Intelligent Information Management, 2023, 15, 180-215
https://www.scirp.org/journal/iim
ISSN Online: 2160-5920
ISSN Print: 2160-5912

Clicking through the Clickstream: A Novel Statistical Modeling Approach to Improve Information Usage of Clickstream Data by E-Commerce Entities

Corban Allenbrand

Analytics, Information, and Operations Management, University of Kansas, Lawrence, USA
Email: callenbrand@ku.edu

How to cite this paper: Allenbrand, C. (2023) Clicking through the Clickstream: A Novel Statistical Modeling Approach to Improve Information Usage of Clickstream Data by E-Commerce Entities. *Intelligent Information Management*, **15**, 180-215.
https://doi.org/10.4236/iim.2023.153010

Received: March 21, 2023
Accepted: May 28, 2023
Published: May 31, 2023

Abstract

Success or failure of an E-commerce platform is often reduced to its ability to maximize the conversion rate of its visitors. This is commonly regarded as the capacity to induce a purchase from a visitor. Visitors possess individual characteristics, histories, and objectives which complicate the choice of what platform features that maximize the conversion rate. Modern web technology has made clickstream data accessible allowing a complete record of a visitor's actions on a website to be analyzed. What remains poorly constrained is what parts of the clickstream data are meaningful information and what parts are accidental for the problem of platform design. In this research, clickstream data from an online retailer was examined to demonstrate how statistical modeling can improve clickstream information usage. A conceptual model was developed that conjectured relationships between visitor and platform variables, visitors' platform exit rate, boune rate, and decision to purchase. Several hypotheses on the nature of the clickstream relationships were posited and tested with the models. A discrete choice logit model showed that the content of a website, the history of website use, and the exit rate of pages visited had marginal effects on derived utility for the visitor. Exit rate and bounce rate were modeled as beta distributed random variables. It was found that exit rate and its variability for pages visited were associated with site content, site quality, prior visitor history on the site, and technological preferences of the visitor. Bounce rate was also found to be influenced by the same factors but was in a direction opposite to the registered hypotheses. Most findings supported that clickstream data is amenable to statistical modeling with interpretable and comprehensible models.

Keywords

Business Intelligence, Intelligent Information Management, Web Analytics,

Web Technology Management, Exit Rate, Bounce Rate, Online Consumer Model, Discrete Choice Model

1. Introduction

Digitally mediated transactions are becoming more ubiquitous as the advantages of speed, convenience, access to personalized options, and quality of online systems increase. In particular, the purchasing and selling of goods electronically, known as E-commerce, continues to demonstrate an increasing share of total sales [1]. Digitalization of economic interactions was stimulated by the COVID-19 pandemic. It is reasonable to predict that usage of E-commerce systems will not abate. Extraction of a benefit from this technology depends on tools that can identify and analyze patterns contained in data generated by the technology, hereafter to be called clickstream data. In the setting of online retail, webometrics, or web analytics (WA), represents a family of methods and tools that can deliver insights to its users. Efficacy and validity of insights from WA are based on many relationships between numerous metrics and user behavior put forth by non-experts.

In the endeavor to construct higher quality web systems, firms become attracted to the many relationships that can be found in WA. This can prove to be inauspicious as conjectured relationships may suffer issues of validity [2]. What is needed is a more disciplined characterization of patterns in the clickstream data so that any informational insights can be validated. When it comes to an E-commerce platform, keeping visitors on the site is a fundamental requirement for success. A determination of which factors are associated with exiting behavior from the website and which factors influence if a visitor buys or not, is paramount for crucial information management.

The paper is organized as follows. In Section 2, background and a survey of relevant literature are offered. In Section 3, a presentation of a clickstream conceptual model based on the supporting literature is provided. Hypotheses regarding the existence, direction, and magnitude of relationships postulated in the conceptual model are also given in Section 3. In Section 4, a description of the data and empirical setting is supplied. In Section 5, the analytic methodologies used to test the hypotheses and evaluate the conceptual model from Section 3 are provided. In Section 6, all results, discussion, and implications are presented. In Section 7, a conclusion is delivered.

2. Literature Review and Background

The Digital Analytics Association defines web analytics (WA) as, "the measurement, collection, analysis, and reporting of internet data for the purposes of understanding and optimizing web usage." This definition is based on several premises: the ability to analyze implies methodologies that can lead from data to

understanding, reporting implies an agreed-upon standard, optimization implies a direct link between technology, people, and context. Its sheer size and popularity as an information source allows limitless possibilities for the measurement and analysis of internet [3] Transaction log analysis and search log analysis were precursors to modern WA [4] [5] [6] [7]. Clickstream data has volume, velocity, and variety that complicate the intuition-based judgment used by humans. To attain business insights from clickstream data, understanding is needed of the diverse types of metrics available for the analysis of user behavior [8] [9].

Analysis and optimization of a website begins with the collection and evaluation of metrics. Standard categories of metrics include,

- Site usage
- Referral and web traffic
- Site content analysis
- Quality Assurance

Site usage encompasses measures such as geographic information and number user visits. Referral and web traffic comprises measures such as source or web traffic onto a site. Site content analysis contains measures for site content effectiveness and top pages regarding exits and value. Quality assurance measures the presence and impact of broken pages and visitor errors [10]. Although metrics are numerous, a sample of a few standard ones are presented in **Table 1**.

Tracking visitor type information is conducive to the personalization of a visitor's experience and identification of emerging customer needs. Tracking errors aids in the identification of troubles experienced by the customer. Direct access to sources of incoming web traffic by referring URLs is a primary way advertising and marketing effectiveness can be measured. Top pages are those areas of a website that receive the most traffic and can be used to confirm and align the website's functionality with the firm's goals. Missing content or areas of confusion may direct visitors away from goal pages of a website resulting in an underperforming website [10].

These metrics come with challenges. For instance, measuring visitor type requires consideration of if a visitor is to be defined at the unit of an individual unique visitor that requires a single IP address and web cookies to track the user or is the visitor to be defined at the unit of a single session which tracks when a user begins and ends his/her interaction with the website [11] [12]. Also consider visit length, which is defined by the subtraction of the time of a user's first

Table 1. Commonly used WA metrics with descriptions and statement of metric category.

Metric	Description	Category
Visitor type	Details of user	Site usage
Errors	Page retrieval errors	Quality assurance
Referring URLs	Sources of traffic to website	Referral and web traffic
Top pages	Pages that receive most views	Site content analysis

activity on the website from the user's final activity; if any of these start and end points are missing, the visit length metric can become zero [13].

Investments in metric tracking and WA infrastructure can be viewed as any business investment. Website data is beneficial only if information gathered from it can be utilized to improve the site. Several methods exist that a firm can follow to become equipped to make good judgements on data collection and goal definition [14] [15]. For E-commerce websites, the objective is to convince visitors to purchase goods or services directly from the site. The principal quantity of interest to an E-commerce site is the conversion rate. The conversion rate quantifies the proportion of visitors to a site that accomplishes the desired goal – to make a purchase. Accurate assessments of conversion rate are crucial for decisions regarding adjustments to web systems. Decisions to filter visitors according to certain exclusion criteria, like intent to purchase or exclusion of web crawlers, would aid firms in allocating resources to the staff and capital necessary to have a functional website [16].

Another valuable WA quantity to examine is traffic quality. Related to the idea of traffic quality are the bounce and exit rates of the website. The essence of these two metrics is that bounce rate quantifies how many arrivals to the site leave immediately whereas exit rate provides information on which pages of a site are contributing to customer departure. Many factors determine bounce and exit decisions, but bounces are usually generated by customers with no interest in the content of the site and are caused by factors external to the website. Exit rate is instead generated by the interaction between internal factors of the website and the user. Exits are inevitable but the page location of the exit and whether this location is suitable for an exit becomes a question of optimization.

Detailed examination of the relationships between online user behavior and commercial activity has been explored by several researchers. Evidence for a positive correlation between site visitation frequency and purchase propensity was found in an online book retailer [17]. Consumers who revisited a site over a longer period were found to be more likely to react to product exposures with greater click proneness than customers that had visited over a shorter interval of time [18]. Difference in and previously unknown and hidden user behavioral types on online systems can be detected with unsupervised clustering techniques with identified behaviors being amenable to visualization tools [19]. A customer's online behavior is influenced by several endogenous and exogenous factors. Access to product information is one such endogenous factor. It was found that visitors who consulted product recommendations exhibited more complex online behavior including showing a greater number of page views per session [20].

A classic model of the purchasing behavior posits that the purchase process begins with an intention to buy [21]. An extension to this allows for the intent to purchase to become activated such as when a customer contacts informative content [22] [23] [24] [25]. This intent to purchase is time dependent in that variation in purchase intention can be expected to occur between days or months.

Like an experience in a physical store, insufficient choice or information are known to be significant predictors of shopping cart abandonment [26] [27]. Not only insufficient information, but any force that raises customer irritation with the online environment predisposes abandonment and departure from the site. Negative correlations between navigational and informational aspects of website design with irritation were found with consumer irritation being particularly sensitive to navigational design [28]. In addition to irritation, any form of mismatch between customer expectation and experience will engender disappointment and an increase probability of abandonment of the site [29]. In the setting of E-commerce, the compatibility between the shopping platform and the customer's prior experiences, expectations, or lifestyle becomes a relevant factor.

This research was undertaken to add discipline to the many conjectured WA relationships. Motivation was initially elicited by whether empirical observations of online user behavior on an E-commerce platform can be regarded as realizations of some learnable data-generating mechanism. If clickstream data is the result of a data-generating process (DGP), then it should be amenable to the tools of statistical modeling. Capabilities for better understanding and predicting variation in online behavior on E-commerce platforms are both informative and necessary, particularly with respect to user preference and decision factors. A related motivation was whether inter-individual heterogeneity in online behavior on E-commerce platforms was the result of more than one DGP and at what level of granularity are these potentially separate DGPs acting. Addressing the second question entails determining if distinguishable groups of online visitors are identifiable in the data.

3. Clickstream Conceptual Model and Hypotheses

Given that the goal of an E-commerce platform is to maximize induction of a purchase or commitment to a product, determinants of these behaviors are important to capture. A graphical representation of the relationships between clickstream data supported by the literature is displayed in **Figure 1** where explanatory variables are in the boxes and outcome variables are in the diamonds. This conceptual model served as the predicate for a set of testable hypotheses.

Demographic and site usage variables include region of user and whether that user is a returning visitor. These are considered important as different geographical regions may select for different activity on E-commerce platforms and past usage of a site should modify a user's interaction with it. Returning users have possession of external information about products offered thereby showing a smaller tendency to be convinced to make an impulse purchase on the site. Site content and quality encompasses the type of pages visited and the duration spent on those pages. Page type is decomposed into administrative, informational, and product related with further discussion of this in Section 4. User flow through a website as measured by page views is reasoned to impact decisions to complete a transaction. Duration can capture several factors such as information processing

mode used by the visitor, level of confusion by the visitor, or prior experience with the page. Technological variables include the browser, operating system, and traffic type of the visitor. Browser and operating systems primarily influence bounce, exit, and purchasing behavior through website compatibility with the browser, errors created due to incompatibility, and difference in search query results returned by different browser search engines. Traffic type incorporates how the user arrived at the site. This is considered relevant because it is postulated that the method by which a visitor arrives at a site might subsume latent motivational or attitudinal features of that visitor. For instance, a user whose arrival at the website occurred by way of a direct typing in of the URL might be expected to have a strong underlying motivation to visit the site whereas a user who accessed the website through social media may not have as compelling a motivation Lastly, temporal factors include the influence that the month, if the day was a weekend or not, and the closeness to a special day may have on decisions on the E-commerce platform.

3.1. Clickstream Conceptual Model

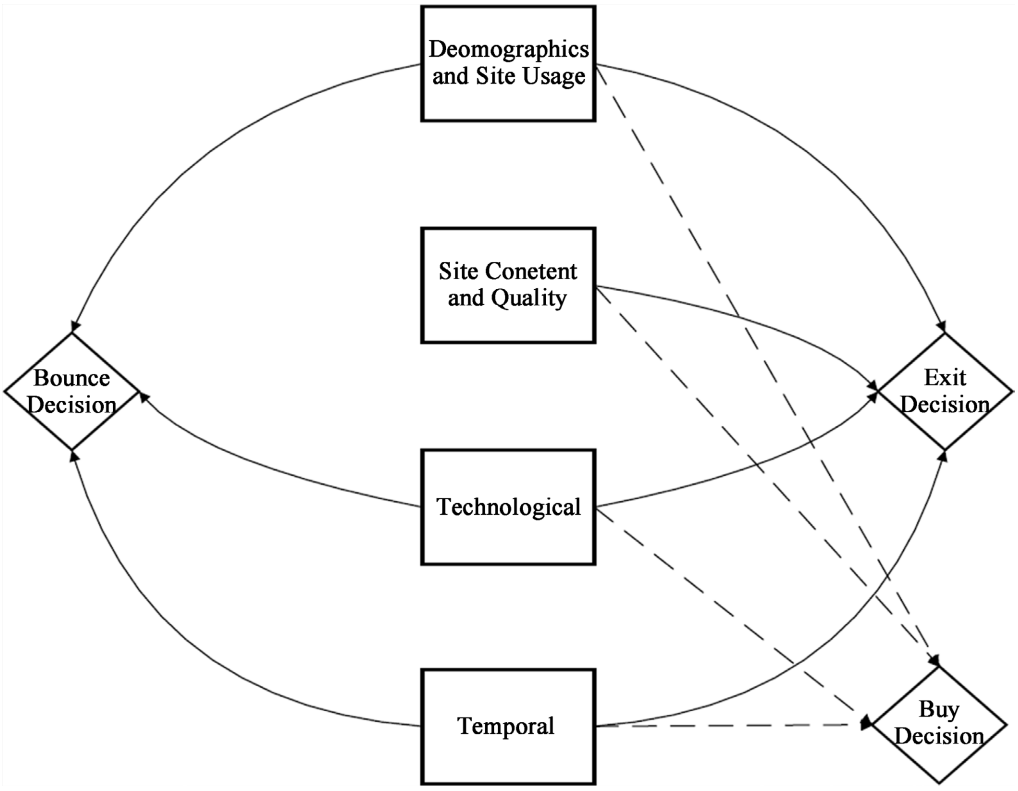


Figure 1. Conceptual model of the interactions hypothesized between clickstream variables. Directed arrows indicate the proposition that the variable which is the origin of the arrow is an independent variable whereas variables that are terminal are considered dependent, response variables.

3.2. Hypotheses

The hypotheses regarding the existence, direction, and magnitude of the rela-

tionships posited in Figure 1:

Hypothesis 1a:

Previous site usage is negatively associated with a decision to purchase.

Hypothesis 1b:

Previous site usage is negatively associated with bounce rate.

Hypothesis 1c:

Previous site usage is positively associated with exit rate.

Hypothesis 2a:

Site content and quality are positively associated with a decision to purchase.

Hypothesis 2b:

Site content and quality are negatively associated with average exit rate.

Hypothesis 3:

Association between site content and exit rate is heterogeneous at the previous site usage group level

Hypothesis 4a:

Previous site usage is negatively associated with variability in exit rate.

Hypothesis 4b:

Previous site usage is negatively associated with variability in bounce rate.

Hypothesis 5a:

Technological factors are associated with variability in exit rate.

Hypothesis 5b:

Technological factors are associated with variability in exit rate.

Hypothesis 6:

The amount of viewed product content is negatively associated with variability in exit rate.

4. Data and Empirical Setting

The dataset is comprised of clickstream data gathered from 12,172 sessions on an online retailer's web platform [30]. Formation of the dataset ensured that each session corresponded to a different user over a one-year time. A primary reason for this sessionization was to avoid any tendency towards a specific advertisement campaign, special day, user profile, or time [30]. Table 2 describes the variables. Previous site usage was indicated by the value of the VisitorType variable. The *Administrative*, *Administrative_Duration*, *Informational*, *Informational_Duration*, *ProductRelated*, and *ProductRelated_Duration* variables represented site content and quality. Browser was chosen as the technological factor to examine. The amount of viewed product was indicated by the value of the *ProductRelated* variable.

5. Analytical Methodology and Framework

5.1. Response Variable Distributional Analysis

Proper analytical treatment of the outcome variables begins with an appropriate selection of the model structure. Exit rate, bounce rate, and revenue variables are

Table 2. Description of the predictors and response variables used in the models. Exit rates, bounce rates, and revenue acted as response variables. The other variables were candidates as predictors.

Variable	Description
Administrative	Number of account management pages visited
Informational	Number of pages visited about website
Product related	Number of pages visited about products
Administrative duration	Total seconds on administrative pages
Informational duration	Total seconds on informational pages
ProductRelated duration	Total seconds on product related pages
Bounce rates	Average bounce rate of pages visited by user
Exit rates	Average exit rate of pages visited by user
Page values	Average page values of pages visited by user
SpecialDay	Closeness of the session to a holiday
Month	Calendar month
Operating systems	Operating systems of user
Browser	Web browser of the user
Region	Geographic region
Traffic type	Traffic source for arrival
Weekend	Site visited on a weekend or not
Visitor type	User is a new or returning visitor
Revenue	Session finalized with a transaction or not

all bounded. To avoid misspecification of the model for bounce and exit rates, empirical distributions were examined to detect which model for the conditional distribution would be most appropriate. Values for exit and bounce rates were scaled according to a min-max scaling so that observed values were bounded in [0, 1]. Densities for both bounce and exit rates across subpopulations of visitors are displayed in Figure 2. The data is bimodal with peaks near zero and a smaller peak near one. This bimodality suggests the possibility of two distinct data generating processes that govern exit and bounce behavior.

To improve the selection of a distributional model, scaled third and fourth moments (skewness and kurtosis) were plotted for a set of different theoretical distributions with maximum-likelihood estimation used to compute these moments. Bounce and exit rate exhibited shapes consistent with a beta distribution. The third and fourth moment evidence is shown for exit rate in Figure 3 but a highly similar result was found for bounce rate. Further maximum-likelihood estimation was used to fit the bounce and exit data to beta distributions. Removal of observations making up the right peak in both the bounce and exit data

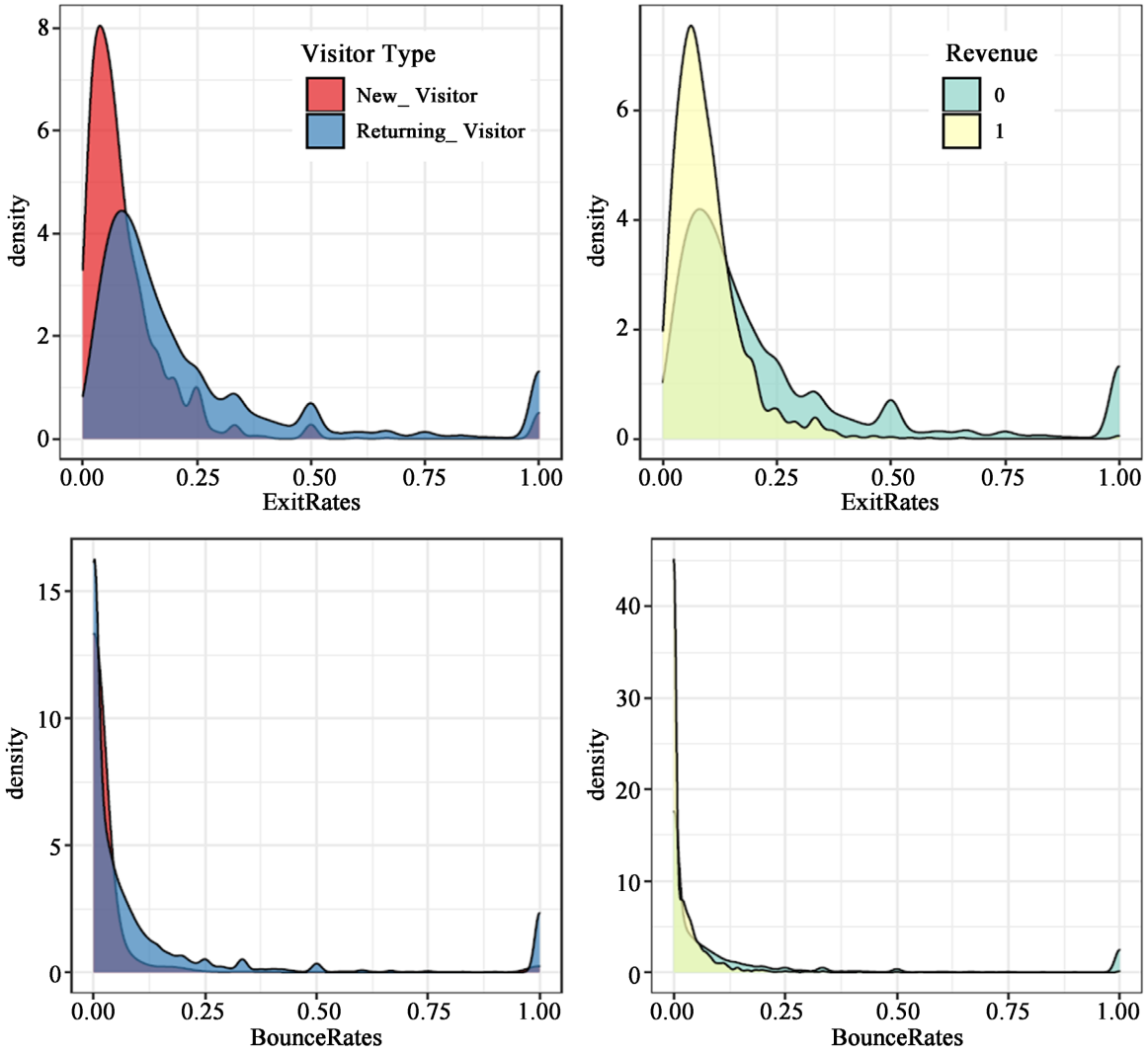


Figure 2. Empirical distributions of the exit rate and bounce rate. The left column shows variation in distributions under different visitor types and the right column shows variation in distributions under different decisions to make a purchase or not on the website.

resulted in beta fits with greater suitability as displayed in **Figure 4** and **Figure 5**. Considering these observations, it was decided to remove observations in the 99th percentile for both the exit and bounce rate data. In both, this amounted to about 5.5% of observations being taken out. This process is equivalent to winsorization. Doing this does ignore a source of information but enables more accurate modeling of the other 94.5% of observations.

5.2. Background on Analytical Model Framework

5.2.1. Logit and Discrete Choice Utility Model

Discrete choice utility models are used to model the choice a decision maker has among a set of alternatives [31]. These models treat the utility or value returned by a decision as a random process and have been show to be applicable to many domains including retail and commerce [32]. When the choice is binary, between

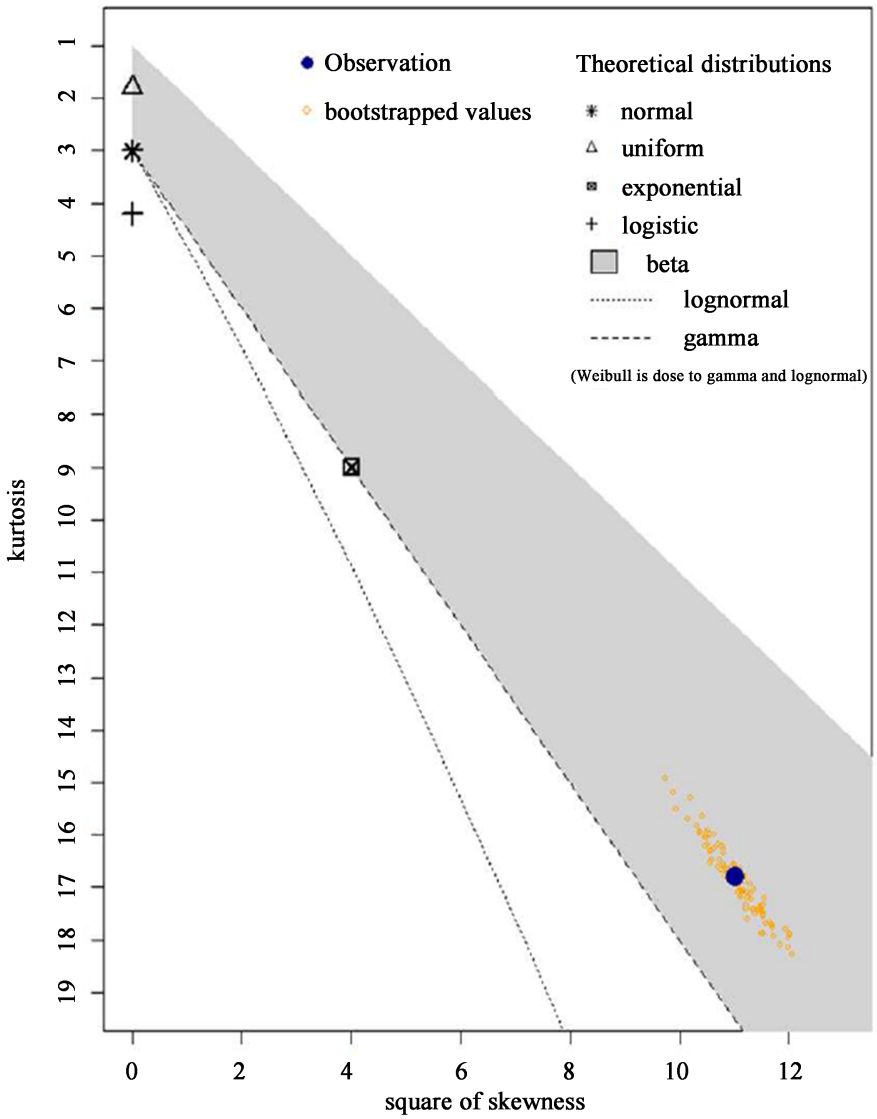


Figure 3. Plot of sample kurtosis and square of sample skewness calculated via maximum-likelihood estimation (MLE) for exit rate (blue dot) and their relation to value of these moments from several theoretical distributions. The left displays results for the data without 99th percentile removed with estimated skewness = 2.169 and estimated kurtosis = 7.13. The right displays results after these observations were removed with estimated skewness = 1.931 and estimated kurtosis = 7.28. MLE of the moments were repeated for 100 bootstrap samples from the data with resulting estimates shown with the yellow points. Observed data has tailedness and skewness consistent with a beta distribution.

"yes" or "no", then the logistic regression model can be used to estimate a discrete choice utility model [33].

Consider n observations of a random variable Y that represents a visitor's decision to purchase or not on the E-commerce platform where,

$$y_i = \begin{cases} 1, & \text{purchase} = \text{yes} \\ 0, & \text{purchase} = \text{no} \end{cases} \quad (1)$$

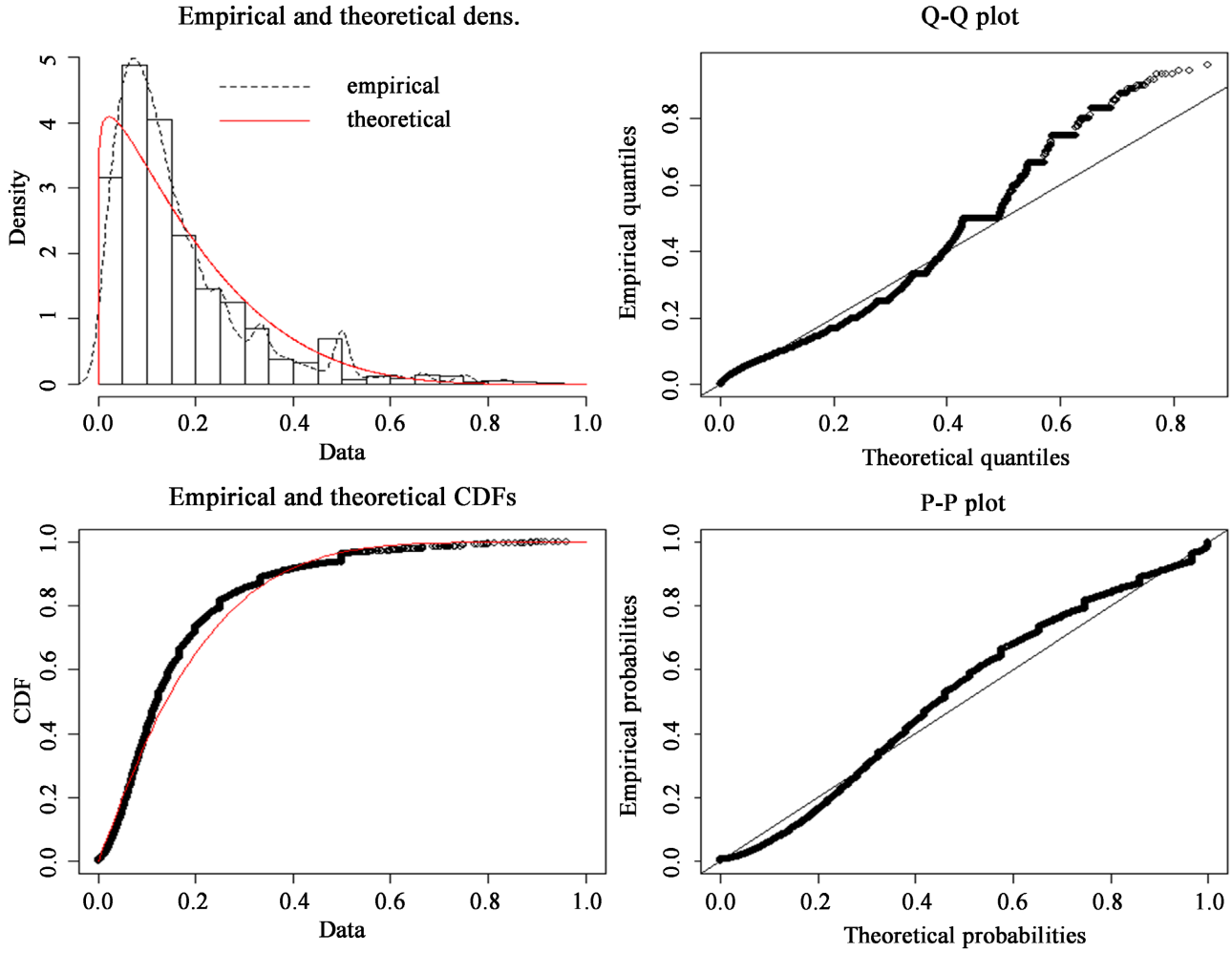


Figure 4. Plots comparing empirical bounce rate data to a maximum likelihood estimation based best fit beta distribution with 99th percentile of observations removed. MLE based estimates for the first shape parameter of the best beta distribution was 1.084 and the estimate for the second shape parameter was 5.166. Inspection of the empirical versus theoretical CDFs (lower left) and empirical versus theoretical quantiles (upper right) clearly show an improved fit of the data to a theoretical beta distribution with the shape parameters. The removal of the extreme observations at the right tail resulted in a better fit.

are the possible values for observation *i*. We can approximate this variable with a linear predictor with *p* explanatory variables,

$$Y = X\beta + \varepsilon \tag{2}$$

where β is a $p \times 1$ matrix of coefficients, X is a $n \times p$ data matrix, and ε is a $n \times 1$ matrix of errors. The expectation of Y for observation i is,

$$E\left[y_i\right] = \hat{p}_i = X_i\beta \tag{3}$$

which is the estimated probability of making a purchase where X_i is a row vector of X that corresponds to the vector of predictor values for visitor *i*. By probability axioms, $\hat{p}_i$ must be bounded in [0, 1] and $\hat{p}_i + \left(1 - \hat{p}_i\right) = 1$. On the scale of the linear predictor, $X\beta$, there is no guarantee estimated probabilities will satisfy these axioms. Furthermore, using definition of variance for Bernoulli random variable, we have,

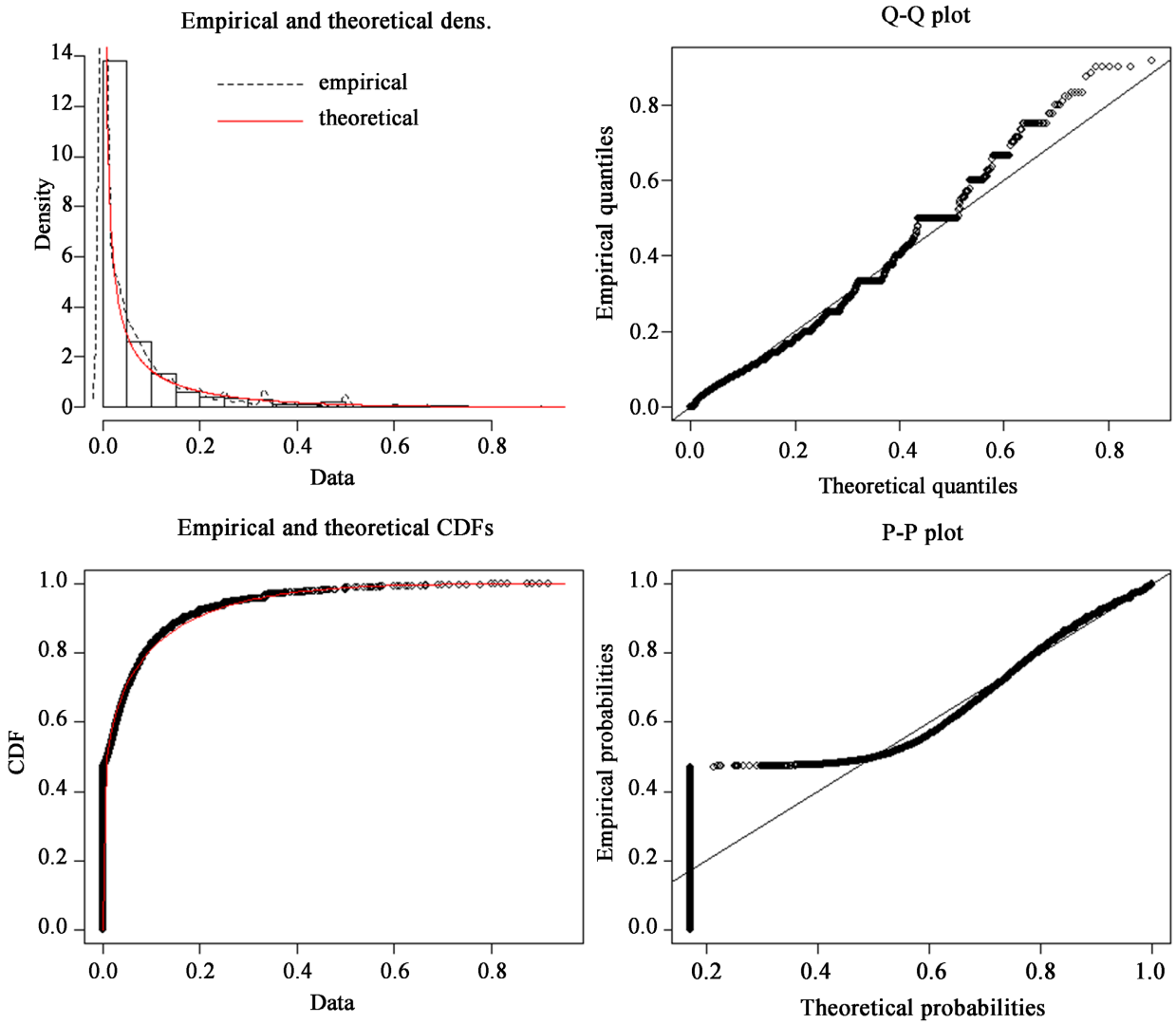


Figure 5. Plots comparing empirical bounce rate data to a maximum likelihood estimation based best fit beta distribution with 99th percentile of observations removed. MLE based estimates for the first shape parameter of the best beta distribution was 0.206 and the estimate for the second shape parameter was 3.336. Inspection of the empirical versus theoretical CDFs (lower left) and empirical versus theoretical quantiles (upper right) clearly show an improved fit of the data to a theoretical beta distribution with the shape parameters. The removal of the extreme observations at the right tail resulted in a better fit.

$$E\left[\varepsilon_i^2\right] = \hat{p}_i\left(1-\hat{p}_i\right) = X_i\beta\left(1-X_i\beta\right) \tag{4}$$

Which indicates that variance in error is heteroskedastic and so standard OLS would yield logically inconsistent estimates that ignore unequal variances.

A visitor on the site is assumed to face a dichotomous choice with a non-zero utility attached to the purchase option and zero utility to the non-purchase option. Three assumptions are made, 1) the binary choice set is mutually exclusive in that no combinations of choices are possible, 2) the choices are exhaustive in that no relevant alternative is available, and 3) the visitor to the website selects the choice alternative that provides the highest utility. With these assumptions, let U_i denote the utility or benefit visitor i achieves when taking the action to

purchase so that,

$$U_i = V_i\left(X^*\right) + E_i \tag{5}$$

where $V_i\left(X^*\right)$ is a deterministic part of the utility and depends on a set of predictors that are not necessarily observed and E_i is the random part of utility. Here, E_i is assumed to take a logistic distribution which allows for larger tails in the error distribution. Utility U_i, is continuous but interest lies in whether a visitor creates revenue or not. Hence, we redefine Y such that,

$$y_i = \begin{cases} 1, & U_{visit} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

A generalized linear model can be used to model the observed decisions to purchase or not,

$$\log_e \frac{\hat{p}_i}{1-\hat{p}_i} = X_i\beta\,. \tag{7}$$

5.2.2. Beta Regression Model

When the outcome variable is constrained to (0, 1), several complications arise. A few of these complications can be ameliorated with a logit transformation of expected mean [34]. The beta regression model was proposed to address these issues more fully [35] [36] [37] [38] [39]. Main characteristics of the beta regression model will be presented. Consider an outcome variable $Y \in (0,1)$ assumed to follow a beta distribution with density,

$$f\left(y;\mu,\varphi\right) = \frac{\Gamma\left(\varphi\right)}{\Gamma\left(\varphi\mu\right)\Gamma\left(\varphi\left(1-\mu\right)\right)} y^{\varphi\mu-1}\left(1-y\right)^{(1-\mu)\varphi-1} \tag{8}$$

where μ is the mean of Y, φ is a precision parameter, and $\Gamma(.)$ is the gamma function. If observed values for Y are not strictly in (0, 1) then the transformation $\left(y*\left(n-1\right)+0.5\right)/n$ can be used where n is the sample size [35] [36]. The conditional mean of Y, $\mu_Y = E\left[Y \mid X\right]$, is related to the predictor variables with an invertible link function g,

$$g_1\left(\mu_Y\right) = X\beta\,. \tag{9}$$

Different links can be used but the logit link retains interpretability in terms of the odds ratio. Beta regression allows a flexible variance,

$$Var\left(Y\right) = \frac{\mu_Y\left(1-\mu_Y\right)}{1+\varphi}\,. \tag{10}$$

This more flexible variance becomes indispensable when modeling proportional data from heterogeneous populations that exhibit overdispersion or excessive variance. If overdispersion is thought to depend on values of predictor variables, then the precision parameter, φ, can regressed to the variance boosting factors,

$$g_2\left(\varphi\left(Z\right)\right) = Z\gamma \tag{11}$$

where g_2 is an invertible link function, Z is the matrix of predictors, and γ is a parameter vector. Choice of the link function for φ is context dependent but a

log link is common.

5.2.3. More Advanced Beta Regression Techniques

Extensions to the beta regression model introduced in the previous section are possible. In real life scenarios exact 0's or 1's may be observed in the data. These boundary values can be handled with data transformations that shift the 0- and 1-point mases to the center of the data. The transformation increases the value of the zero and decreases the value of the one observations marginally but allow the beta model to be used. However, the 0's and 1's may carry meaningful information that would be lost if transformed. The need to directly model the 0's and 1's is addressed with the zero-one inflated beta regression model. Readers interested in greater details are referred to the literature [40] [41] [42].

Data that has been collected across a wide range of individuals might be comprised of unobserved clusters of individuals and ignoring this by imposing a homogenous model will underfit the data. Explicit attention to variability in effects is paramount in trying to understand the complex ways variables interrelate. Heterogeneity between groups of individuals can be tackled with the strategy of model-based recursive partitioning with beta regression tree models [43]. This partitioning is like that of the classification and regression tree (CART) method but differs in its broader aim of capturing differences in the parameters that describe the distribution of the response variable [44] [45].

Finite mixtures of beta regression belong to the larger class of finite mixtures of regressions which attempt to capture systematic differences in the association between variables. Mixtures of regressions introduce latent classes for which the effect of a predictor may differ in magnitude and/or direction from the other latent classes. A mixture model is a probabilistic model for the presence of subgroups within an overall population without the requirements that those subgroups be explicitly labeled [46]. A full exploration of mixture of regressions models and its applications are outside the scope of this paper but an extensive body of literature can be found on it [47] [48] [49] [50].

5.3. Revenue Model

With the assumption that the decision to purchase or not follows the random utility model from Section 5.2.1, the following logistic regression model was used to determine if the available clickstream variables captured the process generating the revenue response data.

$$\begin{aligned}
&\log\frac{Revenue_i}{1-Revenue_i}\\
&= Administrative_i + Administrative_Duration_i + Informational_i\\
&\quad + Informational_Duration_i + ProductRelated_i\\
&\quad + ProductRelated_Duration_i + ExitRate_i + BounceRate_i\\
&\quad + VisitorType_i + Browser_i + TrafficType_i + OpeartingSystem_i\\
&\quad + Region_i + Month_i + Weekend_i + SpecialDay_i
\end{aligned} \tag{12}$$

The first six variables are theorized to convey site content and quality information. Visitors to the website are assumed to be goal-directed utility maximizing agents and so time spent on certain pages should assist the visitor in deciding on what course of action is benefit-producing. Variables *Browser*, *TrafficType*, and *OperatingSystem* are included as they are surmised to encode technological preferences, technical knowledge, and technology-website incompatibility. Variables *Region*, *Month*, *Weekend*, and *SpecialDay* are control variables that should adjust for variability in purchase intention seen over various times and geographical contexts. *ExitRate* and *BounceRate* are included as it is surmised that both contain information independent from the other variables

5.4. Bounce Rate Model

Bounce rate is the probability a visitor traffics to the website but departs immediately. This response varies between pages on the website and its value is derived from previous arrival and departure events of other visitors which are divergent from visitor *i*. A flexible probability model of these probabilities which can incorporate non-symmetries, complex variance structures, and skewness is warranted. The inherent flexibility of the beta distribution with respect to symmetry, skewness, and variance structure motivated the choice of a beta regression model for bounce rate. As bounce rate is theorized to be a process external to the website content and quality, independent variables that correspond to internal aspects of the website were excluded.

$$BounceRate_i = \frac{\exp(\eta)}{1+\exp(\eta)} \tag{13}$$

$$\begin{aligned}\eta = {} & VisitorType_i + Browser_i + TrafficType_i + OpeartingSystem_i \\ & + Region_i + Month_i + Weekend_i + SpecialDay_i\end{aligned} \tag{14}$$

$$\Phi_i = \exp\left(VisitorType_i + Browser_i + TrafficType_i\right) \tag{15}$$

where η is the linear predictor for the mean bounce rate and Φ_i is the precision parameter for visitor *i*

5.5. Exit Rate Model

A beta regression model was chosen to represent dependencies between the exit rate and the explanatory variables. The argument to do so closely matches that made for the bounce rate. Unlike the bounce rate, a visitor's observed exit probability is posited to be more dependent on the user interface and user experience of the website. It was decided to include *ProductRelated* in the precision model as this is seen as a proxy for commitment to a goal on the website and it is reasoned a visitor should exhibit less variable exit rates when more committed.

$$ExitRate_i = \frac{\exp(\eta)}{1+\exp(\eta)} \tag{16}$$

$$\begin{aligned}\eta = & Administrative_i + Administrative\_Duration_i + Informational_i \\ & + Informational\_Duration_i + ProductRelated_i \\ & + ProductRelated\_Duration_i + VisitorType_i + Browser_i \\ & + TrafficType_i + Region_i + Month_i + Weekend_i + SpecialDay_i\end{aligned} \tag{17}$$

$$\Phi_i = \exp\left(VisitorType_i + ProductRealated_i\right) \tag{18}$$

where η is the linear predictor for the mean exit rate and Φ_i is the precision parameter for visitor i

6. Results and Discussion

6.1. Revenue Model

Any model of user behavior should possess the qualities of being interpretable, valid, and useful. Interpretability and practical usefulness are coupled as any loss of interpretability will erode how useable the model is. Validity here refers primarily to the predictive accuracy and support of the model on the available data. It is for these reasons that variable selection was included in the analysis so that the resulting models were as simple as possible yet capable of explaining the appearance of the observed data.

To improve interpretability of the revenue model, the least absolute shrinkage and selection operator (LASSO) method was employed to filter out independent variables (IV's) that do not contribute to model accuracy. Regularization with LASSO was used as a first step for two primary reasons. First, fitting a model against the available IV's would yield non-unique solutions. Second, the model inferred should not include parameters that are not justified by the data. Details of how LASSO performs variable selection can be found in several sources [51] [52]. Cross validation was used to select a value for the lambda parameter for which the revenue logistic model achieved the smallest mean cross validation error. Mean cross validation error was defined in two ways as either binomial deviance or area under the receiver operating characteristic curve. Results of this process are shown in **Figure 6**. Deviance is minimized, and AUC is maximized at the models with optimal complexity fitted to an 80% subset of the data. Model 1 was a full complexity model with all available IV's included whereas Model 2 included the variables identified in the variable selection phase.

To assess the relative quality of the two models, a series of goodness-of-fit (GOF) tests were performed. Examination of GOF must involve determining whether the fitted model's residual variation is minor, shows no systematic patterns, and follows variability assumed by the fitted model. Hosmer and Lemeshow's (HL) C and the le Cessie-van Hourwlignen-Copas-Hosmer (CHCH) unweighted sum of square test were used to evaluate global GOF. Both tests specify a null hypothesis that the tested model does not need to be more complex. Details will not be explored but can be found in the literature [53] [54]. Judgements on the power of each model to explain the phenomena were based on McFadden's adjusted pseudo-R^2 which is based on the log likelihood ratio of the fitted

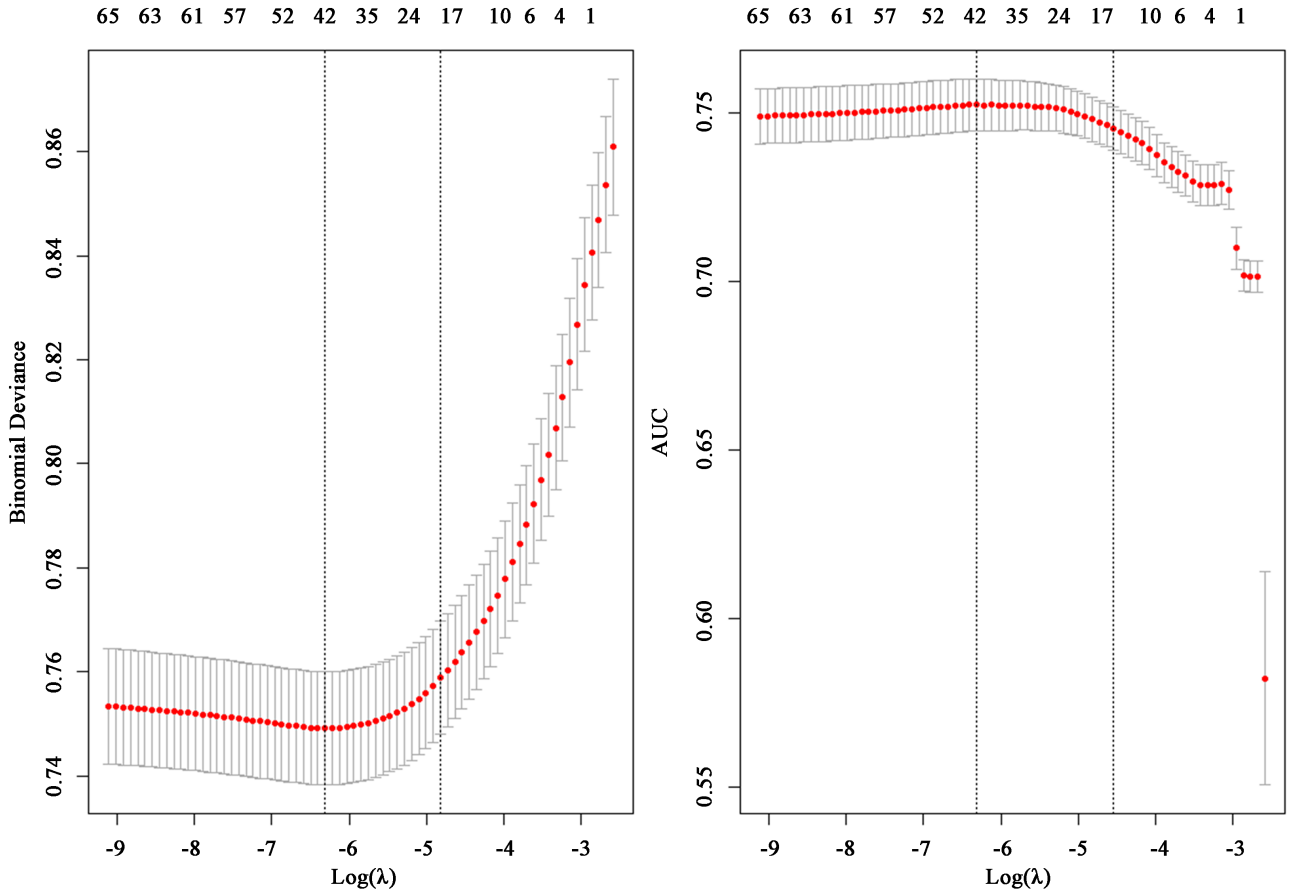
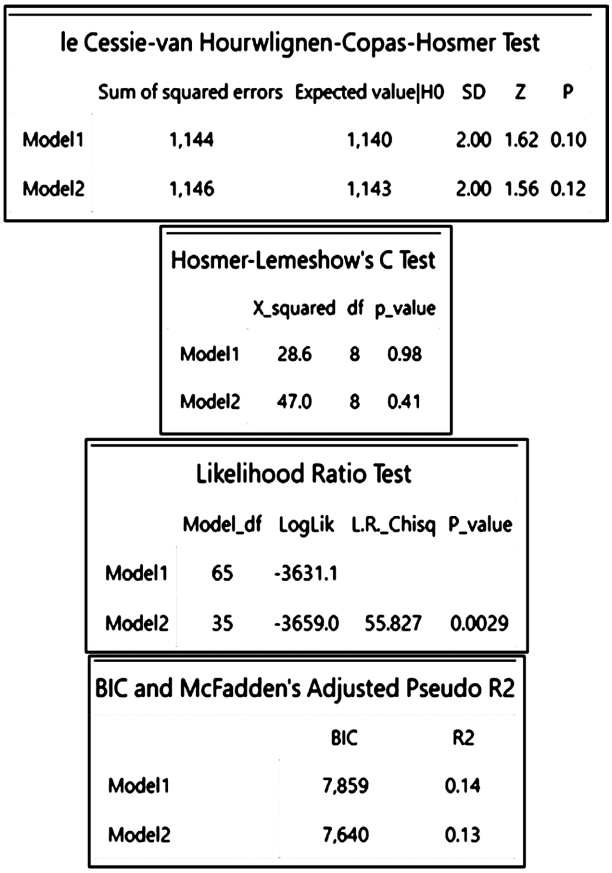


Figure 6. Cross validated LASSO on revenue logistic model. Plots show the number of non-zero predictors (top of plots) in a model for two different values of estimated out-of-sample error. The left column shows binomial deviance with a minimum at the left vertical line and a value at one standard error at the right vertical line. The right column shows AUC with a maximum at the left vertical line and a value at one standard error at the right vertical line.

model to the null, intercept only model. It is a relative merit of how close each model can approximate the DGP [55]. Communication of each model's model fitting and complexity tradeoff is given by reporting the Bayesian information criterion (BIC) and performance of a log likelihood ratio test on the two models. Results for each are in **Figure 7**.

At the 5% significance level, the HL and CHCH tests indicate that there is insufficient evidence to rule out the null hypothesis, thus concluding that the addition of interaction terms or nonlinearity to the models is not supported. The data did support the conclusion that the more complicated model results in a better fit to the data when measured by likelihood. However, the difference in McFadden's R^2 between the two models is not considerable with the BIC for model 2 being 2.8% lower than Model 1. Therefore, it was concluded that the addition of extra variables is not supported by the data and Model 2 was pursued with estimated effects given in **Table 3**. Estimates greater than one on site content variables provides support for Hypothesis 1a whereas the estimate for visitor type less than one provides support for Hypothesis 2a.



le Cessie-van Hourwlignen-Copas-Hosmer Test

	Sum of squared errors	Expected value|H0	SD	Z	P
Model1	1,144	1,140	2.00	1.62	0.10
Model2	1,146	1,143	2.00	1.56	0.12

Hosmer-Lemeshow's C Test

	X_squared	df	p_value
Model1	28.6	8	0.98
Model2	47.0	8	0.41

Likelihood Ratio Test

	Model_df	LogLik	L.R._Chisq	P_value
Model1	65	-3631.1		
Model2	35	-3659.0	55.827	0.0029

BIC and McFadden's Adjusted Pseudo R2

	BIC	R2
Model1	7,859	0.14
Model2	7,640	0.13

Figure 7. Goodness-of-fit and model validation tests on the saturated model (Model1) and the conceptual model (Model2).

Table 3. Logistic discrete choice model results for modeling of revenue. Odds Ratios are reported with standard errors in parentheses. Odds ratios with P-Values less than 0.05 are marked by *, less than 0.01 by **, and less than 0.001 by ***.

Explanatory Variable	Odds Ratio	CI (95%)	P-Value
Administrative	1.0183* (0.0092)	(1.0014, 1.0364)	0.044
Informational	1.0389 (0.0230)	(0.9943, 1.0846)	0.085
ProductRelated	0.9995 (0.0012)	(0.9971, 1.0018)	0.647
ProductRelated_Duration	1.0001** (0.0000)	(1.0000, 1.0001)	0.003
ExitRate	0.0034*** (0.0014)	(0.0015, 0.0073)	<0.001
SpecialDay	0.4430*** (0.1091)	(0.2686, 0.7071)	0.001
Month [Feb]	0.6144* (0.1083)	(0.4368, 0.8724)	0.006
Month [Mar]	0.5544*** (0.0985)	(0.0250, 0.5197)	0.011
Month [May]	0.8753 (0.1446)	(0.6759, 1.5860)	0.838
Month [June]	0.7808 (0.2072)	(0.4582, 1.3014)	0.351
Month [July]	1.0368 (0.2252)	(0.3929,.7894)	0.001
Month [Sep]	0.8656 (0.1787)	(0.6367, 1.2177)	0.420

Continued

Month [Oct]	1.1311 (0.2174)	(1.0977, 2.0633)	0.012
Month [Nov]	1.4951* (0.2404)	(0.7779, 1.6538)	0.521
Month [Dec]	0.6144** (0.1083)	(0.5773, 1.2985)	0.485
TrafficType [2]	1.3020** (0.1213)	(1.0862, 1.5650)	0.005
TrafficType [3]	0.8016 (0.0963)	(0.6326, 1.0133)	0.066
TrafficType [4]	1.2231 (0.1603)	(0.9450, 1.5801)	0.124
TrafficType [5]	1.2798 (0.2617)	(0.8504, 1.8982)	0.228
TrafficType [6]	0.9842 (0.1847)	(0.6745, 1.4097)	0.932
TrafficType [7]	2.5208* (0.9795)	(1.1368, 5.2893)	0.017
TrafficType [8]	1.5198* (0.2723)	(1.0647, 2.1510)	0.020
TrafficType [9]	0.6824 (0.5113)	(0.1079, 2.3808)	0.610
TrafficType [10]	1.5240** (0.2458)	(1.1064, 2.0834)	0.009
TrafficType [11]	1.4857 (0.3060)	(0.9826, 2.2066)	0.055
TrafficType [13]	0.5272** (0.1074)	(0.3487, 0.7753)	0.002
TrafficType [14]	0.9024 (0.9306)	(0.0754, 5.0996)	0.921
TrafficType [16]	6.1621 (9.1434)	(0.2233, 170.9)	0.220
TrafficType [20]	1.7307* (0.4336)	(1.0443, 2.7982)	0.029
VisitorType [Returning_Visitor]	0.7616*** (0.0627)	(0.6484, 0.8957)	0.001

Examination of the odds ratio column in Table 3 provides access to an interpretation of the estimates. For the *Administrative*, *Informational*, and *ProductRelated* variables, an increase in one more page view would correspond to an increase in odds of purchasing by 1.0183, 1.0369, 0.9995. In other words, a one-page increase in *Administrative* would result in about a 1.8% increase in odds to purchase. The number of pages regarding account management (*Administrative*) and web site, communication, and address information (*Informational*) would be important for new visitors to the site. The estimated effect of the duration spent on product related pages appears small, but when considering that the maximum observed value for this variable was 63,973 seconds (about 18 hours) and the standard deviation was 1912 seconds (about 32 minutes), a small effect could easily become magnified. Exit rate exhibited the most influence on revenue probability where a unit increase in it yields a 99% reduction in odds of purchasing. As visitors who visit more pages with higher exit rates can be anticipated to not select the action to purchase. It is important to keep in mind that the exit rate is itself influenced by the other predictors and any statement of causal effect would require mediation analysis. *Month* and *TrafficType* were included as control variables as these are less in control space of a web designer. Surprisingly, the nearness to a special day had a negative impact on the odds ra-

tio of making a purchase as the odds of making a purchase decrease by 56% as the closeness to a special day increases by one day. This may be due to collinearity with the *Month* or the fact that any companies may offer incentives during holidays that elevate competition and lower propensity for any visitor to convert on the website. *VisitorType* had an estimated negative effect on the odds ratio of making a purchase with a 24% reduction in odds of making a purchase for a returning visitor. This is consistent with the supposition that returning visitors are more informed consumers and are not susceptible to the conversion mechanisms that operate on websites to persuade visitors to buy.

Given that the predictors are correlated, a quick check of multicollinearity using variance inflation factors (VIF) was conducted. A large VIF value (>10) for a predictor indicates that much of its variation is predictable from the other predictors. In the presence of high multicollinearity, the precision in estimates is low. Table 4 contains all VIF greater than two and shows that issues of major multicollinearity are absent.

The decision to purchase on the website was framed as discrete choice, utility maximization problem. Under the assumptions that the decision to purchase had a binary choice set, the choice set was mutually exhaustive, and irrelevant alternatives did not impact utility calculus, then the estimates for the logit model can be formulated as contributions to utility those variables confer, conditional on the correctness of the assumptions and the current level of other variables and the visitor's utility. The following were estimated to contribute positively to utility on the website—site content and quality (*Administrative*, *Informational*, *ProductRelated_Duration*)—whereas factors that are estimated to contribute to lower utility—exit decisions, temporal factors (*SpecialDay*, *Month*), previous site usage (*ReturningVisitor*).

6.2. Exit Rate Model

Exit rate for a visitor is hypothesized to be an intermixture of several processes

Table 4. Variance inflation factors (VIF) for explanatory variables used in discrete choice logistic regression model.

Variable	VIF
MonthNov	6.73
MonthMay	5.35
ProductRelated	4.84
ProductRelated_Duration	4.63
MonthDec	3.64
MonthMar	3.53
MonthOct	2.42
TrafficType2	2.42

including ones involved in the prediction of the intent to purchase. Departure behavior is also hypothesized to be a result of forces internal to the website and user experience. It is assumed that visitors that traffic to the website and do not immediately bounce have some goal that could be satisfied from usage of the web site. The web site is tasked with designing and implementing the web system such that any interest held by the visitor is channeled into conversion. The beta regression model will include only those predictors that correspond to site content and quality or are involved in the site-user interaction. The logit transformed expected value for the exit rate and the log transformed precision parameter are analyzed as linear functions of certain predictors. Using $\left(ExitRate*\left(n-1\right)+0.5\right)$ and $\left(ExitRate-\min\left\{ExitRate\right\}\right)/\left(\max\left\{ExitRate\right\}-\min\left\{ExitRate\right\}\right)$ transformation made it so that exit rate was bounded in (0, 1). Testing of hypotheses 4, 5, 6 required *VisitorType*, *Browser*, and *ProductRelated* to be included in the linear predictor for the precision parameter. To ease interpretation, three separate models were examined whereby only one of the three precision parameter predictors was used (Model A: VisitorType, Model B: Browser, Model C: ProductRelated). Results for these three separate fits are provided in **Table 5**.

It is instructive to emphasize that the proposed DGP for exit rate is believed to be primarily a function of factors that are internal to the platform. The E-commerce platform serves as an interface between the visitor and personalized goals. Although there are numerous design features that are considered to contribute positively to human-computer interaction, a central premise of this work is that the usability and perceived usefulness of the web system is of utmost importance to user experience [56] [57]. This focus on a visitor's viewpoint of the quality of a website is justified from three user experience principles: principle of predictive aiding, minimization of interaction cost, principle of multiple resources. Predictive aiding states that digital systems should forecast user needs and provide tools that ease cognitive burden before it occurs. Minimization of interaction cost asserts that a user should be able to accomplish his goal with the smallest investment of effort possible [58] [59]. The principle of multiple resources claims visitors can process multiple streams of information and this can be exploited to design an informative system [60]. Based on this theory, only site content variables, technology variables, and the visitor type variable were included in the model.

Viewing **Table 5**, it can seen that site content and quality variables – *Administrative*, *Informational*, *ProductRelated* - and the duration spent on each had a considerable influence on exit rate. With attention on Model A, an increase in a single page view for the three types of pages results in a reduction in the odds of visiting pages with high exit rate. For instance, if a visitor accesses an administrative page, the reduction in odds of exit rate is about 5.5%. Furthermore, the duration of time spent on administrative and product related pages also yields a reduction in odds of having a higher exit rate. Estimated values for duration variables are small but are measured in seconds and with large standard deviations

Table 5. Estimated results for beta regression models with exit rate as the response. The precsion paramter of the beta model is considerd a function of visitor type, web browser, and product related for Models A, B, and C, respectively.

Explanatory Variable	Odds Ratio (95% CI)		
	Model A	Model B	Model C
Administrative	−0.057***	−0.058***	−0.055***
	(−0.066, −0.047)	(−0.068, −0.049)	(−0.065, −0.045)
Administrative_Duration	−0.0002**	−0.0002**	−0.0002*
	(−0.0004, −0.00003)	(−0.0004, −0.00001)	(−0.0004, 0.00002)
Informational	−0.12	−0.011	−0.007
	(−0.037, 0.013)	(−0.036, 0.013)	(−0.032, 0.018)
Informational_Duration	0.00005	0.0001	(0.00002
	(−0.0002, 0.0003)	(−0.0002, 0.0003)	(−0.0002. 0.0003)
ProductRelated	−0.006***	−0.006***	−0.010***
	(−0.007, −0.005)	(−0.007, −0.05)	(−0.10, −0.009)
ProductRelated_Duration	−0.00003**	−0.00003**	−0.00001
	(−0.0001, −0.000001)	(−0.0001, −0.00001)	(−0.0001, −0.000001)
Browser [2]	−0.98***	−0.172***	−0.078***
	(−0.159, −0.037)	(−0.237, −0.106)	(−0.13, −0.019)
Browser [3]	0.092	0.049	0.074
	(−0.159, 0.343)	(−0.219, 0.317)	(−0.172, 0.321)
Browser [4]	−0.188***	−0.264***	−0.168***
	(−0.296, −0.080)	(−0.380, −0.148)	(−0.274, −0.063)
Browser [5]	−0.212***	−0.353***	−0.179***
	(−0.345, −0.080)	(−0.491, −0.215)	(−0.308, −0.050)
Browser [6]	−0.212**	−0.333***	−0.187*
	(−0.413, −0.011)	(−0.540, −0.126)	(−0.381, 0.007)
Browser [7]	0.067**	0.069***	0.083**
	(−0.333, 0.467)	(−0.359, 0.496)	(−0.303, 0.470)
Browser [8]	0.036	0.038	0.065
	(−0.197, 0.269)	(−0.212, 0.288)	(−0.160, 0.290)
Browser [9]	−0.213	−0.838**	−0.143
	(−2.575, 2.150)	(−0.920, −0.756)	(−2.494, 2.208)
Browser [10]	−0.225	−0.391	−0.213
	(−0.429, −0.021)	(−0.602, −0.180)	(−0.412, −0.015)
Browser [12]	−0.358	−0.778**	−0.304
	(−1.133, 0.418)	(−1.435, −0.122)	(−1.078, 0.470)

Continued

Browser [13]	1.360***	0.936*	1.451***
	0.613, 2.107	(−0.011, 1.882)	(0.729, 2.173)
VisitorType [Returning_Visitor]	1.077***	0.730***	0.679***
	(1.001, 1.153)	(0.662, 0.798)	(0.612, 0.747)
Precision Parameter Predictor Estimate:			
VisitorType	−0.646*		
Browser [12]		1.286*	
Browser [13]		−0.624*	
ProductRelated			0.009

Note: *p < 0.1;**p < 0.05; ***p < 0.01.

in the variables, hence effects are magnified. Hypothesis 2b is supported by the data as increases in site content and quality demonstrates a negative association with average exit rate. This finding is consistent with expectations as pages that are most frequently visited are ideally the ones where the visitor should exit. If a visitor moves to a low frequency page and departs, this page will have a large exit rate. Superior site content and quality should facilitate behavior on the website that agrees with the site's objective. It can also be seen that previous site usage confers a positive influence on the average exit rate. According to Model A, if a visitor has had prior experience with the website, his odds of demonstrating an increase in expected exit rate increased by 108% compared to a user without experience. Hypothesis 1c is therefore supported by the data. Support for Hypothesis 1c and 2b appears to be contradictory but are compatible. Consider that returning visitors are assumed to have less uncertainty about their goals on the website and are therefore less likely to visit unnecessary pages. Because of this mechanism, the pages frequented by returning visitors would correspond to the ones where transactions are completed.

The precision parameter can be estimated using the technique of a generalized linear model (GLM). Examining **Table 5**, previous site usage (*VisitorType*) has a negative influence on the dispersion of exit rate, technological factors (*Browser*) have both positive and negative influences on dispersion in exit rate, and site content has a negative influence on dispersion in exit rate. It can be concluded that history of usage of the website is estimated to have a negative marginal effect, choice of technology in the form of a web browser is estimated to display both positive and negative marginal effects, and site content as represented by the number of product related pages viewed shows a positive marginal effect on exit rate. Thus, Hypotheses 4a, 5a, 6 are supported by the data.

Assessment of goodness-of-fit was performed by examining the pseudo-R^2 proposed by Ferrari and Cibari-Neto [35], which is the squared correlation between the linear predictor for the mean and the logit-transformed response. BIC

and a likelihood ratio test of the fitted models against the null intercept only model which assumes that the precision parameter is not a function of any covariates, were also used. Tests for calibration errors in the model were done with the Hosmer-Lemeshow and le Cessie-van Houwelingen-Copas-Hosmer tests. To further expose model specification issues, deviance residuals were plotted against unscaled predictor variables. Results of these tests can be found in Figure 8 and Figure 9. Model C was used for residual diagnostics as it demonstrated the largest log likelihood and smallest BIC.

Residual diagnostics in Figure 9 indicate the presence of bias in the estimated

Hosmer-Lemeshow's C Test

	X_squared	df	p_value
ModelA	1429.8	8	<< 0.05
ModelB	1312.8	8	<< 0.05
ModelC	1279.7	8	<< 0.05

le Cessie-van Houwelingen-Copas-Hosmer Test

	Z	p_value
ModelA	-115.34	<< 0.05
ModelB	-143.72	<< 0.05
ModelC	-164.68	<< 0.05

Likelihood Ratio Test

	Model_df	LogLik	Test_df	LR_Chisq	P_value
ModelA	21	4,382	19	1620	<< 0.05
ModelB	31	4,369	29	1593	<< 0.05
ModelC	21	5,179	19	3215	<< 0.05

BIC and Ferrari & Cribati-Neto Pseudo R2

	BIC	R2
Null	−7,125	0.000
ModelA	−8,572	0.121
ModelB	−8,452	0.120
ModelC	−10,166	0.115

Figure 8. Goodness-of-fit and model validation test results for the three different beta regression models. The precsion paramter of the beta model is considerd a function of visitor type, web browser, and product related for Models A, B, and C, respectively.

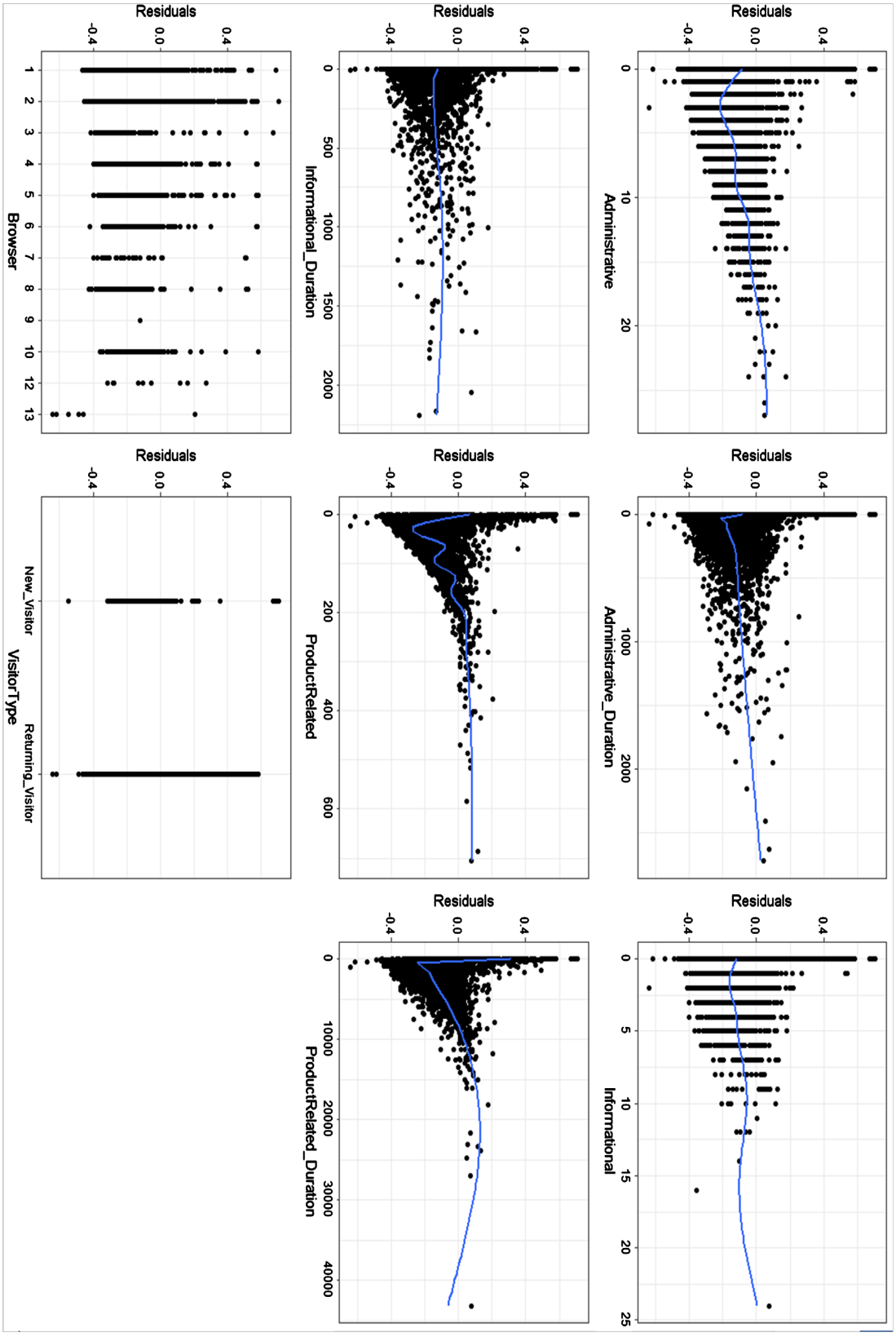


Figure 9. Deviance residual plots showing residuals plotted against the predictor variables used in each model.

model. This is consistent with the results of the le Cessie-van Houwelingen-Copas-Hosmer global miscalibration test which suggested that the model required the addition of bias reducing factors such as non-linear transformations of predictors or interaction terms between explanatory variables.

6.3. Bounce Rate Model

Unlike the exit rate model, the decision to bounce is hypothesized to be the result of factors external to the website. If the observed pattern of bouncing from pages is determined by external factors, then too much investment in mechanisms to decrease it could be wasteful. Three categories of variables were conjectured to have influence on bounce rate. Variables from each of these three categories were selected to test if bounce rate is dependent on them. Testing of relationships was conducted by fitting a beta regression model. Variability in bounce rate was hypothesized to be a function of two variables: technological variables (Browser) and previous site usage (VisitorType). Two separate beta regression models were fitted to the data where a linear predictor for the precision parameter included Browser (Model E) or VisitorType (Model D). Results of estimated effects are provided in Table 6.

Hypothesis 1b posited that as a visitor accumulated greater knowledge of the website and reduced his uncertainty about his goals, then observed bounce rate would decrease. Looking at the estimated for VisitorType in Table 6, the data does not support this hypothesis. The estimate is in the opposite direction. One might consider that visitors with a greater understanding of the layout and organization of a website may know how to traffic directly to the page they are interested in. Other websites or browsers may not link to the pages returning visitors enter at. Consequently, the page first seen by a new visitor may immediately contradict expectations, thus promoting a bounce. Hypothesis 4b is also not supported by the data as it proposed a negative association between previous site usage and dispersion in bounce rate. Looking at the results indicate that the estimate is in the opposite direction with an estimated positive marginal effect of being a returning visitor. It is worth noting that bouncing was hypothesized to be strongly dependent on factors exogenous to the website. It can be assumed that many of these are unobserved and could complicate the findings. Hypothesis 5b is supported by data which shows that technological factors do influence the dispersion in bounce rate. The two browser types with the largest positive and negative marginal effects are reported in Table 6. This is highly consistent with theory and literature as incompatibilities between browser output and visitor intention can engender immediate departure from the website.

Assessment of goodness-of-fit was performed by examining the pseudo-R^2, BIC, and likelihood ratio tests against the null intercept only model. Calibration errors in the model were tested for with the Hosmer-Lemeshow and le Cessie-van Houwelingen-Copas-Hosmer tests. Additional concerns with specification are documented in deviance residual plots. Results of these tests can be

Table 6. Estimated results for beta regression models with bounce rate as the response. Model D considers the precision parameter in the beta model to be a function of visitor type. Model E considers the precision parameter in the beta model to be a function of web browser.

Explanatory Variable	Odds Ratio (95% CI)	
	Model D	Model E
Month [Feb]	0.116 (−0.122, 0.355)	0.143 (−0.098,0.383)
Month [Mar]	−0.137 (−0.279, 0.005)	−0.113 (−0.256,0.030)
Month [May]	0.028 (−0.106, 0.163)	0.047 (−0.089, 0.183)
Month [June]	0.157 (−0.047, 0.361)	0.199 (−0.009, 0.408)
Month [July]	0.087 (−0.093, 0.267)	0.119 (−0.063, 0.301)
Month [Sep]	−0.094 (−0.269, 0.081)	−0.077 (−0.254, 0.100)
Month [Oct]	−0.130 (−0.296,0.037)	−0.119 (−0.288, 0.051)
Month [Nov]	0.061 (−0.076,0.199)	0.064 (−0.075, 0.203)
Month [Dec]	0.013 (−0.131, 0.157)	0.035 (−0.111, 0.181)
Browser [2]	−0.102*** (−0.164, −0.040)	−0.290*** (−0.371, −0.210)
Browser [3]	0.226 (−0.045, 0.497)	0.236 (0.087, 0.559)
Browser [4]	−0.214*** (−0.324, −0.104)	−0.282*** (−0.433,−0.132)
Browser [5]	−0.123* (−0.256, 0.010)	−0.308*** (−0.486, −0.130)
Browser [6]	−0.148 (−0.361, 0.065)	−0.408*** (−0.456, −0.130)
Browser [7]	−0.183 (−0.558, 0.192)	−2.478*** (−3.116, −1.840))
Browser [8]	0.028 (−0.219, 0.276)	0.102 (−0.192,0.397)

Continued

Browser [10]	−0.234**	−0.358**
	(−0.451, −0.017)	(−0.660, −0.057)
Browser [12]	−0.379	−1.813**
	(−1.243, 0.486)	(−3.296, −0.329)
Browser [13]	0.194	0.498
	(−0.942, 1.330)	(−0.802, 1.797)
Region [2]	−0.020	−0.025
	(−0.107, 0.067)	(−0.114, 0.063)
Region [3]	0.011	0.004
	(−0.054, 0.077)	(−0.064, 0.071)
Region [4]	−0.031	−0.029
	(−0.116, 0.055)	(−0.116, 0.058)
Region [5]	0.012	0.025
	(−0.138, 0.162)	(−0.129, 0.179)
Region [6]	0.035	0.039
	(−0.067, 0.137)	(−0.064, 0.143)
Region [7]	−0.042	−0.030
	(−0.143, 0.059)	(−0.133, 0.073)
Region [8]	−0.050	−0.048
	(−0.180, 0.079)	(−0.180, 0.084)
Region [9]	−0.116*	−0.104
	(−0.244, 0.012)	(−0.235, 0.026)
TrafficType [2]	−1.184***	−0.412***
	(−1.280, −1.088)	(−0.485, −0.339)
TrafficType [3]	−0.048	−0.068
	(−0.147, 0.052)	(−0.152, 0.016)
TrafficType [4]	−0.699***	−0.365***
	(−0.836, −0.562)	(−0.470, −0.260)
TrafficType [5]	−0.948***	−0.406***
	(−1.224, −0.671)	(−0.581, −0.231)
TrafficType [6]	−0.291***	−0.203***
	(−0.469, −0.113)	(−0.344, −0.062)
TrafficType [7]	−2.363***	−0.374*
	(−2.995,−1.730)	(−0.767, 0.019)
TrafficType [8]	−0.572***	−0.321***
	(−0.792, −0.351)	(−0.477, −0.164)

Continued

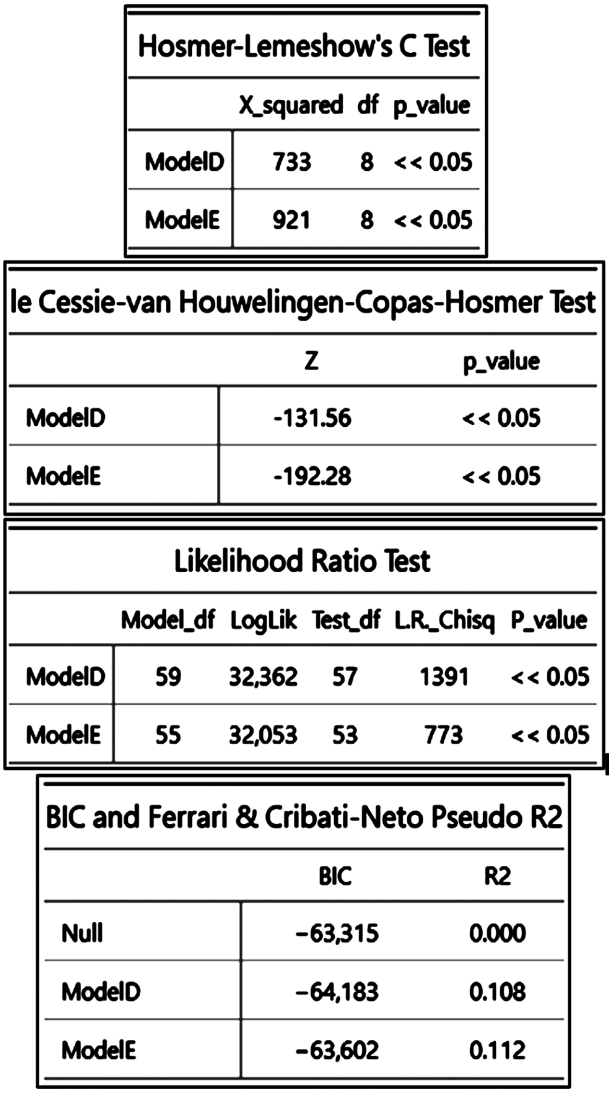
TrafficType [9]	0.172	−0.161
	(−0.335, 0.679)	(−0.572, 0.249)
TrafficType [10]	−0.700***	−0.348***
	(−0.888, −0.511)	(−0.487, −0.210)
TrafficType [11]	−0.362***	−0.212**
	(−0.599, −0.124)	(−0.397, −0.026)
TrafficType [13]	0.240***	0.389***
	(0.107, 0.373)	(0.268, 0.511)
TrafficType [14]	−3.405***	−0.649*
	(−4.717, −2.093)	(−1.332, 0.033)
TrafficType [16]	−9.234***	−0.840
	(−9.646, −8.821)	(−2.401, 0.721)
TrafficType [20]	−0.283*	−0.205*
	(−0.583, 0.017)	(−0.442, 0.033)
VisitorType [Returning_Visitor]	0.452***	0.431***
	(0.382, 0.552)	(0.358, 0.504)
Weekend	−0.022	−0.031
	(−0.078, 0.034)	(−0.089, 0.026)
Precision Parameter Predictor Estimate:		
VisitorType	0.978*	
Browser [7]		2.688*
Browser [13]		−0.598*

found in **Figure 10** and **Figure 11**. Model D demonstrated the smaller BIC and larger log likelihood and hence it was examined in the residual diagnostics.

The presence of bias in the deviance residual plots in **Figure 11** are less pronounced and detectable than in the exit rate model, but the appearance of two clusters of residuals above and below zero conveys a bias that may be addressed with fitting a mixture model on the data. The appearance of gaps in the residual lines is consistent with the skewness of the data and is suggestive of the presence of a second mode in the data's distribution.

7. Conclusions

Digitalization of commerce presents opportunities and challenges to firms including knowing how to best manage information derived from consumer behavior. A central challenge is how to interact with customers unobtrusively but effectively. Design of an effective E-commerce platform crucially depends on knowledge of visitors. Web analytics and clickstream data analysis are two



Hosmer-Lemeshow's C Test

	X_squared	df	p_value
ModelD	733	8	<< 0.05
ModelE	921	8	<< 0.05

le Cessie-van Houwelingen-Copas-Hosmer Test

	Z	p_value
ModelD	-131.56	<< 0.05
ModelE	-192.28	<< 0.05

Likelihood Ratio Test

	Model_df	LogLik	Test_df	L.R._Chisq	P_value
ModelD	59	32,362	57	1391	<< 0.05
ModelE	55	32,053	53	773	<< 0.05

BIC and Ferrari & Cribati-Neto Pseudo R2

	BIC	R2
Null	−63,315	0.000
ModelD	−64,183	0.108
ModelE	−63,602	0.112

Figure 10. Goodness-of-fit test and model validation for the bounce rate beta regression models. Model D considers the precision parameter in the beta model to be a function of visitor type. Model E considers the precision parameter in the beta model to be a function of web browser.

approaches to obtain such knowledge. A haphazard implementation of web analytic techniques represents a risk if it is not grounded in theory and tested analytically. The importance of E-commerce platforms necessitates a well-informed and well-defined evaluation of actionable information in clickstream data. This research investigated the dependencies in clickstream data and if the information patterns are interpretable from perspective of statistical models. Clickstream data from an online retailer was analyzed to ascertain correlation, distributional properties, and the extent to which statistical models could supply usable insights of the data. Exploratory investigation of correlation structure revealed that both categorical and continuous clickstream variables showed correlation. A distributional examination of the bounce and exit rate uncovered that both

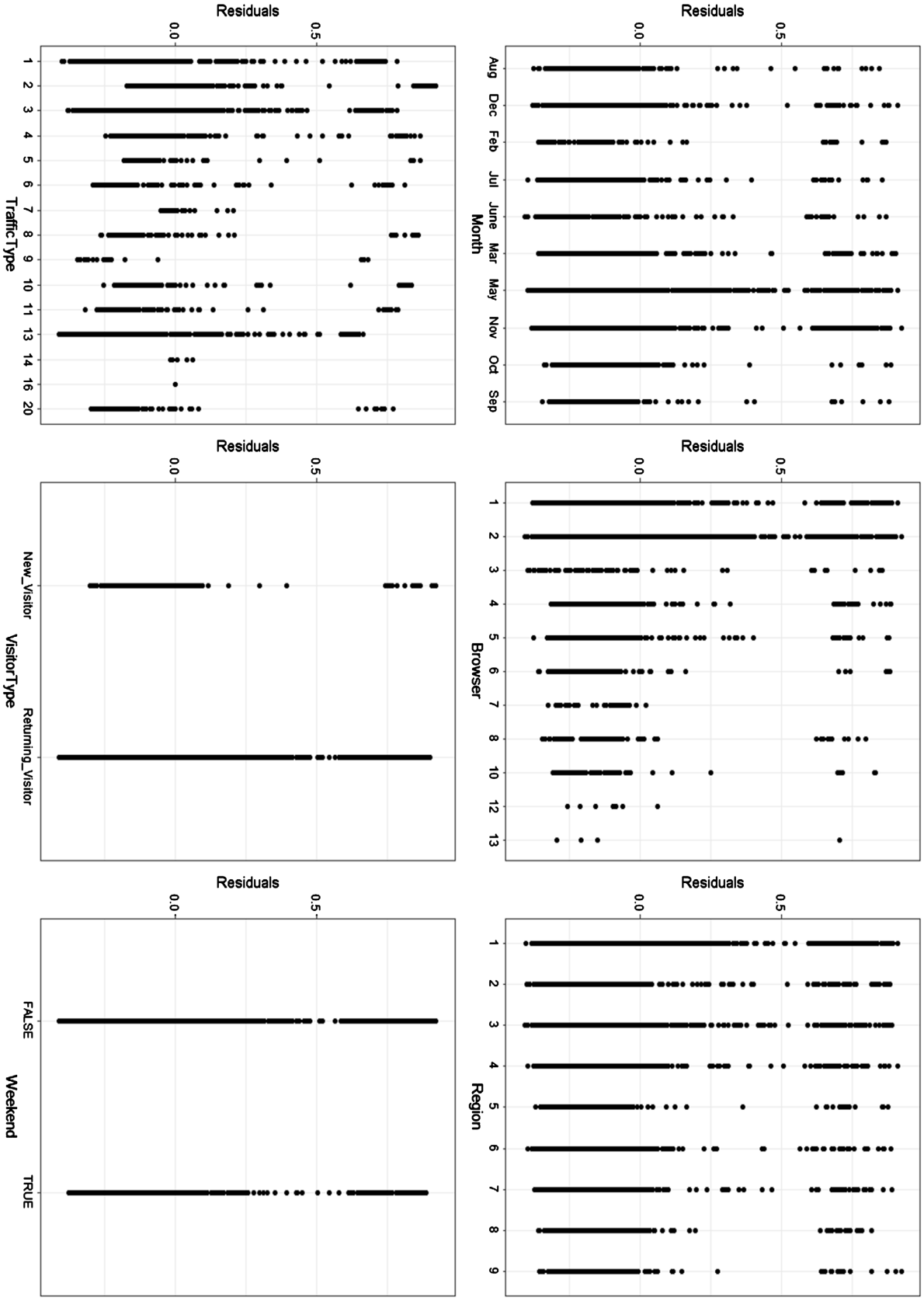


Figure 11. Deviance residual plots against the predictor variables used in the bounce rate beta regression models.

would be amenable to beta regression model analysis. A conceptual model with associated hypotheses was tested by fitting the data to three classes of models: a logistic discrete choice model to capture associations with purchasing behavior, a beta regression model to capture associations with exit rate, a beta regression model to capture associations with bounce rate. All hypotheses were supported by the data besides hypotheses regarding the influence of previous site usage on bounce rate and variability in bounce rate. Previous site usage, site content and quality, and web browser preference were shown to impact the decision to purchase, exit rate, and variability in exit rate. Influence of the previous site usage, site content, and browser preference on exit and bounce rate variability was determined by modeling the precision parameter of the beta regression models as explicit functions of these covariates. Most findings conformed to expectations from theory and literature.

This research was not without its limitations. One of the central assumptions used in this research was that the exit and bounce rates were in the open interval (0, 1). A shortcoming of excluding 0 and 1 is that the data contained several of these boundary observations and a data transformation was used instead. An exit or bounce rate value of 0 or 1 originates from a process separate from the process used to create values in (0, 1). Future work could explore zero and one inflated beta regression models to determine if they offer additional insight. This research was predominantly interested in mathematically tractable and interpretable models. Data exploration showed the potential presence of two subpopulations in the exit and bounce rate data. Only the major subpopulation was concentrated as the inclusion of both would forbid a direct beta model. An approach that would improve the external validity of the models would be to decompose the data into two subgroups and model the heterogeneity with a mixture of beta regressions. Another slightly related method to detect the existence of intergroup heterogeneity is to apply a beta regression tree model to determine if the influence of certain covariates is heterogeneous across different subgroups.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Ramanathan, V., Yellayi, V.S., Karim, S., Funari, P., Wilk, J. and Schneier, C.R. (2018) E-Commerce Trends—A Service Enterprise Engineering Perspective. Penn State College of Engineering. SEE360Initiative. https://see360.psu.edu/files/2018/07/E-Commerce-White-Paper-rf0fxz.pdf

[2] Chaffey, D. and Patron, M. (2012) From Web Analytics to Digital Marketing Optimization: Increasing the Commercial Value of Digital Analytics. *Journal of Direct, Data, and Digital Marketing Practice*, **14**, 30-45. https://doi.org/10.1057/dddmp.2012.20

[3] Thelwall, M. (2009) Introduction to Webometrics: Quantitative Web Research for

the Social Sciences. Morgan and Claypool, San Rafael.
https://doi.org/10.1007/978-3-031-02261-6

[4] Penniman, W.D. (1975) A Stochastic Process Analysis of Online User Behavior. *Annual Meeting of the American Society for Information Science*, Washington DC, 30 March-3 April 1975, 147-148.

[5] Penniman, W.D. (2008) Historic Perspective of Log Analysis. In: Jansen, B.J., Spink, A. and Taksa, I., Eds., *Handbook of Research on Web Log Analysis*, IGI, Hershey, 18-38. https://doi.org/10.4018/978-1-59904-974-8.ch002

[6] Peters, T. (1993) The History and Development of Transaction Log Analysis. *Library Hi Tech*, **42**, 41-66. https://doi.org/10.1108/eb047884

[7] Peterson, E. (2004) Web Analytics Demystified: A Marketer's Guide to Understanding How Your Web Site Affects Your Business. Celilo Group Media, New York.

[8] Booth, D.L. and Jansen, B.J. (2008) A Review of Methodologies for Analyzing Websites. In: Jansen, B.J., Spink, A. and Taksa, I., Eds., *Handbook of Research on Web Log Analysis*, IGI, Hershey, 143-164.
https://doi.org/10.4018/978-1-59904-974-8.ch008

[9] Özmutlu, S., Özmutlu, H.C. and Spink, A. (2008) Topic Analysis and Identification of Queries. In: Jansen, B.J., Spink, A. and Taksa, I., Eds., *Handbook of Research on Web Log Analysis*, IGI, Hershey, 345-358.
https://doi.org/10.4018/978-1-59904-974-8.ch017

[10] Jansen, B.J. (2009) Understanding User-Web Interactions via Web-Analytics. Synthesis Lectures on Information Concepts, Retrieval, and Services. Springer, Berlin.
https://doi.org/10.1007/978-3-031-02264-7

[11] Chen, H.-M. and Cooper, M.D. (2002) Stochastic Modeling of Usage Patterns in a Web-Based Information System. *Journal of the American Society for Information Science and Technology*, **53**, 536-548. https://doi.org/10.1002/asi.10076

[12] Chen, H.-M. and Cooper, M.D. (2001) Using Clustering Techniques to Detect Usage Patterns in a Web-Based Information System. *Journal of the American Society for Information Science and Technology*, **52**, 888-904.
https://doi.org/10.1002/asi.1159

[13] Burby, J. and Atchison, S. (2007) Actionable Web Analytics: Using Data to Make Smart Business Decisions. Wiley, Indianapolis.

[14] Becher, J.D. (2005) Why Metrics-Centric Performance Management Solutions Fall Short. *Information Management Magazine*, March.

[15] Sapir, D. (2004) Online Analytics and Business Performance Management. BI Report.

[16] Ansari, S., Kohavi, R., Mason, L. and Zheng, Z. (2001) Integrating E-Commerce and Data Mining: Architecture and Challenges. *IEEE International Conference on Data Mining*, San Jose, 29 November-2 December 2001, 27-34.

[17] Moore, W.W. and Fader, P.S. (2004) Capturing Evolving Visit Behavior in Clickstream Data. *Journal of Interactive Marketing*, **18**, 5-19.
https://doi.org/10.1002/dir.10074

[18] Chatterjee, P., Hoffman, D.L. and Novak, T.P. (1998) Modeling the Clickstream: Implications for Web-Based Advertising Efforts. *Marketing Science*, **22**, 520-541.
https://doi.org/10.1287/mksc.22.4.520.24906

[19] Wang, G., Zhang, X., Tang, S., Zheng, H. and Zhao, B. (2016) Unsupervised Clus-

tering for User Behavior Analysis. *CHI* 16: *Proceedings of the* 2016 *CHI Conference on Human Factors in Computing Systems*, San Jose, 7-12 May 2016, 225-236. https://doi.org/10.1145/2858036.2858107

[20] Senecal, S., kalczynski, P.J. and Nantel, J. (2005) Consumers' Decision-Making Process and Their Online Shopping Behavior: A Clickstream Analysis. *Journal of Business Research*, **58**, 1599-1608. https://doi.org/10.1016/j.jbusres.2004.06.003

[21] Howard, J.A. and Sheth, J.N. (1969) The Theory of Buyer Behavior. John Wiley & Sons, Inc., New York.

[22] Cialdini, R.B. (2001) Harnessing the Science of Persuasion. *Harvard Business Review*, **10**, 72-79.

[23] Baumeister, R.F. (2002) Yielding to Temptation: Self-Control Failure, Impulsive Purchasing, and Consumer Behavior. *Journal of Consumer Research*, **28**, 670-676. https://doi.org/10.1086/338209

[24] Li, D. and Wang, M. (2015) 60% Purchase Is Impulsive. http://www.bbtnews.com.cn/2015/1208/131385.shtml

[25] Vohs, K.D. and Faber, R.J. (2007) Spent Resources: Self-Regulatory Resource Availability Affects Impulse Buying. *Journal of Consumer Research*, **33**, 537-547. https://doi.org/10.1086/510228

[26] Cho, C.-H., Kang, J. and Cheon, H.J. (2006) Online Shopping Hesitation. *Cyberpsychology and Behavior*, **9**, 261-274. https://doi.org/10.1089/cpb.2006.9.261

[27] Park, C.-H. and Kim, Y.-G. (2010) Identifying Key Factors Affecting Consumer Purchase Behavior in an Online Shopping Context. *International Journal of Retail & Distribution Management*, **31**, 16-29. https://doi.org/10.1108/09590550310457818

[28] Hasan, B. (2016) Perceived Irritation in Online Shopping: The Impact of Website Design Characteristics. *Computers in Human Behavior*, **54**, 224-230. https://doi.org/10.1016/j.chb.2015.07.056

[29] Swinyard, W.R. and Smith, S.M. (2003) Why People (Don't) Shop Online: A Lifestyle Study of the Internet Consumer. *Psychology & Marketing*, **20**, 567-597. https://doi.org/10.1002/mar.10087

[30] Sakar, C.O., Polat, S.O., Katircioglu, M. and Kastro, Y. (2019) Real-Time Prediction of Online Shopper's Purchasing Intention Using Multilayer Pereptron and LSTM Recurrent Neural Networks. *Neural Computing and Applications*, **31**, 6893-6908. https://doi.org/10.1007/s00521-018-3523-0

[31] Moshe, B., Mcfadden, D., Abe, M., Bockenholt, U., Bolduc, D., Gopinath, D., Morikawa, T., Ramaswamy, V., Rao, V., Revelt, D. and Steinberg, D. (1997) Modeling Methods for Discrete Choice Analysis. *Marketing Letters*, **8**, 273-286. https://doi.org/10.1023/A:1007956429024

[32] Horowitz, J.L. (1994) Advances in Random Utility Models: Report of the Workshop on Advances in Random Utility Models. *Marketing Letters*, **5**, 311-322. https://doi.org/10.1007/BF00999207

[33] Abe, M. (2012) A Generalized Additive Model for Discrete Choice Data. *Journal of Business & Economic Statistics*, **17**, 271-284. https://doi.org/10.1080/07350015.1999.10524817

[34] Cover, T. and Thomas, J. (2006) Elements of Information Theory. 2nd Edition, Wiley & Sons, Hoboken.

[35] Reshef, Y.A., Reshef, D.N., Finucane, H.K., Sabeti, P.C. and Mitzenmacher, M. (2016) Measuring Dependence Powerfully and Equitably. *Journal of Machine Learning Research*, **17**, 1-63.

[36] Reshef, D.N., Reshef, Y.A., Finucane, H.K., Grossman, S.R., McVean, G. and Turnbaul, P.T. (2011) Detecting Novel Associations in Large Data Sets. *Science*, **334**, 1518-1524. https://doi.org/10.1126/science.1205438

[37] Szekely, G.J., Rizzo, M.L. and Bajirov, N.K. (2007) Measuring and Testing Dependence by Correlation of Distances. *The Annals of Statistics*, **25**, 2769-2794. https://doi.org/10.1214/009053607000000505

[38] Ferrari, S. and Cribari-Neta, F. (2004) Beta Regression for Modelling Rates and Proportions. *Journal of Applied Statistics*, **31**, 799-815. https://doi.org/10.1080/0266476042000214501

[39] Schmid, M., Wickler, F., Maloney, K.O., Mitchell, R., Fenske, N. and Mayr, A. (2013) Boosted Beta Regression. *PLOS ONE*, **8**, e61623. https://doi.org/10.1371/journal.pone.0061623

[40] Smithson, M. and Verkuilen, J. (2006) A Better Lemon Squeezer? Maximum-Likelihood Regression with Beta-Distributed Dependent Variables. *Psychological Methods*, **11**, 54-71. https://doi.org/10.1037/1082-989X.11.1.54

[41] Simas, A.B., Barreto-Souza, W. and Rocha, A.V. (2010) Improved Estimators for a General Class of Beta Regression Models. *Computational Statistics & Data Analysis*, **54**, 348-366. https://doi.org/10.1016/j.csda.2009.08.017

[42] Cribari-Neto, F. and Zeileis, A. (2010) Beta Regression in R. *Journal of Statistical Software*, **34**, 1-24. http://www.jstatsoft.org/v34/i02 https://doi.org/10.18637/jss.v034.i02

[43] Hatfield, L.A., Boye, M.E., Hackshaw, M.D. and Carlin, B.P. (2012) Models for Survival Times and Longitudinal Patient Reported Outcomes with Many Zeros. *Journal of the American Statistical Association*, **107**, 875-885. https://doi.org/10.1080/01621459.2012.664517

[44] Ospina, R. and Ferrari, S.L. (2012) A General Class of Zero-or-One Inflated Beta Regression Models. *Computational Statistics & Data Analysis*, **56**, 1609-1623. https://doi.org/10.1016/j.csda.2011.10.005

[45] Swearingen, C.J., Melguizo castro, M.S. and Bursac, Z. (2012) Inflated Beta Regression: Zero, One, and Everything in between. *SAS Global Forum* 2012: *Statistics and Data Analysis*, Cary, North Carolina, 2012, Paper 325.

[46] Zeileis, A., Hothorn, T. and Hornik, K. (2008) Model-Based Recursive Partitioning. *Journal of Computational and Graphical Statistics*, **17**, 492-514. https://doi.org/10.1198/106186008X319331

[47] Breiman, L., Friedman, J., Olshen, R.A. and Stone, C.J. (1984) Classification and Regression Trees. Chapman and Hall, New York.

[48] Grün, B., Kosmidis, I. and Zeileis, A. (2020) Extended Beta Regression in R: Shaken, Stirred, Mixed, and Partitioned. *Journal of Statistical Software*, **48**, 1-25. https://doi.org/10.18637/jss.v048.i11

[49] McLachlan, G. and Peel, D. (2000) Finite Mixture Models. Wiley, New York. https://doi.org/10.1002/0471721182

[50] Wedel, M. and DeSarbo, W.S. (1994) A Review of Recent Developments in Latent Class Regression Models. In: Bagozzi, R., Ed., *Advanced Methods of Marketing Research*, Blackwell Pub., Hoboken, 352-388. https://ssrn.com/abstract=2789856

[51] Magidson, J. and Vermunt, J.K. (2004) Latent Class Models. In: Kaplan, D., Ed., *The Sage Handbook of Quantitative Methodology for the Social Sciences*, Sage, Thousand Oaks, 175-198. https://doi.org/10.4135/9781412986311.n10

[52] Muthén, B.O. and Asparouhov, T. (2009) Multilevel Regression Mixture Analysis. *Journal of the Royal Statistical Society*, *Series A*, **172**, 639-657. https://doi.org/10.1111/j.1467-985X.2009.00589.x

[53] Oberski, D.L. (2015) Beyond the Number of Classes: Separating Substantive from Non-Substantive Dependence in Latent Class Analysis. *Advances in Data Analysis and Classification*, **10**, 171-182. https://doi.org/10.1007/s11634-015-0211-0

[54] Hastie, T., Tibshirani, R. and Friedman, J. (2017) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Edition, Springer, Berlin.

[55] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, *Series B*, **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[56] Lemeshow, S. and Hosmer, D.W. (1982) A Review of Goodness of Fit Statistics for Use in the Development of Logistic Regression Models. *American Journal of Epidemiology*, **115**, 92-106. https://doi.org/10.1093/oxfordjournals.aje.a113284

[57] Hosmer, D.W., Hosmer, T., le Cessie, S. and Lemeshow, S. (1997) A Comparison of Goodness-of-Fit Tests for the Logistic Regression Model. *Statistics in Medicine*, **16**, 965-980. https://doi.org/10.1002/(SICI)1097-0258(19970515)16:9<965::AID-SIM509>3.0.CO;2-O

[58] McFadden, D. (1974) Conditional Logit Analysis of Qualitative Choice Behavior. In: Zarembka, P., Ed., *Frontiers in Econometrics*, Academic Press, Cambridge, 105-142.

[59] Shneiderman, B., Plaisant, C. and Cohen, M. (2008) Designing the User Interface: Strategies for Effective Human-Computer Interaction. 5th Edition, Pearson, London.

[60] Wickens, C.D., Lee, J.D., Liu, Y.L. and Gordon-Becker, S. (2003) An Introduction to Human Factors Engineering. 2nd Edition, Pearson Prentice Hall, London.