COMPARATIVE ANALYSIS OF SELECTED PROBABILISTIC CUSTOMER LIFETIME VALUE MODELS IN ONLINE SHOPPING

Pavel JASEK[1] [ID], Lenka VRANA[1] [ID], Lucie SPERKOVA[1],

Zdenek SMUTNY[1*] [ID], Marek KOBULSKY[2]

[1]*Faculty of Informatics and Statistics, University of Economics, Prague,*
W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic
[2]*Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague,*
Brehova 7, 115 19 Prague 1, Czech Republic

Abstract. The selection of a suitable customer lifetime value (CLV) model is a key issue for companies that are introducing a CLV managerial approach in their online B2C relationship stores. The online retail environment places CLV models on several specific assumptions, e.g. non-contractual relationship, continuous purchase anytime, variable-spending environment. The article focuses on empirical statistical analysis and predictive abilities of selected probabilistic CLV models that show very good results in an online retail environment compared to different model families. For comparison, eleven CLV models were selected. The comparison has been made to the online stores' datasets from Central and Eastern Europe with annual revenues of hundreds of millions of euros and with almost 2.3 million customers. Probabilistic models have achieved overall good and consistent results on the majority of the studied transactional datasets, with BG/NBD and Pareto/NBD models that can be considered stable with significant lifts from the baseline Status quo model. Abe's variant of Pareto/NBD have underperformed multiple criterions and would not be fully useful for the studied datasets without further improvements. In the end, the authors discuss the deployment implications of selected CLV models and propose further issues for future research to address.

Keywords: online retail, marketing management, E-commerce, probabilistic model, CEE region, B2C.

JEL Classification: C53, M21, M31.

Introduction

Customer segmentation according to customer lifetime value (CLV) enhances evaluating decisions in the context of customer relationship management (Borle, Singh, & Jain, 2008; Haenlein, Kaplan, & Schoder, 2006) which helps with building the long-term relationships

*Corresponding author. E-mail: *zdenek.smutny@vse.cz*

with customers and efficient investments into marketing tools. In recent decades, many different CLV models were formulated. In parallel, the progress in ICT supported rise to e-commerce where offers are sold directly to consumers through the website and other Internet-based services (Centre for Retail Research, 2017; Ecommerce Europe, 2016). Due to these direct interactions with customers, e-commerce companies boast high data availability. The competitive pressure initiatives companies to efficiently release finances in marketing activities. Implementation of CLV model using available historical data can be helpful for estimation of customer value. The right selection of proper CLV model as a decision-making base for marketing management suitable for the conditions of the given business is the main issue the e-commerce companies face.

The purpose of this article is to subject selected probabilistic CLV models to comparative analysis in an empirical way and use statistical metrics to evaluate their predictive ability and quality in an e-commerce environment. The practical aim is to help online retailers with the selection and application of proven CLV model and introduce recommendations regarding its implementation. From the managerial point of view, this research proposes a research question: *Which of the compared probabilistic CLV models have a good predictive performance of CLV for e-commerce?*

Achieving stable model predictions for all put datasets and overcoming other compared models suggests good predictive performance. Different models selected for comparison (see Section 2.1) and the fact that datasets were obtained from the Central and Eastern European region (see Section 2.3), which imposes cultural limitations of the results, is an implicit restriction of this study. The selected CLV models follow various assumptions covering different underlying customer behaviour, which constitutes the research interest of this paper (W. Chang, C. Chang, & Li, 2012).

There are two primary motivations to perform such research. First, many literature reviews compare the theoretical aspects of selected CLV models based on results from secondary sources (e.g. Estrella-Ramón, Sánchez-Pérez, Swinnen, & VanHoof, 2013; Gupta et al., 2006; Fader & Hardie, 2009; Damm & Monroy, 2011; W. Chang et al., 2012), but only a few comparative papers based on own empirical research of a larger number of selected models have been presented (Platzer & Reutterer, 2016; Donkers, Verhoef, & De Jong, 2007; Batislam, Denizel, & Filiztekin, 2007). Also missing is an empirical verification on robust datasets in the e-commerce field and generalisation of results achieved beyond the dataset. There is no study addressing all the models entirely. In previous research studies (Jasek, Vrana, Sperkova, Smutny, & Kobulsky, 2018, 2019), the authors focused on representative models of different model families suitable in an e-commerce environment and their evaluation and comparison. The authors also stated that the results of Markov chain model and Vector autoregressive model which use additional non-financial data about customer behaviour (e.g. traffic source, delivery address, purchase day) do not justify the effort which is needed to collect such data (Jasek et al., 2019) even though e.g. social media is an important component of marketing for online stores (Vojtko, 2014). From the results, the authors found a good performance of Extended Pareto/NBD model (Fader, Hardie, & Lee, 2005a) in an e-commerce environment. The suitability of Pareto/NBD framework for non-contractual settings, which is typical for online shopping, already assumed (Schmittlein, Morrison, & Colombo, 1987)

and also remind (Donkers et al., 2007) when they state its inappropriateness in contractual settings (insurance industry). In this regard, the authors decided to continue the research of probabilistic models and examine their different modifications and add new insights through newer models.

In this regard, the presented article compared to previous research is unique because it offers an evaluation of ten mainly probabilistic models on ten big datasets of selected Czech and Slovak online stores – in total almost 2.3 million customers, 3.8 million transactions and 146 million euros in profit. Furthermore, this article puts a view based on the non-contractual relation typical in the e-commerce business.

The second motivation of this study is methodological. In the CLV research, the researchers used design-type of research and presented new models in the form of the artefact, thus empirical comparison of models is appropriate. For this reason, the research methodology follows principles applied in design science (Dresch, Lacerda, & Antunes Jr, 2015; Hubka & Eder, 1996). Selected models are designed and investigated artefacts repeatedly tested in different contextual conditions. Any similar artefacts in the future will be treated according to the generated knowledge from this research, which will help to theorise in the area (Hubka & Eder, 1996; Wieringa, 2014).

This type of research also connects practical and theoretical levels because the comparison results are discussed in relation to individual models and their suitability, but also implementation and managerial aspects of their predictive accuracy and robustness. Medium to large-sized e-commerce companies, therefore, do not need to experiment with different probabilistic CLV models. The practical motivation for this article is to facilitate the choice of the most appropriate model in the online shopping environment. Currently, e-commerce is missing more wide-ranging comparative analyses of CLV models performed on more than one datasets with up-to-date empirical data.

The structure of the paper follows: The theoretical basis of CLV is presented in Section 1. An explorative analysis and description of given datasets, as well as selection of performed CLV models for comparative analysis along with the relevant literature concerning these models, are contained in Section 2. This section also includes statistical metrics for the comparative analysis. Section 3 presents the obtained results from the analysis of all performing models which are discussed in Section 4.

1. Background

The customer is central to all its marketing activities as a valuable intangible asset of the company (Gupta, 2009) which increases the company's market value next to generated income. Marketing emphasizes the interconnection among these activities and all processes related to creation, communication and value for customers, including customer relationship management (Kotler & Keller, 2015). The progress in the field of customer relationship management (CRM) enabled a shift from the transaction-centricity to a customer-centric approach (Kumar, Pozza, Petersen, & Denish Shah, 2009; Fader, 2012). The CLV approach plays a vital role in the customer retention and long-term relationship, which are the primary CRM goals (Haenlein, 2017). CLV identifies segments of customers bringing the company

the largest profit in time (Kim, Jung, Suh, & Hwang, 2006). Based on CLV, it is possible to decide on appropriate strategies for activities in the company's CRM.

Various slightly different definitions of CLV exist in research articles (Estrella-Ramón et al., 2013; Hwang, Jung, & Suh, 2004; Singh & Jain, 2010). The standard accepted formula of CLV contains the calculation of the present value of future net cash flows (Pfeifer, Haskins, & Conroy, 2005) associated with a particular customer (Fader, 2012).

CLV shows how changes in customer purchase behaviour can influence future profitability (Chang et al., 2012). The CLV approach connects marketing and financial metrics as they are always related together, allowing space for management and optimization (C. Williams & R. Williams, 2015). The main applications of CLV target to business-to-consumer (B2C) context while ignoring business-to-business applications of customer asset management (Nenonen & Storbacka, 2016). The relevance of CLV applications is primarily utilized at the customer level, which affects customer loyalty (Qi, Qu, Zhou, & Li, 2015) and customer behaviour impacting retention (Abdolvand, Baradaran, & Albadvi, 2015; Óskarsdóttir, Baesens, & Vanthienen, 2018). According to the results of (Kumar & Pansari, 2016) the country's economy has a direct impact on the frequency of purchases and the margins of purchases and attention should be paid to cultural dimensions and economic conditions due to their different impact on CLV. All of these (and other) elements with the given model and dataset influence directly or indirectly the CLV calculation.

Different models and approaches to CLV calculation emerged over the past three decades. They are designated for different types of business, management views or B2C relationship. Often mentioned is the division of CLV models into contractual, semi-contractual and non-contractual relations between customer and companies (Estrella-Ramón et al., 2013). In the case of online shopping environment, CLV models should consider the non-contractual relation of the B2C relationship, non-membership of the customers in the relationship, always-a-share factor, continuity of the purchases and variable-spending environment (see Jasek et al., 2018).

In the literature, the Web of Science lacks comparative analyses of a greater number of CLV models based on their predictive capabilities. Donkers et al. (2007) found the best performance of simple profit regression models performed on the dataset of the contractual insurance company. Batislam et al. (2007) compared the evaluation of three probabilistic models used in grocery retailer's dataset focusing on loyalty cards and their usage in higher customers' purchase frequency. They observed the improved performance of their own modified Beta Geometric/NBD (BG/NBD) model in this business setting over Pareto/NBD and original BG/NBD models. Probabilistic models are also incorporated into their research (Platzer & Reutterer, 2016) and applied to datasets from different environments. Their new flexible Pareto/GGG model shows better performance against the Pareto/NBD model in datasets with regular timing patterns present in the data. The Pareto/NBD and BG/NBD models along with Pareto/GGG are also included in our non-contractual based dataset with seven other probabilistic models. The complex universally applicable models capture the complexity of relationship developments and achieve consistently solid results in various situations. However, even simple models can achieve exceptional prediction results and outperform complex models, especially if they are modified or designed for specific conditions.

This paper focuses on the prediction of selected models in non-contractual e-commerce environment which is the current topic both in local and global context. These types of companies usually possess an extensive database containing the data about their customers and their transactions, which they use for different purposes in financial management or marketing. The online retail market has been growing continuously, so has the competitive pressure in the area of online shopping – see e.g. reports (Statista, 2018; Centre for Retail Research, 2017). The aforementioned facts make this research in a current environment extremely up-to-date.

2. Methodology and data collection

Figure 1 illustrates the six phases of the research methodology used. The initial phase formulated research question and objectives. It vindicated the necessity and suitability of the proposed research – see Sections 1 and 2. In the next phase, probabilistic models appropriate for use in e-commerce according to the results of (Jasek et al., 2018, 2019) were selected. The Data understanding phase defined data requirements based on the selected models. In Data preparation phase was collected data in the required form from several online shops. The datasets are described in Section 2.3. Further, the acquired datasets were pre-processed according to Section 2.2 to meet the needs of the individual models.
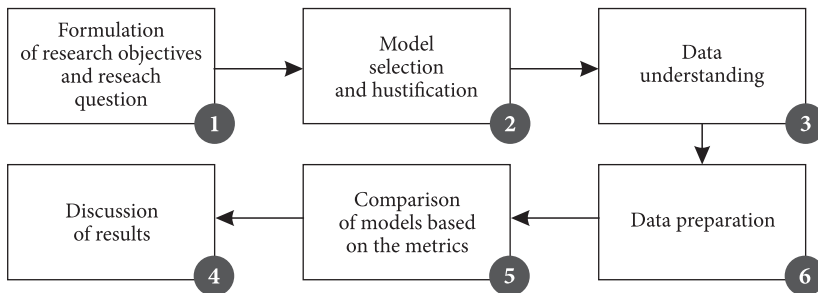

Figure 1. Six primary phases of the research methodology

The next phase is dedicated to the comparison of the models. The performance of the models is evaluated by statistical metrics listed in Section 2.4 the performance of the models was evaluated. Section 2.4 also describes the method of the comparison and definition of a training and testing period. The research question is discussed in Section 3. The obtained results and implications are subjected to a broader discussion in Section 4 including the relevant managerial view.

2.1. Selection of CLV models and their description

Different approaches to modelling CLV was presented by (Gupta et al., 2006): RFM, econometric, persistence, probabilistic, computer science and diffusion/growth models. Previous research (Jasek et al., 2018) focused on evaluation and comparison of these different ap-

proaches – probabilistic with RFM factors and econometric as the most exciting representatives of models to calculate CLV suitable for the e-commerce business conditions. Except for this research, the literature lacks studies comparing CLV models using the same robust dataset(s). The empirical studies (Donkers et al., 2007; Batislam et al., 2007) compare only a handful of selected models concerning a specific area of deployment.

In our previous studies (Jasek et al., 2018, 2019), Extended Pareto/NBD model outperformed other selected models in most of the evaluation criteria. Its results are suitable for e-commerce non-contractual business outputting CLV and probability of a customer being active in the next period.

Thus, the authors decided to focus on the empirical comparison of various modifications of this type of model. Batislam et al. (2007) did compare probabilistic models in his study, but only three of them – Pareto/NBD, MBG/NBD and BG/NBD. Platzer and Reutterer (2016) developed a new Pareto/GGG model and compared it to Pareto/NBD. Both studies also use too small dataset at the order of thousands of customers.

The authors chose probabilistic or also called "buy-till-you-die" (BTYD) models implemented in R package BTYD (Dziurzynski, Wadsworth, & McCarthy, 2015) and BTYDplus (Platzer, 2016) to make a comprehensive comparison of ten types of probabilistic models. The latter model was added on the basis of a naive approach to calculating CLVs as a baseline for easier comparison of individual approaches. The chosen probabilistic models for our comparison are:

- Negative Binominal Distribution (Ehrenberg, 1959) | NBD;
- Pareto/NBD (Schmittlein et al., 1987) | Pareto/NBD;
- Beta-geometric/NBD (Fader et al., 2005a) | BG/NBD;
- Modified Beta-geometric/NBD (Batislam et al., 2007) | MBG/NBD;
- Beta-geometric/NBD with Fixed Regularity (Platzer, 2016) | BG/CNBD-k;
- Modified Beta-geometric/NBD with Fixed Regularity (Platzer, 2016) | MBG/CNBD-k;
- Hierarchical Bayes Pareto/NBD (Ma & Liu, 2007) | Pareto/NBD (HB);
- Abe's variant of Pareto/NBD (Abe, 2009) | Pareto/NBD (Abe);
- Abe's variant of Pareto/NBD with incorporated covariates (Abe, 2009) | Pareto/NBD (Abe M2);
- Pareto/Gamma-Gamma-Gamma (Platzer & Reutterer, 2016) | Pareto/GGG.

The first six models (NBD, Pareto/NBD, BG/NBD, MBG/NBD, BG/CNBD-k, MBG/CNBD-k) belong to maximum likelihood models as they can be efficiently estimated via means of maximum likelihood estimation (MLE). The second group of models (Pareto/NBD (HB), Pareto/NBD (Abe), Pareto/NBD (Abe M2), Pareto/GGG) belong to Markov-Chain-Monte-Carlo (MCMC) models as they rely on MCMC simulation for parameter estimation and thus overcome the restriction of the MLE models from relaxing the underlying behavioural assumptions. As Pareto/NBD (Abe) has the possibility to incorporate covariates of customer behaviour into the computation, a new variant denoted Pareto/NBD (Abe M2) was computed – see Section 2.2 for a list of used covariates.

The NBD model by (Ehrenberg, 1959) assumes a heterogeneous constant purchasing process, with exponentially distributed interpurchase times and Gamma-distributed purchase rate over the customers. One of the essential outcomes of NBD application was the Pareto/NBD model (Schmittlein et al., 1987) describing the repeat-buying behaviour in a

non-contractual setting. Thus, it takes into account the possibility of customers becoming inactive. Pareto/NBD has become a benchmark model in customer-base analysis and empirically studied by other researchers, e.g. (Schmittlein & Peterson, 1994; Reinartz & Kumar, 2000; Batislam et al., 2007; Colombo & Jiang, 1999; Fader & Hardie, 2001; Adomavicius & Tuzhilin, 2005; Singh & Jain, 2013; Knox & van Oest, 2014) and further modified into models mentioned above.

The BG/NBD model adjusts the Pareto/NBD assumptions concerning the dropout process to overcome the computational complexities with parameter estimation (Batislam et al., 2007) while retaining a similar level of data fit and forecast accuracy of Pareto/NBD. The MBG/NBD eliminates the inconsistency of BG/NBD between customers with many purchases and customers without a repeat transaction and assigns a dropout probability after each purchase. This probability does not change until the next purchase. BG/CNBD-k and MBG/CNBD-k models extend previous two for regularity within the transaction timings.

The hierarchical Bayes variant of Pareto/NBD is identical to Pareto/NBD concerning its model assumptions but relies on MCMC for parameter estimation. As it does not take full advantage of the gained flexibility of MCMC simulations, Abe's variant of the Pareto/NBD releases the independence of purchase and dropout process, and moreover, it is capable of incorporating customer covariates, which helps in explaining the heterogeneity within the customer cohort. Finally, the Pareto/GGG is another generalization of Pareto/NBD, which allows a varying degree of regularity within the transaction timings as Gamma-Gamma-Gamma (GGG) mixture distribution.

Spending submodel

As neither one of the BTYD and BTYDplus models use any spending model, authors extend them all by adding the unified spending submodel. The first such extension to Pareto family models was introduced already in (Schmittlein & Peterson, 1994) as Extended Pareto NBD model. Presented data was extended by the monetary value of individual orders independent of their recency and frequency. Results of two submodels, one for the expected average order value and the other for the expected number of transactions, are then multiplied for CLV calculation. Therefore, chosen models consist of the model itself and Gamma-Gamma spending submodel.

This article uses the Gamma-Gamma spending submodel described in (Fader et al., 2005b), further clarified in (Fader & Hardie, 2013). It assumes the following (Jasek et al., 2018):

   – The monetary value of a customer's transaction varies randomly around their average order value.
   – Average transaction values differ across customers but do not differ over time for any given individual.
   – The distribution of average transaction values across customers is independent of the transaction process.

The spending submodel uses 4 fitting parameters derived from the heterogeneities in transaction and dropout rates across all customers. Its calculation and application to Pareto/NBD are described in detail in previous research (Jasek et al., 2018).

Status Quo model

Status Quo model was chosen by authors as a baseline for a comparison of the benefits of tested methods. The necessity of the model should lie in a simple calculation. Given that an appropriate alternative for non-contractual relations could not be found, the authors modified the model of (Donkers et al., 2007) used in the contractual environment. The Status quo model in this paper has three prerequisites:

1. Inactive customer is assumed not to purchase for more than a year.
2. Active customer is assumed to purchase the next week the same value as his average weekly purchase in the last year of that period (52 weeks).
3. If the customer was active in the last year of the period, he is not expected to leave during the forecasted period, which can be written as:

$$Profit_{i,p+j} = \frac{\sum_{t=p-52}^{p-1}Profit_{i,t}}{52} , \; j = 0, 1, 2, 3, \ldots, h - p \; , \tag{1}$$

where $Profit_{i,t}$ is the profit from customer i in time t, p is the threshold of the prediction and h is its horizon. The oneyear period is chosen to keep the described model simple. Also, in practice, the inactivity of customer exceeding one year is considered as leaving the company. Thanks to the results of the Status Quo model, the real benefits of probabilistic models can be better evaluated – see also (Jasek et al., 2018). In comparison to Status Quo model, the performance of other probabilistic models depends on the advancement of the statistical methods or dataset size they use.

2.2. Data collection and its pre-processing

The definition of the required data features and its structure for models from section 2.1 was published in an openly available call for data. Several medium- to large-sized online shops from the Czech Republic and the Slovak Republic were asked to participate in the research. The requirements imposed on data attributes included date, purchase-level identification of customer and purchase status, marketing source of the purchasing visit, shipping costs, item quantity, postcode of purchase delivery country and region, revenue and net profit alongside with a unique purchase identifier. All data was anonymized to avoid infringement of personal data. A minimum of thousands of unique customers and two years purchase range was required.

Data pre-processing included (i) descriptive analysis, (ii) data cleaning and (iii) selection of a feature subset to perform models. All the identified outliers detected in descriptive analysis phase were removed in cooperation with the given online store. Datasets were cleaned to include only purchases from the most frequent country in data. Datasets were aggregated on a weekly level with aggregation details described below and trimmed to whole weeks as a suitable basic unit for the prediction for all models.

The collected data served as a baseline for the creation of the dataset for individual models which shares the same structure among all the models. Table 1 presents randomly selected rows of the given dataset. Letters A-J denote individual online store datasets. Records are

built on weekly purchases of every customer, where customer identifier *customer_id* is unique to each dataset. Weeks (*week_number*) correspond to the dataset's data range counting from number 1. A specific week is characterised by *monday_date*. *Profit_EUR* is the amount of gross profit attributed to every purchase.

Table 1. Five randomly selected rows from the complete dataset

Dataset (online store)	customer_id	week_number	monday_date	profit_EUR
J	1800324	10	9.3.2015	14.9
H	225064	114	30.3.2015	27.7
F	1543194	183	1.7.2013	4.8
I	224346	137	11.8.2014	5.7
J	1460541	42	19.10.2015	31.2

Despite different calculation of methods, all models have the same output – calculated CLV. These outputs are evaluated against the real values of the same metrics. On this base, this paper can be considered a relevant comparative analysis.

As Pareto/NBD (Abe M2) has a possibility to incorporate covariates (Platzer, 2016), following four customer-level variables processed from the minimal dataset: (1) log of inter-transaction times, (2) value of the first transaction, (3) quarter of the first purchase and (4) value of the last transaction.

2.3. Description of datasets

The datasets for the comparative analysis of the models originate in selected Czech and Slovakia online stores. These datasets met the criteria set in Section 2.2. The summary information about the available data of individual stores is listed in Table 2. Not every store has the entire history of data since its existence available, but recent history encompasses years between 2008 and 2017 ranging from 104 to 381 weeks. If a store has also a conventional store, so it was used data only from an online store. Following part concisely summarizes the business verticals of the datasets. All selected companies agreed with their participation in this study with the condition of anonymity and non-dissemination of data to other parties.

The dataset A represents online store operating in the narrow area of games of all sorts. The online store B belongs among the biggest sports equipment stores in the Czech Republic with numerous stores and activities also outside the country. The online store C operates in vertical of health products in the Czech Republic and Slovakia. The online shop D focuses on winter and adrenaline sports with the complementary store. Online store E is one of the biggest erotic online shops with a strong brand in the Czech Republic running also stores. Online shop F focuses on health and beauty products expanding also to other European countries. The company G provided relatively low two-year transaction history focusing on health products and baby care. Online store H focuses on home decoration and interior design products. The online store I focuses on health and beauty products. The company J is the

Table 2. Summary information about individual online stores and available data

Dataset (online store)	Number of transactions	Number of customers	Sum of profit EUR	Average transaction profit EUR	Data range (in weeks)
A	19 433	14 404	158 385	8.37	218
B	136 611	92 234	2 735 974	20.45	151
C	106 129	49 376	592 177	5.88	173
D	119 439	74 303	1 727 439	15.23	364
E	62 744	44 503	1 170 913	18.85	381
F	2 409 019	804 710	19 173 745	8.38	301
G	424 944	138 698	29 530 295	70.83	104
H	603 366	180 699	10 261 410	19.95	241
I	371 178	221 324	4 352 746	11.89	206
J	3 181 038	664 556	76 373 448	30.95	149

largest dataset of this study focusing on beauty and fashion products. Companies C, D, F and H experience strong and steady year-over-year sales growth. The online stores A, B and F are impacted by strong Christmas season with dominated sales and some peaks during the year.

2.4. Evaluation by statistical metrics

After data exploration was defined the training and testing periods for the prediction. One-year (52 weeks) testing period was chosen to cover enough data points, but also to keep a sufficient amount of data in the training period. The 52 weeks from each dataset were separated for the evaluation of the models' predictive performance.

The remaining data interval (the total number of weeks without the final 52 weeks) served as the training period, and its length differs per each data set.

Data in the testing period was cleansed of newcomer customers, as the models serve to estimate the behaviour of the existing customers and not to predict the newly acquired ones. The comparisons were performed both at the individual and the aggregated customer base level consistent with (Donkers et al., 2007).

The aggregated customer base level predicts CLV for all the customers together as an overall perspective. The performance of the models was evaluated with the Forecast vs. Actual metric, which assesses the quality of the models at predicting the total profits. The Forecast vs. Actual (FA) metric is defined as

$$FA = \frac{\sum_{i=1}^{n} F_i}{\sum_{i=1}^{n} A_i} \times 100,$$ (2)

where A_i is the sum of actual profits from the i-th customer over the whole testing period, F_i is the sum of forecasted profits from the i-th customer over the entire testing period, and n is the number of customers.

The other perspective compares the predictions at the individual customer level. These metrics assess how good are the models for selecting the top customers. Several metrics are evaluated: mean absolute error (MAE), Spearman's rank correlation coefficient and sensitivity. The mean absolute error is defined as

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|A_i - F_i\right| \tag{3}$$

and it produces values from zero to infinity. Zero means that all the predicted values are equal to the actual values. The higher the number, the more significant differences there are between the actual and predicted values. The values of MAE in its base form depend on the distribution of the input data and couldn't be compared across several data sets. Therefore, MAE in the percentage of the of the average actual profit is used instead. Donkers et al. (2007) used the MAE as the percentage of the average CLV, but the overestimated values of CLV would also mean the high values of average CLV and that would cause the unreasonably small values of MAE as the percentage of the average CLV.

Another metric is the Spearman's correlation coefficient, which measures the correlations between ranks of two variables (actual and forecasted profit of i-th customer). It was used by Chamberlain, Cardoso, Liu, Pagliari, and Deisenroth et al. (2017) to evaluate CLV prediction for e-commerce ASOS. The coefficient can reach values from –1 to 1. 1 means that there is a monotonic relationship between the two variables. This means, that the customers with the highest forecasted profit are also the ones with the highest actual profit. The Spearman's correlation coefficient is defined as

$$r_s = \frac{\sum_{i=1}^{n}\left(R\left(A_i\right) - \overline{R(A)}\right) \times \left(R\left(F_i\right) - \overline{R(F)}\right)}{\sqrt{\sum_{i=1}^{n}\left(R\left(A_i\right) - \overline{R(A)}\right)^2 \times \sum_{i=1}^{n}\left(R\left(F_i\right) - \overline{R(F)}\right)^2}}, \tag{4}$$

where $R\left(A_i\right)$ and $R\left(F_i\right)$ are ranks of actual and forecasted profits from the i-th customer, respectively.

The success rate of selecting the most profitable customers was evaluated with individual models. 1% and 10% of the most profitable customers were selected and compared to the actual top 1% and 10%. The sensitivity of performance was computed as

$$sensitivity = \frac{TP}{TP + FN}, \tag{5}$$

where TP (true positives) is the number of customers assigned correctly to the top class, and FN (false negatives) is the number of customers who were not assigned to the top class by the CLV calculation but who eventually turned out to belong to this class. Sensitivity gets values from 0 to 1. The higher the number, the better the performance of the model.

3. Results

This study aims to compare the predictive ability and quality of the introduced CLV models using selected evaluation metrics. This section presents the results of each model for every

dataset by performance metrics of FA, MAE on a customer level, Spearman's rank correlation coefficient on a customer level, and sensitivity for identifying 1% or 10% of the most profitable customers, respectively. Results across datasets are weighted by either profit or number of customers, and relative standard deviation is computed to evaluate variability in results. The overview and comparison methodology of evaluation metrics is presented in Section 2.4. Section 4 further offers the implication and discussion of the results. A graphical comparison of results and extended evaluation of sensitivity for identifying 5% or 20% of the most profitable customers can be found in the Appendix (Figures 2–4).

For results interpretation, it is imperative to emphasise that a quality of some results corresponds to the complex prediction subject: the models not only aim at estimating purchase probability, purchase frequency and purchase value, but the final output is the profit in time.

Table 3. Results for Forecast vs. Actual (in %)

Dataset (online store)	Status Quo	BG/NBD	BG/CNBD-k	MBG/NBD	MBG/CNBD-k	NBD	Pareto/NBD	Pareto/NBD (HB)	Pareto/NBD (Abe)	Pareto/NBD (Abe M2)	Pareto/GGG
A	280.70	83.27	83.27	82.94	82.94	95.01	72.96	70.49	44.21	58.72	74.91
B	370.26	142.90	142.90	138.20	138.20	183.35	152.90	136.36	82.12	77.09	136.79
C	175.81	101.42	97.67	97.49	91.80	134.70	113.37	95.66	56.76	31.56	93.56
D	232.93	88.25	88.25	86.18	86.18	153.75	101.38	93.46	54.26	54.84	90.11
E	281.04	86.00	86.00	85.56	85.56	116.96	97.53	90.73	58.06	56.23	89.56
F	114.67	82.17	82.17	81.84	81.84	88.53	85.05	80.79	54.52	31.08	81.08
G	171.27	146.93	121.89	139.90	118.57	163.54	154.27	132.15	98.47	45.63	109.23
H	138.57	85.84	85.84	84.70	84.70	119.58	94.60	86.16	50.30	28.04	85.80
I	194.06	94.25	94.25	90.88	90.88	126.83	104.49	90.48	44.67	38.57	91.01
J	112.09	88.72	88.72	85.64	85.64	108.17	93.37	84.49	54.41	30.53	84.58
Weighted mean (by profit)	130.06	100.44	95.20	96.94	92.48	119.14	105.85	94.61	63.44	34.17	89.89
Relative standard deviation (%)	43.32	18.70	12.95	16.82	11.69	15.42	16.88	13.96	21.99	29.24	8.74

It can be inferred that undervaluation of the customer base offers lower risks for managerial implications than overestimation. According to results in Table 3, the best model for predicting customer base value is BG/NBD model, reaching an average 100.44% of the actual profits and solid relative standard deviation of 19%. Even the other close models (BG/CNBD-k, MBG/NBD and MBG/CNBD-k) have performed well with overall undervaluation.

The most stable results were delivered by Pareto/GGG with 90% average FA and 9% standard deviation. The worst model is SQ with always highly overestimated values (130%)

and according to the standard deviation (43%) also very unstable in predictions. Pareto/NBD (Abe) and its M2 variation with covariates underperformed actual profits in all cases (63% and 34%, respectively).

Datasets B (sports equipment) and G (health products and baby care) have significantly high levels of FA, with an average of 155% or 127%, respectively. The other evaluation metrics for these datasets do not follow the same pattern, so there is no common factor influencing such results.

Table 4. Results for MAE (customer-level, in %)

Dataset (online store)	Status Quo	BG/NBD	BG/CNBD-k	MBG/NBD	MBG/ CNBD-k	NBD	Pareto/NBD	Pareto/NBD (HB)	Pareto/NBD (Abe)	Pareto/NBD (Abe M2)	Pareto/GGG
A	272.10	97.74	97.74	97.28	97.28	102.83	132.60	98.18	136.09	123.00	97.67
B	374.22	177.93	177.93	174.37	174.37	207.58	184.09	173.25	152.99	147.04	173.96
C	156.68	113.05	110.84	110.46	106.76	132.58	118.64	110.54	116.88	101.38	107.68
D	251.55	144.55	144.55	142.38	142.38	196.84	153.71	147.05	141.56	140.95	146.34
E	297.23	155.66	155.66	155.24	155.24	178.51	164.19	158.56	141.86	140.40	157.83
F	104.82	96.03	96.03	95.92	95.92	99.04	97.13	96.30	104.90	94.79	95.49
G	139.38	118.97	106.02	114.12	101.26	129.32	123.05	111.13	110.12	87.00	97.76
H	134.54	104.63	104.63	103.88	103.88	123.25	107.89	104.63	110.68	97.52	104.65
I	193.75	125.36	125.36	123.40	123.40	147.11	131.31	123.77	125.45	118.35	122.65
J	103.59	93.90	93.90	93.01	93.01	103.49	95.56	93.08	102.06	91.36	92.70
Weighted mean (by profit)	117.08	101.31	98.61	99.65	96.96	111.31	103.66	99.17	105.47	92.23	96.04
Relative standard deviation (%)	56.56	19.89	21.27	20.44	22.05	24.20	21.64	21.78	12.92	19.01	23.75

Table 4 summarizes the results for MAE. Given the ideal MAE value of 0%, the majority of models performed poorly, yet delivered quite consistent results. The lowest errors can be observed for Pareto/NBD (Abe M2) (MAE of 92.23% on average) with covariates what has decreased MAE in comparison with Pareto/NBD (Abe) from 105% to 92% but also increased standard deviation from 13% to 19%. Worst results were displayed by SQ model with MAE of 117% and standard deviation of 57%, with a range of results from 104% to 374%).

Table 5. Results for Spearman's rank correlation coefficient (customer-level)

Dataset (online store)	Status Quo	BG/NBD	BG/CNBD-k	MBG/NBD	MBG/CNBD-k	NBD	Pareto/NBD	Pareto/NBD (HB)	Pareto/NBD (Abe)	Pareto/NBD (Abe M2)	Pareto/GGG
A	0.2287	0.2987	0.2987	0.2982	0.2982	0.3007	0.2938	0.2862	0.1207	0.2440	0.2929
B	0.1730	0.2509	0.2509	0.2553	0.2553	0.2513	0.2603	0.2384	0.1151	0.1606	0.2334
C	0.3674	0.4124	0.3948	0.4165	0.4043	0.4014	0.4140	0.3991	0.3357	0.3913	0.3946
D	0.2902	0.2651	0.2651	0.2694	0.2694	0.2509	0.2705	0.2521	0.1661	0.1848	0.2458
E	0.2052	0.2251	0.2251	0.2259	0.2259	0.2143	0.2215	0.1902	0.1367	0.1409	0.1913
F	0.4400	0.4793	0.4793	0.4799	0.4799	0.4702	0.4773	0.4601	0.3940	0.4465	0.4632
G	0.4606	0.5308	0.4968	0.5358	0.5208	0.5226	0.5294	0.5192	0.4952	0.5137	0.5100
H	0.3998	0.4322	0.4322	0.4340	0.4340	0.4406	0.4425	0.4208	0.4091	0.4439	0.4203
I	0.2682	0.3070	0.3070	0.3082	0.3082	0.3064	0.3109	0.2885	0.2155	0.2550	0.2860
J	0.4256	0.4803	0.4803	0.4816	0.4816	0.4799	0.4868	0.4686	0.4494	0.4831	0.4649
Weighted mean (by number of customers)	0.3941	0.4386	0.4361	0.4402	0.4390	0.4345	0.4414	0.4230	0.3735	0.4140	0.4218
Relative standard deviation (%)	18.13	18.13	17.71	18.13	17.92	18.59	18.23	20.12	24.91	24.57	20.01

Table 5 summarizes the results for Spearman's rank correlation coefficient. The aim is to get the highest correlation. The best result could be observed for Pareto/NBD (0.4414), with MBG/NBD delivering very similar results (0.4402). It is notable that all models have their weighted mean of the correlation coefficient in a narrow range of 0.37 to 0.44, while the range of all results is 0.22 to 0.54. This result is very close to Chamberlain et al. (2017) that reported Spearman's rank correlation coefficient of 0.46 in the case of a specific CLV model for ASOS.

Table 6 contains the results for the sensitivity of selecting 1% of the most profitable customers. Given the fact that selecting top 1% of customers in the random process would have a probability of success of 1%, all of these results have significantly overperformed such baseline. The best result was observed for the Status quo model, with a sensitivity of 38.38%. The very close performance was demonstrated by MBG/CNBD-k. model with a sensitivity of 38.21%. This rather surprising result is discussed in detail in Section 4. The worst results were observed for Pareto/NBD (Abe) with a sensitivity of 19.74% and the range of results between 4% to 26%.

Table 6. Results for Sensitivity of selecting 1% of the most profitable customers (customer level, in %)

Dataset (online store)	Status Quo	BG/NBD	BG/CNBD-k	MBG/NBD	MBG/CNBD-k	NBD	Pareto/NBD	Pareto/NBD (HB)	Pareto/NBD (Abe)	Pareto/NBD (Abe M2)	Pareto/GGG
A	45.56	51.11	51.11	51.11	51.11	50.00	47.78	51.11	14.44	44.44	51.11
B	10.00	18.97	18.97	18.28	18.28	18.45	18.10	18.28	3.79	7.59	17.07
C	36.68	38.06	37.02	37.37	38.06	34.95	35.99	37.02	14.53	22.15	36.68
D	21.67	24.70	24.70	24.55	24.55	22.73	23.48	24.09	6.52	7.12	23.18
E	11.87	11.87	11.87	12.19	12.19	12.50	13.13	11.25	6.56	6.88	11.56
F	43.48	40.90	40.90	40.87	40.87	40.16	40.53	40.36	19.57	21.43	40.88
G	27.09	28.84	28.84	28.44	29.25	28.44	28.57	27.76	16.98	20.08	27.90
H	34.26	35.56	35.56	35.47	35.47	33.65	35.13	35.82	20.99	21.16	34.95
I	37.26	38.70	38.70	38.70	38.70	36.63	38.38	37.70	17.41	23.17	38.95
J	43.60	43.02	43.02	43.22	43.22	42.24	43.05	42.95	26.05	27.26	42.90
Weighted mean (by number of customers)	38.38	38.17	38.15	38.15	38.21	37.16	37.86	37.74	19.74	22.04	37.89
Relative standard deviation (%)	45.78	43.61	43.60	43.30	43.24	42.46	41.81	44.47	46.13	46.05	44.42

Table 7. Results for Sensitivity of selecting 10% of the most profitable customers (customer level, in %)

Dataset (online store)	Status Quo	BG/NBD	BG/CNBD-k	MBG/NBD	MBG/CNBD-k	NBD	Pareto/NBD	Pareto/NBD (HB)	Pareto/NBD (Abe)	Pareto/NBD (Abe M2)	Pareto/GGG
A	49.51	70.10	70.10	69.71	69.71	71.07	65.24	64.08	26.80	53.98	65.05
B	22.29	30.10	30.10	30.45	30.45	29.93	31.33	30.10	17.38	20.36	29.69
C	45.77	49.12	49.43	48.98	49.33	46.43	48.05	48.19	36.31	43.25	48.22
D	40.73	46.02	46.02	46.18	46.18	41.19	44.75	44.12	23.99	27.18	43.18
E	22.93	27.11	27.11	27.04	27.04	26.26	26.76	25.17	19.28	17.87	25.39
F	47.55	48.69	48.69	48.75	48.75	47.85	48.45	47.81	33.80	38.96	48.04
G	43.97	48.15	48.65	48.02	48.90	47.24	47.55	47.89	38.34	41.73	48.45
H	46.64	49.85	49.85	49.93	49.93	48.41	50.02	49.38	39.15	43.30	49.19
I	35.92	42.88	42.88	43.01	43.01	42.35	43.84	42.19	28.85	33.89	42.01
J	50.61	51.59	51.59	51.61	51.61	50.55	51.73	50.96	38.43	43.65	50.90

Dataset (online store)	Status Quo	BG/NBD	BG/CNBD-k	MBG/NBD	MBG/CNBD-k	NBD	Pareto/NBD	Pareto/NBD (HB)	Pareto/NBD (Abe)	Pareto/NBD (Abe M2)	Pareto/GGG
Weighted mean (by number of customers)	48.16	50.11	50.22	50.11	50.30	49.10	50.07	49.53	37.56	42.31	49.62
Relative standard deviation (%)	33.40	28.73	28.79	28.76	28.87	29.19	28.92	30.50	31.96	36.67	30.39

Table 7 shows the results for the sensitivity of selecting 10% of the most profitable customers. In comparison with results for the sensitivity of the top 1% of customers, as reported in Table 6, the differences are not dramatic. All models have performed well (and have significantly beaten the random selection baseline of 10%). The overall best result was seen with NBD model on dataset A with a sensitivity of 71% (could be compared with the also good result of 50% in case of top 1% customers in Table 6). Exceptions to the good results belong to Pareto/NBD (Abe), and two datasets B and E, where overall results were significantly worse.

Several datasets have experienced poor results across various models: dataset B (sport equipment) had the worst levels of FA (average of 155%), worst levels of MAE (average of 193%), the second lowest Spearman's correlation coefficient (average of 0.22) and the lowest 10% sensitivity records (average of 27%). Dataset E (erotic and health accessories) has performed poorly in top 1% sensitivity (average of 11%) and moving to the top 10% of the most profitable customers have improved the sensitivity metric to only 25% (compared to the overall average of 48% for all datasets).

Conclusions from achieved results answer the research question from the Introduction: *Which of the compared probabilistic CLV models have a good predictive performance of CLV for e-commerce?* The results demonstrate that no single model has outperformed the rest in all selected evaluation criterions, however, almost all probabilistic models have achieved overall good and consistent results on the majority of the datasets both on customer level and for the whole customer base metrics. As two exceptions, Abe's variant of Pareto/NBD have underperformed multiple criterions and would not be fully useful for the studied datasets without further improvements. Finally, MBG/CNBDk has steadily performed well in comparison with other models in most of the statistical metrics and can be considered as balanced and suitable CLV model for online stores.

4. Discussion and implications

The results in Section 3 propose new insights into the performance of the ten compound models and a very simple one, considering ten different datasets with the same minimal

requirements typical for e-commerce. Pareto/NBD (Abe M2) pre-computes covariates from this minimal dataset. Despite different calculation methods, input parameters and use of spending model described in Section 2.1 for all models but Status Quo, have all selected models the same output: CLV calculation, which makes this study a relevant comparative analysis of CLV calculation results.

Using multiple metrics in Section 3 spurs a stimulating discussion on model performance from various perspectives. FA is a high-level key performance indicator for assessment of CLV prediction on the overall customer base. When customer-level prediction quality is important for a company, MAE provides the key information. In case of the datasets studied, MAE indicated a high error rate in customer-level predictions with an average of 102%. However, combined with good results of FA this demonstrates that the effect of customer-level errors is diminishing in customer base level aggregation. The last two metrics (Spearman's rank correlation coefficient and Sensitivity) assume no specific error value of predicted profit but emphasize the ranking of customers despite absolute errors.

4.1. Discussion of the results for each model

Status quo model overestimates profit value of customer base for all online stores (weighted average of 130%, with a maximum of 370%), as seen in Table 3. It had the worst variability and the highest error rate as measured by MAE. One of the few environments in which Status quo prevails is top 1% classification (Table 6). Inherently, all the competing non-trivial models allow for a fairly fresh customer to achieve higher predicted value than an established customer with more orders who may already be lost. If a newcomer completes 2–3 orders in a short period of time ahead of calibration cut-off, with a high purchasing frequency, she can surpass the expectation of a customer with 5 past orders totalling more profit for the company.

The question is whether the very best (top one per cent) customers' churn rate is different from the average churn rate and by how much. Although this has not been investigated in the analysis, if very top customers had records of proven loyalty with lowest churn rates, it would explain why Status quo performs the best in top 1% classification, while not on par when top 10% classification is concerned. The authors argue and call for a separate analysis to prove that the numeric advantage of the Status quo, in this case, is mainly caused by the uniqueness of the segment itself and its natural retention affinity. By being so rare, when designating the very top customers for the future, Status quo wins by drawing from a favourable set, last-period's top customers with a proven record.

Across **Beta Geometric models** (BG/NBD, BG-CNBD-k, MBG/NBD, MBG/CNBD-k) the best performance was found for MBG/CNBD-k that was consistent in a majority of criterions. Models with Fixed regularity had a lower standard deviation of results in terms of FA. For the majority of the datasets, the regularity component hasn't provided any new signal, so the results of MAE and the selection of the top customers were identical with models without the regularity. FA was improved for dataset G (health products and baby care) with a substantial ratio of customers with specific buying patterns of baby products.

Pareto/NBD model performed consistently well both in individual and aggregated predictions. To some extent, bad results for datasets B, D and E according to MAE correlate with

lower repurchase frequency. This finding is supported by the feature of the spending sub-model used in Pareto/NBD model. As discussed in (Jasek et al., 2018): "whenever a customer features a low number of total transactions, the Gamma-Gamma spending submodel predicts their lifetime average order value considering predominantly the rest of the customer base".

Traditional **NBD model** achieved significantly worse performance than Pareto/NBD, mainly based on FA (the largest difference was seen for dataset D with FA of 154% for NBD model and 101% for Pareto/NBD) and individual error rates (MAE of 197% and 154%, respectively, for the same dataset D). No significant lifts in sensitivity and correlation metrics were found.

Hierarchical Bayes Pareto/NBD improved the performance when compared with Pareto/NBD for FA (with more stable variability) and MAE, while the rest of the evaluation metrics were not significantly changed.

Abe's variant of Pareto/NBD had the lowest values in FA metric, especially when covariates were incorporated (decrease in FA from 63% to 34% when using covariates). There was just one exception of dataset A where covariates helped to improve the results, but this in no circumstance could be seen as an improvement. This result contradicts the expectation of all four selected covariates that should have added an important signal about customer transactional behaviour. The fact that results of MAE were ranked the best for Pareto/NBD (Abe M2) does not improve its evaluation given the level of these individual-level errors (92.23% for this model compared to a maximum of 117.08% whereas the ideal value is 0%).

Recency and frequency analysis revealed the existence of customers with frequent purchase behaviour, which can reach better predictive power. The purchase regularity and inter-transaction timing were researched in (Platzer & Reutterer, 2016). Their results indicate that regularity highly improves predictability. Contrary, high regularity can go against the specified attributes of business settings (see Section 2.1). Comparing the results of Pareto/NBD with **Pareto/GGG** that directly incorporates regularity, it can be seen lower FA for the latter (106% for Pareto/NBD and 90% for Pareto/GGG), improved MAE (104% and 93%, respectively), very similar Spearman's rank correlation coefficient (0.44 and 0.42, respectively), and finally very close quality in sensitivity (50% for top 10% of the most profitable customers), what can be summarized as a success for Pareto/GGG and its approach of regularity.

4.2. Managerial implications

The research findings suggest that outputs from CLV models can be used at several levels of detail with satisfying error rates (overall FA of 93%) according to a specific application: aggregated customer base level or individual customer level. Aggregated customer base level of CLV applications is beneficial in business planning and complete strategic management of customer equity. By calculating what value is to be generated in the coming period from the existing customers, it is easy to obtain an estimate how many new customers the company must acquire in order to meet a business plan. Customer equity may also be an important factor when valuing or selling company shares. The most stable results of overall customer base prediction in this research were delivered by Pareto/GGG with 90% average FA, while BG/NBD achieved the best results of FA metric (100%) with the standard deviation as twice as high in comparison to Pareto/GGG.

Individual customer level of CLV has a wide range of applications from a selection of marketing campaigns to customer support preferences. Individual customer scoring expressed by the top percent of the most profitable customers helps to experiment with their exclusion or inclusion in segmented and personalised campaigns according to business goals. Such customers are expected to bring high value to the company. Sensitivity is an important metric for correct classification into top percent of the most profitable customers. When identification of a narrow top-customer group is an issue, one may achieve favourable results using the simplest of models.

Since each of the calculated CLV gained specific benefits from the evaluated metric in Section 3, it is crucial for managers to assess the main outputs of the CLVs with other advantages and shortcomings offered by the model. Selection of models can be narrowed down according to the most relevant dataset, as described in Section 2, and based on the evaluation metrics that conform to the planned use case.

For the most basic use of a selection of a limited number of customers, Status quo model brings consistent results despite its simplicity, naïve approach, high variability of customer scoring, and always overestimated values of customer base profit. Given the calculation approach of Status quo, top customers in the predicted time horizon are predicted mainly based on the recent performance.

Managers should keep in mind that different models or model families are suitable for different business settings. Thus, it is not appropriate to use Pareto/NBD models in contractual settings, see (Schmittlein et al., 1987). Also, the achieved results can be very different for different business settings. For example, there is a big difference in comparison of results for MAE on customer-level with results in (Donkers et al., 2007). The use of CLV models in contractual settings (insurance industry; MAE min 19.4%, max 31.8%) in the case of (Donkers et al., 2007) brings a smaller error than in non-contractual settings presented in this article, see also (Wübben & Wangenheim, 2008). That is all the online store operators should consider choosing the CLV model for their business, as even the best-performing models have a high MAE compared to the models used in the contractual setting.

Conclusions

This study compared the performance of Pareto family models for CLV calculation. The comparison is focused on the area of online shopping and B2C relationship. The main finding is that no single model has outperformed the rest in all selected evaluation criteria. While not obtaining the ideal model for online stores, good and stable results were achieved by BG/NBD and Pareto/NBD models due to their relatively low standard deviations of results and significant lifts from the baseline Status quo model, despite not ranking the best in all the selected criterions. The used datasets were from Central and Eastern European region, so conclusions have limitations connected with cultural specifics of the region. When choosing a CLV model, it is also important for managers to consider individual online store conditions and the advantages and disadvantages of each.

The main limitation of all selected models is their poor performance for seasonal purchase behaviour caused by Christmas season. Previous empirical studies have not discussed

these seasonal effects. A model taking into consideration monthly cohorts or incorporated seasonal effects of time series could gain improved performance. The covariates in Pareto/NBD (Abe M2) could use the variable with the date of acquisition. From such improvement could benefit mostly the online stores with strong Christmas season and companies keen to estimate not only the sum of lifetime value but also the trended data of such predictions.

For future research, the authors would encourage to work on the concerns and limitations raised in this paper. One of the concerns in the model evaluation was the Pareto/NBD (Abe M2) model that incorporated covariates but hasn't demonstrated improved performance in comparison with Pareto/NBD (Abe). Selection and evaluation of individual covariates could be a focus of future research, especially in the context of seasonality and computational requirements and difficulties. Finally, a further opportunity to better capture the underlying customer behaviour could be seen in a combination of different models. Evaluation of Pareto/NBD with Pareto/GGG, introduced in 2016 to incorporate interpurchase timing as a regularity submodel, led to positive results. Evidently, ensemble learning could improve the model in the next stage.

Funding

Author contributions

PJ defined methodology, formulated the research question, obtained datasets from selected online stores, prepared models from R packages BTYD and BTYDplus for CLV calculation and contributed to data interpretation. LV prepared statistical analysis and contributed to data interpretation. LS pre-processed the data and analyzed the results. ZS defined methodology, prepared literature review and contributed to data interpretation. MK contributed to data interpretation. The manuscript was written equally together.

Disclosure statement

The authors declare no conflict of interest.

References

Abdolvand, N., Baradaran, V., & Albadvi, A. (2015). Activity-level as a link between customer retention and consumer lifetime value. *Iranian Journal of Management Studies*, *8*(4), 567-587.

Abe, M. (2009). "Counting your customers" one by one: A hierarchical Bayes extension to the Pareto/NBD model. *Marketing Science*, *28*(3), 541-553. https://doi.org/10.1287/mksc.1090.0502

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering, 17*(6), 734-749. https://doi.org/10.1109/TKDE.2005.99

Batislam, E. M., Denizel, M., & Filiztekin, A. (2007). Empirical validation and comparison of models for customer base analysis. *International Journal of Research in Marketing*, *24*(3), 201-209. https://doi.org/10.1016/j.ijresmar.2006.12.005

Borle, S., Singh, S. S., & Jain, D. C. (2008). Customer lifetime value measurement. *Management Science*, *54*(1), 100-112. https://doi.org/10.1287/mnsc.1070.0746

Centre for Retail Research. (2017). *Online Retailing: Britain, Europe, US and Canada 2017*. Retrieved from http://www.retailresearch.org/onlineretailing.php

Chamberlain, B. P., Cardoso, A., Liu, C. H., Pagliari, R., & Deisenroth, M. P. (2017). Customer lifetime value prediction using embeddings. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1753-1762). Halifax, Canada. https://doi.org/10.1145/3097983.3098123

Chang, W., Chang, C., & Li, Q. (2012). Customer lifetime value: A review. *Social Behavior and Personality, 40*(7), e2243. https://doi.org/10.2224/sbp.2012.40.7.1057

Colombo, R., & Jiang, W. (1999). A stochastic RFM model. *Journal of Interactive Marketing*, *13*(3), 2-12. https://doi.org/10.1002/(SICI)1520-6653(199922)13:3<2::AID-DIR1>3.0.CO;2-H

Damm, R., & Monroy, C. R. (2011). A review of the customer lifetime value as a customer profitability measure in the context of customer relationship management. *Intangible Capital, 7*(2), 261-279. https://doi.org/10.3926/ic.2011.v7n2.p261-279

Donkers, B., Verhoef, P. C., & De Jong, M. G. (2007). Modeling CLV: A test of competing models in the insurance industry. *Quantitative Marketing and Economics, 5*(2), 163-190. https://doi.org/10.1007/s11129-006-9016-y

Dresch, A., Lacerda, D. P., & Antunes Jr, J. A. V. (2015). *Design science research: A method for science and technology advancement*. Springer. https://doi.org/10.1007/978-3-319-07374-3

Dziurzynski, L., Wadsworth, E., & McCarthy, D. (2015). *BTYD: Implementing buy'til you die models* (R package version, 2.4). Retrieved from https://cran.r-project.org/web/packages/BTYD/BTYD.pdf

Ecommerce Europe. (2016). *European B2C E-commerce Report 2016*. Ecommerce Foundation, Brussels.

Ehrenberg, A. S. (1959). The pattern of consumer purchases. *Applied Statistics, 8*(1), 26-41. https://doi.org/10.2307/2985810

Estrella-Ramón, A. M., Sánchez-Pérez, M., Swinnen, G., & VanHoof, K. (2013). A marketing view of the customer value: Customer lifetime value and customer equity. *South African Journal of Business Management, 44*(4), 47-64. https://doi.org/10.4102/sajbm.v44i4.168

Fader, P. S. (2012). *Customer centricity: Focus on the right customers for strategic advantage*. Wharton Digital Press.

Fader, P. S., & Hardie, B. G. S. (2001). Forecasting repeat sales at CDNOW: A case study. *Interfaces, 31*(3), S94-S107. https://doi.org/10.1287/inte.31.3s.94.9683

Fader, P. S., & Hardie, B. G. S. (2009). Probability models for customer-base analysis. *Journal of Interactive Marketing*, *23*(1), 61-69. https://doi.org/10.1016/j.intmar.2008.11.003

Fader, P. S., & Hardie, B. G. S. (2013). *The Gamma-Gamma model of monetary value*. Retrieved from http://www.brucehardie.com/notes/025/gamma_gamma.pdf

Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2005a). Counting your customers the easy way: An alternative to the Pareto/NBD Modelmodel. *Marketing Science, 24*(2), 275-284. https://doi.org/10.1287/mksc.1040.0098

Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2005b). RFM and CLV: Using iso-value curves for customer base analysis. *Journal of Marketing Research, 42*(4), 415-430. https://doi.org/10.1509/jmkr.2005.42.4.415

Gupta, S. (2009). Customer-based valuation. *Journal of Interactive Marketing*, *23*(2), 169-178. https://doi.org/10.1016/j.intmar.2009.02.006

Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N., Ravishanker, N., & Sriram, S. (2006). Modeling customer lifetime value. *Journal of Service Research*, *9*(2), 139-155. https://doi.org/10.1177/1094670506293810

Haenlein, M. (2017). How to date your clients in the 21st century: Challenges in managing customer relationships in today's world. *Business Horizons, 60*(5), 577-586. https://doi.org/10.1016/j.bushor.2017.06.002

Haenlein, M., Kaplan, A. M., & Schoder, D. (2006). Valuing the real option of abandoning unprofitable customers when calculating customer lifetime value. *Journal of Marketing, 70*(3), 5-20. https://doi.org/10.1509/jmkg.70.3.5

Hubka, V., & Eder, W. E. (1996). *Design science: Introduction to the needs, scope and organization of engineering design knowledge*. Springer. https://doi.org/10.1007/978-1-4471-3091-8

Hwang, H., Jung, T., & Suh, E. (2004). An LTV model and customer segmentation based on customer value: A case study on the wireless telecommunication industry. *Expert Systems with Applications*, *26*(2), 181-188. https://doi.org/10.1016/S0957-4174(03)00133-7

Jasek, P., Vrana, L., Sperkova, L., Smutny, Z., & Kobulsky, M. (2018). Modeling and application of customer lifetime value in online retail. *Informatics*, *5*(1), 2. https://doi.org/10.3390/informatics5010002

Jasek, P., Vrana, L., Sperkova, L., Smutny, Z., & Kobulsky, M. (2019). *Predictive performance of customer lifetime value models in ecommerce and the use of non-financial data.* Prague Economic Papers (Accepted paper). https://doi.org/10.18267/j.pep.714

Kim, S-Y., Jung, T.-S., Suh, E.-H., & Hwang, H.-S. (2006). Customer segmentation and strategy development based on customer lifetime value: A case study. *Expert Systems with Applications*, *31*(1), 101-107. https://doi.org/10.1016/j.eswa.2005.09.004

Knox, G., & van Oest, R. (2014). Customer complaints and recovery effectiveness: A customer base approach. *Journal of Marketing, 78*(5), 42-57. https://doi.org/10.1509/jm.12.0317

Kotler, P., & Keller, K. L. (2015). *Marketing management* (15th ed). New Jersey: Prentice Hall.

Kumar, V., & Pansari, A. (2016). National culture, economy, and customer lifetime value: Assessing the relative impact of the drivers of customer lifetime value for a global retailer. *Journal of International Marketing, 24*(1), 1-21. https://doi.org/10.1509/jim.15.0112

Kumar, V., Pozza, I. D., Petersen, J. A., & Denish Shah, D. (2009). Reversing the logic: The path to profitability through relationship marketing. *Journal of Interactive Marketing*, *23*(2), 147-156. https://doi.org/10.1016/j.intmar.2009.02.003

Ma, S. H., & Liu, J. L. (2007). The MCMC approach for solving the Pareto/NBD model and possible extensions. In *Proceedings of the 3rd International Conference on Natural Computation* (pp. 505-512). Haikou, China. https://doi.org/10.1109/ICNC.2007.728

Nenonen, S., & Storbacka, K. (2016). Driving shareholder value with customer asset management: Moving beyond customer lifetime value. *Industrial Marketing Management*, *52*, 140-150. https://doi.org/10.1016/j.indmarman.2015.05.019

Óskarsdóttir, M., Baesens, B., & Vanthienen, J. (2018). Profit-based model selection for customer retention using individual customer lifetime values. *Big Data*, *6*(1), 53-65. https://doi.org/10.1089/big.2018.0015

Qi, J.-Y., Qu, Q.-X., Zhou, Y.-P., & Li, L. (2015). The impact of users' characteristics on customer lifetime value raising: evidence from mobile data service in China. *Information Technology and Management*, *16*(4), 273-290. https://doi.org/10.1007/s10799-014-0200-6

Pfeifer, P. E., Haskins, M. E., & Conroy, R. M. (2005). Customer life time value, customer profitability, and the treatment of acquisition spending. *Journal of Managerial Issues*, *17*(1), 11-25.

Platzer, M. (2016). *Customer base analysis with BTYDplus*. Retrieved from https://cran.r-project.org/web/packages/BTYDplus/vignettes/BTYDplus-HowTo.pdf

Platzer, M., & Reutterer, T. (2016). Ticking away the moments: Timing regularity helps to better predict customer activity. *Marketing Science*, *35*(5), 779-799. https://doi.org/10.1287/mksc.2015.0963

Reinartz, W. J., & Kumar, V. (2000). On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing. *Journal of Marketing*, *64*(4), 17-35. https://doi.org/10.1509/jmkg.64.4.17.18077

Schmittlein, D. C., & Peterson, R. A. (1994). Customer base analysis: An industrial purchase process application. *Marketing Science*, *13*(1), 41-67. https://doi.org/10.1287/mksc.13.1.41

Schmittlein, D. C., Morrison, D. G., & Colombo, R. (1987). Counting your customers – who are they and what will they do next. *Management Science, 33*(1), 1-24. https://doi.org/10.1287/mnsc.33.1.1

Singh, S. S., & Jain, D. C. (2010). Measuring customer lifetime value. In K. M. Naresh (Ed.), *Review of marketing research,* 6, 37-62. Emerald Group Publishing Limited, Bingley. https://doi.org/10.1108/S1548-6435(2009)0000006006

Singh, S. S., & Jain, D. C. (2013). *Measuring customer lifetime value: Models and analysis.* INSEAD Working Paper No. 2013/27/MKT. https://doi.org/10.2139/ssrn.2214860

Statista. (2018). *Retail e-commerce sales worldwide from 2014 to 2021*. Retrieved from https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/

Vojtko, V. (2014). Rethinking the concept of just noticeable difference in online marketing. *Acta Informatica Pragensia*, *3*(2), 204-218. https://doi.org/10.18267/j.aip.49

Wieringa, R. J. (2014). *Design science methodology for information systems and software engineering.* Berlin: Springer. https://doi.org/10.1007/978-3-662-43839-8

Williams, C., & Williams, R. (2015). Optimizing acquisition and retention spending to maximize market share. *Journal of Marketing Analytics*, *3*(3), 159-170. https://doi.org/10.1057/jma.2015.11

Wübben, M., & Wangenheim, F. (2008). Instant customer base analysis: Managerial heuristics often "Get It Right". *Journal of Marketing*, *72*(3), 82-93. https://doi.org/10.1509/jmkg.72.3.82
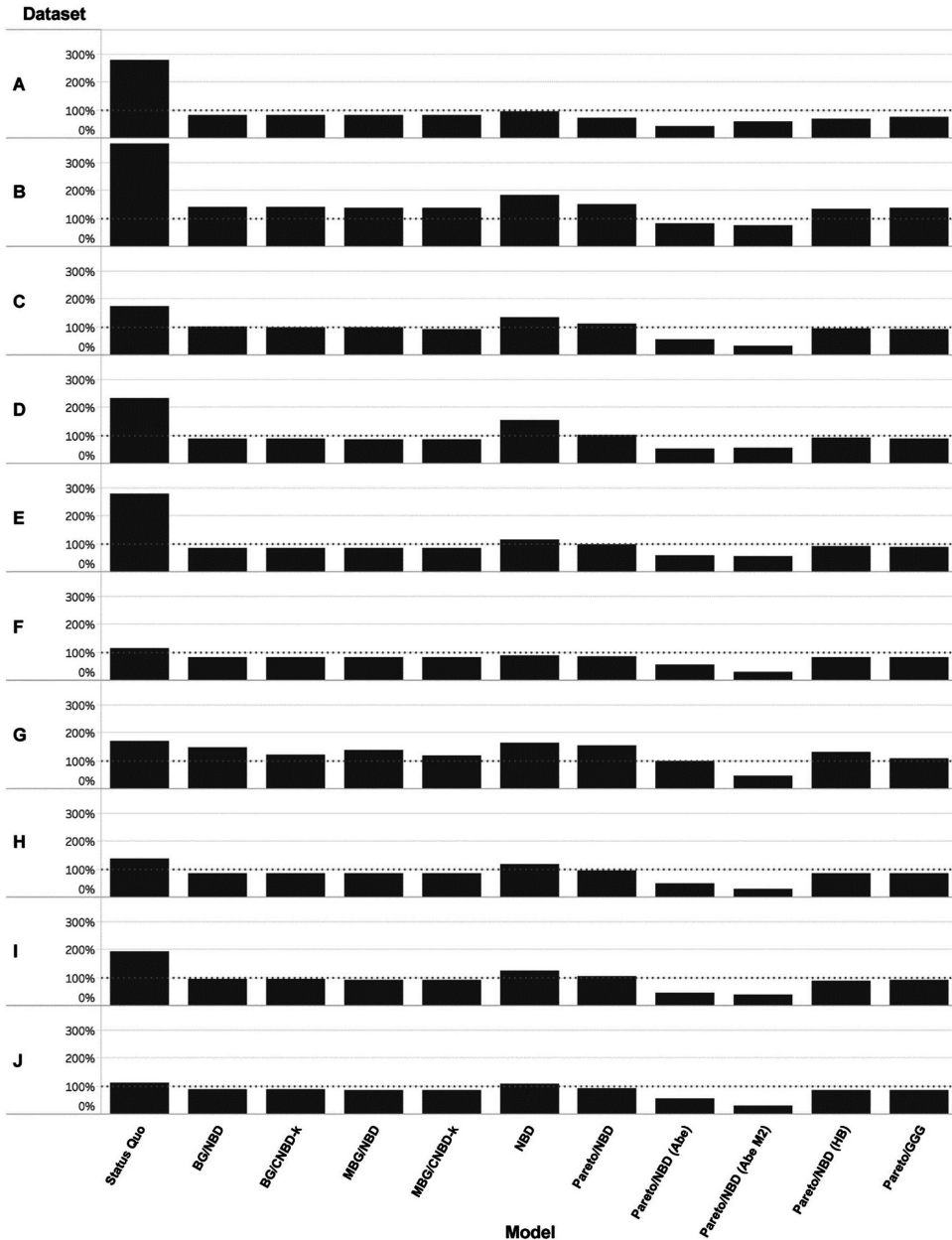
APPENDIX

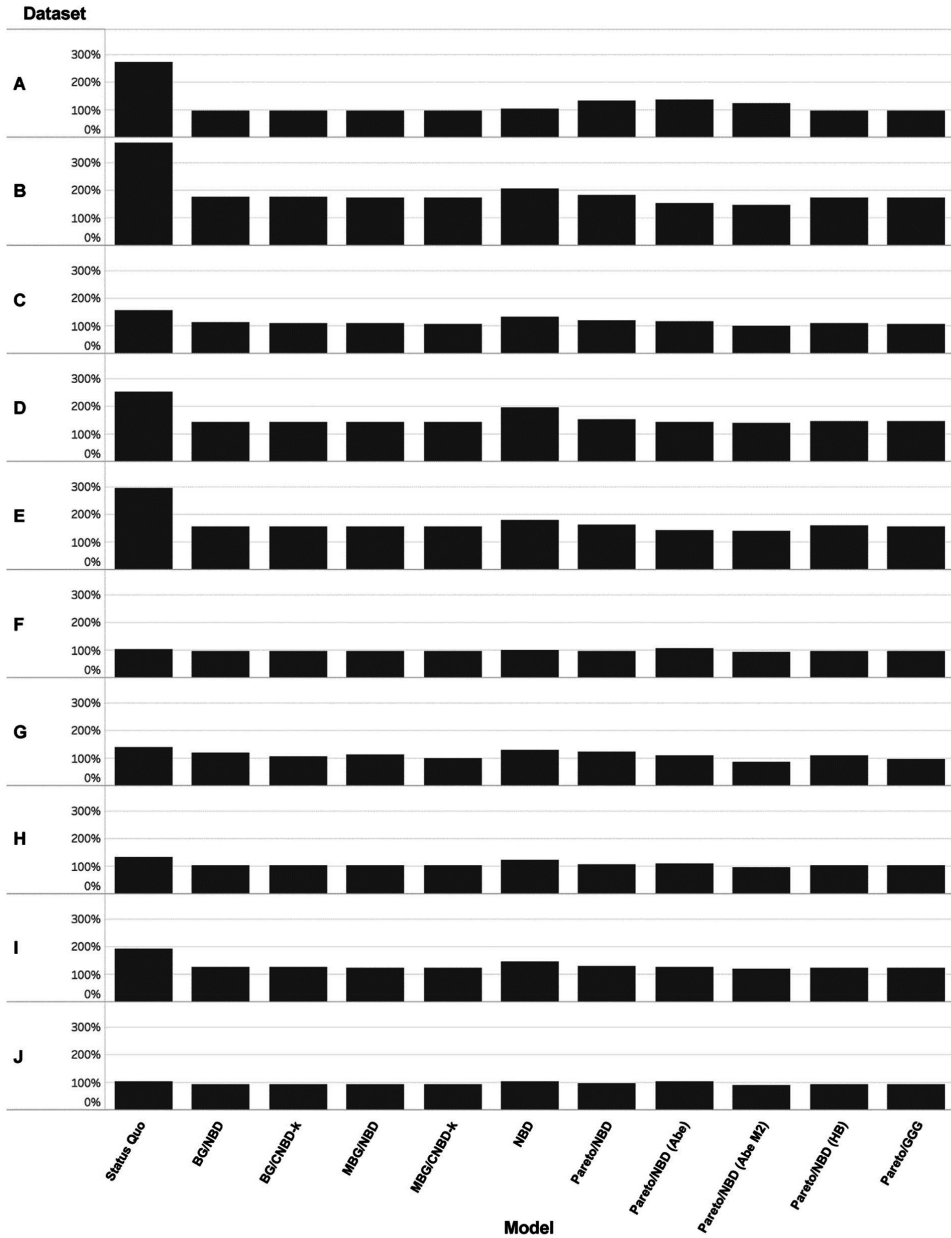Figure 2. Results for Forecast vs. Actual (in %) for online stores A-J

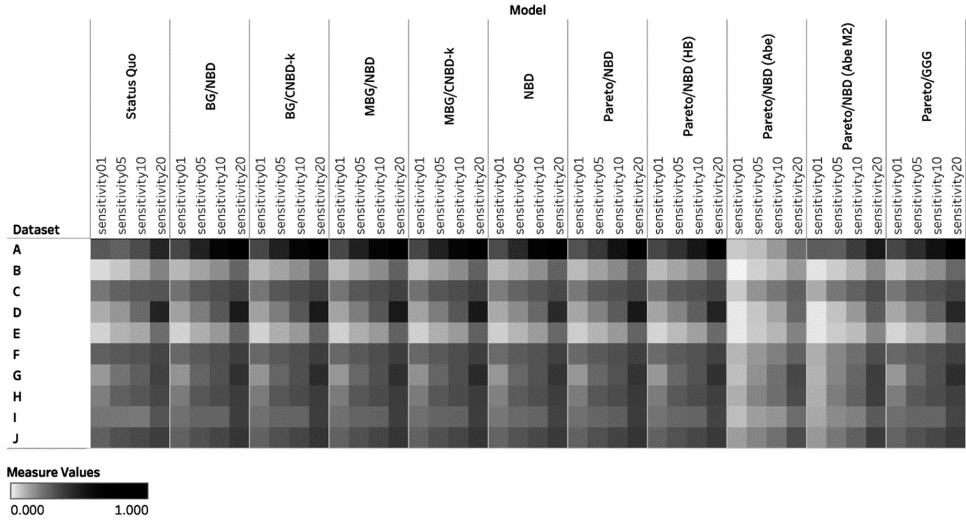Figure 3. Results for MAE (customer-level, in %) for online stores A-J

Figure 4. Results for Sensitivity (customer-level, in %) for online stores A-J