratification should be subsequently applied.

Second, the model parameters of interest are estimated. To evaluate the obtained parameters, the researcher can select between multiple methods to measure the power of the model and its estimated parameters. Although measures such as the ALC provide preliminary insights into the convergence of the parameters (Wang and Gelfand, 2002), these measures do not reveal information about the predictive accuracy of the model. Instead, the estimated parameters should be used to perform an out-of-sample test, which gives insights in the predictive accuracy and reduces the risk for over-fitting. According to previous learning curve studies (Gu et al., 2001; Meek et al., 2002), we know that the estimations for the parameters converge if additional data are considered. This also applies to the confusion matrices obtained from the out-of-sample tests, which can be used to calculate desired indicators to express the model performance, such as the accuracy, precision or the AUC. A confusion matrix is a 2×2 matrix that contains the number of true positive predictions, true negative predictions, false positive predictions and false negative predictions. The convergence of the parameter estimates thereby also determines the maximal utility for different sample sizes and the optimal sample size for a given modeling scenario, which we demonstrate in the next chapter.

Third, based on the results of the estimates from each sample size, the benefits of the prediction are estimated. These benefits can be evaluated by the element-wise multiplication of the cost matrix and the confusion matrix. Like the confusion matrix, the cost matrix is a 2×2 matrix that contains the costs for false predictions and the (negative) costs for true predictions, such as the costs for an advertisement or a lost contribution margin. In addition, the cost of data collection, data storage and data processing should be determined. For example, if in-house servers are provided for these purposes, the costs can be estimated based on the prices and the maintenance costs for these systems. If cloud services such as Amazon S3 are used, the costs are equivalent to the monthly fees for the services and are easier to identify (refer to Amazon, 2014 for exemplary calculations). Compared with online marketing, the cost of data collection may be higher in other fields. As a result, the dependency of the optimal sample size on the data collection costs has been substantially investigated (Brutti et al., 2009; Cohen, 1998).

Last, based on the results from steps 1 to 3, the optimal sample size is selected for the model estimation, i.e., the sample size for which the utility function is maximized. This optimal sample size is obtained when an additional set of data records does not increase the benefit for predictions on the validation set to an amount greater than the associated additional costs of data collecting and storage and computation time, as previously described.

Once the model is deployed and decisions are rendered based on its predictions, the data collection and storage procedures can be adjusted based on the results from step 1 to 4. If the risk of change in the data-generating mechanisms, such as unexpected changes in user behavior, is observed, the predictive accuracy of the deployed model should be monitored to rapidly address these changes. If necessary, the model should be estimated a second time to obtain updated parameters for prediction. If the model requires a complete revision, e.g., due to significantly modified external influences, steps 1 to 4 should be repeated to determine an updated optimal sample size.

3.4 Prediction of Conversion Probabilities

RTA enables advertisers to limit the exposure of ads to users who show a particular tendency to click on an ad (Perlich et al., 2012). As a prerequisite, these companies need to know the individual click and/or conversion probabilities. Our case studies demonstrate how these individual conversion probabilities are determined. We use user journey data from two German online retailers to estimate the model parameters and predict conversion probabilities $Pr(Conv = 1)$. In an RTA setting, our model can be used by a bidding agent. In a simple scenario, the bidding agent should only place a bid if the predicted probability for a conversion is higher than a previously determined threshold probability p_{thres}. Thus, the number of ineffective impressions and the marketing costs can be reduced. In this setting, we apply our previously described procedure to two data sets and three different Bayesian models.

3.4.1 Data Description and Preparation

We use two data sets from two different German online retailers, which we may not disclose here. Both data sets contain user tracking data from a period of one month (December 2013 and March 2013). Most prices of both retailers range from 10 to 100 EUR. The first data set is influenced by the Christmas-trade, which results in shorter user journeys due to spontaneous gift purchases.

Both retailers record each touch point for every user. A touch point may be an interaction on the retailers' websites or an interaction with an advertising channel, such as an organic search, search engine advertising (SEA) or banner advertising. For each touch point, the retailers record the user-id, the time stamp, and the type of interaction. The latter can be a click, an onsite activity or a conversion. The data set from the first case study also includes TV advertising spot data. From the TV data set we only use the time stamp to determine how many TV spots have been shown on television within the last 30 minutes before an interaction.

To clarify how the term user journey is used in this paper, consider the following example of a user journey: first, assume that a user clicks on an search engine ad (SEA). Second, after less than one hour, the user clicks a display ad. Third, after 6 hours, the user returns to the web site by clicking a display ad. Fourth and fifth, still in the current session, the user searches a specific product and returns trough the SEA channel and purchases the product. Last, after 2 hours, the user is exposed to additional display ads. In the next section, we describe how to translate this user journey into a design matrix (Table 3.1).

From the available subsets (80 and 30 million touch points for the first and second case study, respectively), we focus on user journeys with more than two interactions. We divide our subset into two parts of equal size to obtain a training set and a holdout sample. We decide not to use the stratified sampling strategy because the ratio of conversions is relatively high in our data sets (approximately 0.5%), which is sufficient for receiving robust estimations.

3.4.2 Model Description

Based on previous studies (e.g., Chatterjee et al., 2003), we know that the individual conversion probability is influenced by the user's intrinsic conversion proneness and the effects from within sessions and across sessions for each channel. Thus, in our model, each user's design matrix can be subdivided into three parts. First, the intercept terms I are used as covariates to estimate the users' intrinsic conversion proneness per channel. Second, to estimate delayed effects of the individual channels, we model the cumulated previous interactions within the sessions, which are denoted as X, and across sessions, which are denoted as Y. Third, we introduce the respective session number SN, the number of onsite contacts within the session $OCWS$ and in previous sessions $OCPS$, the number of conversions in previous sessions CPS and the intersession time IST as additional control variables, which resemble models from previous studies (e.g., Chatterjee et al., 2003). According to this notation, the user journey from the previous example would be modeled as demonstrated in Table 3.1.

Inter. no.	I_0	I_{SEA}	I_{Ban}	X_{SEA}	X_{Ban}	Y_{SEA}	Y_{Ban}	CPS	IST	SN	$Conv$
1	1	1	0	0	0	0	0	0	0 h	1	0
2	1	0	1	1	0	0	0	0	0 h	1	0
3	1	0	1	0	0	1	1	0	6 h	2	0
4	1	1	0	0	1	1	1	0	6 h	2	0
5	1	1	0	1	1	1	1	0	6 h	2	1
6	1	0	1	0	0	3	2	1	2 h	3	0

Table 3.1: Example of a user journey design matrix D_i. We leave out some of the described covariates, such as the onsite contacts within and across sessions, for convenience.

In addition to this notation, the number of TV Spots within the 30 minutes before the current interaction is denoted as TV. We transform the intersession time in hours and the amount of onsite contacts to the logarithmic scale due to the high variance of these values within the data. Equation 1 and 2 show the j^{th} interaction of the i^{th} user, which is represented as one row $(D_i)_j$ of the users design matrix D_i. Refer to the left hand side of Table 3.2 for the of the subscripts for I, X and Y. The additional covariates used in the design matrix are listed on the right hand side of Table 3.2.

$$\begin{aligned}\text{First case study:}(D_i)_j &= \{I_0, I_{SEO}, I_D, I_A, I_{Ban}, I_{SEA}, I_{EM}, I_R, \\ &\quad X_{SEO}, X_R, X_A, X_{SEA}, X_{EM}, \\ &\quad Y_{SEO}, Y_D, Y_A, Y_{Ban}, Y_{SEA}, Y_C, Y_{EM}, Y_R, \\ &\quad SN, IST, OCWS, OCPS, CPS, TV\} \qquad (3.1) \\ \text{Second case study: } (D_i)_j &= \{I_0, I_{SEO}, I_{SEA}, I_{Ban}, I_{PS}, I_A, I_{EM}, \\ &\quad X_{SEO}, X_{SEA}, X_{Ban}, X_{PS}, X_{EM}, \\ &\quad Y_{SEO}, Y_{SEA}, Y_{PS}, Y_{EM}, \\ &\quad SN, IST, OCWS, OCPS\} \qquad (3.2)\end{aligned}$$

Index	Channel	Index	Additional Variables
SEA	Search engine advertisement	OCWS	Onsite contacts within the curr. session
SEO	Organic search	OCPS	Sum of onsite contacts in prev. sessions
R	Referral from another website	SN	Session Number
A	Affiliate marketing	IST	Time between two sessions
Ban	Display advertisement	CPS	Number of conversions in prev. sessions
D	Direct type-in		
C	Cooperation link		
PS	Price search engine		
EM	Email advertisement		

TABLE 3.2: Indices and corresponding covariates used in the design matrices.

Every user is expected to exhibit a different proneness for conversions and clicks on ads, such as email ads, banner ads or search engine ads. As it was shown by Chatterjee et al. (2003) a model with random intercept and random slopes best fits the users' behavior. However, to predict future behavior of (unknown) users, it is sensible to create user clusters prior to model estimation and apply the same clustering method to new users to predict their conversion probabilities. In both case studies we use the time of the day, i.e., morning/afternoon and evening/night, to build up the two clusters C_1 and C_2. These clusters are feasible, because in both data sets the conversion rates during the day differ from the conversion rates at night, which implies different intercept terms for customers who visit the shops at different times of the day. For both case studies, we use three different Bayesian models for analysis: a simple logit model, a random intercept model and a random intercept/slope model. The models are presented in equations 3.3 through 3.5. In the following, m denotes the m^{th} cluster ($m \in \{1,2\}$) and n denotes the n^{th} interaction within the cluster. We use the non-hierarchical logit model from the `R` package `BayesLogit` (Windle et al., 2014) as shown in Equation 3.3. It allows only one set of β values. Therefore, the simple logit model is only feasible for a data set with little heterogeneity across users.

$$\begin{aligned} Conv_{mn} &\sim \text{Bernoulli}(\theta_{mn}) \\ \text{logit}(\theta_{mn}) &= X_{mn}\beta \end{aligned} \tag{3.3}$$

For the random intercept model we use the function `MCMChlogit` from the `R` package `MCMCpack` (Martin et al., 2011). The model is shown in Equation 3.4.

$$\begin{aligned} Conv_{mn} &\sim \text{Bernoulli}(\theta_{mn}) \\ \text{logit}(\theta_{mn}) &= X_{mn}\beta + I_0 b_m \end{aligned} \tag{3.4}$$

In this equation, b_m is a scalar value. It accounts for the different conversion rates for customers during the day and by night. For the random slope model we use the function `rhierMnlRwMixture` from the `R` package `rpud` (Yau, 2015), which is the parallelized version of the algorithm from the `R` package `bayesm` (Rossi and McCulloch, 2010). We simplify the model here for convenience to obtain the model shown in Equation 3.5.

$$\begin{aligned} Conv_{mn} &\sim \text{Bernoulli}(\theta_{mn}) \\ \text{logit}(\theta_{mn}) &= X_{mn}\beta_m \\ \beta_m &= \beta + \delta_m. \end{aligned} \tag{3.5}$$

A random slope model is feasible in the context of user journey analyses, because it accounts for individual marketing channel effects β_m for each cluster. For instance, the impact of TV advertisement on the conversion rate could differ between during the day and at night. In Equation 3.5, δ_m is a vector with the same length as β including the cluster specific intercept $\delta_m^{I^0}$. For all three models we use vague priors around 0 for the the parameters β and for the cluster parameters b_m and δ_m. Please refer to the above mentioned `R` packages for additional information about the MCMC samplers.

3.4.3 Results

To show the convergence of the predictive accuracy, we execute 6 analyses including 1,000, 2,000, 4,000, 8,000, 16,000 and 32,000 user journeys for each combination of model and case study, resulting in 36 analyses in total. The computation times of the individual analyses are presented in Table 3.3. The computation times of the random slope model are short in comparison with the random intercept model due to the GPU parallelization from the `rpud` package (Yau, 2015).

The results from the first case study are presented in Table 3.4. We report the results based on 32,000 user journeys for the simple logit model and the random intercept model. The results from the computation based on the other sample

Sample Size in 1,000	1	2	4	8	16	32	1	2	4	8	16	32
Simple Logit	14	28	56	108	214	446	11	22	44	87	173	348
Random Intercept	31	70	142	259	593	1,066	41	81	161	323	649	1,303
Random Slope	23	36	66	129	245	478	22	38	68	138	282	557

TABLE 3.3: Computation times in seconds for the first (left) and the second case study (right). The computation was performed on an Intel i7 4820K processor and a GeForce 770 GPU.

sizes and the random slope model are not reported due to space limitations[1]. However, the convergence of the variables is demonstrated in Figure 3.2 by example. We do not discuss the individual results in full detail here, but want to outline some of the major findings. We demonstrate that the impact from TV Spots results in a non-zero value for β_{TV}. However, we do not observe significant effects of TV spots in this case study. This topic should get more attention in future studies.

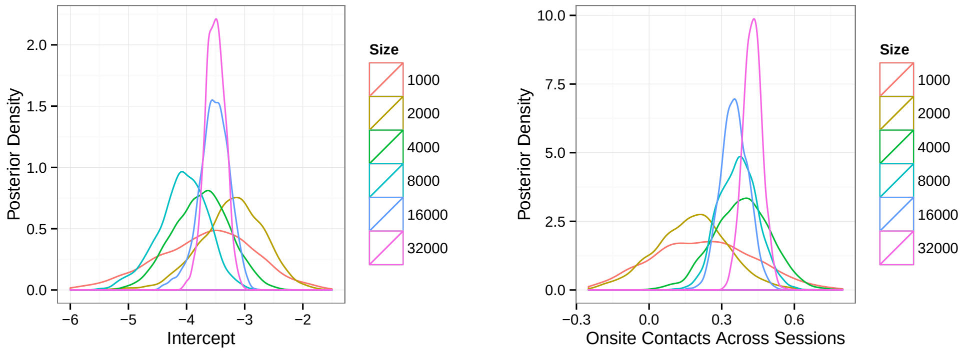

FIGURE 3.2: Density plots for β_{I^0} and β_{OCPS} from the simple logit model from the first case study. The HDI lengths decrease with the increase of the sample size.

The effects from within the sessions ($\beta_{X^{(\cdot)}}$) show that users who use several channels within one session are less likely to convert. The same is valid for the majority of the cross session effects. In summary, the more often a user visits the shop through different channels the less likely is the conversion. The negative impact from the intersession time β_{IST} indicates that the probability for a conversion decreases with the increase of time between the current and the next session. We suppose that these negative effects result from the relatively high amount of spontaneous customers with a short journey, which could be an effect of the Christmas-time. The highest positive effects result from the number of onsite contacts from within and across sessions. This finding is intuitive, meaning the more product pages a user visits within and across sessions, the more likely is the conversion. The values for b_1 and b_2, which can be interpreted as the probability offset for the two clusters C_1 and C_2, show that the tendency to purchase a product in the morning and afternoon is higher than in the evening and at night.

[1]Upon request we are pleased to provide the full result list.

Parameter	2.5%	50%	97.5%	2.5%	50%	97.5%
β_{I^0}	**-5.03**	**-3.80**	**-2.54**	**-3.86**	**-3.51**	**-3.19**
$\beta_{I^{SEO}}$	**0.40**	**0.67**	**1.02**	**0.42**	**0.75**	**1.11**
β_{I^D}	**1.52**	**1.77**	**2.06**	**1.45**	**1.77**	**2.10**
β_{I^A}	**1.41**	**1.63**	**1.99**	**1.33**	**1.65**	**2.00**
$\beta_{I^{Ban}}$	-0.82	-0.40	0.17	-0.80	-0.30	0.18
$\beta_{I^{SEA}}$	**0.71**	**0.94**	**1.29**	**0.69**	**0.99**	**1.32**
$\beta_{I^{EM}}$	**1.29**	**1.54**	**1.89**	**1.26**	**1.58**	**1.93**
β_{I^R}	**0.45**	**0.75**	**1.14**	**0.43**	**0.77**	**1.13**
$\beta_{X^{SEO}}$	**-0.64**	**-0.51**	**-0.42**	**-0.59**	**-0.46**	**-0.34**
β_{X^R}	-0.21	-0.09	0.05	-0.24	-0.10	0.04
β_{X^A}	-0.04	0.01	0.05	-0.04	0.01	0.06
$\beta_{X^{SEA}}$	**-0.47**	**-0.39**	**-0.33**	**-0.46**	**-0.39**	**-0.32**
$\beta_{X^{EM}}$	-0.18	-0.07	0.04	-0.21	-0.09	0.04
$\beta_{Y^{SEO}}$	**-0.22**	**-0.12**	**-0.05**	**-0.23**	**-0.12**	**-0.03**
β_{Y^D}	**-0.38**	**-0.28**	**-0.20**	**-0.38**	**-0.25**	**-0.14**
β_{Y^A}	**-0.29**	**-0.21**	**-0.12**	**-0.32**	**-0.19**	**-0.08**
$\beta_{Y^{Ban}}$	-0.25	-0.04	0.13	-0.25	-0.03	0.16
$\beta_{Y^{SEA}}$	**-0.15**	**-0.08**	**-0.02**	-0.14	-0.06	0.01
β_{Y^C}	-0.17	0.07	0.30	-0.13	0.12	0.34
$\beta_{Y^{EM}}$	-0.17	-0.07	0.01	-0.17	-0.06	0.05
β_{Y^R}	**-0.73**	**-0.54**	**-0.42**	**-0.80**	**-0.51**	**-0.27**
β_{SN}	**-0.47**	**-0.36**	**-0.27**	**-0.49**	**-0.37**	**-0.26**
β_{IST}	**-0.10**	**-0.08**	**-0.06**	**-0.10**	**-0.08**	**-0.05**
β_{OCWS}	**0.34**	**0.39**	**0.44**	**0.32**	**0.37**	**0.43**
β_{OCPS}	**0.36**	**0.46**	**0.52**	**0.35**	**0.43**	**0.51**
β_{CPS}	**-0.60**	**-0.30**	**0.00**	**-0.58**	**-0.29**	**-0.02**
β_{TV}	-0.02	0.02	0.08	-0.01	0.04	0.09
b_1	-1.11	0.19	1.38	-	-	-
b_2	-1.47	-0.14	1.02	-	-	-

Table 3.4: Results of the random intercept model (left) and the simple logit model (right) from the first case study (Sample Size = 32,000). Significant values are printed in boldface.

All the HDI lengths for the individual β values converge with the increase of the sample size. This underlines the convergence of the predictive accuracy as demonstrated in the next section. We present the convergence of two variables in Figure 3.2.

The results from the second case study are presented in Table 3.5. They are mainly consistent with the results from the first case study in terms of the effects of the individual channels and the additional variables, such as the session number and the number of onsite activities across sessions and the cluster variables b_1 and b_2. In contrast to the first case study, the intercept terms $\beta_{I^{(\cdot)}}$ are mainly negative, except the affiliate channel. In summary, customers do not often converge spontaneously, but frequently use the information from search engines and third party websites to make their choices. The positive onsite variables β_{OCWS} and β_{OCPS} show the importance of the onsite session length for the purchase decision.

Parameter	2.5%	50%	97.5%	2.5%	50%	97.5%
β_{I0}	**-4.55**	**-4.05**	**-3.54**	**-4.46**	**-4.05**	**-3.69**
β_{ISEO}	**-0.50**	**-0.33**	**-0.14**	-0.54	-0.18	0.22
β_{ISEA}	**-0.72**	**-0.54**	**-0.34**	-0.76	-0.41	0.00
β_{IBan}	**-1.54**	**-1.14**	**-0.67**	**-1.87**	**-1.23**	**-0.57**
β_{IPS}	-0.30	-0.15	0.10	-0.41	0.01	0.44
β_{IA}	**0.22**	**0.49**	**0.73**	**0.16**	**0.62**	**1.09**
β_{IEM}	**-0.48**	**-0.33**	**-0.14**	-0.50	-0.09	0.35
β_{XSEO}	**-0.18**	**-0.16**	**-0.12**	**-0.31**	**-0.20**	**-0.11**
β_{XSEA}	**-0.15**	**-0.11**	**-0.05**	**-0.19**	**-0.12**	**-0.05**
β_{XBan}	-0.01	0.13	0.34	-0.08	0.17	0.36
β_{XPS}	**0.01**	**0.06**	**0.14**	-0.01	0.07	0.14
β_{XEM}	-0.10	-0.02	0.03	-0.17	-0.07	0.01
β_{YSEO}	**-0.15**	**-0.11**	**-0.07**	**-0.19**	**-0.10**	**-0.03**
β_{YSEA}	**-0.29**	**-0.23**	**-0.14**	**-0.19**	**-0.11**	**-0.03**
β_{YPS}	**-0.23**	**-0.13**	**-0.07**	-0.16	-0.04	0.06
β_{YEM}	-0.35	-0.16	0.07	**-0.39**	**-0.23**	**-0.11**
β_{SN}	**-0.25**	**-0.18**	**-0.13**	**-0.31**	**-0.20**	**-0.10**
β_{IST}	-0.03	-0.01	0.01	-0.03	-0.01	0.02
β_{OCWS}	**0.31**	**0.37**	**0.39**	**0.31**	**0.37**	**0.44**
β_{OCPS}	**0.48**	**0.55**	**0.60**	**0.46**	**0.54**	**0.62**
b_1	-0.35	0.10	0.55	-	-	-
b_2	-0.56	-0.11	0.33	-	-	-

TABLE 3.5: Results of the random intercept model (left) and the simple logit model (right) from the second case study (Sample Size = 32,000). Significant values are printed in boldface.

Like in the first case study, we observe convergence of the β values with increasing sample size. The density plots for β_{I0} and β_{OCPS} are shown in Figure 3.3.

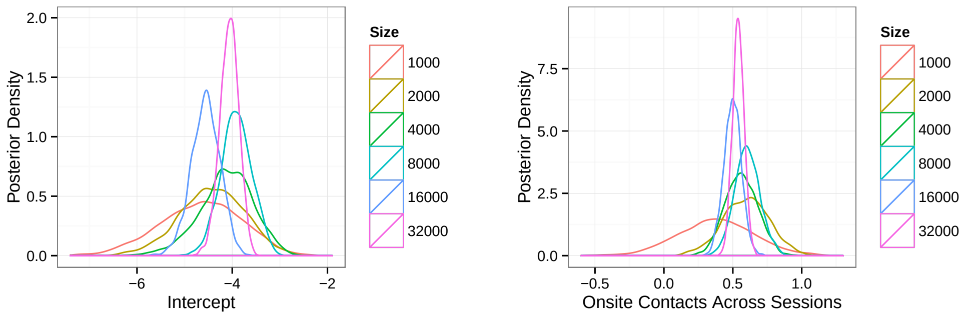

FIGURE 3.3: Density plots of β_{I0} and β_{OCPS} from the simple logit model from the second case study. As in the first case study, the HDI lengths decrease with the increase of the sample size.

3.4.4 Prediction and Benefits

Following the proposed process, we run four out-of-sample tests for each sample size and model. We use the parameter estimates to predict the conversions of each individual user from the holdout sample. We sample from the posterior distribution and use the logit link function to obtain values for $Pr(Conv = 1)$ for each user at each interaction from the holdout sample.

For dichotomous data with rare events, the predicted probabilities should not be misunderstood as true probabilities because the probabilities are typically underestimated (King and Zeng, 2001). Therefore, we introduce a threshold value for the posterior probability p_{thres}, for which the prediction is 0 for all interactions below this value and 1 for all interactions equal to or above this value. For each sample size and model, we iteratively increase the threshold value from 0 to 1 and compare the predicted outputs with the actual user decisions. In this manner, we obtain the confusion matrix for each iteration. These matrices are used to draw the ROC and calculate the AUC.

The left hand side of Figure 3.4 shows the AUC for each sample size for the first case study. Since we split off the holdout sample into four equally sized sets, we are able to report the standard deviations of the AUCs, which is an indicator for the variance in the predictive accuracy. The random intercept model outperforms the simple logit model for all sample sizes (neglecting the standard deviation), because it accounts for the different conversion rates over the day. The random slope model shows lower AUC values than the other models, except for 32,000 user journeys. This shows that a more complex model needs more data to achieve a desired predictive accuracy. We expect, that if even more data would be used for the estimation, the random slope model would outperform the random intercept model, like observed by Chatterjee et al. (2003). The analyses based on sample sizes of $> 8{,}000$ result in AUC values that are close to each other. This result is consistent with previous studies (e.g., Meek et al., 2002), i.e., the predictive accuracy converges when the sampling size is increased.

To calculate the costs of the prediction, we assume typical costs for impressions in the RTA industry. The benefits are given by the difference between the costs that incur by applying the model and the maximum costs, i.e., all interactions are classified as positive and the ad is always shown ($p_{thres} = 0$). For true positive predictions, we define a benefit of 0.15-0.01 EUR (i.e., the contribution margin for a click minus the cost of the impression). For incorrect negative predictions, we assume a loss of 0.15 EUR (the lost contribution margin) and for incorrect positive predictions, we assume a cost of 0.01 EUR for exposing the advertisement without success and we assign no cost to true negatives. We calculate the costs of the prediction for each threshold value p_{thres}. The threshold values that determine the minimum costs are relatively low ($p_{thres} \approx 0.05$), because of the low conversion rate and the relatively high ratio of the contribution margin (CM) and the cost (C) for an impression. The minimum costs, the benefit per decision and the threshold value are highly dependent on this ratio. When the ratio increases, the threshold value converges to zero, i.e., all interactions will be predicted as positive. Therefore, the classifier is only useful in a certain range of CM/C. If the ratio is higher than the maximum value from this range, the classifier will always predict a conversion, and if it is lower, the classifier will never predict a conversion. The ratio of $15/1$ lies within the allowed range for both of our case studies.

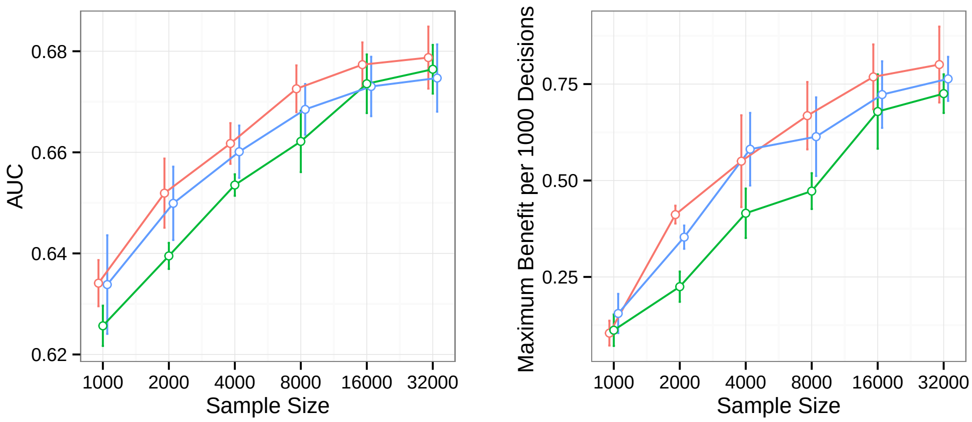

FIGURE 3.4: Results from the first case study: Area under the curve and the maximum benefit per 1,000 decisions of the simple logit model (blue), random intercept model (red) and the random slope model (green). Note the logarithmic scale on the abscissa.

For convenience, we report the benefit per 1,000 decisions. The right hand side of Figure 3.4 shows the maximum benefit per 1,000 decisions for different sample sizes and models for the first case study. The associated maximum benefits vary significantly with sample size. The benefits based on sample sizes of 16,000 and 32,000 user journeys are close to each other because the predictive accuracies of the obtained models are nearly equal. If the advertiser uses the random intercept model with a threshold value of $p_{thres} = 0.05$, the desired economic benefit is achieved at 16,000 user journeys. This is a very small amount in comparison to the complete set of several millions of user journeys.

The results of the predictions for the second case study are presented in Figure 3.5. The AUCs of all three models show convergence when the sample size is increased. The best AUC is already achieved at 8,000 user journeys. Consequently the benefit per 1,000 decisions do not further grow, as the sample size is increased. However, the standard deviations of the maximum benefit per 1,000 decisions is higher as compared to the first case study, which is due to the defined ratio of CM/C. A higher ratio in the second case study would result in a higher benefit per 1,000 decisions.

Despite the mentioned limitations, the results show that the models are valuable in a real-life RTA setting: If a bidding agent only responds to bid requests for which the click probability of the current user is greater than the determined threshold value, the costs of ineffective impressions can be reduced and new customers can be attracted effectively.

In both case studies, the random intercept model is superior over the other models, which matches to our observation that the conversion rates at night and during the day are different (left hand side of Figure 3.4 and 3.5). Overall, the random slope model performs worst (except 32,000 users, first case study). This implies that the effect of the individual marketing channels, such as display ad-

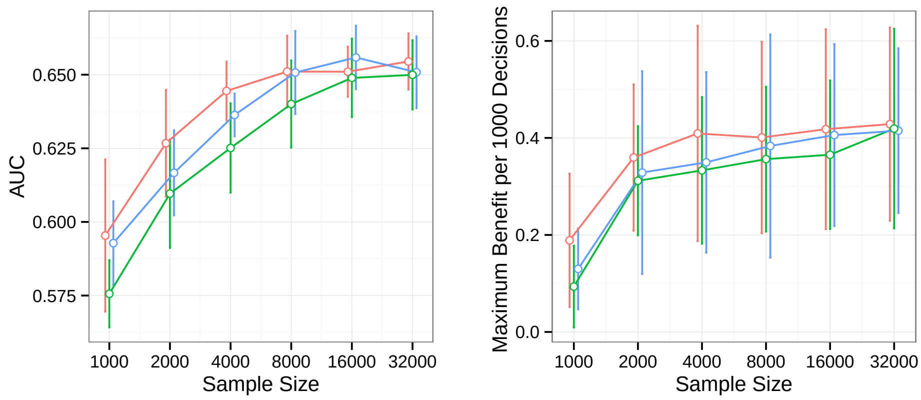

FIGURE 3.5: Results from the second case study: Area under the curve and maximum benefit per 1,000 decisions of the simple logit model (blue), random intercept model (red) and the random slope model (green).

vertisement or TV spots, does not differ significantly between during the day and at night for either case study.

3.5 Concluding Remarks

3.5.1 Limitations

Although our results distinctly reveal the expected saturation effects, the study contains some limitations. First, we only apply the procedure to a specific set of variables. We expect that more complex models, which include additional covariates or non-linear transformations of covariates, will require additional data for model estimation. This expectation may be important because previous studies (Chatterjee, 2008; Yang and Ghose, 2010) have indicated that unclicked ads should also be considered in clickstream modeling. However, the models used here for demonstration purposes ignore unclicked impressions and, thus, the potential enduring effects of ad exposure if no click-through is achieved. Second, the model may require frequent updating based on new data. This updating may result in additional costs with respect to collecting, storing, and analyzing data and may also negatively influence the predictive accuracy. Thus, as the variability increases, the fraction of data used for model estimation will have to increase to achieve optimal results. The more stable is the user behavior, the smaller will be the fraction of data required to estimate the model and the higher the gain from our approach. Third, we only apply simple random sampling in either case study. Applying stratified sampling could result in even smaller sample sizes to obtain appropriate results. However, this sampling strategy needs a correction of the estimates after the calculation (Falk et al., 2004), which requires knowledge about the total number of conversions in the population. Finally, the choice of the ratio of the contribution margin and the costs per impression is set by the authors based on typical costs in the industry. However, the threshold value p_{thres} is highly dependent on this ratio. If the ratio increases, the costs for incorrect negative predictions are much higher than the costs for incorrect positive predictions and, thus, the advertiser should always bid for the impression. On the other hand, if the ratio decreases, the advertiser should eventually never bid, because the costs for the bids become higher than the benefit from the generated conversions. We propose a deeper investigation of this fact for further research.

3.5.2 Conclusion and Outlook

In this paper, we propose a general procedure to determine the optimal sample size in a data analysis, which is applicable to an extensive range of scenarios. In two case studies of German online retailers, we apply this general procedure to user journey data using a non-hierarchical, a random intercept and a random slope logit model and determine the optimal sample size. We obtain this minimal value by including less than 1% of our subsets of user journeys. This is considerably less than the total amount of available data at most online marketing companies. Given the cost of collecting, storing and analyzing the data, an increase in the sample size is not economically beneficial. Although the notion that only a subset of data is required to provide adequate predictions is rather in line with common intuition, we determine how much data is actually needed in an online marketing context; this finding contributes to comparable future analyses in research and practice. In the context of the growing user awareness for tracking activities, our findings can be a driver for rethinking the collection of user specific data towards leaner user journey analyses. In the beginning of this paper, we suggested that Pbytes of data could easily be collected in a short period of time in an RTA setting. Storing a vast amount of data can be expensive. For instance, storing 1 Pbyte of data on an Amazon S3 server currently costs approximately 100,000 EURO/month (Amazon, 2014). Our results show that only a relatively small amount of data is required to provide statistically significant and useful parameters for prediction. Thus, storage costs can be reduced significantly by applying our proposed approach.

To summarize, our research not only contributes to the process of model estimation but also shows that advertisers do not have to collect all data that is produced from user interaction with their advertising campaigns and websites. Moreover, our process serves as an additional opportunity for decision makers from many industries to reduce the cost of infrastructure for data analysis.

3.6 References

Amazon (2014). *Amazon Web Services Simple Monthly Calculator*.

Anderl, E., I. Becker, F. Wangenheim, and J. H. Schumann (2014). "Mapping the Customer Journey: A Graph-Based Framework for Online Attribution Modeling". In: *SSRN 2343077*.

Bradley, A. P. (1997). "The use of the area under the ROC curve in the evaluation of machine learning algorithms". In: *Pattern recognition* 30.7, pp. 1145–1159.

Brutti, P., F. D. Santis, and S. Gubbiotti (2009). "Mixtures of prior distributions for predictive Bayesian sample size calculations in clinical trials". In: *Statistics in medicine* 28.17, pp. 2185–2201.

Bucklin, R. E. and C. Sismeiro (2003). "A Model of Web Site Browsing Behavior Estimated on Clickstream Data". In: *Journal of Marketing Research* 40.3, pp. 249–267.

Bucklin, R. E. and C. Sismeiro (2009). "Click here for Internet insight: Advances in clickstream data analysis in marketing". In: *Journal of Interactive Marketing* 23.1, pp. 35–48.

Chatterjee, P. (2008). "Are unclicked ads wasted? Enduring effects of banner and pop-up ad exposures on brand memory and attitudes". In: *Journal of electronic commerce Research* 9.1, pp. 51–61.

Chatterjee, P., D. L. Hoffman, and T. P. Novak (2003). "Modeling the clickstream: Implications for web-based advertising efforts". In: *Marketing Science* 22.4, pp. 520–541.

Cho, C.-H. (2003). "Factors influencing clicking of banner ads on the WWW." In: *Cyberpsychology & behavior* 6.2, pp. 201–215.

Cohen, M. P. (1998). "Determining sample sizes for surveys with data analyzed by hierarchical linear models". In: *Journal of Official Statistics* 14.3, pp. 267–275.

Da Silva, A. F. (2010). "cudaBayesreg: Bayesian Computation in CUDA." In: *R Journal* 2.2, pp. 48–55.

Deelman, E., G. Singh, and M. Livny (2008). "The cost of doing science on the cloud: the montage example". In: *Proceedings of the 2008 ACM/IEEE conference on Supercomputing*, pp. 50–62.

Falk, E., J. Kim, and W. Rotz (2004). "Minimum Sample Sizes with Rare Events in Stratified Designs". In: *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 1174–1179.

Gu, B., F. Hu, and H. Liu (2001). "Modelling Classification Performance for Large Data Sets". In: *Advances in Web-Age Information Management*, pp. 317–328.

Henriksen, S., A. Wills, T. B. Schön, and B. Ninness (2012). "Parallel implementation of particle MCMC methods on a GPU". In: *System Identification* 16.1, pp. 1143–1148.

Jacob, P., C. P. Robert, and M. H. Smith (2011). "Using parallel computation to improve Independent Metropolis-Hastings based estimation". In: *Journal of Computational and Graphical Statistics* 20.3, pp. 1–19.

Joo, M., K. C. Wilbur, B. Cowgill, and Y. Zhu (2014). "Television Advertising and Online Search". In: *Management Science* 60.1, pp. 56–73.

Joseph, L., D. B. Wolfson, and R. D. Berger (1995). "Sample size calculations for binomial proportions via highest posterior density intervals". In: *The Statistician* 44.2, pp. 143–154.

King, G. and L. Zeng (2001). "Logistic regression in rare events data". In: *Political analysis* 2.1, pp. 137–163.

Klapdor, S. (2013). *Effectiveness of Online Marketing Campaigns*. Wiesbaden: Springer Fachmedien.

Lee, K., B. Orten, A. Dasdan, and W. Li (2012). "Estimating conversion rate in display advertising from past erformance data". In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, Beijing, China*, pp. 768–776.

Lee, K. C., A. Jalali, and A. Dasdan (2013). "Real time bid optimization with smooth budget delivery in online advertising". In: *Proceedings of the 7th International Workshop on Data Mining for Online Advertising*.

Martin, A. D., K. M. Quinn, and J. H. Park (2011). "MCMCpack: Markov Chain Monte Carlo in R". In: *Journal of statistical software* 42.9.

Meek, C., B. Thiesson, and D. Heckerman (2002). "The Learning-Curve Sampling Method Applied to Model-Based Clustering". In: *Journal of Machine Learning Research* 2, pp. 397–418.

Nottorf, F. and B. Funk (2013). "The economic value of clickstream data from an advertiser's perspective". In: *Proceedings of the 21st European Conference of Information Systems, Utrecht, Belgium*.

Perlich, C., B. Dalessandro, R. Hook, O. Stitelman, T. Raeder, and F. Provost (2012). "Bid optimizing and inventory scoring in targeted online advertising". In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China*, pp. 804–812.

Rossi, P. E. and R. McCulloch (2010). *Bayesm: Bayesian inference for marketing/microeconometrics, R package version: 2.2.*

Sahu, S. and T. Smith (2006). "A Bayesian method of sample size determination with practical applications". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 169.2, pp. 235–253.

Santis, F. (2007). "Alternative Bayes factors: Sample size determination and discriminatory power assessment". In: *Test* 16.3, pp. 504–522.

Stange, M. and B. Funk (2014). "Real-Time Advertising". In: *Business & Information Systems Engineering* 56.5, pp. 305–308.

Wang, F. and A. E. Gelfand (2002). "Approach to Bayesian A Simulation-based for Sample Size Determination under a Given Model Performance and for Separating Models". In: *Statistical Science* 17.2, pp. 193–208.

Wilkinson, D. J. (2005). "Parallel Bayesian Computation". In: *Handbook of Parallel Computing and Statistics*. Chapman and Hall/CRC; 1 edition, pp. 477–509.

Windle, J., N. G. Polson, and J. G. Scott (2014). *BayesLogit: Bayesian logistic regression, R package version 0.5.1.*

Yang, S. and A. Ghose (2010). "Analyzing the Relationship Between Organic and Sponsored Search Advertising: Positive, Negative, or Zero Interdependence?" In: *Marketing Science* 29.4, pp. 602–623.

Yau, C. (2015). *RPUD, R Package version: v0.5.0. Retrieved from: http://www.r-tutor.com/content/download*.

4 How Big Does Big Data Need To Be?

Stange, M. and B. Funk (2016). "How Big Does Big Data Need To Be?" In: Enterprise Big Data Engineering, Analytics, and Management. Ed. by M. Atzmueller, S. Oussena, and T. Roth-Berghofer. IGI Global. Chap. 1, pp. 1–15.

4.1 Introduction

Today, collecting and storing of as many data as possible is common practice in many companies (Auschitzky et al., 2014; Beath et al., 2012). One of the purposes is the generation reports, which may contain descriptive statistics about customer acquisition, sales or the supply chain. Although such reports enable managers to make decisions based on past performance data, their ability to predict future company needs is limited. In contrast, predictive analytics use historical data and apply methods from machine learning and data mining to derive predictions about the future (Waller and Fawcett, 2013). For instance, predictive analytics can be used to prevent potential malfunctions in production or to predict future sales to determine the needed number of products in stock. Due to the potential benefits of predictive analytics, the collection of as many data as possible is supposed to be very beneficial for many companies. However, on the other hand, storage of these mass data is combined with significant costs (Stange and Funk, 2014). These may be infrastructure costs or monthly fees for the use of cloud services, such as Amazon S3. A question that arises from this discrepancy is: "How much data do we actually need to make the best predictions about the future?"

To reduce costs of collecting and storing data that is not relevant, it is crucial to first define which predictions are required, i.e., to determine, which analytical questions are to be answered. Afterwards – to prevent storing irrelevant data – the amount of data that is needed to obtain useful predictive results can be determined. In this chapter, we extend a generally applicable framework (Stange and Funk, 2015) to determine this amount, i.e., to find the minimum amount of data that is needed to obtain optimal predictive results. The process can be used to maximize the benefits of predictive analytics with respect to the costs for data collection and storage. Due to the scalability of cloud services, companies that use such services can benefit from the proposed process in particular. We apply the process to a data set from the online marketing field and observe convergence of the predictive accuracy when the sample size is increased. Thereby, we show that – compared to the available amount of data – only a very small sample is needed to achieve a desired predictive accuracy.

4.2 Related Work

Sample size determination is a topic that has often been examined in medical or sociological science (e.g., Brutti et al., 2009; Sahu and Smith, 2006; Santis, 2007), since, in these fields, samples are often expensive in comparison to big data environments. Additionally, the available methods to determine the needed sample size often focus on a specific task, such as to find the needed number of participants of a survey. Therefore, these methods do not seem appropriate for a generally applicable framework in predictive analytics with its variety of machine learning and data mining techniques.

In contrast to the available methods that calculate the needed sample size a priori, the proposed process is based on the evaluation of the predictive accuracy and the calculation of the economic value of the classifier.

The predictive accuracy of classifiers with dichotomous outcomes can be calculated by integrating the receiver operator characteristic (ROC) curve. The obtained value is called the area under the curve (AUC, Bradley, 1997), which represents the probability that a data record with unknown class is classified correctly. In our case study, we employ two logistic regression models with elastic net regularization (Friedman et al., 2010) to estimate the model parameters that we use to predict the dependent variable on the holdout sample. Based on these predictions, we show that increasing the sample size results in convergence the AUC. Other types of dependent variables, such as multinomial outcomes, require other measures, such as the misclassification error.

The so-called learning curve sampling method (Meek et al., 2002) is an approach for obtaining the relation between sample size and predictive accuracy. This generally applicable sampling approach is based on the observation that an increase in the sample size reduces the uncertainty in the parameter estimates of the learned model (Gu et al., 2001; Meek et al., 2002; Stange and Funk, 2015). This observation has been formalized by Meek et al. (2002) who find the optimal sample size by continuously increasing the sample size while observing the predictive accuracy. The optimal sample size is found when additional samples do not further increase the predictive accuracy by a predefined value of $\epsilon > 0$.

However, the predictive performance of a classifier does not provide information about its economic value. How the economic value can be obtained has been shown by Nottorf and Funk (2013b). Based on a clickstream data set from a German retailer for electronic devices, they build a user journey model in order to predict future user behavior. In particular, they predict the users' conversion probability based on their user journey. To measure the economic value of the applied model, they multiply the number of true and false predictions by the benefits and costs that can be assigned to these forecasts. Thus, they show that it is beneficial for an advertiser to apply the proposed model in real-time advertising, where a bidding agent decides whether a given user should be exposed to a display ad or not. Based on this approach and the finding of Meek et al. (2002), Stange and Funk (2015) develop a framework to determine the optimal sample size for a data analysis and apply it to two case studies. The idea behind this framework is to train a model with increasing sample sizes and to monitor the resulting predictive accuracy. As soon as a critical sample size is reached, the predictive accuracy of the model does not further increase. Since additional data records are related to additional computational costs, increasing the sample size to train the model is not recommended when the critical sample size is reached. Based on their approach, this chapter provides a business process model and an algorithm for a more structured view of this framework. In addition, we calculate the costs for additional data records and thereby show that the benefits of an analysis decrease when the sample size is further increased.

Although we focus on dichotomous classification in this chapter, the presented process is not limited to this kind of classification. The economic value of classifiers with multinomial outcomes can be obtained by applying the same methods as for the binomial case.

4.3 The Optimal Sample Size

Before the proposed process can be applied, it is crucial to thoroughly define the questions that ought to be answered by the analysis. It is clear that these questions highly influence the actually needed amount of data. In this chapter, however, we suppose that the purpose of the analysis has already been defined and do not further discuss this issue.

Figure 4.1 provides the business process model of the proposed framework, which is explained step by step in this section. The process is split up into data collection and data analysis, which are both represented by a single lane in the process diagram. The data collection step can be an arbitrary automatic mechanism that systematically collects data. The succeeding filtering and storing step persists the generated data.

How much data is required depends on the following: First, the model updating frequency determines how often the analysis needs to be repeated based on new data to obtain "fresh" results. Second, the sampling strategy influences the filter settings in the data collection step. For instance, if stratified sampling is chosen, only "interesting" data is stored. Third, the model complexity influences which types of data need to be stored and which can be omitted.

These three steps result in a set of requirements for the filtering mechanism, which are indicated by the dashed arrows that end in the filter object in Figure 4.1.

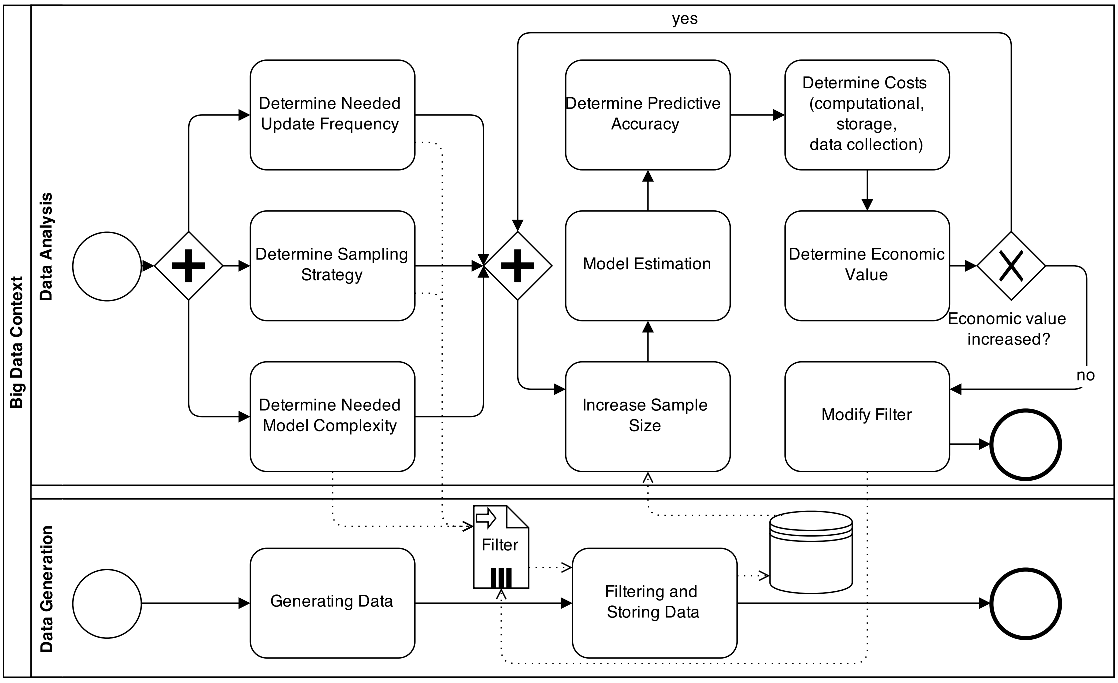

FIGURE 4.1: Finding the optimal sample size.

The sub-process beginning with the increase of the sample size and ending with the determination of the economic value can be transferred into an algorithm (Algorithm 1) to find the optimal sample size. This optimal sample size N_{opt} is obtained when the additional benefits ΔB related to the increase of the sample size by m samples do not outweigh the additional costs ΔC for the related data storage and analysis.

Algorithm 1 Determination of optimal sample size N_{opt}

$C, B, \Delta C, \Delta B, C_{old} \leftarrow 0$
$N = N_0$
$H \leftarrow \text{getSample}(N_H)$
while $\Delta B \geq \Delta C$ **do**
  $S \leftarrow \text{getSample}(N, m)$
  $C_{old} \leftarrow C$
  $P \leftarrow \text{estimateAndPredict}(S, H)$
  $C \leftarrow \text{calculateCosts}(S)$
  $\Delta C = C - C_{old}$
  $B_{old} \leftarrow B$
  $B \leftarrow \text{calculateBenefits}(P, H)$
  $\Delta B = B - B_{old}$
  $N = \text{length}(S)$
end while
$N_{opt} \leftarrow \text{length}(S)$

In Algorithm 1, the function getSample(N, m) returns a sample S of size $N + m$, estimateAndPredict(S, H) estimates the parameters based on the training sample S and returns the predictions P based on the holdout sample H with sample size N_H, calculateCosts(S) returns the costs C that are related with storing, preparing and analyzing the sample S, and calculateBenefits(P, H) returns the benefits B that are related to the application of the classifier. N_0 is the initial sample size. Although Algorithm 1 is rather intuitive, its containing functions can be very complex. We focus on the individual functions in greater detail.

- getSample$(N[, m])$: The selection of the new sample is depending on the sample size increment and the sampling strategy. In addition, it should be decided whether the complete data is re-sampled or a sample of size m is added to the previous sample. In addition, the number of additional samples m may vary between two succeeding iterations. In our case study, for instance, we exponentially increase the number m of the additional data records in each iteration.

- estimateAndPredict(S, H): The parameters of the given model are estimated based on the sample S. Various machine learning methods are available that enable classifying new data. Cross-validation can be used to avoid over-fitting. The holdout sample H is used to evaluate the classifier, i.e., predictions are made for each data record from the holdout sample H. For each data record from the holdout sample the function returns the probability that the data record belongs to a certain class.

- calculateCosts(S, P): The costs of the analysis are given by the costs for data collection, storage and computational efforts. These costs grow with the amount of data, because faster CPUs and larger storages are needed. In general, it might be difficult to determine these costs exactly.

- calculateBenefits(P, H): The benefits that result from data analyses grow with the amount of available data, because the applied algorithm has more data to learn from. The benefits that result from a data analysis can be calculated as follows: First, the costs for false predictions have to be determined (e.g., an undetected downtime in production, a competitor who is about to leave the company). Second, the benefits (or saved costs) for true predictions have to be determined (e.g., the detection of a malfunction of a tablet press in advance, the prevention of churn). Third, these values have to be multiplied by the number of true positive, true negative, false positive, and false negative predictions (based on the probabilities P) of the classifier to obtain the benefits from the prediction.

The while-loop in Algorithm 1 stops, when the additional benefits ΔB are smaller than the additional costs ΔC. After the model has been deployed to predict new data, N_{opt} can be used to modify the filter settings of the data collection step. This results in a smaller amount of collected data and thus in reduced costs.

Once the model has been deployed into the productive environment, monitoring the predictive accuracy is required to quickly address changes in the data generation process, such as changed user behavior or the go-live of a new production line. If significant changes are observed, the complete process should be re-executed. Consequently, the data collection mechanism needs to be adjusted so that enough data is available to re-execute the process.

4.4 Case Study

The case study selected for this paper belongs to the real-time advertising field. In this field of online advertising, every second thousands of records are generated that can be used to analyze and predict user behavior.

We use two different models to predict the users' purchasing behavior based on web tracking data and show the convergence of the predictive accuracy with increasing sample size.

The section is structured as follows: First, we provide a brief overview over the real-time advertising field. Second, we describe the tracking data as the basis for the proposed models. Third, we describe the modeling approach before we present the results from the analysis.

4.4.1 Real-Time Advertising

In real-time advertising (RTA) free advertising spaces on publisher websites are sold through auctions. In the moment a user visits a website using this form of advertising an auction is issued and so-called bid requests are broadcast to all potential bidders (the advertisers). Storing these messages from multiple ad exchanges quickly results in Tbytes or Pbytes of data, which can be associated with significant costs (Stange and Funk, 2014).

On the other hand, RTA enables advertisers to optimize their campaigns, by only targeting users who show a particular tendency to click on an ad (Perlich et al., 2012). As a prerequisite, these companies need to know the individual click and/or conversion probabilities. Our case study demonstrates how these individual conversion probabilities can be determined. This information can be used to optimize marketing campaigns and to increase economic benefit (Nottorf and Funk, 2013a). However, considering the costs for collection and storage, advertisers should carefully assess payoffs from related analyses (Stange and Funk, 2014).

4.4.2 Data Description

We use the data set from a German online retailer as it has been used by Stange and Funk (2015). The data set contains user tracking data (approx. 60 million records) from one month (December 2013) and, thus, includes seasonal effects, which results in shorter user journeys due to spontaneous gift purchases. Approximately 90% of the overall data is related to user behavior on the retailers website. The remaining 10% contain information about the advertising channels used to access the website. We only focus on user journeys with more than four interactions to reduce noise in our signal. Thereby, we remove 651,991 user journeys and finally obtain 280,459 user journeys that we split into a training set of 180,459 journeys and a holdout set of 100,000 user journeys. The average user journey length is 7.45.

The data set contains every interaction with the retailers website for every user. An interaction is meant to be the contact of a user with a certain advertising channel, such as a search engine advertising (SEA) or a newsletter. Each data record also contains the type of the interaction which can be a click (e.g. on a newsletter or a banner ad) or a conversion, i.e., the purchase of a product.

In addition to the online tracking data, the data set contains TV advertising data. This enables us to model the spillover effects from offline advertising to the online purchasing behavior.

4.4.3 Model Description

The set of touch points for each user is converted into their user journey according to Stange and Funk (2015). Each entry in the user journey contains information about the type of the current interaction and the number of previous touch points for each advertising channel within the current web session and across previous sessions (Chatterjee et al., 2003).

The classifiers presented in this section are based on the logistic regression model. We compare two models $M1$ and $M2$: Model $M1$ is only based on online tracking data, model $M2$ additionally involves TV advertising data. In the following, we explain the independent variables. The dependent variable for each touch point is a binary variable which is 1 if the user purchases a product, and 0 otherwise.

We adopt the modeling approach from previous studies (Nottorf and Funk, 2013a; Stange and Funk, 2015). The conversion probability is depending on the advertising channel through which the current touch point has been established as well as short term advertising effects from the current session (last 60 minutes) and from previous sessions. In our model the design matrix D consists of three parts First, the intercept terms I can be interpreted as a measure for the baseline probability for a conversion after using a certain channel to interact with the retailers website. Second, we include the number of previous interactions with the online shop within one session, denoted as X, the number of interactions in previous sessions, denoted as Y. Third, we introduce additional control variables, i.e., the number of the current session SN, the number of purchases in previous sessions CPS and the time between two sessions IST (Stange and Funk, 2015). The second model $M2$ additionally contains variables that are set to 1 if a certain TV spot Sp has been broadcast on a certain TV station St

within the last 30 minutes before the touch point, and 0 otherwise. This modeling approach can be used to measure the spillover effect of TV advertising on the online customer journey.

Equation 4.1 and 4.2 show the j^{th} interaction of the i^{th} user, which is represented as one row $(D_i)_j$ of the users design matrix D_i. Refer to the left hand side of Table 4.1 for possible values for C, i.e., the subscripts for I, X and Y. The additional covariates used in the design matrix are listed on the right hand side of Table 4.1.

$$M1 : (D_i)_j = \{I_C, X_C, Y_C, SN, IST, CPS\}_{ij} \tag{4.1}$$

$$M2 : (D_i)_j = \{I_C, X_C, Y_C, SN, IST, CPS, TV_{St}^{Sp}\}_{ij} \tag{4.2}$$

Advertising channels (C)	Index	Additional control variables
Search engine advertisement	SN	Session Number
Organic search	IST	Time between two sessions
Referral from another website	CPS	Number of conversions in previous session
Affiliate marketing	TV_{St}^{Sp}	Spot Sp was aired on Station St
Display advertisement		
Direct type-in		
Cooperation link		
Price search engine		
Email advertisement		

TABLE 4.1: Advertising channels and additional control variables used to model the design matrix as stated in Equation 4.1 and 4.2 (Stange and Funk, 2015).

We use the elastic net regularized logistic regression of the `R` package `glmnet` (Friedman et al., 2010) as classification method with $\alpha = 0.5$. This method is feasible, because it automatically selects important parameters by shrinking unimportant parameters towards 0. Hence, in the context of finding the optimal sample size, this method also determines which types of data can be deleted while maintaining high predictive accuracy.

4.4.4 Results

To determine the optimal sample size, we execute 9 analyses including $N = 500$, 1,000, 2,000, 4,000, 8,000, 16,000, 32,000, 64,000, 128,000 user journeys for both models.

The elastic net model selects different numbers of features for both models. We report the number of features selected for each sample size in Table 4.2. The model $M2$ contains more variables due to the inclusion of TV advertising effects. Therefore, the number of selected features is higher compared to model $M1$.

Using the results from the elastic net regression, we run five out-of-sample tests ($N_H^{1,\dots,5} = 10{,}000$) for each sample size and model. For each observation from the

Sample Size	$M1_{min}$	$M2_{min}$	$M1_{sd}$	$M2_{sd}$
500	13	7	10	6
1,000	21	42	10	13
2,000	27	23	12	11
4,000	22	21	14	11
8,000	23	24	14	12
16,000	21	32	14	21
32,000	19	26	15	14
64,000	22	34	15	17
128,000	21	49	15	18

TABLE 4.2: Different sample sizes and number of selected features for M1 and M2. The index *min* denotes the number of features that results in the highest AUC value. The index *sd* represents the lowest number of features that result in an AUC that lies within the standard error of the maximum AUC value.

holdout sample we predict the conversion probabilities P. Based on the actual values of the target variables from the holdout sample and the probabilities P, we are able to measure the AUC for each sample size and model, which are presented in 4.2. The variance in the predictive accuracy is indicated by the error bars in the plot. The simpler model $M1$ outperforms the model including offline advertising data $M2$ at lower sample $N < 8{,}000$. For greater sample sizes, model $M2$ shows higher predictive accuracies. This indicates that model complexity and required sample sizes are positively correlated.

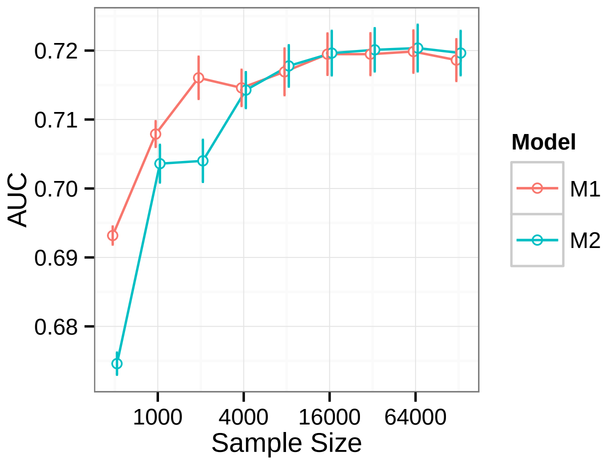

FIGURE 4.2: Area under the curve for both models and different sample sizes. The predictive accuracy converges with increasing sample sizes.

Although the AUC can be used to as a measure for the predictive performance of the classifier, it cannot be used to estimate the economic value of the classifier. The predictions have to be weighted with the costs for true and false predictions. We assume the economic values of the models $M1$ and $M2$ by applying the two different cost matrices given in Table 4.3.

Type	Description	CM_1 (EUR)	CM_2 (EUR)
TP	Ad impression costs minus contribution margin	-0.14	-0.19
TN	No loss	0.00	0.00
FP	Ad impression costs	0.01	0.01
FN	Lost contribution margin	0.15	0.20

TABLE 4.3: Two different cost matrices (CM_1 and CM_2) for true positive predictions (TP), true negative predictions (TN), false positive predictions (FP) and false negative predictions (FN).

By multiplying the costs for true and false predictions with the number of true and false predictions, we obtain the overall costs of the classifiers. To calculate the benefits of the classifier, its costs C_{min} have to be compared to the costs (C_{max}) of a trivial classifier that would always predict a conversion. The benefit of the advanced classifier is given by the costs of a trivial classifier minus the costs of a advanced classifier ($C_{max} - C_{min}$).

To include infrastructure costs we assume 0.0001 EUR for collecting and storing a single user journey. The difference between the benefit of the advanced classifier and the infrastructure costs C_I for collection and storage of the training sample divided by the number of data records N_H in the holdout sample results in the effective benefits per decision (Equation 4.3).

$$B = \frac{C_{max} - C_{min} - C_I}{N_H} \tag{4.3}$$

Figure 4.3 shows the benefits per decision for both models and cost matrices. The difference in the two diagrams results from the different ratios of contribution margin and costs for an impression in CM_1 and CM_2. This shows that the benefits of a classifier are highly dependent on this ratio.

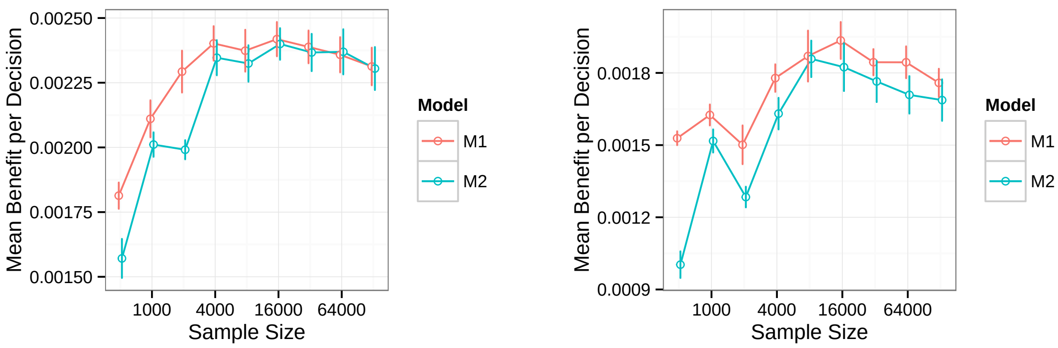

FIGURE 4.3: Mean benefits per decision for both models and different sample sizes and different benefit costs ratios (in EUR). Left hand side ratio: 15/1; Right hand side ratio: 20/1.

Figure 4.3 shows that the maximum benefit is obtained at $N = 16{,}000$ user journeys for both ratios. If the sample size is further increased, the costs ΔC for additional samples become greater than the related benefits ΔB. Thus, the while-loop in Algorithm 1 would stop at 16,000 samples in this scenario.

4.5 Conclusion

In this chapter, we extend the framework proposed by Stange and Funk (2015) to determine the optimal sample size in a predictive analytics application, which can be applied in a broad range of scenarios. We apply the developed process to a case study and show that a maximum benefit is obtained with a sample size of 16,000 user journeys, which is far less than typically available user journeys in the online marketing field. Thus, companies in this field should consider to reduce the amount of data in order to save costs from collecting and storing unnecessary data. For instance, the proposed process can be very beneficial in the area of real-time advertising where thousands of bid requests per second are sent to advertisers. Only a very small fraction of this data would be required to predict user behavior accurately. Due to the users' growing privacy concerns, applying the proposed process could also have a positive effect on the companies' image. Although we only apply our process to data from the online marketing field, it can be used by other companies generating and collecting big data for predictive analytics, for instance for demand planning, replenishment or predictive maintenance.

Although we show that the application of the process can be beneficial, our study includes some limitations: First, we neither investigate the need for model updates nor discuss the sampling strategy. Frequent model updates require collecting and storing new data, which leads to additional costs, especially when third party data is used. We use random sampling for our analyses. However, even fewer data might be required, if we decided to use stratified sampling. Second, the costs for true and false predictions as well as the costs per user journey are set by the authors according to typical costs from the industry. However, as Figure 4.3 shows, the benefit of a classifier is highly dependent on these values. In general, these values might not be constant over time and might also vary between different users. Hence, the results from the case study can only be seen as an approximation and a starting point for more advanced analyses. Third, user journey creation requires the complete set of touch points of a user. Hence, to predict the outcome (conversion or no conversion) for future touch points, historical data is needed – data that has to be collected and stored. Consequently, in addition to the needed amount of data for the training sample, data for the prediction of future user behavior has to be stored as well, which is combined with additional costs.

In summary, the proposed process can be beneficial for companies collecting and storing data in an unsystematic manner and often do not even use a small fraction of the overall available data (Stange and Funk, 2015). The process can be a starting point for further investigations in this field and for discussions among researchers and practitioners who face a dramatically growing amount of data without knowing exactly how much data they have to keep to obtain appropriate results from analyses.

4.6 References

Auschitzky, E., M. Hammer, and A. Rajagopaul (2014). *How big data can improve manufacturing, McKinsey, http://digitalstrategy.nl/ wp-content/uploads/ 2014.01-A-How-big-data-can-improve-manufacturing-_-McKinsey-Company.pdf, retrieved on 20 Jul 2015.*

Beath, C., I. Becerra-Fernandez, J. Ross, and J. Short (2012). "Finding value in the information explosion". In: *MIT Sloan Management Review* 53.4, pp. 18–20.

Bradley, A. P. (1997). "The use of the area under the ROC curve in the evaluation of machine learning algorithms". In: *Pattern recognition* 30.7, pp. 1145–1159.

Brutti, P., F. D. Santis, and S. Gubbiotti (2009). "Mixtures of prior distributions for predictive Bayesian sample size calculations in clinical trials". In: *Statistics in medicine* 28.17, pp. 2185–2201.

Chatterjee, P., D. L. Hoffman, and T. P. Novak (2003). "Modeling the click-stream: Implications for web-based advertising efforts". In: *Marketing Science* 22.4, pp. 520–541.

Friedman, J., T. Hastie, and R. Tibshirani (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software* 33.1, pp. 1–22.

Gu, B., F. Hu, and H. Liu (2001). "Modelling Classification Performance for Large Data Sets". In: *Advances in Web-Age Information Management*, pp. 317–328.

Meek, C., B. Thiesson, and D. Heckerman (2002). "The Learning-Curve Sampling Method Applied to Model-Based Clustering". In: *Journal of Machine Learning Research* 2, pp. 397–418.

Nottorf, F. and B. Funk (2013a). "A cross-industry analysis of the spillover effect in paid search advertising". In: *Electronic Markets* 23.3, pp. 205–216.

Nottorf, F. and B. Funk (2013b). "The economic value of clickstream data from an advertiser's perspective". In: *Proceedings of the 21st European Conference of Information Systems, Utrecht, Belgium*.

Perlich, C., B. Dalessandro, R. Hook, O. Stitelman, T. Raeder, and F. Provost (2012). "Bid optimizing and inventory scoring in targeted online advertising". In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China*, pp. 804–812.

Sahu, S. and T. Smith (2006). "A Bayesian method of sample size determination with practical applications". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 169.2, pp. 235–253.

Santis, F. (2007). "Alternative Bayes factors: Sample size determination and discriminatory power assessment". In: *Test* 16.3, pp. 504–522.

Stange, M. and B. Funk (2014). "Real-Time Advertising". In: *Business & Information Systems Engineering* 56.5, pp. 305–308.

Stange, M. and B. Funk (2015). "How Much Tracking Is Necessary - The Learning Curve in Bayesian User Journey Analysis". In: *Proceedings of the 23rd European Conference on Information Systems, Muenster, Germany*.

Waller, M. A. and S. E. Fawcett (2013). "Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management". In: *Journal of Business Logistics* 34.2, pp. 77–84.

5 The Impact of TV Ads on the Individual User's Purchasing Behavior

Stange, M. (2015). "The Impact of TV Ads on the Individual User's Purchasing Behavior". In: Proceedings of the 36th International Conference on Information Systems, Fort Worth, United States.

Abstract

The importance of a well-balanced cross-channel marketing strategy has increased over the past few years. The synergies caused by the interdependencies of different online channels, such as e-mail advertising, search engine and banner advertising, have also drawn the attention of many researchers. However, relatively little is known about the impact of offline marketing, such as TV advertising, on online user behavior. In this article, a model commonly used in clickstream analysis is extended by adding several TV advertising variables. Based on this model, a hierarchical Bayesian logistic model is developed to estimate the cross-channel effects of both offline and online channel contacts. By applying this model to a case study, it is shown which online channels are most supported by television ads. The findings of this paper have managerial implications for practitioners in the field, in particular because of the increasing use of a so-called "second screen" while watching TV.

5.1 Introduction

Data sets produced in online marketing make it possible to analyze advertising effects on the individual user's level (Bucklin and Sismeiro, 2009). Based on such analyses, marketing budgets can be attributed to individual online channels, such as affiliate marketing, search engine advertising (SEA) or e-mail advertising (Shao and Li, 2011). In practice, however, rather simple heuristics are typically used to determine the marketing success of the individual advertising channels (Dalessandro and Perlich, 2012). Thus, such simple methods do not necessarily measure the actual effects of the channels. Consequently, analytical solutions for advertisers are becoming increasingly popular, in particular in the emerging field of real-time advertising, where the probability for a "conversion" or "click" can be used to determine the size of the bid for a given advertising slot (Stange and Funk, 2014). To estimate the probability for a conversion or a click, many publications use clickstream data, which is usually then transformed into user journeys (Bucklin and Sismeiro, 2009; Chatterjee et al., 2003; Nottorf, 2014). The results of user journey analyses can be used to apportion the marketing budget across the channels and to predict the user behavior of future user contacts.

However, when it comes to the interaction of offline marketing campaigns and online user behavior, very few analytical approaches are available. The problem is clear: It is almost impossible to track whether a user visiting a website has been watching a television ad or listening to radio ads in the past hour, day or week. This makes it difficult to apply analytical methods and to attribute the conversions that are achieved online to offline ads.

Nowadays, an increasing number of customers use more than one media device at a time (Courtois and D'heer, 2012). For instance, many users surf on the Internet using their tablet or smart phone while watching TV. This use of a so-called "second screen" presents new opportunities for marketers, whose aim it is to reach their customers in a complementary way to maximize marketing effects. Thus, a detailed analysis of a combined online and offline marketing strategy is extremely desirable for many advertisers.

The central research question of this study is how television advertising affects the online purchasing behavior of customers. More specifically, the study investigates how the time dependency of television advertising effects can be modeled and how these time dependent television advertising effects can be included in a commonly used clickstream model to reveal cross-media advertising effects. The study contributes to IS research by providing insights into users' tendency to use different information sources to make decisions in e-commerce contexts. In addition, it encourages IS researchers to focus on this dynamic research field at the intersection of IS and marketing, and for instance, investigate users' multi-device usage or develop strategic decision support systems and real-time bidding agents using probabilistic models (Stange and Funk, 2014) such as the one presented in this study. In addition, they can develop new business models for marketing agencies and other service providers in the field (Veit et al., 2014). These models are important for companies, because they often focus only on either online or offline decision support services (Joo et al., 2014).

The modeling approach presented in this paper is structured as follows: First, the clickstream model introduced by Chatterjee et al. (2003) is extended by introducing non-linearity terms to model saturation effects. These changes allow for a more detailed answer to the research questions. Second, a model to estimate the time dependency of the TV advertising effects is developed. Third, a hierarchical Bayesian logistic regression model is developed to estimate the interdependencies of online and offline channels. Applied to a case study, it is shown that these measures decrease the residual deviance of the fit and increase the predictive accuracy in comparison to the original model. The analysis shows different effects for each TV station and advertised product and delivers the interdependencies of the individual online channels and the TV advertising effects.

The remainder of this paper is structured as follows: First, relevant work on clickstream data analysis in the literature is summarized and different approaches for the analysis of offline-online advertising effects are presented. Second, the data set of the case study and the applied modeling process are described. Third, the results of the analyses are presented in detail. Finally, the findings are summarized and their managerial implications are outlined.

5.2 Related Work

This article uses the achievements of two research streams. For the case study presented in this paper, models from previous clickstream analyses are used and combined with modeling approaches from studies dealing with the effects of TV advertising.

5.2.1 Clickstream and Cross-Channel Advertising

Clickstream data consists of data records produced when users interact with an advertiser on the Internet (Bucklin and Sismeiro, 2003). This kind of contact might be a click on a display ad, a search request, or any activity on the advertiser's website. The complete set of contacts of a user with the marketing channels of an advertiser is referred to as their user journey.

More than a decade ago, Chatterjee et al. (2003) developed a model to predict a user's individual click proneness based on clickstream data. They derived their model from several theoretical considerations about the effect of display advertisement and used several models to explain the dependent variable $Pr(Click = 1|UserJourney)$, i.e., the probability for a click given the user's current contact and the previous contacts. They show that a logistic model with heterogeneity terms across users and user sessions best describes user behavior and discuss how the results from the logistic regression can be interpreted as the different effects of marketing activities (Cho, 2003; Lee et al., 2012; Nottorf, 2014; Richardson et al., 2007). The resulting knowledge about the effect of the individual channels is valuable for distributing the appropriate amount of money across the marketing channels, i.e., for the budget allocation.

The clickstream modeling approach by Chatterjee et al. (2003) is based on counting the channel contacts within and across user sessions. This procedure assumes a linear relationship between the marketing effect and the logit of the click probability. To account for the short-term and long-term effects of the contacts of the customer with the advertiser, such as display views, Chatterjee et al. (2003) introduce two sets of variables containing the total number of contacts in the current session and in earlier sessions. To account for wear-in effects (the customer growing more and more aware of the product) and wear-out effects (customer awareness becoming saturated and decreasing with the time), they introduce the quadratic form of some channel variables. However, they do not observe significant effects. In this paper, the channel contacts are modeled as non-linear, i.e., instead of counting the channel variables, a function $f(x) = x^\gamma$

for each covariate is introduced and the exponents $\gamma_{(\cdot)}$ are estimated in a separate step before conducting the main hierarchical analysis. It is shown that the predictive accuracy can be increased significantly by adding this step.

Ever since the work of Chatterjee et al. (2003), this topic has drawn the interest of many other researchers, not least because often, in actual practice, rather simplistic heuristics are used to determine the budget for each channel (Anderl et al., 2014; Jordan et al., 2011; Kitts et al., 2010). An important finding from many studies using different methods is that effects of marketing activities are interdependent (Anderl et al., 2014; Chatterjee, 2010; Ghose and Yang, 2009; Klapdor, 2013; Nottorf, 2014; Piercy, 2012). That is to say, marketing channels cannot be analyzed separately, but have to be seen in the context of other channels. For this reason, analyses should not only focus on individual channel effects but also on cross-channel effects. This has been done by Ghose and Yang (2009) for the search engine advertising channel and the organic search channel, but to the best of the author's knowledge no published study has investigated the complete set of interdependencies across all advertising channels yet.

However, cross-channel marketing is not limited to online channels (Dinner et al., 2014; Kitts et al., 2010; Olbrich and Schultz, 2014). Joo et al. (2014) show that offline data from TV advertising spots can be used to predict customers' online search behavior. The more brand-related TV spots are broadcast, the more users tend to search for these brands. However, modeling offline advertising effects on the individual user's level is not as simple as modeling online advertising contacts, since it is almost impossible to determine if a user, in fact, was exposed to the offline advertisement. However, in this study, a new approach to include the offline variables into the user journey is provided. This approach has not been taken in the literature before.

This paper contributes to the field of clickstream analysis in two ways. First, it presents a new way to model and estimate non-linearity parameters and demonstrates how to include TV advertising data in user journeys. Second, it shows how to use a hierarchical Bayesian logistic model to determine cross-channel effects for all marketing channels for a given user journey data set.

5.2.2 Effect of TV Advertising on Online Behavior

In principle, there are two possible approaches to measuring the effectiveness of TV advertising on online user behavior. First, the so-called advertising stock for a product can be calculated as a function of the frequency of the spots and their reaches. This advertising stock can be used as a measure of the long-term (awareness) effects of TV advertising (Lodish et al., 1995). Second, the direct, performance-oriented impact of a spot can be measured by observing the uplift of page impressions and conversions in the few minutes after the broadcasting of the spot (Lewis and Reiley, 2013; Zigmond and Stipp, 2010). This study focuses on the latter.

As stated above, the main difficulty in measuring the effect of TV advertising on online behavior is the lack of an indication if a user actually was exposed to the advertisement or not. A way to gain this information is described by Kitts et al. (2010) who propose displaying discount codes during TV spots, codes which can be entered online to receive a discount. Conversions that result from this technique can easily be attributed to TV advertising. However, most campaigns have not followed this approach.

The simultaneous use of two devices, e.g., television or a mobile device, has been the subject of increasing interest over the past five years. A recent survey showed that approximately 30% of television consumers use a mobile device at some point during the day (Bolten, 2014). This number indicates that the phenomenon of second-screen use is a significant new opportunity for advertisers to optimize their cross-channel budget allocation and to increase profit due to marketing campaigns.

Today, the analysis of television advertising effects on online user behavior is often driven by very simple heuristics. For instance, website traffic is observed in the minutes before and after a spot is broadcast on TV. The uplift in user traffic resulting from the spot, i.e., the difference of the number of page impressions, can then be assigned to the impact of TV advertising. It is readily apparent that this heuristic can only approximate true TV impact, because it ignores any other effects that might cause the uplift at that moment, effects which cannot be neglected, particularly by bigger online shops. A more sophisticated method is to observe website traffic over longer periods of time, including periods with and without TV advertising broadcasts. The resulting website traffic can be decomposed using a time-series analysis to obtain the periodic effects (i.e., daily and weekly traffic patterns) and the trend effects indicating the long-term effects from TV advertising and the residual effects indicating the short term effects, which can be attributed to TV advertising.

Investigations of TV advertising on online user behavior are not frequent in the literature. However, there are some researchers who analyze the impact of TV ads on an aggregated level. Lewis and Reiley (2013) and Zigmond and Stipp (2010) analyzed the effect of TV ads during the Super Bowl and the opening show of the Olympic Games, respectively. They see a significant increase in terms of in page impressions and conversions for advertised brands in the minutes after the TV spot. The diagram representing the uplift of visits and conversions after the spot has roughly the form of a Γ distribution, which is relatively intuitive: In the first few seconds after the spot, only a few users use their second screen to search for the product, but after a few minutes a maximum is reached. Afterwards, the direct effect of the advertisement decreases. The limitation that a significant uplift can only be observed for spots with a very high reach is clearly evident here, and the application of this approach to smaller reaches is limited.

A more analytical approach is used by Joo et al. (2014), who measured the effect of TV spots on online search behavior. In contrast to the studies discussed in the previous paragraphs, they also accounted for long-term effects using exponential decay to explain the advertising stock. They observed an increase of brand search requests after a TV advertising broadcast. Liaukonyte et al. (2015) measured the impact of TV ads on the purchasing behavior of customers on an aggregated level. They find that the TV advertising effects vary from spot to spot. However, in contrast to their analysis, this paper focuses on the conversion probability on the level of the individual users and also includes other marketing channels such as display and e-mail advertising. This appears to be the first study that analyzes the cross-channel effects of television advertising on the level of the individual user.

5.3 Data Description and User Journey Modeling

In this paper, a data set from a German retailer who operates both online and offline stores is used. The available data set contains approximately 60 million records. Each record included in the raw data represents an instance of a user contacting the website and most of these instances (approximately 90%) are generated by user activity on the retailer's website. These activities include, but are not limited to, viewing product pages, or shipping or payment information, or purchasing of products. The advertising channel that customers used to reach the retailer's website is identified in the first record generated for each session in the data set (10% of the complete data set). This channel may be a display banner, a link in an e-mail, a search engine ad, a social media ad, or an affiliate ad. Additional advertising channels are organic search requests, price engine search requests, direct type-ins of the shop URL in the browser, and direct referrals from another website. The case study period includes the 24 days before Christmas Eve, 2013. This period is most likely characterized by many instances of spontaneous gift shopping. The sample was chosen, because a relatively high spillover effect from TV ads to the online user behavior is expected during that time. The clickstream data and the TV advertising data were collected by two different service providers and provided by the retailer. More information about the experimental setting cannot be given due to non-disclosure agreements.

Using this data and closely following Chatterjee et al. (2003), user journeys are built. These consist of three different types of channel variables and additional control variables. First, the intercept terms $I_{(\cdot)} \in \{0, 1\}$ indicate the type of the current contact. For instance $I_{SM} = 1$ indicates, that the current contact is caused by a click on a social media ad. Second, the session variables $J_{(\cdot)} \in \mathbb{N}_0$ indicate the previous number of channel contacts within the current user session. User activity after one hour of no activity defines a new user session (Chatterjee et al., 2003). For example $J_{SEA} = 2$ indicates that the user has clicked twice on a search engine ad within the current session. Third, the variables $K_{(\cdot)} \in \mathbb{N}_0$ indicate the number of contacts with certain channels in previous user sessions. For instance, $K_{OS} = 5$ indicates that a user has searched five times for the product or brand in previous user sessions. The additional control variables are the number of previous conversions within the current session (CWS) and in previous sessions (CAS), the session number (SN), the inter-session time (IST), the time of the day terms (t, t^2, t^3 and t^4) and an indicator whether the contact has occurred on a weekend or not ($WE \in \{0, 1\}$). The variance of the inter-session time is reduced by calculating the logarithm of the time between two sessions (Chatterjee et al., 2003). The dependent variable for each contact indicates a conversion ($Conv = 1$) or no conversion ($Conv = 0$). An example of a user journey is shown in Table 5.1 (Stange and Funk, 2015).

Contact No.	I_0	I_{SEA}	I_D	J_{SEA}	J_D	K_{SEA}	K_D	CAS	IST	SN	$Conv$
1	1	1	0	0	0	0	0	0	0 h	1	0
2	1	0	1	1	0	0	0	0	0 h	1	0
3	1	0	1	0	0	1	1	0	6 h	2	0
4	1	1	0	0	1	1	1	0	6 h	2	0
5	1	1	0	1	1	1	1	0	6 h	2	1
6	1	0	1	0	0	3	2	1	2 h	3	0

TABLE 5.1: User journey example.

The set of variables for the j^{th} contact of the i^{th} user is given by the j^{th} entry of the design matrix X_i for user i as stated in Equation 5.1. An overview of the indices and variable types is given in Table 5.2.

$$\begin{aligned} X_{ij} \;=\; & \{I_0, I_{OS}, I_{TI}, I_A, I_D, I_{SEA}, I_{SM}, I_{EM}, I_{PS}, J_{TI}, J_A, J_D, J_{SEA}, J_{SM}, \\ & J_{EM}, J_{PS}, J_R, K_{OS}, K_{TI}, K_A, K_D, K_{SEA}, K_{SM}, K_{EM}, K_{PS}, K_R, SN, \\ & IST, CWS, CAS, t, t^2, t^3, t^4, WE\}_{ij} \end{aligned} \tag{5.1}$$

Index	Meaning	Variable	Meaning
A	Affiliate marketing	CWS	Number of conversions in current session
D	Display advertising	CAS	Number of conversions across sessions
EM	E-mail advertising	IST	Time between two sessions
OS	Organic search	$OffB_L$	Offline Brand Spot Nr. L
TI	Direct type-in	$OffP_K$	Offline Product Spot Nr. K
PS	Price search	OnB_M	Online Brand Spot Nr. M
R	Direct referral	SN	Session Number
SEA, SA	Search-engine advertising	t	Hour of the Day
SM	Social media advertising	TV_S	TV Station S
-	-	WE	Weekend (Yes/No)

TABLE 5.2: Overview of variables and indices.

To extend the clickstream model of Chatterjee et al. (2003), the impact of the TV advertisement is modeled by additional variables TV_S^K that hold the time (in minutes) since the last spot of a certain type K was broadcast on a given TV station S. Only spots that have been run in the ten hours previous to the contact are included, since short-term effects from spots ran even earlier are unlikely (Zigmond and Stipp, 2010). Individual TV stations are distinguished, as are individual spots (Liaukonyte et al., 2015). In addition, TV ads are categorized into brand-related spots (B) and product-related spots (P) as well as offline-related spots (Off) and online-related spots (On). Spot placement or the reaches of the spot cannot be modeled, because this information is unavailable. This is discussed in detail in the limitation section. The design matrix for the TV effect Y_i holds the TV variables for each contact j of user i as stated in Equation 5.2. As an example, the variable $TV_6^{OnB_1}$ holds the time difference in minutes between the time of the contact and the time of the previous broadcast of the online-related brand-related spot #1 on station #6.

$$\begin{aligned} Y_{ij} \;=\; \{ & TV_2^{OnB_1}, TV_4^{OnB_1}, TV_6^{OnB_1}, TV_{11}^{OnB_1}, TV_5^{OffP_1}, TV_7^{OffP_1}, TV_8^{OffP_1}, \\ & TV_1^{OffB_1}, TV_2^{OffB_1}, TV_3^{OffB_1}, TV_4^{OffB_1}, TV_5^{OffB_1}, TV_6^{OffB_1}, TV_9^{OffB_1}, \\ & TV_{10}^{OffB_1}, TV_{11}^{OffB_1}, TV_1^{OnB_2}, TV_2^{OnB_2}, TV_3^{OnB_2}, TV_5^{OnB_2}, TV_{11}^{OnB_2}\}_{ij} \end{aligned} \tag{5.2}$$

Table 5.3 presents the descriptive statistics of the transformed user journey data sample. These numbers suggest the relative frequency of the channel contacts. All user journeys including only one onsite contact and those longer than 50 contacts were removed from the data set to exclude click robots or other Internet fraud. The statistic is based on a random sample of 3,314,920 advertising channel contacts by 898,796 users. The total number of conversions in this set is 138,437.

Parameter	μ	σ	max.	Parameter	μ	σ	max.
I_{OS}	0.464	0.499	1.000	IST	3.797	3.966	11.000
I_{TI}	0.219	0.414	1.000	CWS	0.005	0.073	8.000
I_A	0.080	0.271	1.000	CAS	0.094	0.367	23.000
I_D	0.037	0.189	1.000	WE	0.293	0.455	1.000
I_{SEA}	0.068	0.252	1.000	$TV_2^{OnB_1}$	35.792	86.364	599.000
I_{SM}	0.004	0.060	1.000	$TV_4^{OnB_1}$	42.331	95.597	599.000
I_{EM}	0.080	0.272	1.000	$TV_6^{OnB_1}$	41.818	86.259	599.000
I_{PS}	0.008	0.087	1.000	$TV_{11}^{OnB_1}$	31.971	71.381	599.000
J_{TI}	0.047	0.282	30.000	$TV_5^{OffP_1}$	85.018	125.020	599.000
J_A	0.111	0.772	46.000	$TV_7^{OffP_1}$	61.819	133.435	599.000
J_D	0.018	0.213	22.000	$TV_8^{OffP_1}$	64.053	140.807	599.000
J_{SEA}	0.052	0.398	31.000	$TV_1^{OffB_1}$	101.830	147.150	599.000
J_{SM}	0.003	0.111	29.000	$TV_2^{OffB_1}$	88.921	138.856	599.000
J_{EM}	0.049	0.380	27.000	$TV_3^{OffB_1}$	90.356	142.837	599.000
J_{PS}	0.007	0.151	20.000	$TV_4^{OffB_1}$	56.931	130.628	599.000
J_R	0.011	0.150	21.000	$TV_5^{OffB_1}$	79.487	135.967	599.000
K_{OS}	0.822	2.464	49.000	$TV_6^{OffB_1}$	56.879	128.091	599.000
K_{TI}	0.982	3.184	49.000	$TV_9^{OffB_1}$	49.614	116.594	599.000
K_A	0.225	1.327	47.000	$TV_{10}^{OffB_1}$	59.392	138.278	599.000
K_D	0.097	0.646	37.000	$TV_{11}^{OffB_1}$	95.170	136.942	599.000
K_{SEA}	0.171	1.238	47.000	$TV_1^{OnB_2}$	63.755	129.567	599.000
K_{SM}	0.018	0.320	41.000	$TV_2^{OnB_2}$	80.297	121.517	599.000
K_{EM}	0.265	1.408	48.000	$TV_3^{OnB_2}$	42.329	101.559	599.000
K_{PS}	0.011	0.192	24.000	$TV_5^{OnB_2}$	61.446	109.081	599.000
K_R	0.099	0.790	44.000	$TV_{11}^{OnB_2}$	81.003	129.876	599.000
SN	2.991	3.976	49.000	-	-	-	-

TABLE 5.3: Descriptive statistics of the complete sample of user journeys.

The average user journey length in this data set is rather small, which might be due to it representing the Christmas period. The number and lengths of the user journeys are presented in Table 5.4.

Length	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-50
# Users	765,980	90,872	22,595	9,305	4,663	2,363	1,296	780	544	398

TABLE 5.4: User journey lengths.

5.4 Analysis Process

The analysis of the user journey data conducted in this study consists of three steps: First, a model for the estimation of non-linearity terms is developed. This step is important, because it allows for a comparison in terms of predictive accuracy between the effects of different model extensions presented in this paper. Second, as it has been observed by Lewis and Reiley (2013), TV effects are modeled as Γ distributed over time to estimate the parameters k and θ of the Γ distribution $\Gamma(k, \theta)$. Finally, the results of the first and second steps are used to estimate the β parameters by applying a hierarchical logistic regression model. These parameters represent the valence and size of the effects resulting from online channel contacts and TV advertising. The following sections describe each step in greater detail. The results are discussed in the next chapter.

5.4.1 Modeling Non-Linearity

The first step of the analysis is to estimate the non-linearity of the covariates from the design matrix (Equation 5.1). This step is, of course, not a prerequisite to model the impact of TV ads on user journeys. It puts, however, the gain in predictive accuracy for each of the extensions into perspective. Based on the simple multivariate logistic regression model, a Bayesian model is built up using `JAGS` (Plummer, 2003) following Equation 5.3. The TV covariates are not included in this step to reduce computation times and model complexity.

$$\begin{aligned}
Conv_{ij} &\sim \text{Bernoulli}(Pr_{ij}) \\
Pr_{ij} &= \phi(\hat{X}_{ij}\beta_X) \\
\hat{X}_{ij} &= \{I_0, \ldots, I_{PS}, (J_A)^{\gamma_{J_A}}, \ldots (J_R)^{\gamma_{J_R}}, \ldots, (CAS)^{\gamma_{CAS}}\}_{ij} \\
\gamma &\sim \text{Multivariate Normal}(\mu_\gamma, \Omega_\gamma) \\
\beta_X &\sim \text{Multivariate Normal}(\mu_{\beta_X}, \Omega_{\beta_X})
\end{aligned} \tag{5.3}$$

In this equation, the only difference to the non-hierarchical multivariate logistic regression model is the substitution of X_{ij} with $\hat{X}_{ij}$, in which each covariate of the design matrix is raised to the $\gamma_{(\cdot)}^{th}$ power, except for the intercept terms and the time of day variables t, t^2, t^3, t^4, and WE. The vector γ has the same length as $X_{ij} \setminus \{I_{(\cdot)}, t, t^2, t^3, t^4, WE\}$, denoted as q. The term $\phi(w)$ is the sigmoid function given by $\phi(w) = 1/(1 + e^{-w})$, which returns a value between 0 and 1 for a given w. This value is interpreted as the probability Pr for a conversion. In combination with the `R` package `runjags` (Denwood, 2015), the software package `JAGS` (Plummer, 2003) is used to obtain the exponents $\gamma_{(\cdot)}$ as well as the β_X parameters. The parameters μ_{β_X} and Ω_{β_X} are the multivariate normal priors for β_X. The μ_{β_X} is a vector of length p, i.e., the length of the row vector X_{ij}, with all entries equal to 0. The prior precision matrix Ω_{β_X} is a $p \times p$ union matrix multiplied by 0.1. The prior μ_γ is a vector of length q with all entries equal to 1. The prior precision matrix Ω_γ is a $q \times q$ union matrix multiplied by 2.5 and, thus, it is acting as a strong prior around 1. In this first analysis, the main interest is the estimation of the γ parameters. They are used in the third step, the hierarchical logistic model, in which the complete set of β parameters is obtained, i.e., $\beta = \{\beta_X, \beta_Y\}$.

5.4.2 Modeling TV Effects

In the second step, a model to estimate the time dependency of the TV advertising effect is developed. There are several ways to model this effect. In this paper, two different approaches are presented. First, the TV variable is set to 1 if a spot was broadcast in the last T minutes prior to the online user contact. This approach is easy to implement, but it has also two disadvantages: First, it is very unlikely that the impact of a TV spot in the first few seconds after the broadcast is identical with the impact after 5 minutes, especially if the goal is to explain conversions. Second, the effect instantly collapses after T minutes, which seems very unrealistic. The second approach proposed here is more realistic, as it takes into account the probability that a customer uses her or his second screen to search for the advertised product grows in the first few minutes after the TV spot has been broadcast. Additionally, it is also considered that the direct impact of a TV spot on the purchasing behavior decreases after a while. This modeling approach is feasible, because, according to previous studies, the uplift of page views after a TV spot has been broadcast looks similar to a Γ distribution (Lewis and Reiley, 2013). For this reason, it is proposed to model the TV advertising effect as Γ distributed over time. The Γ distribution can be parameterized by a shape parameter k and a scale parameter θ. The goals of this step is to estimate these values and to obtain the time dependency of the TV effects.

To estimate the parameters, the results from the first modeling step are used to reduce the number of degrees of freedom of the Bayesian model and to reduce computation time. Thus, the β_X values are not estimated in this step. Instead, $z_{ij} = \hat{X}_{ij}\bar{\beta}_X$ is calculated for each contact. The term $\bar{\beta}_X$ represents the median estimate of a simple logistic regression. The Bayesian model is presented in Equation 5.4.

$$
\begin{aligned}
Conv_{ij} &\sim \text{Bernoulli}(Pr_{ij}) \\
Pr_{ij} &= \phi(z_{ij} + \hat{Y}_{ij}\beta_Y) \\
\hat{Y}_{ij} &= \{\Gamma(TV_2^{OnB_1}, k, \theta), \ldots, \Gamma(TV_{11}^{OnB_2}, k, \theta)\}_{ij} \\
k, \theta &\sim \text{Uniform}(0.01, 10) \\
\beta_Y &\sim \text{Multivariate Normal}(\mu_Y, \Omega_Y)
\end{aligned}
\tag{5.4}
$$

In this equation, $\Gamma(TV_S^{(\cdot)}, k, \theta)$ returns the value of the Γ distribution at point $x = TV_S^{(\cdot)}$ with shape parameter k and scale parameter θ. The Γ parameters are sampled from a relatively narrow uniform distribution, because the short term effect of TV advertising is expected to have a maximum in the range between 1 to 60 minutes after the broadcast[1].

5.4.3 Modeling Cross-Channel Effects

The goal of the third modeling step is to estimate a complete set of β parameters for each contact type. The contact type is represented by the intercept terms $I_{(\cdot)}$ from the design matrix X. The obtained results can be interpreted as cross-channel effects. For instance, the effects of previous contacts with certain channels on conversions that result from search engine requests are represented by β^{OS}, i.e., the β parameters for the organic search channel. To achieve this goal, each contact has to be assigned to one of 9 groups representing different contact types. Note that there are only 8 different intercept terms in the design matrix X. The ninth group is related to direct referrals from other websites. A contact of this type is indicated by all other intercept terms being equal to 0, except I_0.

The modeling approach is driven by several considerations. First, it is known from previous studies that the individual marketing channels influence each other in different ways (Ghose and Yang, 2009; Klapdor, 2013). Second, it is expected that returning customers behave in a different manner than new customers. For instance, new customers are more likely to click on a search engine advertisement, whereas returning customers, indicated by a higher session number, tend to type-in the URL directly (Rutz et al., 2011). Third, TV ads are expected to have a higher impact on direct type-ins or search requests than on display or affiliate clicks (Naik and Peters, 2009).

The hierarchical logistic model with random slopes used here is described by Rossi et al. (2006). The corresponding algorithm `rhierRwMNLogit` from the `R` package `bayesm` involves high computational costs when, as in this case, large sample sizes are used. Therefore, the `R` package `rpud` by Yau (2015) is used. It

[1]A detailed analysis conducted after publication of this article showed that the model is unable to identify TV advertising effects in case of a strong correlation of effects related to the time of the day and those related to TV ads. This issue is discussed in greater detail in Section 5.8.

contains a parallelized version of the model from the `bayesm` package that uses Graphical Processing Units (GPUs) for computation. This parallelization reduces the computation time by the order of a magnitude. A simplified version without the multinomial part of the original model is presented in Equation 5.5.

$$\begin{aligned}
Conv_{ij} &\sim \text{Bernoulli}(Pr_{ij}) \\
Pr_{ij} &= \phi(\hat{X}_{ij}\beta_X^G + \hat{Y}_{ij}\beta_Y^G) \\
\{\beta_X, \beta_Y\}^G &= \Delta + \epsilon^G \\
\Delta &\sim \text{Multivariate Normal}(\mu_\Delta, \Omega_\Delta) \\
\epsilon^G &\sim \text{Multivariate Normal}(\mu_\epsilon, \Omega_\epsilon)
\end{aligned} \tag{5.5}$$

In Equation 5.5, $\hat{X}_{ij}$ and $\hat{Y}_{ij}$ represent the transformed covariates from X_{ij} and Y_{ij} according the first two steps of the analysis. The index $G \in \{A, D, EM, OS, PS, R, SEA, SM, TI\}$ indicates the type of the contact, the vector Δ represents the grand mean of the parameters from all groups, and the residual term ϵ^G represents the deviation of the individual β^G values from the grand mean Δ.

5.5 Results of the Analyses

This chapter is structured as follows: First, the results generated during the first two modeling steps, which were described in the previous chapter, are presented. Second, a comparison of the different modeling approaches is made. It is shown that a model that includes both non-linearity terms as well as TV parameters is the best one to describe the data. Third, the results from the hierarchical logistic regression are presented.

5.5.1 Non-Linearity Parameters

To estimate the γ parameters, the software package `JAGS` in combination with the `R` package `runjags` is used to run three MCMC chains with 60,000 burn-in iterations and 125,000 sampling iterations. The sample size is n=20,000 in this step. The Gelman-Rubin diagnostic shows convergence of the chains. The results from the non-linearity analysis are presented in Table 5.5.

The results clearly show that only counting channel contacts is not an appropriate method for user journey analyses. Most median values are < 1, which implies saturation effects for each channel. For instance, γ_{J_A}, the exponent of the number of previous affiliate marketing contacts during a given session, is close to zero, which means that it makes hardly any difference whether a customer contacted the website once, twice, or ten times through this channel during a session. In this case study, the γ parameter for the additional control variable SN, which was introduced by Chatterjee et al. (2003), shows that this variables has a very small impact since it is always very close to 1.

Var.	2.5%	25%	50%	75%	97.5%	Var.	2.5%	25%	50%	75%	97.5%
$\gamma_{J_{TI}}$	0.005	0.058	0.125	0.211	0.401	γ_{K_D}	0.042	0.343	0.635	0.984	1.740
γ_{J_A}	0.008	0.081	0.162	0.270	0.565	$\gamma_{K_{SEA}}$	0.314	0.874	1.097	1.319	1.822
γ_{J_D}	0.049	0.336	0.555	0.791	1.410	$\gamma_{K_{SM}}$	0.048	0.375	0.675	1.018	1.750
$\gamma_{J_{SEA}}$	0.003	0.030	0.070	0.128	0.291	$\gamma_{K_{EM}}$	0.064	0.407	0.622	0.844	1.305
$\gamma_{J_{SM}}$	0.022	0.186	0.355	0.590	1.497	$\gamma_{K_{PS}}$	0.047	0.393	0.700	1.079	2.004
$\gamma_{J_{EM}}$	0.006	0.060	0.128	0.221	0.424	γ_{K_R}	0.033	0.305	0.553	0.839	1.483
$\gamma_{J_{PS}}$	0.070	0.455	0.738	1.011	1.574	γ_{SN}	-0.388	0.083	0.144	0.252	0.698
γ_{J_R}	0.101	0.490	0.714	0.937	1.477	γ_{IST}	0.001	0.016	0.039	0.078	0.185
$\gamma_{K_{OS}}$	0.041	0.307	0.566	0.855	1.476	γ_{CWS}	0.091	0.581	0.994	1.443	2.293
$\gamma_{K_{TI}}$	0.089	0.435	0.617	0.792	1.288	γ_{CAS}	0.504	0.720	0.820	0.913	1.080
γ_{K_A}	0.165	0.430	0.574	0.725	1.070	-	-	-	-	-	-

TABLE 5.5: Quantiles of the sampled γ parameters.

5.5.2 Time-Dependent TV Effect

The densities of the sampled Γ distribution parameters θ and k are presented on the left side and in the middle of Figure 5.1. They are obtained using 60,000 burn-in steps and 125,000 sampling iterations with n=10,000 samples. The graph on the right shows the resulting time-dependent TV spot effect based on the median values for θ and k. In the minutes after the spot broadcast, the effect strength grows very quickly. It is strongest after approximately 20 minutes and decreases exponentially afterwards. The resulting Γ shaped diagram is in line with the observations from previous TV uplift studies[2].

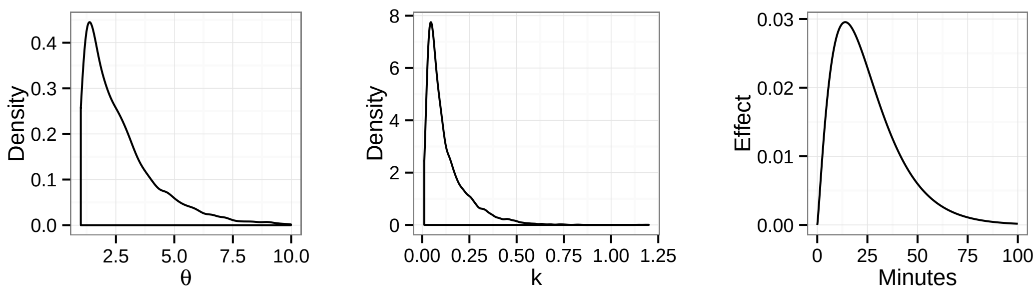

FIGURE 5.1: Densities of θ, k and the resulting time-dependent TV advertising effect.

5.5.3 Model Comparison

Before the final results are presented, alternative versions of the user journey model discussed above are compared (Table 5.6). Each model is estimated with and without the TV variables Y and $\hat{Y}$, respectively. The first two models are

[2]Indeed, a Γ shaped TV effect is in line with other studies. However, the spike in website traffic is usually located in the few minutes after an ad was aired (not after 20 minutes). This contradiction can be explained with the inconsistencies in the original data set and the strong correlation of effects related to time and effects related to TV ads. This issue is described in greater detail in Section 5.8.

based on the approach proposed by Chatterjee et al. (2003) and presented in Table 5.1. In the first case, the TV variables are set to 1 if the time difference between the spot broadcast and the contact is less than $T = 30$ minutes. In the second case, the Γ parameters θ and k obtained during the second analysis to calculate the time-dependent TV effects are applied. In the third model, the time of day variables and the indicator for weekends WE are added to control for time-dependent conversion rates over the day. If these variables were not included, time-dependent conversion rates might be assigned to the effect of TV spots[3]. In the fourth model, the X variables are transformed into $\hat{X}$ using the γ parameters obtained during the first analysis. Finally, a quadratic term for each covariate is added to account for decreasing awareness effects from individual marketing channels. Table 5.6 shows the analysis of variance for the ten different models. The residual deviance is given with and without TV variables and the difference of the residual variance to the next smaller model. In addition, the area under the ROC curve (AUC) is reported as a measure for the predictive accuracy of the model. The values are obtained with 2,000,000 contacts for the training set and 1,000,000 contacts for the holdout set.

Model	R without TV	R with TV	Dev. Delta	Sign.	Dev. Diff.	AUC w. TV
Plain	642409.3	641833.6	575.7	***	-	0.7314
Gamma	642409.3	641612.5	796.8	***	221.1	0.7315
Times	641805.6	641080.1	725.5	***	532.4	0.7326
Non-Linearity	630168.2	629627.7	540.5	***	11452.3	0.7545
Quadratic	625877.2	625368.4	508.8	***	4259.3	0.7624

TABLE 5.6: Residual Deviances R for 10 different models and AUC measures.

The results of the model comparison show that the addition of TV variables reduces residual deviance in all cases. Additionally, they show that the time-dependent TV effects reduce more deviance than the simpler approach to model TV effects. The addition of variables to describe the time of the day also reduces deviance. However, most of the deviance is reduced when the γ parameters are applied to transform X into $\hat{X}$. Even the addition of a quadratic term for each covariate has a much higher impact on residual deviance than the TV effects. This shows the importance of the feature selection process in user journey modeling, which is, however, not the focus of this study.

5.5.4 Results from the Non-Hierarchical Logistic Regression

The results from the hierarchical model cannot be presented in great detail here, because one set of β values for each contact type was obtained. However, to offer an overview of the size of the effects, the complete set of β parameters for the non-hierarchical logistic model in Table 5.7 are presented. The used model is the one with included TV effects and non-linearity from the fourth row in

[3]The correlation between the effects of TV ads and effects that are related to the time of the day when the ads were aired is discussed in greater detail in Section 5.8

Table 5.6. Note that the $\hat{X}$ and $\hat{Y}$ covariates were standardized to allow for easy comparison between individual effect sizes. Thus, the values listed in Table 5.7 show the increase of the logit of Pr by $\beta_{(\cdot)}$ resulting from adding one standard deviation of the respective covariate.

$\beta_{(\cdot)}$	Est.	SD	z	Sig. Lv.	$\beta_{(\cdot)}$	Est.	SD	z	Sig. Lv.
0	-3.461	0.004	-769.985	***	CWS	0.006	0.002	2.487	*
I_{OS}	0.182	0.012	15.367	***	CAS	0.218	0.003	72.476	***
I_{TI}	0.413	0.013	31.175	***	t	1.266	0.090	14.004	***
I_A	0.489	0.009	54.591	***	t^2	-4.812	0.387	-12.441	***
I_D	-0.016	0.008	-1.998	*	t^3	6.262	0.548	11.424	***
I_{SEA}	0.308	0.015	19.906	***	t^4	-2.783	0.252	-11.061	***
I_{SM}	0.017	0.005	3.088	**	WE	-0.030	0.004	-7.482	***
I_{EM}	0.273	0.009	29.778	***	$TV_2^{OnB_1}$	0.008	0.004	2.084	*
I_{PS}	0.039	0.005	7.291	***	$TV_4^{OnB_1}$	0.017	0.004	4.464	***
J_{TI}	0.195	0.002	80.345	***	$TV_6^{OnB_1}$	0.010	0.004	2.458	*
J_A	0.122	0.003	41.561	***	$TV_{11}^{OnB_1}$	0.027	0.004	6.211	***
J_D	0.061	0.003	17.897	***	$TV_5^{OffP_1}$	0.008	0.004	2.089	*
J_{SEA}	0.223	0.004	56.486	***	$TV_7^{OffP_1}$	0.024	0.004	5.773	***
J_{SM}	0.006	0.004	1.542		$TV_8^{OffP_1}$	-0.008	0.004	-1.875	.
J_{EM}	0.146	0.003	51.579	***	$TV_1^{OffB_1}$	-0.002	0.004	-0.617	
J_{PS}	0.029	0.003	9.241	***	$TV_2^{OffB_1}$	-0.009	0.004	-2.095	*
J_R	0.055	0.003	16.093	***	$TV_3^{OffB_1}$	-0.016	0.004	-4.044	***
K_{OS}	-0.056	0.005	-11.882	***	$TV_4^{OffB_1}$	-0.013	0.004	-3.259	**
K_{TI}	-0.118	0.009	-13.263	***	$TV_5^{OffB_1}$	-0.049	0.004	-11.469	***
K_A	-0.168	0.004	-40.109	***	$TV_6^{OffB_1}$	0.035	0.004	9.697	***
K_D	0.006	0.004	1.466		$TV_9^{OffB_1}$	-0.001	0.004	-0.302	
K_{SEA}	-0.078	0.006	-12.301	***	$TV_{10}^{OffB_1}$	0.009	0.004	2.312	*
K_{SM}	-0.015	0.005	-3.060	**	$TV_{11}^{OffB_1}$	0.002	0.004	0.496	
K_{EM}	-0.028	0.005	-6.058	***	$TV_1^{OnB_2}$	0.015	0.004	4.037	***
K_{PS}	-0.033	0.004	-7.619	***	$TV_2^{OnB_2}$	0.011	0.004	2.809	**
K_R	-0.053	0.005	-10.299	***	$TV_3^{OnB_2}$	0.009	0.004	2.220	*
SN	-0.334	0.014	-24.062	***	$TV_5^{OnB_2}$	0.040	0.004	10.317	***
IST	0.728	0.006	116.306	***	$TV_{11}^{OnB_2}$	0.014	0.004	3.769	***

TABLE 5.7: Results from the simple logistic regression with transformed covariates (Signif. codes: *** p < 0.001; ** p < 0.01; * p < 0.05; . p < 0.1).

The intercept β_0 and the term β_{SN} indicate that the baseline conversion probability is, at approximately $p \approx \phi(-3.461 - SN^{\gamma_{SN}} \cdot 0.334) \approx 0.02$, very low. As stated before, this value corresponds to conversion rates for direct referrals. This baseline probability for conversions matches typical conversion rates found in the industry and is in line with previous studies (Ghose and Yang, 2009). The probabilities for conversions through affiliate channel, search engine channel, e-mail channel, and direct type-ins are higher than this baseline probability, which is represented by the positive value of the respective $\beta_{I_{(\cdot)}}$ terms.

The effect of the number of previous channel contacts during one session $\beta_{J_{(\cdot)}}$ is positive in all cases, although the value for contacts through the social media channel is not significant. This indicates that customers who are more likely to convert tend to use multiple marketing channels during a given session before they make their purchase decision.

The effect of the number of channel contacts in previous sessions $\beta_{K_{(\cdot)}}$, however, is negative for all channels except for the display channel effect β_{K_D}, which is not significant. At first, this finding is unexpected, because it is generally assumed that returning customers with more than one session are more likely to convert than customers with shorter user journeys. As the data set has been generated during the Christmas season, one could argue that customers shopping during the season tend to use the shop for spontaneous gift purchases that can be completed during one user session.

The effects of TV spots that advertise for the online shop (i.e., OnB_1 and OnB_2) are positive in all cases. In addition, the offline and product-related spots have a positive impact on the conversion probability, except $TV_8^{OffP_1}$ which does not fulfill the significance criteria.

However, the effect from the TV spots is rather low in comparison to the effect from the online channel contacts. This is because it is not known whether a user in fact watched a certain TV program. Additionally, information about the reach and the placements of the spots within the advertising blocks is unavailable. Consequently, to develop the model it is assumed that the TV spots have an effect on each contact of every customer. Thus, the β_{TV} values represent the mean TV effect for all customers.

Depending on the TV station, there are positive or negative effects resulting from offline brand spots, i.e., $OffB_1$. Significant negative effects from TV spots are rather unintuitive, because one would expect they either have a positive effect, if the customer watched the spot, or no effect at all, if the customer did not watch it. However, negative effects can be explained with customers who make last-minute decisions. They tend to search online for a certain product to read reviews, for instance. However, these customers tend to buy the product in a retail store, because they might worry that the product cannot be delivered in time. This hypothesis is realistic, because a multitude of the brand-related offline spots have been broadcast in the few days before Christmas Eve. Furthermore, there might be lots of customers who visit the online shop to read reviews about a certain product and, nevertheless, decide to visit the retail store to try out the product before purchasing it (Cheema and Papatla, 2010).

5.5.5 Cross-Channel Effects

Running the hierarchical logistic model from the `R` package `rpud` with 2,000,000 samples and 250,000 iterations results in $9 \cdot 48 = 432$ posterior densities of the $\beta^G_{(\cdot)}$ parameters. These values cannot be reported in full detail here. Instead, the results are discussed by example. Some significant effects of the covariates from $X \setminus I_{(\cdot)}$ on the contact types represented by $I_{(\cdot)}$ are shown in Figure 5.2 and 5.3.

The graph included on the left side of Figure 5.2 shows the effect of clicks on display ads in previous sessions on search engine advertising contacts and direct type-in contacts. The non-hierarchical logistic regression did not reveal significant results for β_{K_D}. However, the hierarchical analysis shows that $\beta^{SEA}_{K_D}$ and $\beta^{TI}_{K_D}$ are significant. The result is relatively intuitive, because previous contacts with display ads increase the awareness for the brand and, hence, increase the probability for direct type-ins (Chatterjee, 2008). The probability for conversions after clicks on search engine ads decreases with increasing K_D. The negative value for $\beta^{SEA}_{K_D}$ is sensible, since one of the major goals of search engine advertising is the acquisition of new customers (Rutz et al., 2011).

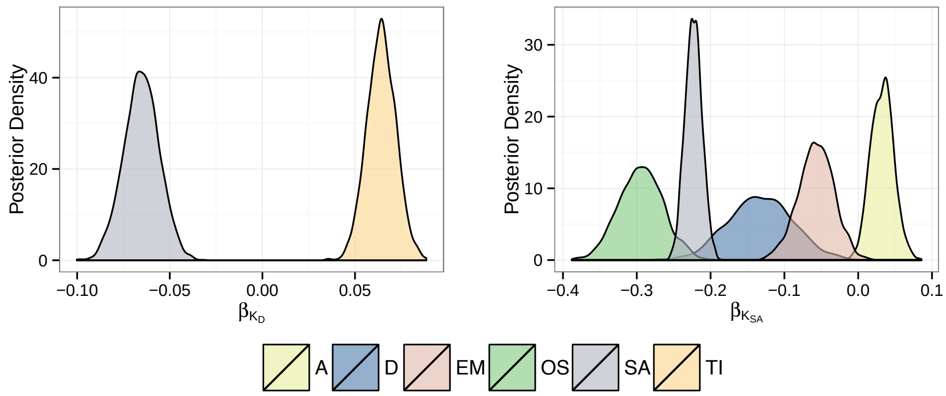

FIGURE 5.2: Significant densities of β_{K_D} and $\beta_{K_{SEA}}$ per contact type.

The graph included on the right side of Figure 5.2 shows the effect of the number of search engine advertising contacts in previous sessions K_{SEA} on five different channels. It has a positive impact on the probability for conversions through the affiliate channel and negative impacts on the probability for conversions through the organic search, the search engine, the display, and the e-mail marketing channels. A negative value for $\beta_{K_{SEA}}$ is in line with the results from the non-hierarchical analysis. However, the hierarchical analysis reveals different effects across contact types. The affiliate channel, for instance, benefits from the increasing value for K_{SEA}. One reason for this might be the growing awareness of the customers for discount coupons that are often offered in affiliate marketing.

The graph included on the left side of Figure 5.3 shows the effect of the intersession time IST on eight different channels. The hierarchical analysis shows that the deviation from the mean effect $\beta_{IST} = 0.728$, calculated in the non-hierarchical logistic regression, is relatively high. The inter-session time has a high impact on the probability for conversions after a direct type-in and a low impact for a conversion after a display click. The effect of the number of conversions in previous sessions CAS on eight different channels is shown in on the right side of Figure 5.3. The CAS value has the highest impact on the probability for conversions after an organic search request and the less impact for a conversion after an e-mail advertising click.

Significant effects of TV spots are shown in Figure 5.4 and 5.5 by example. Non-significant values are omitted for clarity. The graph included on the left side

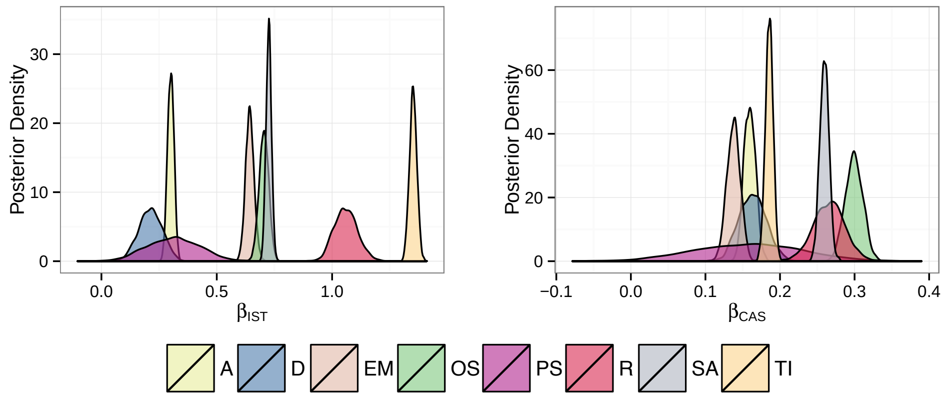

FIGURE 5.3: Significant densities of β_{IST} and β_{CAS} per contact type.

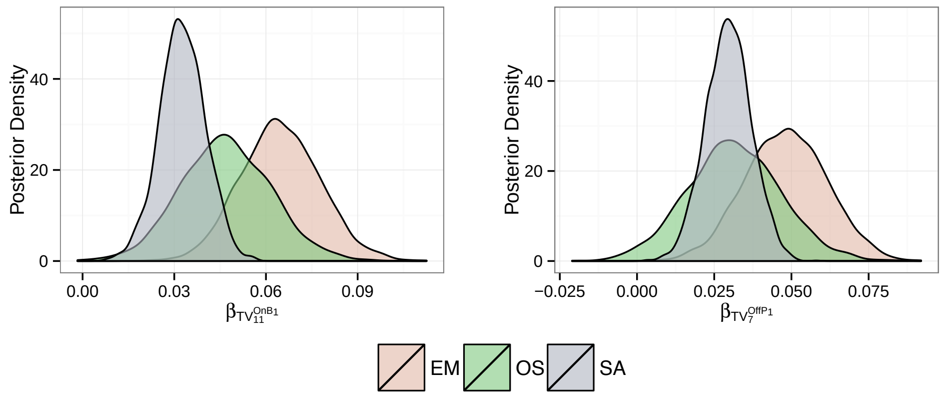

FIGURE 5.4: Significant densities of $\beta_{TV_{11}^{OnB_1}}$ and $\beta_{TV_7^{OffP_1}}$ per contact type.

of Figure 5.4 shows the impact of the online brand spot #1, station #11, on the probability for e-mail conversions, organic search conversions and search engine advertising conversions. The increase of the probability for search engine conversions after a TV spot is relatively intuitive and in line with previous studies (e.g., Kitts et al., 2010; Zigmond and Stipp, 2010). The interpretation of the value for $\beta^{EM}_{TV_{11}^{OnB_1}}$ needs more explanation, because the increase of the probability for e-mail conversions after a TV spot is rather surprising. The reason might be the concurrent e-mail marketing campaign that was dominated by providing discounts coupons for online purchases. Apparently, customers who had received such a coupon have been likely to reopen that e-mail and use the coupon to purchase a product after watching a brand-related TV spot. This finding is consistent for all TV spots and has managerial implications that are discussed in the next chapter. The effect of an offline product-related spot is presented on the right side of Figure 5.4 and shows a similar picture as the online-related brand spot #1. The effect of direct type-ins is not significant in this case because it is rather uncommon to directly type-in the product-related URL in the browser.

The left side of Figure 5.5 shows the cross-channel effects for the offline brand spot on station #6. As seen before, the highest positive impact of the TV spot can

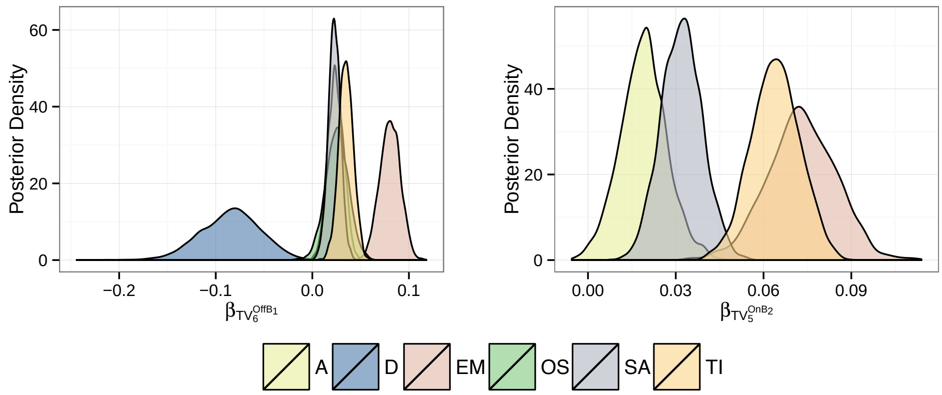

FIGURE 5.5: Significant densities of $\beta_{TV_6^{OffB_1}}$ and $\beta_{TV_5^{OnB_2}}$ per contact type.

be observed for e-mail conversions, followed by the impact on organic search conversions and search engine advertising conversions. Additional significant positive effects are observed for affiliate conversions and direct type-ins. The increase of the probability of affiliate conversions might be caused by the tendency of some customers to search for discount vouchers before visiting the website and purchasing a product. The graph included on the right side of Figure 5.5 presents the effect of the second online brand spot on station #5. It draws a very similar picture. The only difference is the missing negative effect on display conversions. This might be again due to the fact that the advertiser also operates retail stores and cross-channel saturation effects (Piercy, 2012).

In summary, the application of the hierarchical model provides new insights into cross-channel effects that determine the user behavior. The AUC for this model is 0.7606 and thereby higher than the analog model without the quadratic terms from the non-hierarchical analysis. In particular in the context of real-time advertising, where millions of decisions are made every second, such a model is valuable to determine the size of the bids on the basis of the predicted probability for a conversion.

5.6 Implications, Limitations and Outlook

In this paper, a model commonly used for clickstream analyses is extended by introducing offline advertising effects. Such an extended model outperforms the simple clickstream model in terms of predictive accuracy. Adding a hierarchical structure, the cross-channel effects of both online and TV advertising activities are revealed.

5.6.1 Implications

The approach presented in this study has several implications for researchers in the fields IS and marketing. First, future research needs to be done on decision support systems in e-commerce that integrate tracking data and offline advertising data from other data sources to improve the quality of strategic decisions, for example the allocation of parts of the marketing budget. Since customer behavior is influenced by many external factors such as seasonal effects (as in this case study) or new brands or products, analyses based on user data should be updated frequently to obtain reliable results. Therefore, efficient tools to manage and analyze these high-volume tracking data and data from offline sources are required to make the model viable in practice, for example in strategic decision support systems or bidding agents in real-time advertising (Stange and Funk, 2014). Second, for future research at the intersection of IS and marketing, data from tracking systems that identify users across different devices should be used to arrive at a better understanding of users' behavior in e-commerce contexts. In combination with data from traditional offline advertising channels such as print and television advertising, the model presented in this paper can provide even more precise insights into cross-device and cross-media user behavior. Third, marketing researchers can use this approach to analyze the impact of additional properties of offline advertising, such as information about television stations, the positions of the spots within the advertising blocks, or the context in which the spot was aired. These model extensions can yield better predictive accuracies and improve strategic decision support, for example for future advertising campaigns.

In addition, the study has several managerial implications: First, a model to estimate time dependency of the short-term TV spot effects is proposed. The results show that the strongest effect is observed 15 minutes after a spot has been broadcast. Bidding agents in real-time advertising can use this information to select ads related to the product advertised within that time frame to optimize profits from cross-channel marketing. Second, the model proposed here can be used in practice to plan future online advertising campaigns, i.e., the allocation of future marketing budgets. For instance, advertising activities on channels that show a relatively high conversion rate (in the case study, affiliate marketing, search engine advertising) could be expanded. The same is valid for TV advertising activities: The spendings for TV ads on stations that show a high positive effect on the purchase probability could be increased whereas the number of spots on stations without a noticeable effect might be reduced. To optimize the budget the estimated effects per channel have to be weighted in light of the costs per acquisition for each channel. Third, the results of the hierarchical analysis show that the impact of TV advertisement is different for each online channel. Companies who consider a TV advertising campaign should coordinate advertising activities on online channels that greatly benefit from TV advertising activities, for example search engine advertising, organic search, and e-mail advertising. For instance, if an advertiser runs a campaign for a certain product on television, it would make sense to run a search engine advertising campaign during that time to increase the spillover effect from the TV ad. Surprisingly, in this study, the highest impact of TV ads on the conversion probability can be observed for the e-mail channel. During an advertising campaign for a certain product on TV, an advertiser should also conduct a complementary e-mail campaign to optimize the spillover effect from the TV channel. This is but one example emphasizing the importance of a well-balanced cross-channel marketing strategy.

5.6.2 Limitations

Although it clearly reveals the expected TV effects, the analysis has primarily seven limitations. These are mostly related to incomplete data. First, to model the impact of television ads, it is crucial to have an idea of how many customers might have been watching a certain TV station at a given time. While it is possible to determine when a spot was aired and where individual spots where placed in the advertising blocks, it is hard to acquire reliable data on how many people actually saw these ads. This information should be still included in future analyses. Second, the device information of each contact was unavailable. However, this information is important, because it is expected that the probability for a conversion on a mobile tablet, which is used as a second device while watching TV, is higher than for a conversion on a standard PC, which is usually not instantly available when watching TV. Future studies in this area need to include information about user devices to build a model that can validate this hypothesis. Third, to estimate the TV impact, a simple logistic regression and a hierarchical logistic regression model were used. These models do not restrict the impact of TV spots in any form, which might be useful in some cases. For instance, it is rather unlikely that the impact of TV ads is less than zero. Therefore, it could make sense to change the prior information of the TV effects β_{TV} from multivariate normal to truncated multivariate normal. However, efficient algorithms to estimate the parameters for this kind of model are currently unavailable[4]. Fourth, the focus was on conversions here and the activities of a user on the retailer's website were ignored. However, it would also be very important to analyze onsite user behavior with a focus on products and purchase intentions depending on TV advertising activities. Fifth, the increase in the AUC value resulting from the addition of the parameters for TV spots and from the hierarchical model structure were small in comparison to the increase that resulted from the introduction of the γ exponents and the quadratic terms. This fact emphasizes that the feature selection process is more important than a sophisticated model structure. Sixth, the data was collected during Christmas time and, thus, includes many shorter user journeys suggesting the spontaneous purchase of gifts. Although the modeling approach presented here is generally applicable for offline-online studies, the results from the case study can not be transferred to other periods without limitation. Seventh, the focus of this study

[4] In fact, TV ads can have a negative effect on conversion probabilities as it is shown in Chapter 6. By contrast, a negative effect of TV ads on the total number of visits or conversions is rather unlikely.

was on the short-term effects of TV advertising activities. As indicated by previous studies, TV advertising focuses, however, on the awareness of potential customers (Liaukonyte et al., 2015). It was not possible to model the awareness effect, because user journeys without the influence of a TV campaign are not contained in the data set. For experimental settings in future studies, it would be desirable to measure both short-term and long-term TV advertising effects.

5.6.3 Outlook

The analysis of cross-channel marketing effects is often limited to online marketing channels, because they enable advertisers to track every customer activity. However, online marketing is only one part of the broad range of marketing activities that, ideally, are seamlessly coordinated. The analysis of the effect of online-offline marketing activities is a chance for practitioners to improve their marketing activities. For instance, an important question for practitioners is how to allocate parts of the budget to individual channels. The estimations of the proposed model can be used to answer this question. In addition, the measurement of online-offline advertising effects offer many opportunities for future scholarship at the intersection of IS and marketing. This analysis represents a first step in this direction.

5.7 References

Anderl, E., I. Becker, F. Wangenheim, and J. H. Schumann (2014). "Mapping the Customer Journey: A Graph-Based Framework for Online Attribution Modeling". In: *SSRN 2343077*.

Bolten, B. (2014). *SCREEN LIFE 2014; http://www.ipsos.de/assets/files/presse/2014/publikationen/Ipsos_Screen%20Life%202014_Praesentation_final.pdf (retrieved on 17 January 2015).*

Bucklin, R. E. and C. Sismeiro (2003). "A Model of Web Site Browsing Behavior Estimated on Clickstream Data". In: *Journal of Marketing Research* 40.3, pp. 249–267.

Bucklin, R. E. and C. Sismeiro (2009). "Click here for Internet insight: Advances in clickstream data analysis in marketing". In: *Journal of Interactive Marketing* 23.1, pp. 35–48.

Chatterjee, P. (2008). "Are unclicked ads wasted? Enduring effects of banner and pop-up ad exposures on brand memory and attitudes". In: *Journal of electronic commerce Research* 9.1, pp. 51–61.

Chatterjee, P. (2010). "Multiple-channel and cross-channel shopping behavior: role of consumer shopping orientations". In: *Marketing Intelligence & Planning* 28, pp. 9–24.

Chatterjee, P., D. L. Hoffman, and T. P. Novak (2003). "Modeling the clickstream: Implications for web-based advertising efforts". In: *Marketing Science* 22.4, pp. 520–541.

Cheema, A. and P. Papatla (2010). “Relative importance of online versus offline information for Internet purchases: Product category and Internet experience effects”. In: *Journal of Business Research* 63.9, pp. 979–985.

Cho, C.-H. (2003). “Factors influencing clicking of banner ads on the WWW.” In: *Cyberpsychology & behavior* 6.2, pp. 201–215.

Courtois, C. and E. D’heer (2012). “Second screen applications and tablet users: constellation, awareness, experience, and interest”. In: *Proceedings of the 10th European conference on Interactive tv and video*, pp. 153–156.

Dalessandro, B. and C. Perlich (2012). “Causally motivated attribution for online advertising”. In: *Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy*.

Denwood, M. J. (2015). *R Package runjags; http://cran.r-project.org/web/packages/runjags/index.html (V2.0.1-2, retrieved on 18 Apr 2015)*.

Dinner, I., H. J. van Heerde, and S. A. Neslin (2014). “Driving Online and Offline Sales: The Cross-channel Effects of Traditional, Online Display, and Paid Search Advertising”. In: *Journal of Marketing Research* 51.5, pp. 527–545.

Ghose, A. and S. Yang (2009). “An Empirical Analysis of Search Engine Advertising: Sponsored Search in Electronic Markets”. en. In: *Management Science* 55.10, pp. 1605–1622.

Joo, M., K. C. Wilbur, B. Cowgill, and Y. Zhu (2014). “Television Advertising and Online Search”. In: *Management Science* 60.1, pp. 56–73.

Jordan, P., M. Mahdian, S. Vassilvitskii, and E. Vee (2011). “The multiple attribution problem in pay-per-conversion advertising”. In: *Algorithmic Game Theory*. Springer Berlin Heidelberg, pp. 31–43.

Kitts, B., L. Wei, D. Au, A. Powter, and B. Burdick (2010). “Attribution of conversion events to multi-channel media”. In: *Proceedings - IEEE International Conference on Data Mining, ICDM*, pp. 881–886.

Klapdor, S. (2013). *Effectiveness of Online Marketing Campaigns*. Wiesbaden: Springer Fachmedien.

Lee, K., B. Orten, A. Dasdan, and W. Li (2012). “Estimating conversion rate in display advertising from past erformance data”. In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, Beijing, China*, pp. 768–776.

Lewis, R. A. and D. H. Reiley (2013). “Down-to-the-Minute Effects of Super Bowl Advertising on Online”. In: *14th ACMConference on Electronic Commerce* 9.4, pp. 1–22.

Liaukonyte, J., T. Teixeira, and K. C. Wilbur (2015). “Television Advertising and Online Shopping”. In: *Marketing Science* 34.3, pp. 311–330.

Lodish, L., M. M. Abraham, J. Livelsberger, B. Lubetkin, B. Richardson, and M. E. Stevens (1995). “A Summary of Fifty-Five In-Market Experimental Estimates of the Long-Term Effect of TV Advertising”. In: *Marketing Science* 14.3, pp. 133–140.

Naik, P. and K. Peters (2009). “A Hierarchical Marketing Communications Model of Online and Offline Media Synergies”. In: *Journal of Interactive Marketing* 23.4, pp. 288–299.

Nottorf, F. (2014). "Modeling the clickstream across multiple online advertising channels using a binary logit with bayesian mixture of normals". In: *Electronic Commerce Research and Applications* 13.1, pp. 45–55.

Olbrich, R. and C. D. Schultz (2014). "Multichannel advertising: does print advertising affect search engine advertising?" In: *European Journal of Marketing* 48.9/10, pp. 1731–1756.

Piercy, N. (2012). "Positive and negative cross-channel shopping behaviour". In: *Marketing Intelligence & Planning* 30.1, pp. 83–104.

Plummer, M. (2003). "JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling". In: *Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna*.

Richardson, M., E. Dominowska, and R. Ragno (2007). "Predicting clicks: estimating the click-through rate for new ads". In: *Proceedings of the 16th international conference on World Wide Web*, pp. 521–530.

Rossi, P. E., G. M. Allenby, and R. McCulloch (2006). *Bayesian statistics and marketing*. Wiley, pp. 136–142.

Rutz, O. J., M. Trusov, and R. E. Bucklin (2011). "Modeling Indirect Effects of Paid Search Advertising: Which Keywords Lead to More Future Visits?" In: *Marketing Science* 30.4, pp. 646–665.

Shao, X. and L. Li (2011). "Data-driven multi-touch attribution models". In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 258–264.

Stange, M. and B. Funk (2014). "Real-Time Advertising". In: *Business & Information Systems Engineering* 56.5, pp. 305–308.

Stange, M. and B. Funk (2015). "How Much Tracking Is Necessary - The Learning Curve in Bayesian User Journey Analysis". In: *Proceedings of the 23rd European Conference on Information Systems, Muenster, Germany*.

Veit, D., E. Clemons, A. Benlian, P. Buxmann, T. Hess, D. Kundisch, J. M. Leimeister, P. Loos, and M. Spann (2014). "Business Models". In: *Business & Information Systems Engineering* 6.1, pp. 45–53.

Yau, C. (2015). *R Package RPUD 0.5.0; Retrieved from http://www.r-tutor.com/gpu-computing, January 07, 2015.*

Zigmond, D. and H. Stipp (2010). "Assessing a New Advertising Effect Measurement of the Impact of Television Commercials on Internet Search Queries". In: *Journal of Advertising Research* 50, pp. 162–168.

5.8 Erratum

During the data preparation for the study presented in Chapter 6, it was found that the results of the article presented here are based on an inconsistent sample of user journeys: First, user journeys of length 1 that did not result in a conversion have not been included in the original analysis. User journeys of length 1 that resulted in a conversion remained in the sample but the dependent variable was falsely set to zero. This issue resulted from a programming error in the Java code used to generate the sample. As a consequence, the AUC values reported in the original version of the article (Table 5.6) are greater than the ones achieved with a corrected sample (Table 5.9). In addition, the values of several model parameters are biased (Table 5.7). Second, and more problematic, the two data sets used to measure the effect of TV ads stem from two servers using two different time zones. More precisely, the TV ads were aired one hour before the time suggested by the time stamps. Consequently, the measured short-term effect of TV ads on website traffic reflects the effect of TV ads on website traffic after one hour. Figure 5.6 shows the aggregated number of visits over time around each ad. The spike in website traffic is shifted by one our.

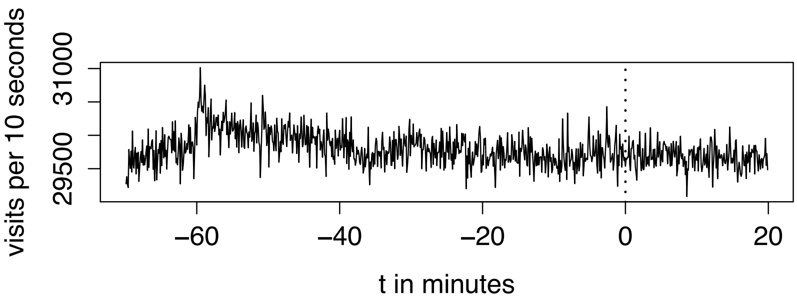

FIGURE 5.6: Illustration of the time shift between the two data sets.

These inconsistency issues, however, do not limit the main contribution of this article, since it focuses on developing and evaluating a novel tool to measure the effect of TV ads on online behavior at the level of the individual user that could be used to improve cross-channel campaigns. For this reason and to preserve the originality of the published article, it was decided to include it into this dissertation without changes.

Nevertheless, using a corrected sample, the approach was tested again for the purpose of this dissertation. In particular, it was investigated why the inconsistencies were not noticed in the first place. During this process, many findings included in the original article could be confirmed. Table 5.8 (the reproduction of Table 5.6) confirms, for instance, that the variance is reduced when TV ads are included into the model and that the inclusion of the non-linearity terms results in the highest increase of the AUC values.

However, the investigation also revealed a major methodological drawback of the original method in combination with the data set used. This drawback concerns the correlation of effects related to TV ads and effects related to the time

Model	R without TV	R with TV	Dev. Delta	Sign.	Dev. Diff.	AUC w. TV
Plain	222899.6	222819.6	80.0	***	-	0.6402
Gamma	222899.6	222843.8	55.8	***	-24.2	0.6403
Times	222647.0	222606.5	40.4	***	237.3	0.6417
Non-Linearity	220380.5	220342.4	38.2	***	2264.2	0.6616
Quadratic	219924.9	219886.2	38.7	***	456.2	0.6631

TABLE 5.8: Residual Deviances R for 10 different models and AUC measures using a corrected sample ($N = 500{,}000$).

when ads were aired (Section 5.8.1). This correlation cannot be captured sufficiently using the modeling approach proposed in the original article and emphasizes the need for a thorough feature selection process. To reduce this correlation and the associated omitted-variable bias, Section 5.8.2 discusses an alternative modeling approach that is more robust in this regard, even without including additional time-related variables into the model.

5.8.1 Correlation of Effects Related to Time and TV Ads

Although they have been generated based on an inconsistent sample, the results presented in the published version of the article show statistical significance in many cases. This contradiction suggests that the data includes unobserved time-dependent effects on conversion probabilities that correlate with the time when ads were aired on TV. To provide deeper insight concerning this correlation, Table 5.9 shows cross-validated AUC values of multiple logistic regression analyses using different parameterizations of the Γ-shaped TV effects. The table is ordered by AUC_{Γ}, which represents the AUC generated with the Γ-shaped transformation of TV variables as described in Section 5.4.2. The value AUC_T is generated with uniform parameterizations where the TV effect is set to 1 if the time difference between TV ad and visit is less than T, and 0 otherwise. The value for T is determined by the 99% quantile of the respective Γ distribution floored to the next integer that is divisible by 5. The numbers show that the highest AUC can be achieved with a uniform parameterization of TV effects with $T = 225$ minutes (first row column AUC_T). The value of AUC_{Γ} from the same row is smaller. Consequently, the Γ-shaped parameterization is less accurate than the uniform parameterization in this case. This finding suggests that the results do not represent short term TV effects but effects that relate to the time of the day in which the ads were aired, and, consequently, that only 5 control variables with respect to time (see Equation 5.1) are not sufficient to explain all time-related effects.

The values in the second row of Table 5.9 draw a different picture: Both AUC values are smaller than in the first row. However, the value of $AUC_{\Gamma(2,0.67)}$ is greater than $AUC_{T=5}$, which suggests that TV advertising effects dominate time-dependent effects in this case. However, the complete separation of effects related to time and effects related to TV ads requires more effort. Obviously, higher AUC values can be achieved by considering larger time frames after spots were

$E(\Gamma(\alpha,\beta))$	Mode$(\Gamma(\alpha,\beta))$	α	β	AUC$_\Gamma$	T (min)	AUC$_T$
50.000	1.170	1.02	0.02	0.6429	225	0.6443
3.000	1.500	2.00	0.67	0.6412	5	0.6402
2.000	0.500	1.33	0.67	0.6411	5	0.6401
4.715	1.170	1.33	0.28	0.6407	15	0.6401
10.000	1.500	1.18	0.12	0.6406	40	0.6414

TABLE 5.9: Cross-validated AUC values for different parameterizations of the Γ-shaped TV effects.

aired. This is the reason for the statistical significance of the results presented in the published version of the article. The results might involve effects that are in fact caused by TV ads which endure for over 90 minutes; however, Table 5.9 suggests that the TV ad variables mainly mediate effects that relate to the time when they were aired.

As additional challenges, the effects of TV ads on conversion probabilities are very small in comparison to the effects of online ads, heterogeneous across referral channels, and not constant over the course of a day. In light of this finding and the correlation between time-related effects and TV advertising effects, the original form of the Bayesian model presented in Section 5.4.2 is not able to adequately find the shape and scale of the Γ-shaped TV effect. Consequently, highly informative priors need to be used to identify the time-dependent TV ad effects more accurately. Figure 5.7 presents the results of this approach with a maximum effect at $t = 70$ seconds.

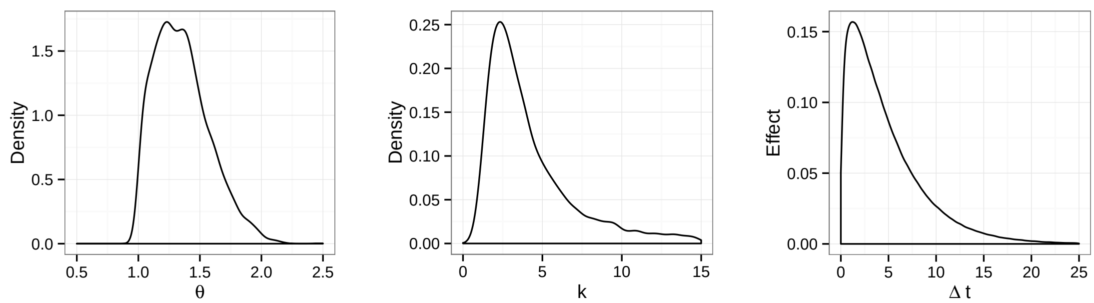

FIGURE 5.7: Densities of θ, k and the resulting time-dependent TV advertising effect generated with a corrected sample and informative priors ($1/\theta \sim$ Exponential(1); $k \sim N(\mu = 1.1, \sigma = 0.1)$).

In summary, the data used here is highly influenced by unobserved time-dependent effects, which is why the inconsistencies of the original sample were not noticed in the first place. These time-dependent effects can be explained by including TV variables, since TV ads are not aired randomly over time and, thus, correlate with effects related to time. However, a distinct causal effect of TV ads on online customer behavior cannot be derived from this kind of analysis. As discussed in Sections 1.4.2, 5.5.3, and 5.6.2, this finding emphasizes that a thorough feature selection process is crucial to clearly identify the effects of offline ads on online purchasing behavior.

5.8.2 Modeling Alternative

In light of the drawbacks described above, this section provides an alternative approach to model TV ads at the level of the individual user, which is closely related to the model developed in Section 6.4.1. The model is based on the probability that a given user session is a response to a TV ad. This probability is then used as independent variable in the user journey model. However, to apply the modeling approach discussed in Section 6.4.1 to the data set used here, three changes are required: First, the data set used here does not contain information on TV ad expenditures, which can be used as proxy variable to describe the reach of ads. Instead of expenditures, information on TV stations and time of the day, which also correlate with reach, are included into the model presented here. Second, information on device types is not available in the data set used here and, therefore, cannot be included into the model. Third, the signal to noise ratio is much smaller in the data used here, because it was recorded during the Christmas Time and stems from an online shop with a much higher baseline traffic. Therefore, the uplift in website traffic is modeled by only two parameters to reduce the degrees of freedom of the model.

Identifying the Spike in Website Traffic

To clearly identify the spike in website traffic caused by a TV ad, a constant baseline traffic is required in the minutes before and after an ad was aired, i.e., no other factors, such as the time of the day, should influence the number of visits per time interval around each ad. Figure 5.8 shows the aggregated website traffic over time for two different time frames (red lines). In the morning hours, the number of visits increases over time, while, in the evening hours, the number of visits decreases over time. This biased baseline traffic needs to be corrected to identify the uplift of website traffic caused by TV ads. This can, for instance[5], be achieved by fitting a linear model and subtracting the fitted website traffic without the intercept (blue lines in Figure 5.8).

[5]Alternatively, this could be achieved by using a time series model to extract seasonal (daily), random, and trend components.

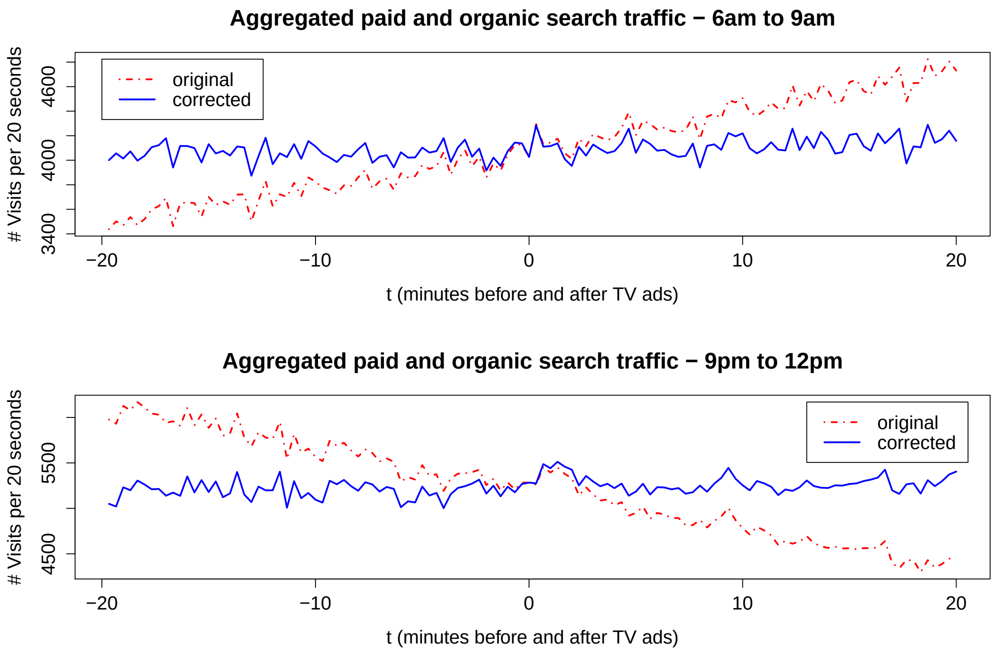

FIGURE 5.8: Measured and corrected aggregated website traffic.

Modeling the Increase in the Number of Sessions

The aggregated number of sessions Δt minutes after an ad was aired is described with Equation 5.6. It considers the station S on which an ad was aired, the referral channel r and the time of the day τ. The model is a variant of the model described with Equation 6.2 (Section 6.4.1).

$$y^{r,\tau,S}(\Delta t) \quad = \quad a^{r,\tau,S} \cdot [1 + b^r \cdot c^\tau \cdot d^S \cdot \Gamma(\Delta t, \alpha, \beta)] + \epsilon^{r,\tau,S}(\Delta t) \tag{5.6}$$

In this equation, $y^{r,\tau,S}(\Delta t)$ represents the measured aggregated website traffic established via referral r, at time of the day τ, Δt minutes after an ad was aired on TV station S. The variable $a^{r,\tau,S}$ represents the respective baseline traffic. An example on how to aggregate website traffic with respect to TV ads is provided in Table 6.7. The product of the parameters b^r, c^τ, d^S, and the Γ term determine the relative uplift of website traffic. The term $\Gamma(\Delta t, \alpha, \beta)$ returns the value of the Γ distribution parameterized with shape α and rate β at $x = \Delta t$.

The baseline traffic $a^{r,\tau,S}$ is sampled from an uninformative Γ distribution. The parameters b^r, c^τ, and d^S are sampled from uniform distributions stemming from different value ranges. This approach assures that the model has only one solution. Otherwise the parameters b^r, c^τ, and d^S would not be identifiable, i.e., the MCMC algorithm would result in inconsistent parameter distributions. This is because only the product of the three parameters is identifiable. The parameters α and β are sampled from two Γ distributions with an expected value of $1.5/0.5 = 3$ and a mode of $(1.5 - 1)/0.5 = 1$:

$$\begin{aligned} a^{r,\tau,S} &\sim \Gamma(1, 0.02) \\ \alpha, \beta &\sim \Gamma(1.5, 0.5) \\ b^r &\sim \text{Uniform}(0, 1) \\ c^\tau &\sim \text{Uniform}(1, 10) \\ d^S &\sim \text{Uniform}(10, 100) \\ \epsilon^{r,\tau,S}(\Delta t) &\sim \text{Normal}(0, 0.1) \end{aligned} \tag{5.7}$$

The highest density intervals of the posterior densities of b^r, c^τ, and d^S are presented in Table 5.10.

Parameter	2.5%	50%	97.5%	Parameter	2.5%	50%	97.5%
α	1.08283	1.09515	1.10777	c^{3-6pm}	2.00362	2.08984	2.17297
β	0.05807	0.05977	0.06159	c^{6-9pm}	9.73400	9.94600	9.99787
b^A	0.00028	0.00041	0.00054	c^{9-12am}	1.00022	1.00520	1.02767
b^D	0.00000	0.00005	0.00023	d^1	83.13984	87.44710	92.13300
b^{SEA}	0.00262	0.00270	0.00279	d^2	42.12270	43.80610	45.59487
b^{OS}	0.00156	0.00164	0.00172	d^3	99.02144	99.81170	99.99460
b^{SM}	0.00000	0.00009	0.00045	d^4	44.64786	46.29695	48.04272
b^{EM}	0.00063	0.00079	0.00096	d^5	64.94450	67.16835	69.51593
b^{PS}	0.00005	0.00062	0.00157	d^6	41.15369	42.65200	44.11290
b^{TI}	0.00152	0.00157	0.00163	d^7	10.00060	10.02150	10.11810
b^R	0.00193	0.00207	0.00221	d^8	53.07410	54.77690	56.55617
c^{12-3am}	1.10511	3.52036	9.20395	d^9	57.38218	61.91110	66.54130
c^{3-6am}	1.01731	1.42753	2.68978	d^{10}	10.24854	14.10075	20.65910
c^{6-9am}	2.55406	2.74914	2.93975	d^{11}	49.21148	50.91740	52.82062
c^{9-12pm}	2.71238	2.84095	2.97217	d^{12}	10.00520	10.16470	10.89951
c^{12-3pm}	1.16096	1.27482	1.38628	-	-	-	-

TABLE 5.10: Highest density intervals of the posterior distributions.

Several posterior distributions reported in Table 5.10 are located at the edges of the allowed value ranges defined by the prior distributions. This finding suggests that the true relative difference in uplift is larger than the ratio of the minimum and the maximum of the respective value range. Adjusting the prior value ranges of the model could resolve this issue.

Figure 5.9 illustrates the marginal relative uplift of different referrals, times of the day, and TV stations. It shows, for instance, that the uplift of paid search referrals caused by ads aired on Station 3 between 6pm and 9pm is high as compared to other combinations of referral, time, and station. In combination with information on the TV ad expenditures, which is not included in the data set used here, the results could be used to determine the cost per lead per TV station.

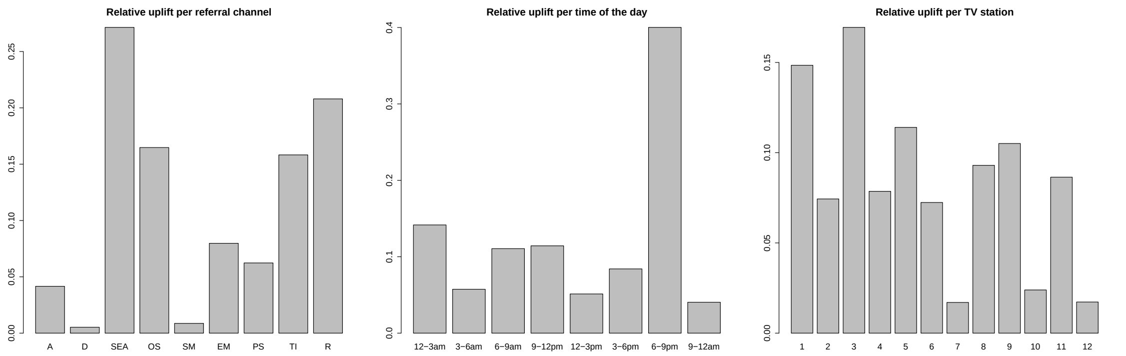

FIGURE 5.9: Relative uplift per channel, time of the day, and station.

Calculating the Probability for Being TV-Induced

In addition to providing insights concerning the uplift of visits, the posterior estimates presented in Table 5.10 can be used to calculate the probability that a given user session is a direct response to a certain TV ad (Equation 5.8).

$$\begin{aligned} p^{r,\tau,S}(\Delta t) &= \frac{\text{Uplift}^{r,\tau,S}(\Delta t)}{\text{Baseline}^{r,\tau,S} + \text{Uplift}^{r,\tau,S}(\Delta t)} \\ &= \frac{a^{r,\tau,S} \cdot [b^r \cdot c^\tau \cdot d^S \cdot \Gamma(\Delta t, \alpha, \beta)]}{a^{r,\tau,S} + a^{r,\tau,S} \cdot [b^r \cdot c^\tau \cdot d^S \cdot \Gamma(\Delta t, \alpha, \beta)]} \\ &= \frac{b^r \cdot c^\tau \cdot d^S \cdot \Gamma(\Delta t, \alpha, \beta)}{1 + b^r \cdot c^\tau \cdot d^S \cdot \Gamma(\Delta t, \alpha, \beta)} \\ &= \frac{1}{1 + [b^r \cdot c^\tau \cdot d^S \cdot \Gamma(\Delta t, \alpha, \beta)]^{-1}} \end{aligned} \tag{5.8}$$

This equation can be used to calculate the probability that a user opened the website in response to a TV ad even without knowing the ads' expenditures. Given a TV ad was aired on Station 3 at 7pm, the probability that a user who opens the website at 7:01pm by clicking a search engine advertising link has seen the ad is $p^{SEA,\text{6-9pm},3}(\Delta t = 1 \text{ min}) \approx 0.108$. Sessions established from 3pm to 6pm via paid search one minute after an ad was aired on Station 3 are less likely to be TV-induced ($p^{SEA,\text{3-6pm},3}(\Delta t = 1 \text{ min}) \approx 0.025$).

The total probability that a given user opening the website via referral r has seen an ad of a certain type can be approximated using Equation 5.9:

$$p^r(t) = 1 - \prod_i \left\{ 1 - p_i^{r,\tau,S}(t - t_i) \right\} \tag{5.9}$$

Non-Hierarchical Analysis

The probability described with Equation 5.9 can be used as independent variable in a user-level model. The data set used here contains five different TV ads.

Therefore, five different variables are generated using Equation 5.9. For instance, the probability that a given visit is a direct response to an offline-related ad for product #1 is denoted as p_{OffP_1}. The results of a non-hierarchical logistic regression that includes the non-linearity terms from Equation 5.3 are presented in Table 5.11.

Var.	Est.	SD	z	Sig. Lv.	Var.	Est.	SD	z	Sig. Lv.
I_0	-2.866	0.003	-861.941	***	K_{SEA}	0.037	0.005	7.868	***
I_{OS}	-0.268	0.006	-45.817	***	K_D	0.014	0.004	3.940	***
I_{TI}	0.082	0.006	13.826	***	K_{SM}	0.003	0.004	0.643	
I_A	0.142	0.004	35.869	***	K_{EM}	0.030	0.003	8.744	***
I_D	-0.178	0.005	-35.811	***	K_{PS}	-0.011	0.004	-2.758	**
I_{SEA}	-0.085	0.006	-13.146	***	K_R	-0.094	0.007	-13.567	***
I_{SM}	-0.086	0.006	-15.479	***	SN	-0.405	0.010	-40.374	***
I_{EM}	0.056	0.004	13.874	***	IST	0.397	0.005	86.008	***
I_{PS}	-0.091	0.005	-18.294	***	CWS	0.058	0.002	24.994	***
J_{TI}	0.091	0.002	43.113	***	CAS	0.112	0.003	40.232	***
J_A	0.054	0.002	24.082	***	t	1.022	0.083	12.317	***
J_D	0.018	0.003	6.665	***	t^2	-3.258	0.352	-9.255	***
J_{SEA}	-0.048	0.004	-13.732	***	t^3	3.952	0.493	8.017	***
J_{SM}	0.007	0.003	2.797	**	t^4	-1.764	0.223	-7.905	***
J_{EM}	0.041	0.002	18.144	***	WE	-0.017	0.003	-5.597	***
J_{PS}	0.024	0.003	8.499	***	p_{OnB_1}	0.016	0.004	4.407	***
J_R	-0.061	0.004	-16.511	***	p_{OffP_1}	0.013	0.004	3.423	***
K_{OS}	-0.010	0.004	-2.441	*	p_{OffB_1}	-0.011	0.004	-3.051	**
K_{TI}	-0.009	0.007	-1.406		p_{OnB_2}	0.012	0.004	3.373	***
K_A	-0.062	0.003	-18.708	***	p_{OffP_2}	-0.017	0.004	-4.437	***

TABLE 5.11: Results generated with the alternative model (N = 2,000,000). Signif. codes: *** p < 0.001; ** p < 0.01; * p < 0.05; . p 0.1.

The results presented in Table 5.11 show that the probabilities $p_{(\cdot)}$ significantly affect the probability to convert. Online-related ads seem to positively affect the probability to convert, whereas the offline-related ad has a negative effect. To validate these results, an 8-fold cross-validation was conducted. The mean AUC of the model that includes the probabilities $p_{(\cdot)}$ is 0.6613, while the AUC of the model that does not include these parameters is 0.6612. However, this difference is smaller than a standard deviation of the cross-validated values of the AUC ($\sigma_{AUC} \approx 0.001$). This finding suggests that the short-term TV advertising effect does not play an important role in predicting behavior at the level of the individual customer in this study. This finding also emphasizes the need to spend more effort on feature selection in this context, as it was mentioned before. Considering interactions between different advertising activities, for instance, could improve predictive accuracy as suggested by the results of the hierarchical analysis conducted here (Section 5.5.5).

Hierarchical Analysis

This paragraph discusses the results of a hierarchical analysis that was conducted to provide insights concerning cross-channel advertising effects using the probability variables described above. The model closely follows the one presented in Section 5.4.3. Figure 5.10 illustrates the effects of different TV ads on the conversion probability after opening the shop's website via different online channels.

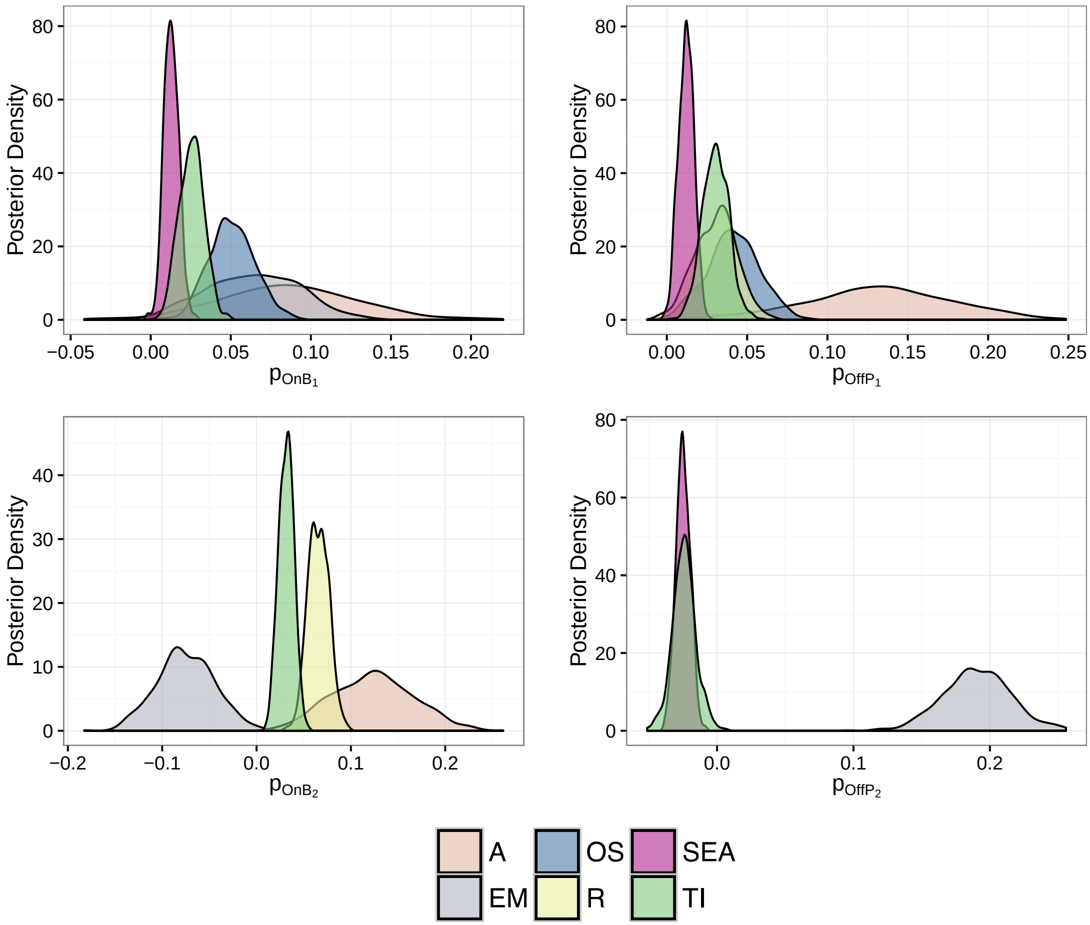

FIGURE 5.10: Significant densities (95% HDI) of $\beta^{G}_{p(\cdot)}$ per ad. The missing ad, $OffB_1$, only shows a significant effect on conversions after clicks on e-mail links ($\bar{\beta}^{EM}_{p_{OffB_1}} \approx -0.16$).

These results clearly show that some online channels are more and others are less strongly affected by TV ads in terms of the individual user's conversion probability. For instance, the probability to convert after clicking an affiliate link seems to benefit more from TV ads than the probability to convert after typing-in the URL directly (left side and top right side of Figure 5.10). However, in light of the correlation of time and TV effects described above, this finding should more be regarded as a suggestion for further research than as a distinct result. For instance, the different valence of the TV effects on the probability to convert after clicking an e-mail link (positive for OnB_1, $OffP_2$, and negative for $OffB_1$, OnB_2) needs further investigation. An even more precise separation of TV effects and time dependent effects could be useful in this regard. One way to close this gap could be to omit the aggregation step in preparation to the analysis described above and to use Equation 5.6 to estimate the increase in website traffic caused by each individual ad. This would lead to more accurate probability values. In addition, more variables related to time (e.g., day of the month and day of the week) should be included, which is particularly important for analyses of e-commerce data stemming from the Christmas season.

In summary, the approaches described in this chapter can be used to measure the effects of TV ads on online behavior at a new level. These kinds of analyses could answer open questions on the behavior of users who opened a shop's website in response to TV ads, for instance, in terms of conversion probability, shopping baskets, device usage, loyalty, or customer lifetime values. This kind of research could also lead to more robust theories on the specific synergies of offline and online advertising activities which could to involve many important managerial implications.

6 The Reduced Customer Revenue of TV-Induced Online Shoppers

Stange, M., N. Abou Nabout, C. Uhl, and B. Funk (2016). "The Reduced Customer Revenue of TV-Induced Online Shoppers". Working Paper (September 2016), Vienna University of Economics and Business.

Abstract

Second screen usage has dramatically increased within the last years. For instance, many customers spontaneously decide to visit an online shop in the few minutes after watching the shop's TV ad. In this study, we examine the behavior of these TV-induced customers along the sales funnel (including conversions, shopping baskets, and repeat purchase behavior) over a 90-day period. The empirical study is based on a unique data set provided by an international online retailer that runs extensive TV advertising. To measure the difference between TV-induced customers and their non-induced counterparts, we develop a Bayesian model to estimate the probability that a given visit to the online shop is a spontaneous reaction to a TV ad. Our results indicate that conversion probabilities and shopping baskets of TV-induced customers tend to be lower than the ones of their non-induced counterparts. As a consequence, we find that their 90-day customer revenue is approximately 10% lower. Advertisers are, thus, well advised to consider the difference in customer revenue when planning their marketing activities or calculating the return on investment of TV advertising campaigns.

6.1 Introduction

Second screen usage has dramatically increased within the last years (Statista, 2016) leading customers to spontaneously visit online shops after watching their ads on TV. In this article, we examine the online shopping behavior of these customers over a 90-day period. While prior literature suggests that TV ads increase the absolute number of visits to an online shop (Joo et al., 2014) and lead to an increased number of conversions (Liaukonyte et al., 2015), it is an open question as to what type of customers are attracted to spontaneously visit an online shop after a TV ad has been aired. More specifically, the shopping behavior of customers who open a shop's website in response to TV ads vs. the behavior of those that visit the shop for another reason (e.g., searching for a specific product using a price search engine, or clicking on a newsletter link) has yet to be investigated.

We call customers who open the website in response to TV ads TV-induced customers and investigate their purchasing decisions, including their probability to convert, the value of items in their shopping baskets, their repeat purchase behavior, and 90-day customer revenue (i.e., shopping baskets $\times$ repeat purchases over a 90-day period). More precisely, we aim to answer the following research questions:

- Are TV-induced customers more or less likely to convert than those who have not been induced by a TV ad to visit the shop?
- Do their basket sizes differ from the ones of their non-TV-induced counterparts?
- Do they return to the online shop as frequently as those customers that found their way to the online shop without seeing a TV ad?
- Does their 90-day customer revenue differ from those customers who have not been attracted to spontaneously visit an online shop after a TV ad has been aired?

To answer these research questions, we utilize a unique sample obtained from merging two data sets that were provided by an international online retailer. The first data set contains a total number of 4,403,866 visits to the retailer's website generated by 3,471,532 unique customers. The total number of conversions is 489,884. The second data set contains information on over 3,500 TV ads including their expenditures and the time when they were aired exact to the second. We transform the sample into customer journeys using a common approach for modeling clickstream data (Chatterjee et al., 2003). Based on this transformation, we estimate four different models: (1) conversion probabilities (logistic model), (2) shopping baskets (linear model), (3) repeat purchase behavior (negative-binomial model), and (4) the 90-day customer revenue (linear model). In contrast to most studies concerning the impact of TV ads on online behavior (Joo et al., 2014; Joo et al., 2015; Liaukonyte et al., 2015), we model the spontaneous effect of TV ads on online shopping at the level of the individual customer, which allows us to identify customers across multiple visits. To do so, we first model the probability that a visit is a spontaneous response to a TV ad and, second, use this probability as an independent variable in customer journey models, which primarily aim to explain differences in conversion probabilities, basket sizes, repeat purchases and customer revenues of TV-induced and non-TV-induced customers. We find that TV-induced customers are characterized by lower individual conversion probabilities, lower shopping baskets, and, consequently, lower customer revenues (over a 90-day period). We also show that the main driver behind their lower customer revenue is that their shopping baskets tend to be lower than the ones of their non-TV-induced counterparts. Lower shopping baskets could have different reasons: (1) TV-induced customers might want to try out the shop before engaging in more extensive buying later on, (2) they might be very price-sensitive, always looking for the best price which reduces their loyalty to a single shop, (3) they might be characterized by less purchase power than their non-TV-induced counterparts.

The remainder of this article is organized as follows: In Section 6.2, recent literature on measuring the effects of TV ads on browsing behavior and attribution modeling is reviewed and research gaps are identified. In Section 6.3, we describe our sample that was obtained from merging two distinct data sets, namely data about (1) TV ads and (2) online shopping behavior. In Section 6.4, we first present our approach to model the probability that a customer has been induced by a TV ad to visit the shop. Second, we introduce the models we use to explain the differences in conversion probability, shopping baskets, repeat purchases, and customer revenue of TV-induced and non-TV-induced customers. We present and discuss the results in Section 6.5 and summarize our contributions and managerial implications in Section 6.6.

6.2 Related Work

This section provides an overview of recent literature on measuring the effect of TV ads on search requests, number of visits, and conversions. Specifically, we review literature dealing with the effect of TV ads on online behavior, which gained in popularity with the emergence of second screen usage since 2010. In addition, we review literature concerning approaches to attribute conversions to TV advertising efforts. Table 6.1 provides an overview of recent studies dealing with TV ads and online behavior and illustrates key differences to our study. Previous studies mainly focus on the short-term influence of TV ads on customer behavior (Joo et al., 2014; Joo et al., 2015; Kitts et al., 2014; Liaukonyte et al., 2015). By contrast, we investigate the influence of TV ads on customer behavior over a 90-day period. For this purpose, we model TV ads at the level of the individual customer. In contrast to previous approaches based on aggregated data, this approach allows for estimating differences in customer behavior of TV-induced and non-TV-induced customers.

6.2.1 Effect of TV Ads on Online Behavior

One of the first articles that aim to make the effect of TV ads on online behavior measurable has been published by Zigmond and Stipp (2010). They present multiple empirical studies (concerning, for instance, TV ads during the opening ceremonies of the Olympic Games) and show that a clear increase in search queries related to the advertised products can be observed after TV ads have been aired. The authors propose using this uplift as one additional measure to better understand the effect of TV ads on customer behavior. This measure, however, can only be applied if TV ads reach a large audience (Lewis and Reiley, 2013). Since the majority of TV ads in our data set only has a small audience, though, their modeling approach is not applicable in our case.

Lewis and Reiley (2013) investigate the uplift in search requests caused by TV ads aired during the Super Bowl 2011 and observe that the tendency to go online and search for an advertised product or brand differs across brands and ads. For

Authors (Year)	Aim	Customer-level	Visit/search	Conversion	Shopping baskets	Repeat purchases	Customer revenue
This study	Customer revenue of TV-induced online shoppers	✓	✓	✓	✓	✓	✓
Liaukonyte et al. (2015)	Influence of TV advertising on online shopping behavior		✓	✓			
Joo et al. (2015)	Influence of TV advertising on online search behavior		✓				
Stange (2015)	Influence of TV ads on conversion probability	✓		✓			
Joo et al. (2014)	Influence of TV advertising on online search behavior		✓				
Kitts et al. (2014)	Attribution modeling for TV-induced conversions.		✓	✓			
Lewis and Reiley (2013)	Effect of different ads on brand search.		✓				
Zigmond and Stipp (2010)	Measuring the effect of TV ads on brand search		✓				

TABLE 6.1: Overview of related work.

instance, they find that the increase in the number of visits tends to be greater in response to movie trailers than to ads for consumer goods. In addition, they observe causal effects between TV ads and the increase in the number of visits by controlling for other brands that did not advertise during that period. However, the authors acknowledge that purchase intention and search requests do not perfectly correlate, which is particularly true for an event such as the Super Bowl, during which many customers are interested in the commercials as such and not necessarily in the related brand or product. This thought suggests that customers who open a website in response to a TV ad might just be curious to see what the shop is like and might, thus, have lower conversion rates than customers who have not been induced to visit the shop by TV ads.

In contrast to previous studies, Liaukonyte et al. (2015) measure the impact of TV ads at three different levels, going beyond search requests. The authors also include the absolute number of direct type-ins of the URL of an online shop and transactions (i.e., conversions/purchases) into their model. Thus, they have been the first to model customer shopping behavior in response to TV ads. They find that the effect of TV ads on direct type-ins and transactions highly depends on the kind of TV ad, whether it is information-focused, emotion-focused or imaginary-focused. They use aggregated data to estimate their models and, therefore, their approach cannot be used to extract characteristics of individual customers. We take a different approach and model the probability that a customer's visit was induced by a TV ad and use this probability as independent variable to model shopping baskets, repeat purchase behavior and customer revenue generated by TV-induced customers.

6.2.2 Attribution Modeling

Kitts et al. (2014) use the data set of Lewis and Reiley (2013) to develop a method that can be used to identify TV-induced visits even if the increase in the number of visits is not significant. Thereby, they show that it is possible to measure web activity bursts after the end of traditional TV ad broadcasts with smaller audiences. They argue that their method could be used to attribute conversions to TV ads. In contrast to the difference in difference approach proposed by Liaukonyte et al. (2015), they identify TV-induced visits by considering heterogeneous user responses with respect to time, geographic location, active device type, and referral channel. A similar approach is used in our empirical study, which considers heterogeneous reactions to TV ads using different device types and referrals. All of the above studies use aggregated data (number of search requests, number of direct visits or purchases) to estimate the effect of TV ads. Since, in this study, we aim to model the impact of TV ads on customer behavior over the entire customer journey, using aggregated data is not possible. Instead, answering our research questions requires a modeling approach that allows to measure the impact of TV ads on customer behavior over the entire customer lifetime. Therefore, we use a modeling approach that is based on an a commonly used clickstream model proposed by Chatterjee et al. (2003). This model has been extended by Stange (2015) to measure the impact of TV ads on conversion probabilities at the level of the individual customer. By using and extending this method, we are able to measure the difference of TV-induced customers and their non-TV-induced counterparts in terms of conversion probability, shopping baskets, repeat purchases, and customer revenue.

6.3 Data Description

This section provides detailed information on the unique sample we use to model shopping behavior of TV-induced customers and their non-induced counterparts. Note that, due to non-disclosure agreements, the numbers presented here have been altered using a constant factor.

6.3.1 Data Sources

The data is obtained from merging two distinct data sets provided by an international online retailer who only operates online shops. Both data sets cover a period from January 1 till April 30, 2015. The first data set is a very common visit-based online shop data set covering customer activity on the shop's website. The sample contains a total number of 4,403,866 visits generated by 3,471,532 unique users. This number contains 375,762 visits of 177,727 customers who registered during the four months, and 686,716 visits of 213,217 customers who registered before January 1. The total number of conversions is 489,884. To obtain a homogeneous set of customers, we only focus on those who registered for the online shop between January and April because the data set does not contain information on previous activities of customers who registered before January 1, 2015. Table 6.2 provides descriptive statistics of these customers related to the number of visits, repeat purchases and customer revenue. It shows that during January and April 2015 most customers visited the website only once.

	2.5%	50%	Mean	97.5%
Number of visits	1.00	1.00	2.11	7.00
Repeat purchases	0.00	0.00	0.17	2.00
Customer revenue	6.77	27.95	36.99	120.96

TABLE 6.2: Descriptive statistics of number of visits, repeat purchases, and revenue generated by customers who registered between January and April 2015.

The second data set includes information on the online shop's TV campaign from January 1 until March 31, 2015. The company broadcasted one TV ad over 3,500 times during that period. No ad was aired in April 2015. The ad includes information on the online shop in general, therefore, rather focuses on branding than on promotion. It was aired on 12 different TV stations and each record from the data set contains an exact time stamp (which enables us to merge the two data sets), the station at which the TV ad was aired, and the corresponding expenditures for each ad. The length of the ad is 15 seconds. The exact time stamps have been provided by a tracking company offering solutions to identify when TV ads were aired exact to the second. Figure 6.1 shows the frequency of TV ad expenditures in our data set. It shows that the majority of TV ad expenditures lies between 100$ and 1,000$. In our model, we use these expenditures as a proxy variable for the reach of ads.

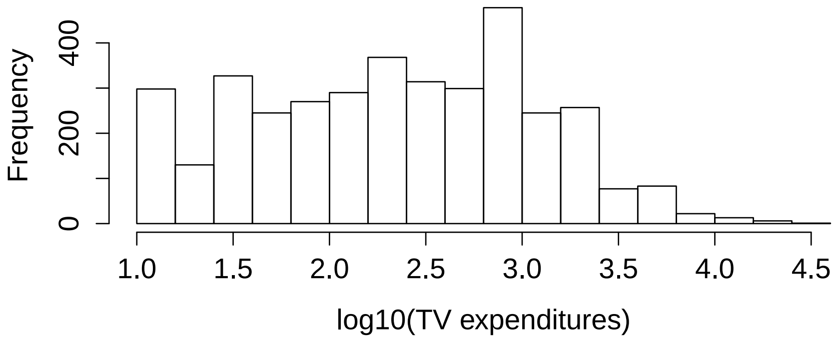

FIGURE 6.1: Frequency of TV ad expenditures on logarithmic scale.

6.3.2 Description of Variables

Each record of the first data set contains the customer's current activity on the website including indicators for registration and conversion. In addition, the data includes the size of the shopping basket for each conversion. For each touch point, it also includes the corresponding referral (i.e., whether a given customer was referred to the website by a direct type-in, an organic or paid search link, a display ad, etc.) as well as the device type (i.e., smart phone, tablet, and desktop/laptop) that the customer used for visiting the shop. Table 6.3 and Table 6.4 show the number of visits and conversions per combination of referral and device. The majority of customers visit the website by clicking paid or organic search engine links. Most of these customers use a desktop device; the number of visits and conversions via tablets and smart phones is much smaller.

	Direct type-in	Organic search	Paid search	Newsletter	Other
Desktop	31,090	42,491	149,250	17,036	36,971
Smartphone	8,030	10,064	22,090	2,154	3,767
Tablet	3,750	5,490	35,034	2,911	5,635

TABLE 6.3: Number of visits for different combinations of device and referral.

	Direct type-in	Organic search	Paid search	Newsletter	Other
Desktop	13,758	22,591	74,070	4,284	13,418
Smartphone	2,033	3,530	6,604	288	1,161
Tablet	1,387	2,814	13,788	462	1,835

TABLE 6.4: Number of conversions for different combinations of device and referral.

In addition, the data set includes demographic information, which is collected during the registration process, i.e., it is only available for those customers that have registered with the shop because they either receive its newsletter or already bought from the shop. This information includes gender and age – variables that are used as control variables in our models.

6.4 Model and Estimation

We model online shopping behavior using four different models to identify whether customers, who are attracted by TV ads behave differently than their non-TV-induced counterparts. More precisely, we aim to identify whether TV-induced customers are characterized by different (1) conversion probabilities, (2) shopping baskets, (3) repeat purchase behavior, or (4) 90-day customer revenues (shopping baskets $\times$ repeat purchases). Since it is impossible to know for sure whether a customer watched a certain TV ad before visiting the online shop, we first need to model the probability that a given visit is TV-induced (Section 6.4.1). Then, we use this probability as an independent variable in each of the four models (Section 6.4.2).

6.4.1 Modeling the Probability of Being TV-Induced

We model the probability that a given visit was induced by a TV ad using four consecutive steps. The following list provides an overview of the four steps. They are described in greater detail in the following paragraphs.

- Step 1: We count the number of new visits on three different device types (desktop, tablet, and smart phone) and three different referrals (direct type-in, organic search, paid search ads[1]) per 5-second interval around each TV ad.
- Step 2: We sum up the number of new visits per 5-second interval over all time intervals around all TV ads and are left with 3x3 sets of aggregated new visits per 5-second interval, i.e., we obtain one set per combination of device and referral.
- Step 3: Using these sets, we estimate the aggregated increase in the number of visits for each combination of device and referral by applying a hierarchical Bayesian model.
- Step 4: For each visit, we calculate the probability that it was induced by a TV ad by weighing the estimated uplift per TV ad with the corresponding TV ad expenditures[2].

The modeling steps described above require different variables that are described in this section. An overview of all variables used in our models is given in Table 6.5.

Step 1: Counting Visits Over Time

To estimate the increase in the number of visits to the online shop, we count visits on device d using referral channel r within 5-second time frames over the 3 minutes before and 9 minutes after the start of each TV ad. This counting step results in the number of visits per 5 seconds in the minutes before and after each TV ad. We use 5-second windows, because they allow for a precise measurement of the time-dependent uplift in the number of visits, and, at the same time, are computationally tractable. Using this data, we are able to estimate the uplift in visits per TV ad. We denote the 5-second time windows around each ad i as Δt_i^j with $j \in \{-36, ..., -1, 1, ..., 108\}$. Negative values of j represent time windows before an ad was aired, positive values represent time windows after an ad was aired. For instance, Δt_i^{-1} represents the time window starting 5 seconds before ad i starts and ending with the start of ad i.

To provide an example of this counting process, Table 6.6 shows the number of new visits established by clicking paid search links on smart phones and organic search links on desktops per 5-second window around two different ads aired

[1] The uplift in the number of visits via other referrals can be neglected.

[2] We use TV expenditures as proxy variable for the reach of ads.

Δt_i^j	Time difference between new visit and TV ad i; time interval j before/after ad i
$y_i^{d,r}(\Delta t_i^j)$	Number of visits around TV ad i on device d and referral r in time interval Δt_i^j
$y^{d,r}(\Delta t^j)$	Aggregated number of visits on device d and referral r in time interval Δt^j
$\alpha^{d,r}$, $\hat{\alpha}^{d,r}$	Aggregated baseline number of visits on device d and referral r
$\hat{\alpha}_i^{d,r}$	Baseline number of visits on device d and referral r around ad i
$\phi_1^{d,r}$, $\hat{\phi}_1^{d,r}$	Weight of the Gamma-shaped uplift component (hat = posterior median)
$\phi_2^{d,r}$, $\hat{\phi}_2^{d,r}$	Shape of the Gamma-shaped uplift component (hat = posterior median)
$\phi_3^{d,r}$, $\hat{\phi}_3^{d,r}$	Rate of the Gamma-shaped uplift component (hat = posterior median)
$\phi_4^{d,r}$, $\hat{\phi}_4^{d,r}$	Weight of the exponential decay component (hat = posterior median)
$\phi_5^{d,r}$, $\hat{\phi}_5^{d,r}$	Exponential decay constant (hat = posterior median)
ψ_k	Hyper prior for $\phi_k^{d,r}$
$\pi^{d,r}$	Probability that a given visit (device d, referral r) is a direct response to a TV ad
π^{any}	Probability that at least one visit of a customer was TV-induced
π^{prev}	Probability that at least one previous visit was TV-induced
N	Number of TV ads
r	Indicator variable for the referral channel used to open the website (one of organic search, paid search, direct type-in, newsletter, other)
d	Indicator variable for the device type used to open the website (one of desktop, smart phone, tablet)
$u^{d,r}$, $u_i^{d,r}$	Aggregated increase in number of visits / increase in number of visits with respect to ad i
c_i, c_{tot}	Expenditures of ad i, total TV ad expenditures

TABLE 6.5: Variables used to determine the probability that a given visit is TV-induced.

at different times. The baseline number of visits around Ad 2 is higher than the number of visits around Ad 1, since Ad 2 has been aired at a time of the day where more customers use the online shop. The example also shows that the relative uplift of the number of visits on smart phones is higher than on desktops due to the lower baseline number of visits on smart phones.

Step 2: Aggregating Visits Over All Ads

To reduce the computational effort of our estimation algorithm, we sum up the number of new visits per 5-second time interval over all TV ads (Equation 6.1).

$$y^{d,r}(\Delta t^j) = \sum_i y_i^{d,r}(\Delta t_i^j) \tag{6.1}$$

	Number of visits via	Δt_i^{-2}	Δt_i^{-1}	Δt_i^{1}	Δt_i^{2}	Δt_i^{3}
TV ad 1	Paid search / smart phone	5	4	6	10	14
(9am)	Organic search / desktop	30	29	35	37	39
TV ad 2	Paid search / smart phone	10	9	12	20	22
(5pm)	Organic search / desktop	51	52	56	58	62

TABLE 6.6: Example of the number of new visits on smart phone/paid search and desktop/organic search over time for two TV ads.

In this equation, $y_i^{d,r}(\Delta t_i^j)$ represents the number of visits established via device type d, using referral r in the 5-second time interval Δt^j seconds before/after TV ad i was aired. Consequently, $y_i^{d,r}(\Delta t^j)$ is the sum of the number of new visits using device d and referral r in the 5-second time interval Δt^j. Table 6.7 illustrates this aggregation approach for two TV ads with respect to the combinations of desktop/organic search and smart phone/paid search.

	Number of visits via	Δt_i^{-2}	Δt_i^{-1}	Δt_i^{1}	Δt_i^{2}	Δt_i^{3}
TV ad 1	Paid search / smart phone	5	4	6	10	14
(9am)	Organic search / desktop	30	29	35	37	39
TV ad 2	Paid search / smart phone	10	9	12	20	22
(5pm)	Organic search / desktop	51	52	56	58	62
Aggre-	Paid search / smart phone	15	13	18	30	36
gated	Organic search / desktop	81	81	91	95	101

TABLE 6.7: Aggregation example on smart phone/paid search and desktop/organic search.

The aggregated number of visits for each combination of device and referral is illustrated in Figure 6.2. For each combination, the uplift reaches a maximum approximately 1 minute after the start of the ads and decreases exponentially afterwards. The size of the uplift and the shape of the spike varies from combination to combination.

Step 3: Estimating the Aggregated Number of Visits

Next, we use the aggregated number of visits to estimate the following uplift model which aims to fit the spike in the number of visits over time. It consists of four components: a baseline component, a Gamma function component to account for the dominating spike in the number of visits after approximately 1 minute, an exponential decay component to compensate for the fast decaying shape of the Gamma function in the few minutes after the maximum uplift, and an error term (Equation 6.2).

$$y^{d,r}(\Delta t^j) = \alpha^{d,r} + \phi_1^{d,r}\ \Gamma(\Delta t^j, \phi_2^{d,r}, \phi_3^{d,r}) + \phi_4^{d,r}\ \Delta t^j\ \mathrm{Exp}(-\phi_5^{d,r}\ \Delta t^j) + \epsilon^{d,r}(\Delta t^j) \quad (6.2)$$

In this equation, $y^{d,r}(\Delta t^j)$ represents the aggregated number of visits on device d and referral r in the 5-second time interval after Δt^j seconds. The variable $\alpha^{d,r}$

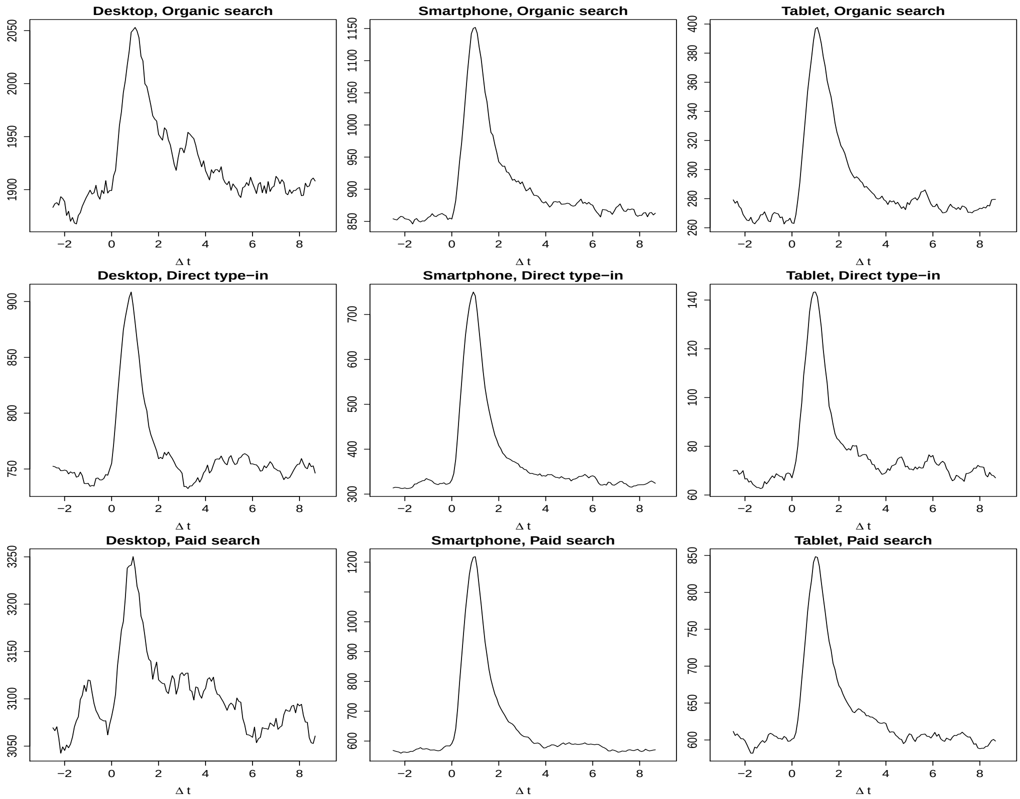

FIGURE 6.2: Aggregated uplift for each combination of referral and device per 5 seconds interval. As reported by previous studies (Kitts et al., 2014; Stange, 2015), the uplift in website traffic has roughly the form of a Gamma distribution.

represents the aggregated baseline number of visits on device d and referral r. The parameter $\phi_1^{d,r}$ indicates the size of the spike in the number of visits, while $\phi_2^{d,r}$ and $\phi_3^{d,r}$ parameterize the Gamma-shaped spike concerning device d and referral r in the first few minutes after a TV ad was aired. The parameter $\phi_4^{d,r}$ indicates the weight of the exponential decay component (relative to the size of the Gamma-shaped spike in the number of visits), whereas $\phi_5^{d,r}$ represents the decay parameter on device d and referral r. This exponential decay component is multiplied with Δt, i.e., the more time elapses since a spot was aired, the more important becomes the exponential decay component. As a result, the uplift in the number of visits in the first few minutes after an ad was aired is dominated by the Gamma-shaped component, and as Δt increases, the exponential decay component becomes more dominant. The weight of the exponential decay component $\phi_4^{d,r}$ defines how dominant this component becomes. The error term $\epsilon^{d,r}(\Delta t^j)$ captures the difference in visits between the model and the measured number of visits.

We use a hierarchical Bayesian model to estimate these parameters using the following prior distributions (Equation 6.3). A hierarchical model is feasible, because it allows for learning parameters on group level (for instance, the parameters that represent the uplift on tablets) from other groups (for instance, the parameters that represent the uplift on desktops). This is achieved by using shared hyper parameters, that allow for the propagation of beliefs among the group level parameters, denoted as $\psi_{1,\ldots,5}$.

$$\begin{aligned}
\psi_{\{1|2|3|4|5\}} &\sim \text{Exponential}(0.01) \\
\phi^{d,r}_{\{1|2|3|4|5\}} &\sim \text{Gamma}(10, 10/\psi_{\{1|2|3|4|5\}}) \\
\alpha^{d,r} &\sim \text{Normal}(\mu^{d,r}, \tau^{d,r}) \\
\epsilon^{d,r}(\Delta t) &\sim \text{Normal}(0, 0.01)
\end{aligned} \tag{6.3}$$

A hierarchical modeling approach is more robust with respect to outliers. In addition, it leads to statistically more significant results than estimating separate models per combination of device and referral. The posterior distributions of the hyper parameters can be interpreted as the grand mean of parameters $\phi^{d,r}_{1,\ldots,5}$. The prior for $\alpha^{d,r}$ is drawn from a normal distribution parameterized with $\mu^{d,r}$ as the mean number of visits on device d using referral r and precision $\tau^{d,r}$. The values of $\mu^{d,r}$ and $\tau^{d,r}$ are obtained from our sample.

Step 4: Calculating the Probability That a Visit Was Induced by TV Ads

Based on the estimates obtained from the uplift model described in the previous paragraph, we calculate the probability that a given visit is a direct response to a TV ad. To do so, we switch to continuous time and calculate the estimated aggregated uplift on device d using referral r after t seconds, denoted as $u^{d,r}(t)$, in accordance with Equation 6.4. This uplift results from subtracting the baseline number of visits $\hat{\alpha}^{d,r}$ from the aggregated number of visits ($y^{d,r}$, Equation 6.2) on device d and referral r:

$$u^{d,r}(t) \quad = \quad \hat{\phi}^{d,r}_1 \, \Gamma(t, \hat{\phi}^{d,r}_2, \hat{\phi}^{d,r}_3) + \hat{\phi}^{d,r}_4 \, t \, \text{Exp}(-\hat{\phi}^{d,r}_5 \, t) \tag{6.4}$$

To estimate the probability, we then need to calculate the uplift caused by one specific TV ad, which can be approximated by weighing the aggregated uplift $u^{d,r}(t)$, with the ratio of costs of TV ad i and the total costs of all TV ads (c_i/c_{tot}). We do so since the costs of ads can be assumed to be an appropriate proxy for their reach. As the aggregated uplift was caused by all TV ads with total costs of c_{tot}, the uplift in the number of visits at time t caused by a single TV ad with individual costs c_i is assumed to be given by:

$$u^{d,r}_i(t) \quad = \quad \frac{c_i}{c_{tot}} u^{d,r}(t - \tau_i) \tag{6.5}$$

In Equation 6.5, the parameter τ_i represents the time when ad i was aired. The uplift of a single TV ad does not correspond to a probability yet. To calculate the probability that a given visit has been induced by TV ad i, one needs to further consider the estimated ad-specific uplift $u^{d,r}_i$ and the baseline number of visits around a single TV ad, denoted as $\hat{\alpha}^{d,r}_i$. Therefore, we divide the aggregated baseline number of visits $\hat{\alpha}^{d,r}$ by the number of ads N:

$$\hat{\alpha}_i^{d,r} = \frac{\hat{\alpha}^{d,r}}{N} \tag{6.6}$$

For simplicity, this approach assumes that the baseline number of visits is constant over time and does not fluctuate over the course of a day. The probability that a given visit has been induced by TV ad i is given by the ratio of $u_i^{d,r}$ and the sum of $u_i^{d,r}$ and $\hat{\alpha}_i^{d,r}$, i.e., the uplift divided by the sum of the uplift plus the baseline number of visits with respect to TV ad i, device d and referral r. Since the uplift in number of visits depends on the time difference between TV ad and the time of a visit, the probability necessarily depends on t:

$$\pi_i^{d,r}(t) = \frac{u_i^{d,r}(t)}{u_i^{d,r}(t) + \hat{\alpha}_i^{d,r}} \tag{6.7}$$

Equation 6.7 represents the time-dependent probability that a given visit at time t is a direct response to TV ad i. The opposite probability, $1 - \pi_i^{d,r}(t)$, represents the probability that a given visit has not been induced by TV ad i. The probability that a visit observed at time t has not been induced by any TV ad, which was aired previously, is given by the product of all these opposite probabilities, namely $\Pi_i(1 - \pi_i^{d,r}(t))$. Consequently, the probability that a given visit established via device d and referral r has been induced by at least one TV ad that was aired previously is given by:

$$\pi^{d,r}(t) = 1 - \prod_i [1 - \pi_i^{d,r}(t)] \tag{6.8}$$

While we are sure that a visit has not been induced by a TV ad when no TV ad has been aired (π is very close to 0), we cannot be sure that a visit has really been induced by a TV ad since we cannot directly observe the customer in his/her living room – even if the visit was established directly after a TV ad.

In our four models to explain explain conversion probability, shopping baskets, repeat purchases and customer revenue, we need three different variables concerning customers' probabilities of being induced by TV ads. First, the probability described with Equation 6.8 represents the probability that the current visit is a direct response to a TV ad. Second, we include the probability that a previous visit has been induced by a TV ad. This probability is defined for all visits j after the first visit, i.e., $j > 1$, and is described with Equation 6.9:

$$\pi_j^{prev} = 1 - \prod_{l=1}^{j-1} (1 - \pi_l) \tag{6.9}$$

For example, if visit 1 and 2 show a 20% probability that they have been induced by a TV ad, the resulting value for π_3^{prev} in visit no. 3 is $1 - 0.8^2 = 36\%$. This variable can be regarded as a proxy for the ad stock.

Third, we define the probability that at least one visit j of a customer k was induced by a TV ad (Equation 6.10).

$$\pi_k^{any} = 1 - \prod_j (1 - \pi_j) \tag{6.10}$$

In Equation 6.10, the variables π_j are the visit-related probabilities described with Equation 6.8. Given a customer with two visits with a probability of 50% to be induced by a TV ad, the resulting probability that the customer has been induced by a TV ad at least once equals $\pi_k^{any} = 1 - 0.5^2 = 0.75$.

6.4.2 Models to Explain Conversion Probability, Shopping Baskets, Repeat Purchases, and Customer Revenue

We would like to investigate the differences of TV-induced customers and their non-induced counterparts in terms of conversion probability, shopping baskets, repeat purchases, and 90-day customer revenue. As discussed before, our data set does not include historical information on existing customers. Therefore, we only include customers who registered within the time period covered by our data set. This selection criterion allows us to compare different customers. Our four models are based on the operationalization of customer journeys proposed by Chatterjee et al. (2003). Their approach allows for estimating the effects of different ad exposures over the course of a customer journey (Bucklin and Sismeiro, 2009; Chatterjee et al., 2003). An overview of the models used here is given in Table 6.8.

Model	DV	Sample	Level
Probability to convert - Logistic regression (Equation 6.11)	$p(conv)_j$	All visits from all new customers	Visit
Shopping basket - Log-linear regression (Equation 6.12)	SB_j	All conversions of all new customers	
Repeat purchases - Negative-binomial regression (Equation 6.13)	RP_k	Customers registered in January. At least one conversion. Max Journey length 90 days.	User
Online customer revenue - Log-linear regression (Equation 6.14)	CR_k		

TABLE 6.8: Overview of models used here.

Conversion Probability

We model conversion probabilities by using a logistic regression that aims to explain the purchasing decision of registered customers for each of their visits.

$$p(conv = 1)_j = \frac{1}{1 + \exp\left\{-(\delta_0 + \delta_1 \pi_j + \delta_2 \pi_j^{prev} + \delta_3 X_j + \epsilon_j)\right\}} \tag{6.11}$$

Here, π_j represents the probability that a given visit j has been induced by a TV ad (Equation 6.8). The variable π_j^{prev} represents the probability that any previous visit has been induced by a TV ad (Equation 6.9). The row vector X_j represents the set of control variables: We account for referral channel (i.e., SEA, newsletter, organic search, direct type-in, others) and device type used to open the website (desktop, smart phone, tablet). These variables take on binary values (0/1). In addition, we account for the number of previous contacts with the website using different devices and referrals. Furthermore, we control for demographics, time of day, day of the week, and month, the total number of visits (independent of device and referral), the time between two visits, and time since the first visit of a customer.

Shopping Baskets

The value of all items in the shopping basket (SB) at checkout is also modeled per visit. To gain deeper insight concerning shopping baskets that result from TV-induced visits and non-TV-induced visits, we focus only on visits that resulted in a conversion and exclude all visits that did not result in a conversion. We define the following log-linear model.

$$\log(SB_j) = \delta_0 + \delta_1 \pi_j + \delta_2 \pi_j^{prev} + \delta_3 X_j + \epsilon_j \tag{6.12}$$

We use the logarithm of shopping baskets as a dependent variable because of the non-normal distribution of shopping baskets. Aside from these differences, we include the same control variables as in the model used to explain conversion probabilities (Equation 6.11).

Repeat Purchases

In contrast to the models used to estimate conversion probabilities and shopping baskets, the number of repeat purchases is modeled on a per customer basis (i.e., over multiple visits). To obtain a homogeneous set of customer journeys, we first exclude customers who registered after January 2015 and truncate every customer journey after 90 days. Since the number of repeat purchases corresponds to the number of purchases after the first purchase, we exclude all customers without conversion in our data set. To model the number of repeat purchases (RP) on a per customer basis, we need to aggregate the visit-specific variables by cumulating the number of contacts established by certain referral channels and device types per customer. We control for the total number of visits, demographics and time represented by X_k. We use a negative binomial regression, because we deal with over-dispersed count data. The variable π_k^{any}

represents the probability that any visit of customer k was induced by a TV ad (Equation 6.10). The model is given by:

$$RP_k = \mathrm{NB}(\delta_0 + \delta_1 \pi_k^{any} + \delta_3 X_k + \epsilon_k) \tag{6.13}$$

Customer Revenue

Similar to the number of repeat purchases, customer revenue is modeled on a per customer basis. We define the revenue per customer as the sum of shopping baskets at checkout within the first 90 days and exclude customers without conversions during that time. Since the customer revenue is non-normally distributed, we use its logarithm as dependent variable. We use the same control variables X_k as in Equation 6.13. The model is described with Equation 6.14:

$$\log(CR_k) = \delta_0 + \delta_1 \pi_k^{any} + \delta_3 X_k + \epsilon_k \tag{6.14}$$

6.5 Results and Discussion

In this section, we first discuss the results of the model used to calculate the time-dependent probability that a given visit was induced by a TV ad (Equation 6.2 – 6.8). Afterwards, we present the results regarding differences in online shopping behavior between TV-induced customers and their non-TV-induced counterparts. Specifically, we focus on differences in conversion probability, shopping baskets, repeat purchases, and 90-day customer revenue.

6.5.1 Probability That a Visit Is TV-induced

To estimate the Bayesian model (Equation 6.2 and 6.3) we run 10,000 burn-in iterations and 10,000 sampling iterations and take every 10th sample to obtain the posterior distributions. Based on these results, we calculate the probability that a given visit is TV-induced. Figure 6.3 shows the distribution of probabilities on logarithmic scale. It shows, that the majority of probabilities of being TV-induced is smaller than 1% ($\pi = 10^{-2}$).

Figure 6.4 illustrates the probability that a visit was induced by a TV ad, given that it was established using a certain device and referral channel: (1) within the two minutes after an ad was aired and (2) after two minutes of an ad's airing. The diagrams show that the effect of TV ads is largest for direct type-ins followed by paid search requests. Since the baseline number of visits on mobile devices is rather small, the probability that a visit observed on these devices is induced by a TV ad is greater than on desktop devices. This finding shows that customers tend to use mobile devices as a second screen while watching TV.

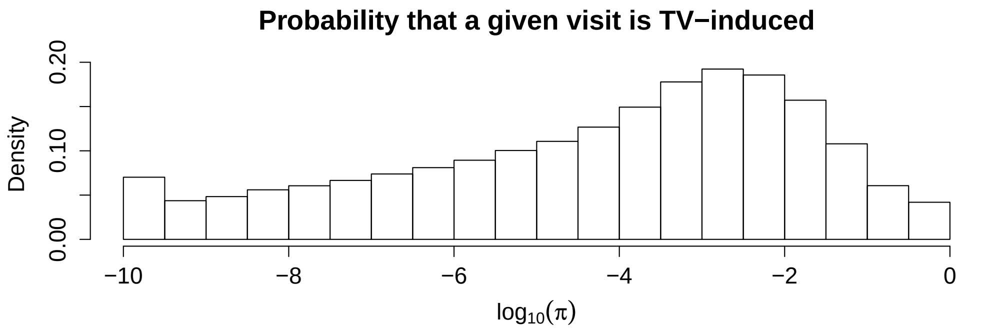

FIGURE 6.3: Distribution of probabilities.

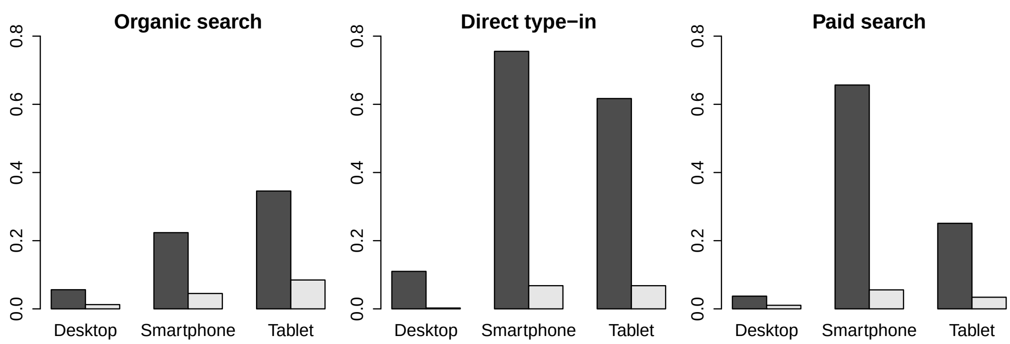

FIGURE 6.4: Probability that a given visit was induced by a TV ad with mean expenditures (c_{tot}/N). The dark grey bar represents the probability within the first two minutes after a spot was aired. The light grey bar represents the probability after two minutes of an ad's airing.

Figure 6.5 shows the time-dependent probability that a visit is TV-induced after broadcasting TV ads at three different points in time (i.e., at $t = 0$, $t = 7.5$, $t = 15$, and $t = 25$ minutes). In addition, it shows three distinct probabilities over time that a given visit has been induced by a TV ad. For example, a visit established with a desktop PC using an organic search result, which is observed at $t = 3$, has a probability of 0.04 of having been induced by a TV commercial. Figure 6.5 illustrates that the probability is largest in the minute after an ad was aired, and that it decays exponentially afterwards. In addition, the size of the probability is highly dependent on the expenditures of a given TV ad. In the example, the third ad is the most expensive one followed by the fourth ad.

6.5.2 Results Regarding Conversion Probabilities, Shopping Baskets, Repeat Purchases, and Customer Revenue

As discussed above, we are interested in the type of customers who are attracted by TV ads and spontaneously visit the shop's website in response to TV ads. In this section, we present the results of the four models (Equations 6.11 – 6.14) to analyze TV-induced customers' probability to convert, their shopping baskets

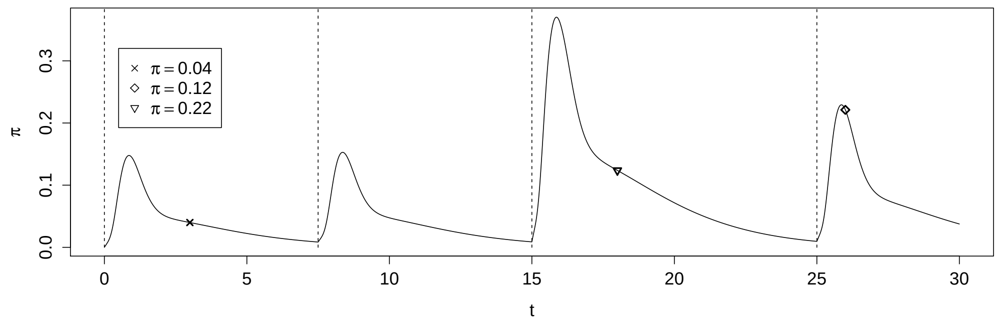

FIGURE 6.5: Probability over time (in minutes) that a new visit established via organic search on a desktop/laptop has been induced by different TV ads (1,500$, 1,500$, 5,000$, 2,000$), and values of π_j for three visits.

and repeat purchase behavior as well as their 90-day customer revenue (Table 6.9).

The first column of Table 6.9 shows that the probability to convert is lower for TV-induced visits, i.e., customers who spontaneously respond to a TV ad are less likely to convert than customers who visit the shop for another reason. The probability of TV-induced customers to convert is by $\exp(-0.32) = 0.73$ lower than the one of their non-TV-induced counterparts[3]. The probability to convert on smart phones is lower than on tablets, which is, in turn, lower than the probability to convert using a desktop or laptop. This finding suggests that customers still tend to use mobile devices to gather information on certain products rather than to directly purchase them. The probability that a previous visit was induced by a TV ad (π^{prev}) does not affect the probability to convert in the current visit.

The results presented in the second column of Table 6.9 suggest that shopping baskets of TV-induced visits are approximately 10% lower than shopping baskets of non-TV-induced visits. In addition, the probability that a previous visit was induced by an ad, π^{prev}, leads to 5% smaller basket sizes in the current visit. This finding suggests that customers who visited the website in response to a TV ad tend to spend less money in the future. In addition, the results suggest that shopping baskets tend to be lower on smart phones than on tablets or desktops. Finally, customers who type in the URL directly tend to spend more money indicating that these customers might be particularly loyal to the shop, doing all their related shopping there instead of chasing the best deal at multiple shops.

The results presented in the third column of Table 6.9 suggest that the number of repeat purchases within the first 90 days after a registration are not influenced by the probability that a customer was induced by a TV ad at least once. However,

[3]The marginal probability of the mean TV-induced customer is by 7.5% lower than the one of his/her non-induced counterpart (41% vs. 33.5%)

Var.	$p(conv)$		SB		RP		CR	
TV variables								
Current/any visit TV-induced	-0.32	***	-0.10	***	-0.00		-0.10	**
Any previous visit TV-induced	-0.07		-0.05	*				
Device indicators								
Smart phone contact	-0.69	***	-0.22	***	-0.02	***	-0.05	***
Tablet contact	-0.31	***	0.01		-0.02	***	-0.02	***
Referral indicators								
SEA contact	0.05	***	0.01	**	0.02	***	-0.00	
Direct contact	0.03	*	0.14	***	-0.01	*	0.03	***
Newsletter contact	-0.64	***	0.16	***	-0.04	***	-0.02	***
Other contact	-0.25	***	0.14	***	-0.05	***	-0.01	*
Customer history variables								
Visit number	-0.33	***	0.01	***				
Total number of visits					0.12	***	0.17	***
Num. prev. tablet contacts	0.01	**	0.00	.				
Num. prev. smart phone contacts	0.04	***	0.01	***				
Num. prev. SEA contacts	0.04	***	0.01	*				
Num. prev. direct contacts	0.07	***	-0.01	.				
Num. prev. newsletter contacts	0.19	***	-0.01	*				
Num. prev. other contacts	0.04	***	0.02	***				
Last basket value	-0.01	***	0.01	***				
Time between two visits	-0.10	***	0.00	*				
Time since last conversion	-0.06	***	-0.03	***				
Time since first visit	0.02	***	0.00	***				
Demographic information								
Age < 18	-0.81	***	-0.28	***			-0.57	***
Age: 18-29	0.14	***	-0.17	***	-0.03		-0.17	***
Age: 30-39	0.27	***	-0.02	***	-0.00		0.01	
Age: 40-49	0.19	***	-0.00		0.02		0.02	.
Gender	-0.15	***	0.04	***	-0.00		0.01	
Time variables								
Monday	-0.05	***	-0.04	***	-0.01		-0.02	**
Tuesday	-0.15	***	-0.05	***	-0.01		-0.02	**
Wednesday	-0.09	***	-0.05	***	0.01		-0.03	***
Thursday	-0.13	***	-0.04	***	-0.01	.	-0.01	*
Friday	-0.20	***	-0.04	***	-0.01		-0.04	***
Saturday	-0.16	***	-0.02	***	-0.02	*	-0.02	*
January	-0.30	***	0.04	***	-0.05	***	-0.08	***
Februrary	-0.15	***	0.00		-0.01		-0.05	***
March	-0.06	***	0.00		-0.02	***	-0.01	
0-5 hours	-0.51	***	-0.06	***	-0.02		-0.06	***
6-11 hours	-0.07	***	-0.03	***	0.01	*	-0.01	**
12-17 hours	-0.10	***	-0.02	***	0.00		0.00	
Intercept	-0.37	***	3.42	***	0.01		3.61	***

TABLE 6.9: Results generated with our four different models. Signif. codes: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; . $p < 0.1$. For all models, the following baselines are used: organic search referral; desktop; age: ≥ 50; day: Sunday.

the average number of repeat purchases equals 0.17, which is why it is hard to observe any differences between TV-induced and non-TV-induced customers.

Finally, the 90-day customer revenue of TV-induced customers is 10% lower than the revenue of non-TV-induced customers (fourth column in Table 6.9). This finding is in line with the findings from modeling shopping baskets (10% lower shopping baskets) and conversion probabilities (27% lower conversion probabilities). Although their probability to convert is lower, TV-induced customers who buy at least once buy as frequent as other customers, i.e., they are characterized by the same number of repeat purchases. Therefore, the main driver for the reduced customer revenue of TV-induced customers is their lower shopping baskets.

6.6 Conclusion

In this paper, we examine the online shopping behavior of customers who spontaneously visit an online shop after seeing the shop's ad on TV. Modeling these decisions at the level of the individual, we find that TV-induced customers are characterized by lower conversion probabilities, lower shopping baskets and, consequently, lower 90-day customer revenues. Since we only observe the customer's behavior, we can only speculate about the reasons for these differences: First, these customers might just want to gather information on the online shop (resulting in lower conversion probabilities) or try it out by shopping a product of lower value (lower shopping baskets). Second, since these customers seem to be very spontaneous they seem to spend less time on the website and, consequently, do not take the time or do not have the time – due to the limited length of the commercial break – to browse through many different products (also resulting in lower shopping baskets). Third, these customers might be very price-sensitive, always chasing after a good deal, which would lead them to buy only some products at the focal shop and cheaper ones elsewhere (also resulting in lower shopping baskets).

The reduced customer revenues are strongly driven by reduced shopping baskets. This insight could be used by the online shop to find measures to increase them. For instance, they could grant discounts related to the number of items in a shopping basket for customers who show a high probability to be TV-induced. In addition, advertisers are well advised to consider differences in online customer revenues when calculating the return on investment of TV advertising campaigns. In practice, the modeling approach proposed here can further be extended to evaluate audiences of different TV stations and measure the effectiveness of different TV ads in terms of long-term customer behavior.

6.7 References

Bucklin, R. E. and C. Sismeiro (2009). "Click here for Internet insight: Advances in clickstream data analysis in marketing". In: *Journal of Interactive Marketing* 23.1, pp. 35–48.

Chatterjee, P., D. L. Hoffman, and T. P. Novak (2003). “Modeling the click-stream: Implications for web-based advertising efforts”. In: *Marketing Science* 22.4, pp. 520–541.

Joo, M., K. C. Wilbur, B. Cowgill, and Y. Zhu (2014). “Television Advertising and Online Search”. In: *Management Science* 60.1, pp. 56–73.

Joo, M., K. C. Wilbur, and Y. Zhu (2015). “Effects of TV advertising on keyword search”. In: *International Journal of Research in Marketing, Available online 17 April 2015, In Press*.

Kitts, B., M. Bardaro, D. Au, A. Lee, S. Lee, J. Borchardt, C. Schwartz, J. Sobieski, and J. Wadsworth-Drake (2014). “Can Television Advertising Impact Be Measured on the Web? Web Spike Response as a Possible Conversion Tracking System for Television”. In: *Proceedings of 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, New York, NY, USA*.

Lewis, R. A. and D. H. Reiley (2013). “Down-to-the-Minute Effects of Super Bowl Advertising on Online”. In: *14th ACMConference on Electronic Commerce* 9.4, pp. 1–22.

Liaukonyte, J., T. Teixeira, and K. C. Wilbur (2015). “Television Advertising and Online Shopping”. In: *Marketing Science* 34.3, pp. 311–330.

Stange, M. (2015). “The Impact of TV Ads on the Individual User’s Purchasing Behavior”. In: *Proceedings of the 36th International Conference on Information Systems, Fort Worth, United States*.

Statista (2016). *Activities undertaken by smartphone users while watching TV in the United States as of 3rd quarter 2015; http://www.statista.com/statistics/276129/ us-smartphone-user-activities-while-watching-tv/ (retrieved on 22 Aug 2016)*.

Zigmond, D. and H. Stipp (2010). “Assessing a New Advertising Effect Measurement of the Impact of Television Commercials on Internet Search Queries”. In: *Journal of Advertising Research* 50, pp. 162–168.

7 Predicting Online User Behavior Based on Real-Time Advertising Data

Stange, M. and B. Funk (2016). "Predicting Online User Behavior Based on Real-Time Advertising Data". In: Proceedings of the 24th European Conference on Information Systems, Istanbul, Turkey.

Abstract

Generating economic value from big data is a challenge for many companies these days. On the Internet, a major source of big data is structured and unstructured data generated by users. Companies can use this data to better understand patterns of user behavior and to improve marketing decisions. In this paper, we focus on data generated in real-time advertising where billions of advertising slots are sold by auction. The auctions are triggered by user activity on websites that use this form of advertising to sell their advertising slots. During an auction, so-called bid requests are sent to advertisers who bid for the advertising slots. We develop a model that uses bid requests to predict whether a user will visit a certain website during his or her user journey. These predictions can be used by advertisers to derive user interests early in the sales funnel and, thus, to increase profits from branding campaigns. By iteratively applying a Bayesian multinomial logistic model to data from a case study, we show how to constantly improve the predictive accuracy of the model. We calculate the economic value of our model and show that it can be beneficial for advertisers in the context of cross-channel advertising.

7.1 Introduction

As more and more data is generated by customers, sensors, or governments, business intelligence and analytics become increasingly important. Many practitioners and researchers have been focusing on this topic and have developed methods to measure the impact of big data in recent years (Chen et al., 2012). In this paper, we focus on the impact of a fairly new source of big data generated during the real-time advertising process on the Internet.

In real-time advertising, advertising slots on a website are sold by auction in the 200 milliseconds after a website is called by a user. In the first few milliseconds after the call, website context information (such as content, language, quality) and an anonymous user id is sent to a so-called ad exchange. The ad exchange, which is a marketplace for advertising slots, sends out so-called bid requests to advertisers and their service providers, who employ bidding agents, which instantaneously select the advertising media which best fits to the current user's interests. In addition, these agents determine the maximum price the advertiser is willing to pay for the ad impression at auction. This information is bundled to a so-called bid response and sent back to the ad exchange, which forwards the advertising media of the highest bidder to the publisher's website and charges the highest bidder the second highest price (second price auction). The whole process is completely invisible for the user because, as the website is completely loaded, the auction is already closed. As the process happens tens of thousands of times per second, it is a distinct source of big data (Stange and Funk, 2014). In this paper, we develop a model to gain economic value from the massive amount of data that is generated during this process.

Bidding agents in real-time advertising usually make use of information from recent customer activity (i.e., cookie data) on the advertiser's website to determine appropriate advertising media for the current user. Users often recognize this by being exposed to advertising material of products they were searching for recently (re-targeting). Of course, real-time advertising is not limited to this rather simplistic form of advertising. However, there is often not enough information about the current user and his or her interests available to make better decisions (Perlich et al., 2012). In this context, we propose a new approach to derive users' interests based on the stream of bid requests that were generated by their browsing activity, and show that users' interests can be accurately predicted by only using this data. In our approach, a user's interest in a certain product is assumed if he or she visited a website related to this product during his or her journey. The method enables advertisers to expose ads only to users that exhibit a certain probability to be interested in their products. At the same time, users who will most likely never be customers can be ignored. In our case study, we focus on the users' interests in certain TV programs.

We contribute to IS research by determining the impact of bid request data from an advertiser's perspective (Chen et al., 2012). To determine this impact, we calculate the economic value of a person-centered model that can be used to understand and predict users' behaviors on the Internet based on bid request data. We develop an iterative Bayesian model that enables us to update once trained parameters with new data according to Bayes' rule. This model can be used by researchers to develop and extend decision support systems on the Internet and to develop new business models for predictive analytics in the field (Veit et al., 2014). The approach is not intended to replace well established methods to target users in online marketing contexts. Instead, our method is supposed to be a conceptual extension of the landscape of methods and tools to target users with proper advertisements. In practice, our approach may be valuable for TV stations or their agencies to coordinate TV and online advertising campaigns, for instance (Joo et al., 2015; Stange, 2015). In this context, the proposed method could be used to expose ads online only to users who show a high probability for having watched a certain TV program recently. We apply the approach to bid request data from a major ad exchange and show how benefits from cross-channel advertising activities can be increased.

The remainder of this paper is structured as follows: First, we review recent literature on using tracking data to analyze user behavior. Second, we describe our modeling approach and show how to integrate the model into bidding agents. Third, we describe data collection and preparation. Fourth, we present the results of the analysis and calculate its economic value from an advertiser's perspective. Finally, we discuss the implications of our study.

7.2 Related Work

In our study, we make use of several results from IS and marketing research, which are going to be outlined in the following paragraphs.

For many companies in e-commerce, it is crucial to identify their customers' interests for products and services. It is clear that the more information companies have available about their customers, the better they can customize their products to the clients' individual needs. For this reason, researchers have analyzed the users' click and purchase behavior on the Internet to make better marketing decisions (Bucklin and Sismeiro, 2009) in order to achieve, for instance, an optimal fit between advertising materials and users or to offer customized products.

Chatterjee et al. (2003) developed a user journey model consisting of variables representing long term and short term advertising effects based on clickstream data that was generated in advertising campaigns. They conducted a hierarchical logistic regression to estimate the variables' effects on the users' click probability. With the results, the impact of individual advertising channels on the customers' click behavior can be extracted and thus, predictions can be made about the click probabilities of future users. This outcome can be used to increase the effects of display advertising campaigns.

However, users' click probability does not necessarily correlate with their probability to purchase a product or to register for a newsletter, i.e., the users' conversion probability (Lee et al., 2012; Pandey et al., 2011). In the context of real-time advertising, these conversion probabilities can be used to determine the size of bids. Many researchers (Perlich et al., 2012; Zhang et al., 2014) have developed methods to determine the most profitable bids from the perspective of a demand-side platform – a service provider who places bids on behalf of the advertisers. Approaches in the literature often focus on using real-time advertising for performance-oriented displaying of advertising materials, i.e., on selecting ads that will most likely lead to direct purchases (Chen and Berkhin, 2011), and on placing optimal bids for these ads (Zhang et al., 2014). In addition, published studies have focused on optimal selection of ads with respect to budget constraints, or recency and frequency capping (Yuan et al., 2013). Lee et al. (2012) showed how to use past performance data to effectively determine the right advertisement to be exposed to the right user on the right publisher website. The proposed models can be used to improve advertising effects or to reduce costs. However, performance-oriented ad selection aims to target users at a late stage in the sales funnel, where often sufficient information about a given user is available. By contrast, our approach aims to target users at an early stage in the sales funnel and focuses on predicting user interests only based on bid request data.

A growing number of advertisers use real-time advertising also for branding campaigns. In these campaigns, exposing the right ads to the right user is a challenge because in this early phase of the sales funnel only little information about users is available. In the context of branding campaigns, many authors have pointed out the importance of a cross-channel advertising strategy to increase users' awareness of certain products (Dinner et al., 2014; Duan and Zhang, 2014; Joo et al., 2015; Yang and Ghose, 2010). The idea behind this strategy is that advertising activities should not only be focused on individual advertising channels but should also consider synergies that can be observed when advertising activities on different channels are seamlessly coordinated. For instance, it might be very sensible to combine a TV advertising campaign with a complementary search engine advertising campaign or an e-mail advertising campaign instead of treating these advertising activities individually (Liaukonyte et al., 2015; Stange, 2015). In contrast to the effect of online advertising, however, the possibilities to measure the effect of offline advertising are limited (Kitts et al., 2014), and therefore it is often challenging to coordinate online and offline advertising campaigns effectively. We address this challenge by developing a method to increase benefits from awareness-related offline-online advertising campaigns using bid request data. We use the method to predict whether a user has watched a certain TV program recently.

Chen et al. (2012) demonstrated the increasing impact of big data analytics that can be observed in many industries. To identify the impact of bid request data on companies in e-commerce and marketing, we propose to measure its economic value as it is proposed by Nottorf and Funk (2013). In the context of online advertising, they determine the economic value of clickstream data by assigning (negative) costs to true and false predictions of a classifier that was trained using the data. We apply this method to a multinomial classifier that is trained using bid request data. Thereby, we show that the analysis of this kind of big data can be particularly beneficial for companies conducting awareness-oriented advertising campaigns.

7.3 Model Development

This section first discusses the general framework of the modeling approach used here. Second, it describes the proposed model in detail. Third, it presents a process that shows how the model can be used by bidding agents in real-time advertising.

7.3.1 Modeling Approach

The goal of our analysis is to calculate the probabilities that a user is going to visit certain websites during his or her journey. We interpret these probabilities as an indication of the user's interests. In our case study, we focus on the users' interest in a certain TV program. To identify whether a user is interested in a certain TV program or not, we use the websites of five different TV stations. For instance, we assume that a user who visits rtl.de is most likely interested in the TV program aired on RTL. Our method could, for instance, be used by TV stations or their agencies to enrich the TV advertising campaigns of their customers (i.e., the advertisers) with complementary online advertising campaigns. Of course, it is impossible to fully understand users' actual TV consumption behaviors only based on their online user journeys. However, since the goal is to demonstrate the possibilities of using bid request data to determine users' interests, we consider this to be only a minor limitation.

The modeling approach presented in this paper addresses the challenge of the high volume and velocity of bid request data. We handle the high velocity by iteratively applying a multinomial Bayesian logistic model, which uses prior knowledge about its parameters as follows: Initially, no information about the regression parameters is available. After the first iteration, the model returns parameter distributions based on the first batch of data. We extract the means and standard deviations of these distributions to then use these values as prior information for the subsequent run. Thus, the precision of the parameter distributions and the predictive accuracy of the model increase with the number of iterations.

We handle the high volume of bid request data by applying stratified sampling. As many analytical tasks in the context of e-commerce, such as the prediction of purchases or clicks, have to deal with rare events (i.e., conversions or clicks are very rare compared to the number of views), this approach can save computational costs. In our data set, over 99% of the overall data set contains user journeys that are irrelevant for our analysis. Stratified sampling makes the learning algorithm more efficient, since less data is required to estimate the parameters.

7.3.2 Model Description

The dependent variable of our analysis is denoted as $K \in \{0, \ldots, 5\}$ and indicates whether a user opened the website of a specific TV station during the entire user journey or not. We regard this variable as a proxy variable that indicates users' interest in a certain TV station and their potential interest in certain products advertised on this station. We include the following target URLs: rtl.de ($K = 1$), rtl2.de ($K = 2$), vox.de ($K = 3$), sat1.de ($K = 4$), and prosieben.de ($K = 5$). If a user did not open one of these websites, K is set to 0. The independent variables are all other URLs from which bid requests can be triggered. For instance, if a user i visits an URL U_j such as amazon.de and ebay.de during her or his user journey J_i, the set of independent variables for this user is 1 for both of these variables, and all other variables are set to 0. In Equation 7.1, X_{ij} is the j^{th} covariate from the design matrix X at row i, and U_j is the j^{th} entry from a list of URLs.

$$X_{ij} = \begin{cases} 1 & \text{if } U_j \in J_i \\ 0, & \text{otherwise} \end{cases} \tag{7.1}$$

We use a Bayesian multinomial logistic regression model, which we estimate using JAGS (Plummer, 2003) based on Equation 7.2:

$$\begin{aligned} K_i &\sim \text{Multinomial}(p_i) \\ p(K_i = l | X_i) &= \frac{\exp(\alpha_l + X_i \beta_l)}{\sum_k \exp(\alpha_k + X_i \beta_k)} \\ \beta_{jk} &\sim \text{Normal}(b_{jk}, \sigma_{jk}) \\ \alpha_k &\sim \text{Normal}(a_k, s_k) \end{aligned} \tag{7.2}$$

In this equation, α_k represents the intercept for class k, and β_{jk} represents the j^{th} entry of the parameter vector β_k, i.e., the slope for the j^{th} URL in the list of parameters of class k. The values a_k and b_{jk} represent the prior knowledge of α_k and β_{jk}. The terms s_k and σ_{jk} represent the prior knowledge of the standard deviation of α_k and β_{jk}. The term $p(K_i = l|X_i)$ represents the probability that the row vector X_i is of class l. In each iteration, the prior values a_k, b_{jk}, σ_{jk}, and s_k are updated with the posterior means and standard deviations from the previous steps of the analysis. The initial values for α_k and β_{jk} are defined as 0, whereas σ_{jk} and s_k are set to 10.

In our case study, the number of users labeled with $K = 0$, i.e., users who never visited a website of a TV station, is high in comparison to the other classes. For this reason, stratified sampling for training the model has been recommended (King and Zeng, 2001). However, after parameter estimation, the sampling bias needs to be considered before using the results for prediction in accordance with Equation 7.3:

$$p(K_i = l|X_i) = \frac{\exp(\alpha_l + X_i \beta_l) \cdot \tau_l / \bar{y}_l}{\sum_k \exp(\alpha_k + X_i \beta_k) \cdot \tau_k / \bar{y}_k} \tag{7.3}$$

In this equation, τ_k represents the ratio of instances of class k in a random sample, and $\bar{y}_k$ represents the ratio of instances of class k in the training set (King and Zeng, 2001).

For each iteration, we only include variables of URLs that are contained in the data of the current training batch. The parameters of the other variables remain unchanged. This is possible due to the assumed statistical independence between the parameters and enables the model to include a large number of parameters.

7.3.3 Updating the Decision Engine

We propose to include the model in a real-time advertising decision support system as presented in Figure 7.1 and described in the following.

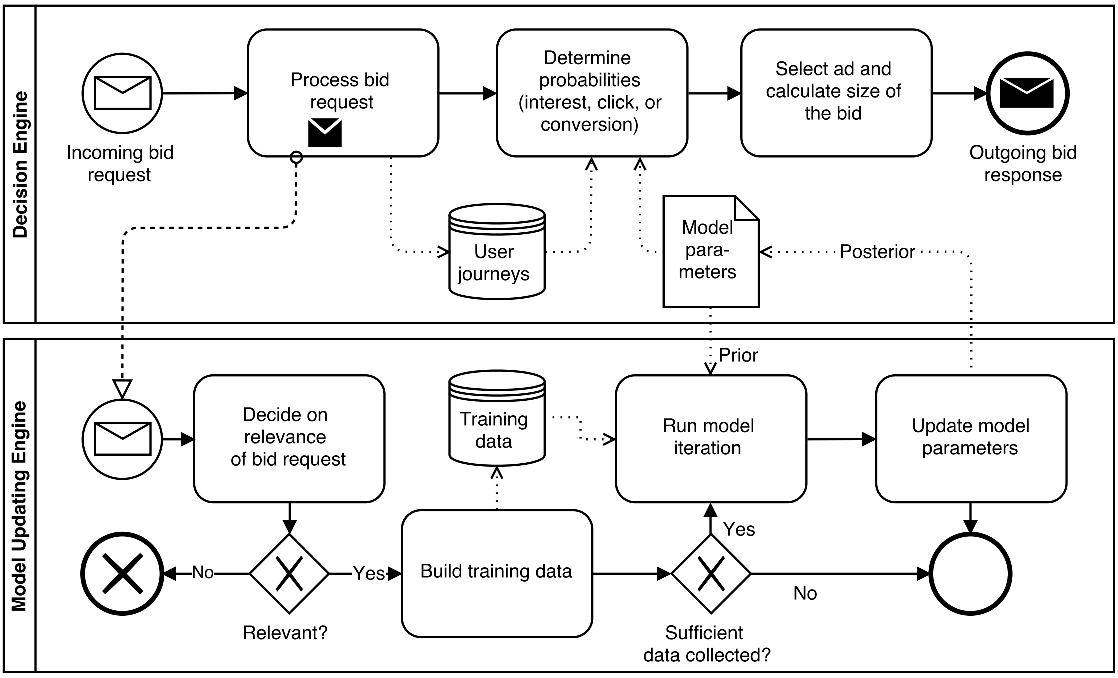

FIGURE 7.1: Bidding and model updating process. Each Bayesian model iteration uses prior information from the previous iteration.

The process begins with the incoming bid request, which is analyzed and stored by the decision engine. Based on prior knowledge of the current user's interests, the decision engine selects the advertising material and the size of the bid and sends a bid response to the ad exchange. Simultaneously, the bid request is forwarded to the model updating engine, which first assesses the relevance of the bid request for the analysis. Depending on its relevance, the bid request is discarded or stored in a training database. The Bayesian analysis is executed when the amount of data reaches a predetermined minimum sample size. Afterwards, the posterior information is sent to the decision engine, which then uses the updated parameters for prediction as new bid requests are processed. Since the model is trained with a stratified sample, the prediction algorithm must rescale the true probabilities of the individual classes in accordance with Equation 7.3.

7.4 Data Description and Preparation

Our data set was provided by a German cross-media online marketing agency and consists of bid requests from a major ad exchange in the form of URL query strings. A query string contains the URL of the website triggering the bid request, the anonymized ID of the current user, location information on the user, and the timestamp of the visit (Figure 7.2)[1].

The data set contains 3 Tbytes of bid request data for a time period of 4 days. During this period, over 1.4 billion bid requests were triggered by over 35 million unique users. The data set contains 35,058,383 users who never visited the websites of the TV stations during their journeys, 275,167 users who visited rtl.de, 56,416 users who visited rtl2.de, 3,978 users who visited vox.de, 3,529 users who visited sat1.de, and 6,738 users who visited prosieben.de.

```
h=http%3A%2F%2Fwww.jetztspielen.de%2Fvda%2friendlyiframe_html_
40.2.1&t=1396894441.691&id=7358864011747200610&ip=93.84&s=DE&c
=Ludwigsburg&a=Mozilla%2F5.0+%28compatible%3B+MSIE+9.0%3B+Wind
ows+NT+6.0%3B+Trident%2F5.0%29

h=http%3A%2F%2Fwww.ebay.de%2F&t=1396894441.692&id=130762252527
275372&ip=88.65&s=DE&c=Munich&a=Mozilla%2F5.0+%28iPad%3B+CPU+O
S+7_1+like+Mac+OS+X%29+AppleWebKit%2F537.51.2+%28KHTML%2C+like
+Gecko%29+Version%2F7.0+Mobile%2F11D167+Safari%2F9537.53%2Cgzi
p%28gfe%29
```

FIGURE 7.2: Examples of bid requests from the raw data. The variable h represents the triggering URL, t the timestamp, id the ID of the current user. In our approach we do not include information on user agents (a) and geographical information (s, c). However, in a real-life situation, these variables may lead to more accurate predictions.

The process of data preparation is often not discussed in the literature, but we would argue that this process deserves critical attention because it may have implications for practitioners who intend to use our model. For this reason, we share our experience with the community and briefly describe how we transformed the data into user journeys. First, we stored the raw data (text files, each capturing one minute of traffic) in a Hadoop file system and then accessed it with Apache Spark. To reduce the number of features, we stripped the URL data after the top-level domain and aggregated the resulting list. We ranked the URLs by visit and encoded the URLs with their positions in the list. We removed all websites with fewer than 5 visits in the data set, resulting in over 500,000 base URL entries. Second, we grouped the data set by user ID and defined the dependent variable for each user by labeling the user with one of the six classes. Our sample does not contain users who visited more than one websites of a TV station, because we truncated all users journeys after the first visit of a TV station website to avoid a bias originating from bid requests that are directly related with the TV station website. Finally, we wrote each user journey into one line of an output file, starting with the class followed by a list of numbers separated with commas. These numbers represent the URLs a user visited during his or her journey. This process resulted in a text file of 4 Gbytes.

[1]Note, that it was not possible to draw conclusions about users' personal information from the data at any time.

7.5 Results

In this chapter, we first describe the results of the iterative parameter estimation. Second and third, we report the misclassification error of the model on a stratified holdout sample and on a random holdout sample. Fourth, we evaluate the model by calculating its economic value.

7.5.1 Parameter Estimation

We split the data into a training set (75% of the available data set) and a holdout set (25% of the available data set). To train the model iteratively, we split the training set into 75 separate batches, each containing 1,600 data records. Each batch contains 500 records of Class 0, 400 records of Class 1, 300 records of Class 2, 100 records of Class 3, 100 records of Class 4, and 200 records of Class 5. This ratio of classes is loosely based on the number of each classes' instances in the complete data set on logarithmic scale. We run the analysis with 15,000 parameters, i.e., we included the 15,000 most frequently visited websites into the model.

Each run of the MCMC sampler consists of 4 chains with 3,000 burn-in iterations and 3,000 sampling iterations. We keep every tenth record from the posterior sample to avoid auto-correlation of the Gibbs sampler. Each iteration took approximately 40 minutes on an Intel i7 4820K processor with 3.7 GHz. Figures 7.3 and 7.4 show the densities of four selected parameters for different iterations. Both figures indicate that the precisions of the parameters' posterior densities increase with the number of iterations. A comparison of Figure 7.3 and 7.4 shows that the less frequent a variable is included in the data, the less precise the estimation of its parameter, and the more iterations are required to obtain a desired parameter precision.

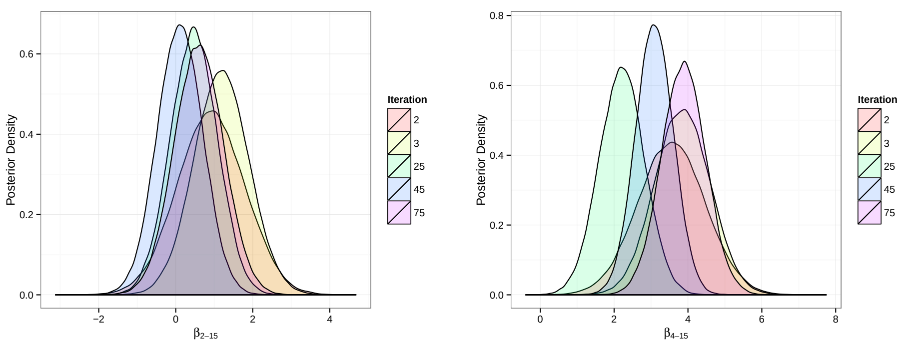

FIGURE 7.3: Posterior densities of $\beta_{2,15}$ and $\beta_{4,15}$.

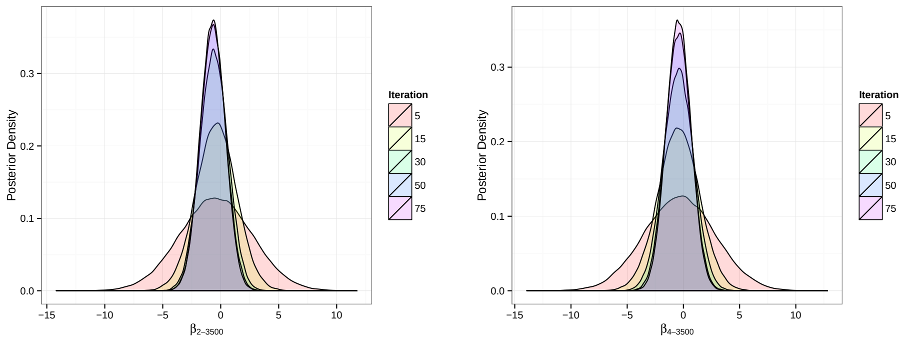

FIGURE 7.4: Posterior densities of $\beta_{2,3500}$ and $\beta_{4,3500}$.

7.5.2 Misclassification Error

After each iteration, we calculate the misclassification error of the model on a stratified holdout sample. The baseline of the misclassification error is established by making a random guess. This approach results in an error of 83.3% when, as in this case, an equal number of instances of the six classes is included in the holdout set. Figure 7.5 shows that the misclassification error decreases with the number of iterations, i.e., the more data is considered for training the model, the more likely it is that a data record is classified correctly.

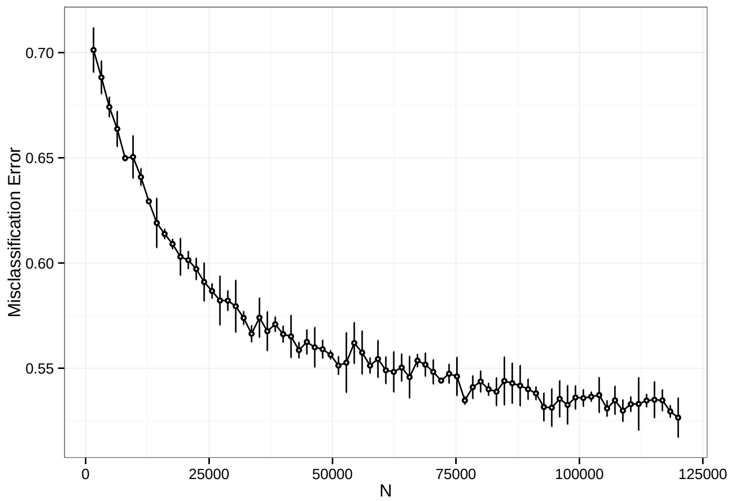

FIGURE 7.5: Misclassification error on the holdout sample depending on the number of training samples. The misclassification error seems to converge to a minimum as more and more iterations of the analysis are performed.

Figure 7.6 presents the confusion matrix describing the misclassification error of the model using the parameters obtained from the 75^{th} model iteration.

To determine how many user contacts have to be observed to achieve a desired misclassification error, we examine the relationship between the misclassification error and the observed user journey lengths. Figure 7.7 shows that the misclassification rate decreases as user journey length increases. This result is expected:

True Class \ Predicted Class	0	1	2	3	4	5
0	**1249**	532	421	224	75	113
1	709	**858**	612	237	84	114
2	574	524	**976**	226	113	201
3	494	454	441	**1133**	45	47
4	164	179	304	67	**1632**	268
5	198	201	363	70	204	**1578**

FIGURE 7.6: Confusion matrix of the prediction on the stratified holdout set after the last iteration of the analysis (misclassifcation error: 52.7%).

The more data on users is available, the more accurate the prediction. The blue line in Figure 7.7 shows, for instance, that users who have been observed at least on 10 different websites are classified correctly in 65% of all cases (i.e., the misclassification error is approximately 35%). The red line shows that users who have been observed 8 to 10 times are classified correctly in more than 60% of all cases (misclassification error of 40%). The green line shows that users who have been observed less than 6 times will be classified correctly in 40% of all cases (misclassification error of 60%).

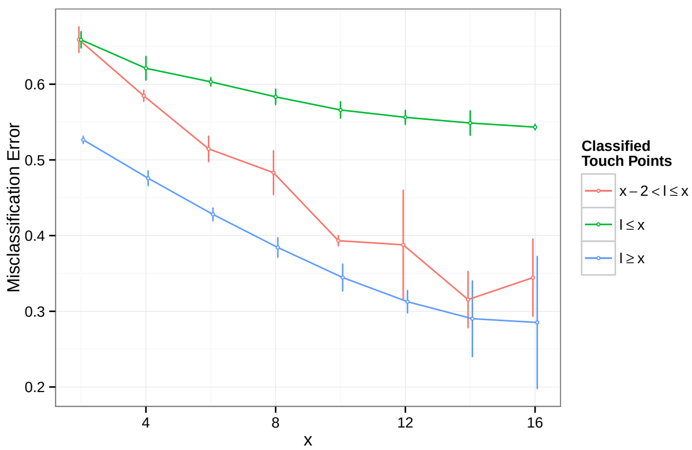

FIGURE 7.7: Misclassification error vs. user journey length. The more information on users is available, the more accurate the prediction of user classes.

7.5.3 Predictions Based on a Random Sample

Stratified sampling is a feasible way to benchmark a model such as the one described here. However, in a real-life scenario, the distribution of user classes is greatly unbalanced. For this reason, we rescale the parameter estimations to predict user classes on the random sample in accordance with Equation 7.3.

We calculate the probabilities $p(K_i = l|X_i)$ to classify each record of a random holdout set. We obtain a rather small misclassification error of 2.5% (Figure 7.8). This value is so small, because of the high number of records of Class 0. Thus, classifying a record with $K = 0$ is nearly always correct. Even a model that yields $p(K = 0) = 1$ for each test record would nearly always classify correctly. For this reason, using the misclassification rate is not a meaningful means to evaluate the model. Instead, we calculate the economic value of applying the model, and, thereby determine the economic value of bid request data.

True Class \ Predicted Class	1	2	3	4	5	6
1	**584751**	6457	1528	58	496	1981
2	3467	**129**	18	0	4	18
3	719	13	**24**	1	0	5
4	45	0	1	**1**	0	0
5	69	2	1	0	**19**	0
6	152	4	0	0	0	**42**

FIGURE 7.8: Confusion matrix obtained from applying the model to a random holdout set containing 600,000 records. The misclassification rate is 2.5%.

7.5.4 Economic Value of Bid Request Data

We determine the economic value of bid request data from the perspective of an advertiser who employs a bidding agent that places bids based on predictions made by our model. The bidding agent would not place a bid for predictions of Class 0, i.e., it would not answer to bid requests triggered by users who are unlikely to be interested in one of the TV programs. In case the prediction of Class 0 is correct, the behavior of the bidding agent produces no costs (true negative prediction). Otherwise, i.e., the current user is in fact interested in one of the TV programs (false negative prediction), the behavior of the bidding agent produces opportunity costs. These costs are determined by the lost contribution margin for not exposing a user to an ad, who would have clicked on the ad or at least have been attracted by it. For predictions of Class 1 through 5, the bidding agent would always place a bid. Consequently, it would produce costs that are equal to the costs of the ad impressions. In addition, it would generate benefits that can be derived by exposing an interested user to an ad that matches his or her interests (true positive prediction).

Based on the aforementioned scheme, we estimate the economic value of bid request data by applying our model to a random holdout sample. We assume the typical costs in the industry for ad impressions and typical benefits from user clicks on display ads at the time of writing. For false negative predictions, we assume costs ranging from 0.01 EUR through 0.40 EUR. We define a range of costs here, because the contribution margin may vary for different advertising scenarios. For false positive predictions, we assume a value of 0.001 EUR, i.e., the costs for an ad impression. For true positive predictions we assume a benefit equal to the contribution margin (i.e., 0.01 EUR through 0.40 EUR) minus the costs for the ad impression (0.001 EUR). A true negative prediction, i.e., no banner is shown to a user who would not have clicked, has no costs at all. For false predictions concerning the Classes 1 through 5, we also assume no costs because users belonging to these classes might, in general, be interested in products they see advertised on TV, but at another TV station. We assume that the costs for the ad impression is annulled by the benefit through the branding effect.

To obtain the best balance between true and false predictions and their costs and benefits we seek for the optimal cutoff value p_{cut} for $p(K > 0)$ that minimizes the costs of the classification by iteratively classifying the data from the holdout set with different values for p_{cut}. The cutoff value is used as follows: All user contacts X_i that show a probability $p(K = 0)$ less than p_{cut} are classified based on the probabilities of the other classes, i.e., $p(K \in \{1, \ldots, 5\})$. All other user contacts X_i are classified as class $K = 0$. Thus, when p_{cut} is very small, nearly all contacts are classified as one class between 1 and 5, and, as p_{cut} is increases, more and more contacts X_i are categorized as class $K = 0$.

To calculate the benefits related to a given cutoff value, we define the maximum benefit per decision of a model B as the minimum of the costs for showing no impressions at all (C_{NI}) and the costs for showing impressions to all users (C_{AI}) minus the minimum costs of applying the model (C_M) using the optimal cutoff value (Equation 7.4):

$$B = \min(C_{NI}, C_{AI}) - \min(C_M) \tag{7.4}$$

Figure 7.9 presents the benefits per 1,000,000 decisions based on different benefit/cost ratios. The diagram indicates that using the model for classification is valuable in a relatively narrow range of benefit/cost ratios. On the left side of the ratio range (below a benefit/cost ratio of 40), the advertiser would never bid for a free ad slot, because the overall costs of impressions are higher than the sum of contribution margins. Above the ratio of 40/1, the advertiser is best advised to always use the model for classification. The highest benefit can be achieved when the benefit/cost ratio is at 75/1. As the ratio increases, the benefit that can be generated by applying the model decreases.

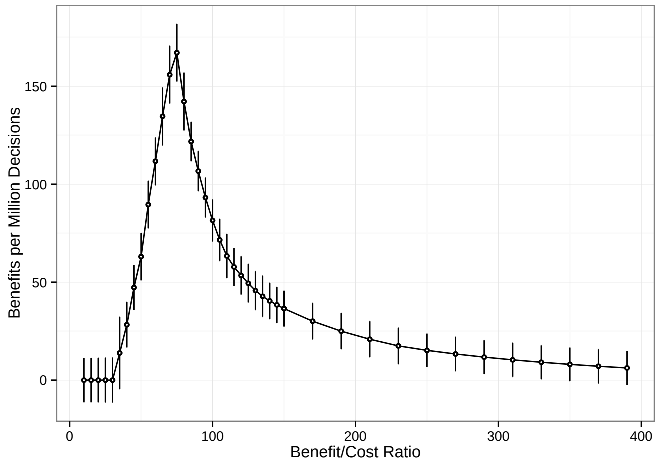

FIGURE 7.9: Max. benefits vs. cost benefit ratio. The model is most beneficial when the benefit/cost ratio is 75/1, i.e., when the costs for a lost contribution margin are 75 times greater than the costs for an impression. The optimal cutoff value for this benefit/cost ratio is $p_{cut} = 0.007898$.

7.6 Conclusion

In this paper, we develop a model employing bid request data that can be used to derive users' interests in certain products or services. Our approach has implications for researchers and practitioners from IS and marketing (Spann et al., 2013).

7.6.1 Implications

The model developed here enables researchers and practitioners to make predictions about users' interests for arbitrary products or services by using corresponding websites as dependent variables. A company offering smart phone contracts, for instance, can use bid requests triggered by a mobile phone product website to identify users who are interested in a new smart phone very early, namely before they visit the website of the mobile phone company. In this way, our approach extends the landscape of methods and techniques to target users based on cookie data or third-party data. In marketing research, the method can be used to measure the impact of television ads on the users' online shopping behavior more accurately because it allows for identifying users who have watched a certain TV program recently (Stange, 2015). In addition, we contribute to IS by developing a framework to extend and improve existing decision support systems employed in e-commerce and marketing to target users based on cookie data. Furthermore, we provide a method to measure the impact of bid request data that is based on the valuation of the analysis conducted in our case study.

In practice, the method can be used to increase profits from cross-channel advertising campaigns. In particular, awareness-oriented campaigns can benefit from the proposed model. These campaigns aim to reach a broad audience early in the sales funnel where relatively little is known about a given user's interests. Thus, using bid request data streams in addition to cookie data and other third-party data is a promising possibility to improve effects from branding campaigns. The integration into existing decision engines employed by ad exchanges or demand side platforms is relatively easy because of the iterative approach of the analysis. They could use the proposed method to extend their portfolio of targeting services. Aside from cross-channel advertising, the probability of a user being interested in certain content can be used by e-commerce companies to customize their products and services.

7.6.2 Limitations

Although the approach proposed here suggests successful possibilities for using bid request data to predict user behavior, it also has primarily five limitations: First, our bid request data was not filtered for any data from background processes such as ad servers, which are not necessarily directly related to the user journey. In addition, the data contains websites in different languages. These websites should eventually be excluded from the data because is very likely that a user is not interested in a German TV program if he or she has only visited French or Polish websites during his or her journey. In addition, we expect a bias in our training data because it is likely that not all (sub) websites of the TV stations we focused on use real-time advertising to sell their advertising inventory. Hence, users who exclusively visited such (sub) websites are falsely labeled with Class 0 in our training data. Second, the Gibbs sampler used here to estimate the parameter values during the iterative analysis is relatively slow. It takes about 40 minutes to run 4 chains simultaneously with 6.000 sampling iterations on an Intel i7 4820K processor. This duration might be unsuitable for some applications, especially when the number of relevant bid requests is much higher than in our case study. However, because of the statistical independence of the parameters, it is possible to parallelize sampling of parameters on GPUs using programming languages such as CUDA-C. These improvements would reduce computational costs to a great extent. Another opportunity to speed up computation time is to employ variational Bayesian methods. These methods can be used to approximate parameters of simple models such as the one presented here. However, the complexity of models that can be estimated with these methods is limited. Third, we assume statistical independence between the URLs to make the model computationally tractable. However, website visits by a given group of users with the same interests are generally not statistically independent. In addition, using TV station websites as proxy variables to determine users' TV consumption behavior can only be an approximation because many users might tend to watch the program of a certain TV station but never visit its website. Fourth, the valuation of the model is simplified, because it assumes that every impression can be bought at the same price. In real-time advertising, however, the general approach is to pay individual prices for individual users. Thus, the calculated benefits based on true and false predictions is only an approximation of the true benefit that could be achieved in a real-life system. Fifth, our approach is not intended to replace current methods applied in behavioral targeting because only relying on bid request data streams could lead into a 'big data trap' (Lazer et al., 2014). Therefore, we recommend to use the proposed method as an extension to techniques and tools used in behavioral targeting. More conceptual research is required on how to integrate predictions from our model into existing bidding agents in RTA.

7.6.3 Outlook

Unlike typical bid request data from other major ad exchanges, our data set neither contains information on free advertising slots nor any contextual information. The model does not contain time-dependent variables that are commonly used in the user journey analyses, such as the number of contacts with a certain website or the time difference between contacts (Chatterjee et al., 2003). However, due to the flexible approach using Bayesian estimation that employs MCMC sampling, it is rather simple to extend the model by new variables and hierarchy levels. For instance, website languages or user locations could be used to determine users' interests depending on their location or spoken languages. In addition, text-mining techniques could be used to derive context information from websites, which could be used as additional independent variables. This kind of information could be very useful for predicting user behavior and would enable e-commerce companies to understand their customers' behavior even better. The high number of parameters could be reduced by applying regularized Bayesian regression that employs non-normal prior distributions Kyung et al., 2010. However, further investigation is needed on how to use non-normally distributed posteriors as prior information in an iterative modeling approach such as the one presented here.

We present an approach to predict the probability for a user being interested in a certain product. It is clear that these predicted probabilities are not the only values that bidding agents require to place a bid. The integration of predictions from our model into real-life bidding agents, which also need to consider all other available information about customers, publisher websites, advertisers' budgets, and advertising materials is still a challenge (Lee et al., 2013).

In summary, we demonstrate that bid request data is a very promising source of big data on the Internet that is worth further investigation by researchers at the intersection of IS and marketing.

7.7 References

Bucklin, R. E. and C. Sismeiro (2009). “Click here for Internet insight: Advances in clickstream data analysis in marketing”. In: *Journal of Interactive Marketing* 23.1, pp. 35–48.

Chatterjee, P., D. L. Hoffman, and T. P. Novak (2003). “Modeling the clickstream: Implications for web-based advertising efforts”. In: *Marketing Science* 22.4, pp. 520–541.

Chen, H., R. H. L. Chiang, and V. C. Storey (2012). “Business Intelligence and Analytics: From Big Data to Big Impact”. In: *Management Information Systems Quarterly* 36.4, pp. 1165–1188.

Chen, Y. and P. Berkhin (2011). “Real-time bidding algorithms for performance-based display ad allocation”. In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1307–1315.

Dinner, I., H. J. van Heerde, and S. A. Neslin (2014). “Driving Online and Offline Sales: The Cross-channel Effects of Traditional, Online Display, and Paid Search Advertising”. In: *Journal of Marketing Research* 51.5, pp. 527–545.

Duan, W. and J. Zhang (2014). “The Impact of Referral Channels in Online Customer Journey”. In: *Proceedings of the 35th International Conference on Information Systems, Auckland, New Zealand*.

Joo, M., K. C. Wilbur, and Y. Zhu (2015). “Effects of TV advertising on keyword search”. In: *International Journal of Research in Marketing, Available online 17 April 2015, In Press*.

King, G. and L. Zeng (2001). “Logistic regression in rare events data”. In: *Political analysis* 2.1, pp. 137–163.

Kitts, B., M. Bardaro, D. Au, A. Lee, S. Lee, J. Borchardt, C. Schwartz, J. Sobieski, and J. Wadsworth-Drake (2014). “Can Television Advertising Impact Be Measured on the Web? Web Spike Response as a Possible Conversion Tracking System for Television”. In: *Proceedings of 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, New York, NY, USA*.

Kyung, M., J. Gill, M. Ghosh, and G. Casella (2010). “Penalized regression, standard errors, and Bayesian lassos”. In: *Bayesian Analysis* 5.2, pp. 369–412.

Lazer, D., R. Kennedy, G. King, and A. Vespignani (2014). “The Parable of Google Flu: Traps in Big Data Analysis”. In: *Science* 343, pp. 1203–1205.

Lee, K., B. Orten, A. Dasdan, and W. Li (2012). “Estimating conversion rate in display advertising from past erformance data”. In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, Beijing, China*, pp. 768–776.

Lee, K. C., A. Jalali, and A. Dasdan (2013). “Real Time Bid Optimization with Smooth Budget Delivery in Online Advertising”. In: *Proceedings of the Seventh International Workshop on Data Mining for Online Advertising, Chicago, IL, USA*, pp. 1–13.

Liaukonyte, J., T. Teixeira, and K. C. Wilbur (2015). “Television Advertising and Online Shopping”. In: *Marketing Science* 34.3, pp. 311–330.

Nottorf, F. and B. Funk (2013). “The economic value of clickstream data from an advertiser's perspective”. In: *Proceedings of the 21st European Conference of Information Systems, Utrecht, Belgium*.
Pandey, S., M. Aly, and A. Bagherjeiran (2011). “Learning to target: what works for behavioral targeting”. In: *Proceedings of the 20th ACM international conference on Information and knowledge management, Glasgow, United Kingdom*, pp. 1805–1814.
Perlich, C., B. Dalessandro, R. Hook, O. Stitelman, T. Raeder, and F. Provost (2012). “Bid optimizing and inventory scoring in targeted online advertising”. In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China*, pp. 804–812.
Plummer, M. (2003). “JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling”. In: *Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna*.
Spann, M., O. Hinz, and V. Ramachandran (2013). “Business and Information Systems Engineering and Marketing”. In: *Business & Information Systems Engineering* 55.3, pp. 119–120.
Stange, M. (2015). “The Impact of TV Ads on the Individual User's Purchasing Behavior”. In: *Proceedings of the 36th International Conference on Information Systems, Fort Worth, United States*.
Stange, M. and B. Funk (2014). “Real-Time Advertising”. In: *Business & Information Systems Engineering* 56.5, pp. 305–308.
Veit, D., E. Clemons, A. Benlian, P. Buxmann, T. Hess, D. Kundisch, J. M. Leimeister, P. Loos, and M. Spann (2014). “Business Models”. In: *Business & Information Systems Engineering* 6.1, pp. 45–53.
Yang, S. and A. Ghose (2010). “Analyzing the Relationship Between Organic and Sponsored Search Advertising: Positive, Negative, or Zero Interdependence?” In: *Marketing Science* 29.4, pp. 602–623.
Yuan, S., J. Wang, and X. Zhao (2013). “Real-time bidding for online advertising: measurement and analysis”. In: *Proceedings of the Seventh International Workshop on Data Mining for Online Advertising, Chicago, IL, USA*.
Zhang, W., S. Yuan, and J. Wang (2014). “Optimal real-time bidding for display advertising”. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining New York, NY, USA*.