HOW MUCH TRACKING IS NECESSARY? – THE LEARNING CURVE IN BAYESIAN USER JOURNEY ANALYSIS

Complete Research

Stange, Martin, Leuphana University, Lueneburg, Germany, martin.stange@uni.leuphana.de
Funk, Burkhardt, Leuphana University, Lueneburg, Germany, funk@uni.leuphana.de

Abstract

Extracting value from big data is one of today's 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.

Keywords: User Journey Analysis, Learning Curve, Big Data, Bayesian Models.

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 publisher's 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 customer's 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.

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.

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.

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; Jacob et al., 2011; Henriksen, 2012) 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.

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 user's individual click proneness based on clickstream data. In their study, a random effects logit model was employed to predict a consumer's 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., 2013).

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 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 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.

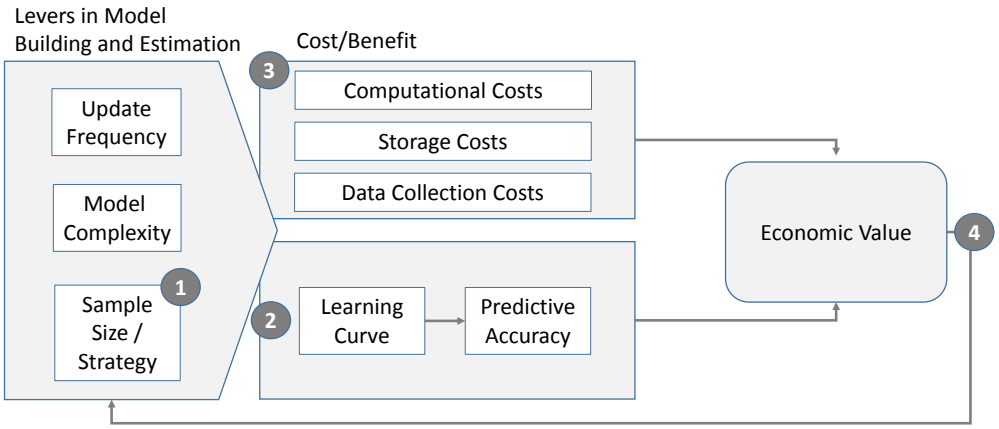

Figure 1. Levers in model building and estimation

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.

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.

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 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 .5%), which is sufficient for receiving robust estimations.

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 1.

Interaction Number	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 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 user's design matrix D_i. Refer to the left hand side of Table 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 2.

$$\begin{aligned} \text{First case study:}(D_i)_j \quad = \quad & \{I_0, I_{SEO}, I_D, I_A, I_{Ban}, I_{SEA}, I_{EM}, I_R, \\ & X_{SEO}, X_R, X_A, X_{SEA}, X_{EM}, \\ & Y_{SEO}, Y_D, Y_A, Y_{Ban}, Y_{SEA}, Y_C, Y_{EM}, Y_R, \\ & SN, IST, OCWS, OCPS, CPS, TV\} \end{aligned} \tag{1}$$

$$\begin{aligned}\text{Second case study: } (D_i)_j &= \{I_0, I_{SEO}, I_{SEA}, I_{Ban}, I_{PS}, I_A, I_{EM},\\ &\quad X_{SEO}, X_{SEA}, X_{Ban}, X_{PS}, X_{EM},\\ &\quad Y_{SEO}, Y_{SEA}, Y_{PS}, Y_{EM},\\ &\quad SN, IST, OCWS, OCPS\} \end{aligned} \tag{2}$$

Index	Channel	Index	Additional Variables
SEA	Search engine advertisement	OCWS	Onsite contacts within the current session
SEO	Organic search	OCPS	Sum of onsite contacts in previous 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 previous sessions
D	Direct type-in		
C	Cooperation link		
PS	Price search engine		
EM	Email advertisement		

Table 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 through 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, 2014) as shown in equation 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.

$$\begin{aligned} Conv_{mn} &\sim \text{Bernoulli}(\theta_{mn})\\ \text{logit}(\theta_{mn}) &= X_{mn}\beta \end{aligned} \tag{3}$$

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 4.

$$\begin{aligned} Conv_{mn} &\sim \text{Bernoulli}(\theta_{mn})\\ \text{logit}(\theta_{mn}) &= X_{mn}\beta + I_0 b_m \end{aligned} \tag{4}$$

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 5.

$$\begin{aligned} Conv_{mn} &\sim \text{Bernoulli}(\theta_{mn})\\ \text{logit}(\theta_{mn}) &= X_{mn}\beta_m\\ \beta_m &= \beta + \delta_m. \end{aligned} \tag{5}$$

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 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.

4.3 Results

To show the convergence of the predictive accuracy, we execute 6 analyses including 1000, 2000, 4000, 8000, 16000 and 32000 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. 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).

Sample Size in 1000	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	1066	41	81	161	323	649	1303
Random Slope	23	36	66	129	245	478	22	38	68	138	282	557

Table 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.

The results from the first case study are presented in Table 4. We report the results based on 32000 user journeys for the simple logit model and the random intercept model. The results from the computation based on the other sample sizes and the random slope model are not reported due to space limitations[1]. However, the convergence of the variables is demonstrated in Figure 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.

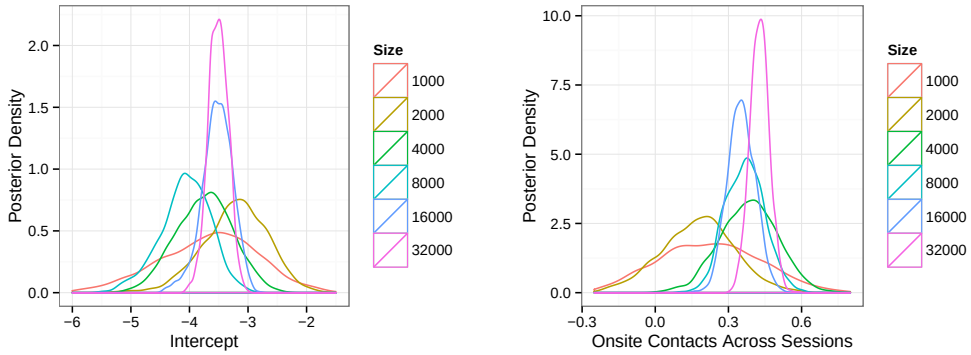

Figure 2. Density plots for β_{I^0} and β_{OCPS} from the simple logit model from the first case study. The HDI lengths decrease with the increase of the sample size.

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

[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 4. Results of the random intercept model (left) and the simple logit model (right) from the first case study (Sample Size = 32000). Significant values are printed in boldface.

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.

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 2.

The results from the second case study are presented in Table 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

Parameter	2.5%	50%	97.5%	2.5%	50%	97.5%
β_{I^0}	**-4.55**	**-4.05**	**-3.54**	**-4.46**	**-4.05**	**-3.69**
$\beta_{I^{SEO}}$	**-0.50**	**-0.33**	**-0.14**	-0.54	-0.18	0.22
$\beta_{I^{SEA}}$	**-0.72**	**-0.54**	**-0.34**	-0.76	-0.41	0.00
$\beta_{I^{Ban}}$	**-1.54**	**-1.14**	**-0.67**	**-1.87**	**-1.23**	**-0.57**
$\beta_{I^{PS}}$	-0.30	-0.15	0.10	-0.41	0.01	0.44
$\beta_{I^{A}}$	**0.22**	**0.49**	**0.73**	**0.16**	**0.62**	**1.09**
$\beta_{I^{EM}}$	**-0.48**	**-0.33**	**-0.14**	-0.50	-0.09	0.35
$\beta_{X^{SEO}}$	**-0.18**	**-0.16**	**-0.12**	**-0.31**	**-0.20**	**-0.11**
$\beta_{X^{SEA}}$	**-0.15**	**-0.11**	**-0.05**	**-0.19**	**-0.12**	**-0.05**
$\beta_{X^{Ban}}$	-0.01	0.13	0.34	-0.08	0.17	0.36
$\beta_{X^{PS}}$	**0.01**	**0.06**	**0.14**	-0.01	0.07	0.14
$\beta_{X^{EM}}$	-0.10	-0.02	0.03	-0.17	-0.07	0.01
$\beta_{Y^{SEO}}$	**-0.15**	**-0.11**	**-0.07**	**-0.19**	**-0.10**	**-0.03**
$\beta_{Y^{SEA}}$	**-0.29**	**-0.23**	**-0.14**	**-0.19**	**-0.11**	**-0.03**
$\beta_{Y^{PS}}$	**-0.23**	**-0.13**	**-0.07**	-0.16	-0.04	0.06
$\beta_{Y^{EM}}$	-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 5. Results of the random intercept model (left) and the simple logit model (right) from the second case study (Sample Size = 32000). Significant values are printed in boldface.

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.

Like in the first case study, we observe convergence of the β values with increasing sample size. The density plots for β_{I^0} and β_{OCPS} are shown in Figure 3.

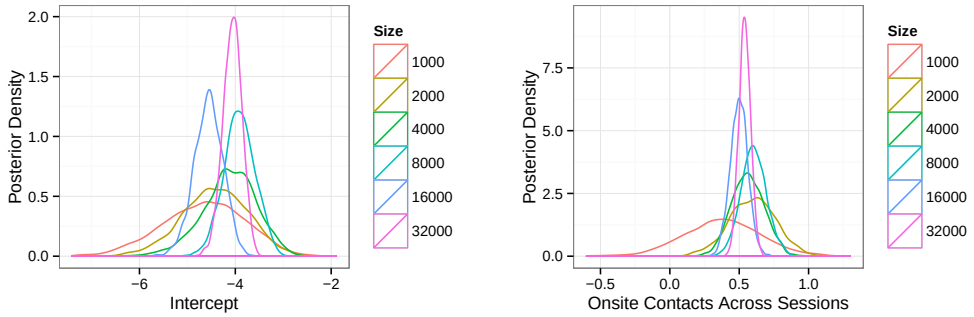

Figure 3. Density plots of β_{I^0} and β_{OCPS} from the simple logit model from the second case study. As in the first case study, the HDI lengths decrease with the increase of the sample size.

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 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 32000 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 > 8000 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.

For convenience, we report the benefit per 1000 decisions. The right hand side of Figure 4 shows the maximum benefit per 1000 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 16000 and 32000 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 16000 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 5. The AUCs of all three models show convergence when the sample size is increased. The best AUC is already achieved at 8000 user journeys. Consequently the benefit per 1000 decisions do not further grow, as the sample size is increased. However, the standard deviations of the maximum benefit per 1000 decisions is higher as

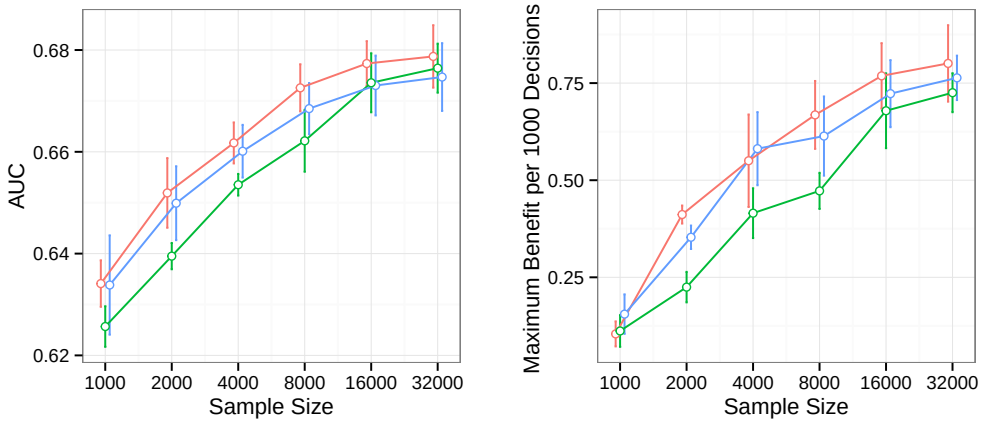

Figure 4. Results from the first case study: Area under the curve and the maximum benefit per 1000 decisions of the simple logit model (blue), random intercept model (red) and the random slope model (green). Note the logarithmic scale on the abscissa.

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 1000 decisions.

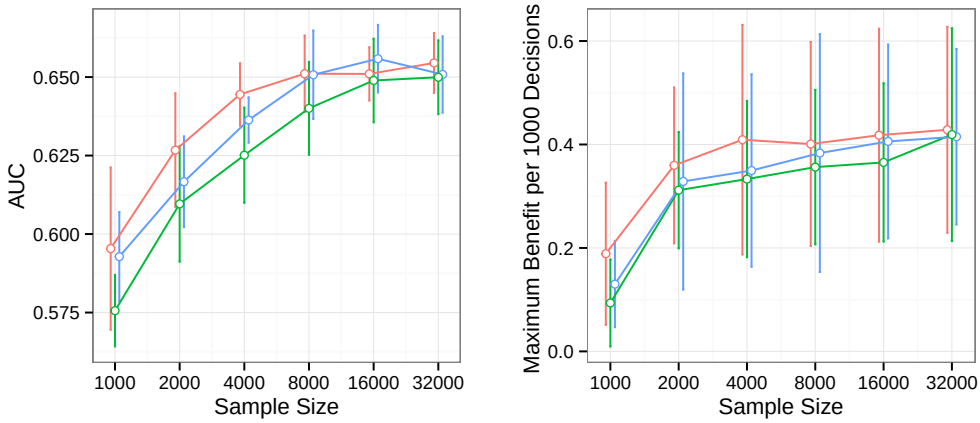

Figure 5. Results from the second case study: Area under the curve and maximum benefit per 1000 decisions of the simple logit model (blue), random intercept model (red) and the random slope model (green).

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 4 and 5). Overall, the random slope model performs worst (except 32000 users, first case study). This implies that the effect of the individual marketing channels, such as display advertisement or TV spots, does not differ significantly between during the day and at night for either case study.

5 Concluding Remarks

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.

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.

References

Amazon. (2014). "Amazon Web Services Simple Monthly Calculator," *http://calculator.s3.amazonaws.com*

Anderl, E. and Becker, I., v. Wangenheim, F. and Schumann, J.H. (2014), "Mapping the Customer Journey: A Graph-Based Framework for Online Attribution Modeling", http://dx.doi.org/ 10.2139/ssrn.2343077

Bradley, A. (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., Santis, F. De, and Gubbiotti, S. (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., and Sismeiro, C. (2003). "A model of web site browsing behavior estimated on clickstream data," In: *Journal of Marketing Research* (40:3), pp. 249–267.

Bucklin, R., and Sismeiro, C. (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., Hoffman, D., and Novak, T. (2003). "Modeling the clickstream: Implications for web-based advertising efforts," In: *Marketing Science* (22:4), pp. 520–541.

Chen, Y., and Berkhin, P. (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.

Cho, C.-H. (2003). "Factors influencing clicking of banner ads on the WWW," In: *Cyberpsychology & behavior* (6:2), pp. 201–215.

Cohen, M. (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*, 48-55.

Deelman, E., Singh, G., and Livny, M. (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., Kim, J., and Rotz, W. 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., Hu, F., and Liu, H. (2001). "Modelling Classification Performance for Large Data Sets," In: *Advances in Web-Age Information Management*, pp. 317–328.

Henriksen, S., Wills, A., Schön, T., and Ninness, B. (2012). "Parallel implementation of particle MCMC methods on a GPU," In: *System Identification*, (16:1), 1143–1148.

Jacob, P., Robert, C. P., and Smith, M. H. (2011). "Using parallel computation to improve independent Metropolis-Hastings based estimation," In: *Journal of Computational and Graphical Statistics*, (20:3), 616-635.

Joo, M., Wilbur, K. C., Cowgill, B., and Zhu, Y. (2014). "Television Advertising and Online Search," In: *Management Science*, (60:1), 56–73.

Joseph, L., Wolfson, D., and Du Berger, R. (1995). "Sample size calculations for binomial proportions via highest posterior density intervals," In: *The Statistician* (44:2), pp. 143–154.

King, G., and Zeng, L. (2001). "Logistic regression in rare events data," In: *Political analysis* (2:1), pp. 137–163.

Klapdor, S. (2013). "Effectiveness of Online Marketing Campaigns", Vasa, Springer Gabler.

Lee, K., Jalali, A., and Dasdan, A. (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*.

Lee, K., Orten, B., Dasdan, A., and Li, W. (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* , pp. 768–776.

Martin, A. D., Quinn, K. M., and Park, J. H. (2011). "MCMCpack: Markov Chain Monte Carlo in R," In: *Journal of statistical software* (42:9)

Meek, C., Thiesson, B., and Heckerman, D. (2002). "The learning-curve sampling method applied to model-based clustering," In: *The Journal of Machine Learning Research* (2), pp. 397–418.

Nottorf, F., and Funk, B. (2013). "The economic value of clickstream data from an advertiser's perspective," In: *Proceedings of the 22nd European Conference of Information Systems*.

Perlich, C., Dalessandro, B., and Hook, R. (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 ACM Press*, 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*.

Sahu, S., and Smith, T. (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.

Rossi, P., and McCulloch, R. (2010). "Bayesm: Bayesian inference for marketing/micro-econometrics," R package version: 2.2.

Santis, F. (2007). "Alternative Bayes factors: Sample size determination and discriminatory power assessment," In: *Test* (16:3), pp. 504–522.

Stange, M., and Funk, B. (2014). "Real-Time Advertising," In: *Business & Information Systems Engineering*, 56(5), pp. 335–338.

Wang, F., and Gelfand, A. E. (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., Polson, N. G., and Scott, J. G. (2014). "BayesLogit: Bayesian logistic regression," R package version: 0.5.1.

Yang, S., and Ghose, A. (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*