Bus Inf Syst Eng 60(2):151–166 (2018)
https://doi.org/10.1007/s12599-017-0485-1

RESEARCH PAPER

Next-Purchase Prediction Using Projections of Discounted Purchasing Sequences

Katerina Shapoval · Thomas Setzer

Received: 27 May 2016 / Accepted: 23 February 2017 / Published online: 27 June 2017

Abstract A primary task of customer relationship management (CRM) is the transformation of customer data into business value related to customer binding and development, for instance, by offering additional products that meet customers' needs. A customer's purchasing history (or sequence) is a promising feature to better anticipate customer needs, such as the next purchase intention. To operationalize this feature, sequences need to be aggregated before applying supervised prediction. That is because numerous sequences might exist with little support (number of observations) per unique sequence, discouraging inferences from past observations at the individual sequence level. In this paper the authors propose mechanisms to aggregate sequences to generalized purchasing types. The mechanisms group sequences according to their similarity but allow for giving higher weights to more recent purchases. The observed conversion rate per purchasing type can then be used to predict a customer's probability of a next purchase and target the customers most prone to purchasing a particular product. The bias–variance trade-off when applying the models to target customers with respect to the lift criterion are discussed. The mechanisms are tested on empirical data in the realm of cross-selling campaigns. Results show that the expected bias–variance behavior well predicts the lift achieved with the mechanisms. Results also show a superior performance of the proposed methods compared to commonly used segmentation-based approaches, different similarity measures, and popular class predictors. While the authors tested the approaches for CRM campaigns, their parameterization can be adjusted to operationalize sequential features of high cardinality also in other domains or business functions.

Keywords Customer relationship management · Campaign management · Feature generation · Purchasing sequence · Next purchase prediction

Accepted after two revisions by Prof. Dr. Suhl.

K. Shapoval (✉) · Prof. Dr. T. Setzer
Karlsruhe Institute of Technology (KIT), Institute of Information Systems and Marketing (IISM), Information and Market Engineering (IM), Fritz-Erler-Straße 23, 76133 Karlsruhe, Germany
e-mail: katerina.shapoval@kit.edu

Prof. Dr. T. Setzer
e-mail: thomas.setzer@kit.edu

1 Introduction

A primary aim of customer relationship management (CRM) is to improve customer understanding based on data available in enterprise information systems. One of such tasks is the selection or targeting of customers for a product-selling campaign that are likely to need (and buy) a specific product. This is often referred to as providing the next-best offer.

By now, a customer is typically modeled by so-called *RFM* attributes related to *R*ecency, *F*requency, and *M*onetary value of purchases [see, among other recent applications, Daoud et al. (2015)]. More precisely, a customer is characterized by the number of days since the last purchase, the total number of purchases, and the total monetary value of the purchases. The respective values are then used as features for customer segmentation as well as for predicting his future behavior. Several approaches successfully use additional features such as the customer lifetime value or particular products purchased in the past [see, amongst others, Chan (2008) and Khajvand and

Tarokh (2011)]. A detailed survey on the process and data used within the marketing context is provided in Bose and Chen (2009).

There is also evidence that the temporal relation between purchases can provide additional predictive value in cases where logical orders of purchases exist. For instance, typical paths of purchases have been determined for financial products (Li et al. 2005), home appliances (Prinzie and Van den Poel 2007), or in recommender systems, e.g., to suggest articles to read (Sahoo et al. 2012). These papers will be discussed in Sect. 2 and we will also benchmark the models we introduce against the proposed approaches in our evaluation.

The general problem is that in larger settings the integration of purchasing histories on a fine-grained level in predictive models, e.g., logistic regression, is challenging. That is because a purchasing sequence can easily become a high-cardinality feature– a non-quantitative feature with many categories but little support (few observations) per category – prohibiting reliable inference from individual sequences. As a consequence, in today's practice such data is typically used only in a limited fashion by considering features such as the number of a customer's purchases (Van den Poel and Buckinx 2005), possession of products in certain categories (Back et al. 2011), or customer value (Han et al. 2012).

The problem is that even for moderate levels of products and the number of past purchases taken into account raw purchasing vectors cannot be operationalized directly within a supervised prediction model when only few observations per sequence exist. This is illustrated in Table 1. The table shows an excerpt of empirical purchasing sequences observed in the customer data base of a leading multinational telecommunications company.[1] Each row corresponds to one purchasing sequence (the last five purchases) of products from ten product categories P offered by the company (P_1, ..., P_{10}). For instance, the first row displays the sequence $\langle P_1, P_1, P_3, P_1, P_4 \rangle$ (ID 168) with a purchase of a product P_1 as the most recent one. Rows are sorted by the observed conversion rate (CR) calculated as the percentage of customers exhibiting a certain sequence that then bought a particular product of interest, in this example P_2. ID displays the relative position in the set, *Support* quantifies the number of observations per unique sequence. A naive approach to select the customers that are most likely to buy a product would assign new customers with the CR of the corresponding sequence, then taking those with the highest CR until a predefined number of customers (campaign size) is reached. The quality of a selection is then evaluated on customer level based on the quotient of actual CR and baseline purchasing probability (the so-called lift criterion that we will formally define in Sect. 3.3).

Table 1 Excerpt of our available data

ID	Last	2nd L	3rd L	4th L	5th L	Support	CR (%)
...	...	...	...	...	...	...	...
168	P1	P1	P3	P1	P4	4	75
176	P1	P1	P5	P1	P3	2	50
...	...	...	...	...	...	...	...
380	P9	P4	P4	P4	P4	23	22
436	P9	P4	P1			56	18
...	...	...	...	...	...	...	...
656	P4	P4	P1			2956	9
...	...	...	...	...	...	...	...

Rows are sorted by conversion rate (CR) per sequence– the percentage of target product purchases following a particular sequence. ID displays the relative position in the set sorted by CR. The baseline purchasing frequency (CR with random customer selection) is 8.8%. Sequences with high CR tend to have little support, while frequent sequences are close to the baseline

Unfortunately, CR estimates with top-ranked sequences, such as sequence 168 or 176, are not only unstable because of the low support, but CR will be systematically overestimated and customer targeting will be biased. That is because with small segments the randomness in observed customer samples will heavily impact CR and therefore the segments' relative ranks, resulting in a systematic overestimation of the top-segments' CR and, consequently, biased targeting. Unfortunately, sequences with higher support, e.g., sequence 656, approach the baseline CR of 8.8% and are of little value as targeting this segment would improve only slightly over a random selection.

Obviously, a supervised aggregating of sequences with high but individually overestimated CR would be a spurious solution due to instability and selection bias. Unsupervised techniques to aggregate sequences to more generalized sequence types, applied prior to any supervised approach, might provide a solution to this problem. The aggregation must, however, find sequence types/segments large enough (for reasons of statistical support) to draw inference with respect to CR, while the discriminatory power of the resulting segments regarding a next purchase should be as high as possible.

In this paper, we make the following contributions to the literature. First, we propose two unsupervised mechanisms to reduce large amounts of purchasing sequences to

[1] The company operates on the US and European markets of telecommunication services. The company achieved annual revenues in the double-digit billion Euro range. The product portfolio ranges from basic starter-products like Internet domains, to various hosting solutions up to professional server solutions for large-scale businesses, mobile telephony, as well as access products such as digital subscriber lines.

generalized and interpretable purchasing types. The latter aspect is of high importance to ensure acceptance and provide further valuable customer insights from the segmentation. Both methods are based on the intuition that (1) there is predictive value in sequential purchasing orders [as, for instance, found in Miguéis et al. (2012)], and (2) the predictive value of information increases with its recency [as, for instance, found in Dunlavy et al. (2011)]. With both proposed mechanisms, we discount the weights of past purchases to address the trade-off between considering the most recent purchases and the whole sequential information.

Second, we propose and discuss the bias–variance estimates of (unachieved) lift when targeting customers according to the CR of their respective segments in descending order until a predefined campaign size is reached. To our knowledge, this has not been studied in the context of customer targeting.

Third, we present empirical results with the approaches based on over 200 thousand purchasing sequences. Results show a superior performance of the proposed methods compared to commonly used segmentation-based approaches such as Hidden Markov Models or Self-Organizing Maps, different similarity measures such as the Levenshtein distance and popular class predictors like Logistic Regression. Furthermore, in contrast to classifiers and common segmentation models, well-interpretable customer segments are derived with the mechanism introduced. Additionally, results show that the expected bias–variance behavior well predicts the lift achieved out-of-sample.

This article is structured as follows. Section 2 outlines prior work on sequence aggregation. Section 3 introduces our models for projecting and aggregating purchasing sequences to determine target customers. Also, estimates for bias–variance trade-off for lift are provided. In Sect. 4 we introduce the available data and the evaluation design, and present empirical outcomes in Sect. 5. Finally, in Sect. 6 we conclude and discuss the impact of our work on information systems and management together with its limitations.

2 Related work on sequence aggregation

Several streams of research exist to efficiently search and identify frequent sequence patterns (Mooney and Roddick 2013). However, these approaches identify the most frequent patterns but do not group sequences together, and thus do not address our problem of sequence generalization.

With respect to sequence aggregation approaches for next purchase prediction, we identified two general types of approaches. The first group considers purchasing histories and applies data mining methods belonging to the field of recommender systems [for a review, see Park et al. (2012)], such as Association Rules (Wong et al. 2005) and different types of matrix factorizations, partly including temporal weighting (Dunlavy et al. 2011).

As a second group we see data mining methods for classification such as Logistic Regression, or Hidden Markov Models (see Ngai et al. 2009 for an extensive survey), incorporating sequence features in their potentially broader features sets to derive a prediction model. As mentioned before, given the exponential growth of the number of potential purchasing sequences, utilizing 'sequence' as a feature in such models requires aggregation as already 100 categories are stated as impracticable within a prediction model (Moeyersoms and Martens 2015).

Approaches have been proposed to transform high-cardinality features into attributes with lower dimensionality, which can be categorized into proximity-based, feature-based and model-based approaches (Bicego et al. 2003).

Proximity-based approaches are widely used in business intelligence and CRM. They utilize a distance (similarity) measure for sequences that allows for clustering customers such that within a segment customers are rather homogeneous and purchasing types can be identified for further marketing endeavors. Several studies, especially in the marketing context, have shown that clustering based on the Levenshtein distance metric (Levenshtein 1966)[2] leads to a superior performance compared to other distance measures such as Euclidean distance, in particular when using Ward clustering afterwards. For instance, this approach was applied for clustering sequences of store visits (Joh et al. 2003) or to compute the dissimilarity of customer contact sequences, which are then aggregated to clusters representing their 'typical' behavior (Steinmann and Silberer 2010). Levenshtein distance and comparable measures consider similarity but do not consider the position in a purchasing vector where a particular editing operation occurs. In contrast, our approaches will consider the diminishing importance of elder purchases and weight the purchasing order respectively.

Feature-based methods to sequence clustering aim at finding and constructing features from sequence information to represent sequences more concisely. Some approaches consider the number of categories of purchased products or the last purchasing category as a binary feature (e.g., Li et al. 2005). For instance, Moon and Russell (2008) propose to encode customer product portfolios so that the typical product combinations can be retrieved and used for prediction. Although a major part of portfolio

[2] Such approaches are often used within a broader category of methods such as the Sequence Alignment Method (SAM; Kruskal 1983).

information is considered with their approach, the temporal order of purchases is not incorporated. Research by Miguéis et al. (2012) has successfully used dummy-coded aggregation of sequences, but neither a systematical framework nor a theoretical rationale was provided. Overall, these approaches consider sequence similarity but do not explicitly account for a decrease in importance of previous purchases when deriving generalized sequence types.

The third stream of research, model-based approaches, employ a set of models corresponding to sequences or sequence segments learned. Besides application of Markov Models, e.g., by Prinzie and Van den Poel (2007), where sequence complexity is reduced to a transition matrix, Hidden Markov Models are used for prediction in Sahoo et al. (2012), and Self-Organizing Maps (SOM) have been used to segment portfolio information (Kohonen 2001). Cho et al. (2005) uses a dummy-coded sequence (one variable per time period with $n-1$ levels for n products) that is then clustered using SOM. These methods do, however, lack interpretability of the resulting segments. For instance, SOM display the relationships between variables only implicitly (Kaski et al. 1998) and latent states of HMM are mostly post-profiled with external data aimed at interpreting the result (Shirley et al. 2010).

In summary, several techniques exist to cluster portfolios and sequences based on their similarity and use the resulting clusters for response prediction later on. The methods we propose will differ from existing approaches in following aspects. First, we assign a higher weight to more recent purchases by explicitly modeling a diminishing importance of elder subsequences. The motivation stems from several studies in various industries given in the introduction that have shown that not only a certain logical order of purchases exists, but that the categories of the last purchases had the strongest predictive value. Second, a drawback of many existing techniques is the missing interpretability of the resulting clusters (e.g., Kaski et al. 1998). Clusters with our models can be easily visualized and intuitively linked to specific types of purchasing behavior. Third, in contrast to existing approaches, we propose a bias–variance estimation of the (unachieved) lift when targeting customers based on the CR-ranked segments. Our empirical analysis shows that the bias–variance components derive sound estimates of the out-of-sample performance for our mechanisms and serve as a guideline for the level of aggregation.

3 Sequence Similarity Considering Recency

In line with findings in other industries (e.g., Li et al. 2005), tests on our telecommunications dataset indicate a decreasing predictive value of older purchases. Table 2 shows the CR when considering different combinations of the last three purchases for predicting the purchase probability of P_2 as target using the observed CR in the training data.

Table 2 Conversion rate (CR) with the best 1 and 10% of customers with a particular segment building type corresponding to a row (sorted by CR)

Segmentation feature	CR Best 1%	CR Best 10%
Combination of the three last purchases	28.4%	14.6%
Combination of the two last purchases	24.3%	13.8%
Last purchase	18.4%	12.9%
Second and third last purchase	15.7%	11.1%
Second last purchase	13.6%	10.6%
Third last purchase	12.8%	9.6%

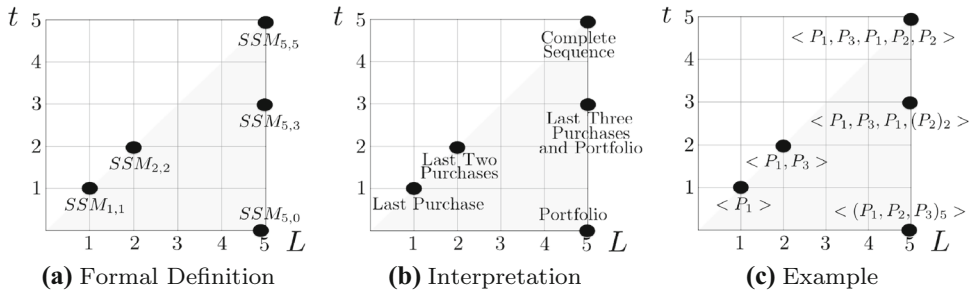
Considering the last purchase only (third row) outperforms elder purchases (rows four, five and six). Considering the last purchase and previous purchases improves in-sample CR (rows one and two). The baseline purchasing frequency (CR with random selection) is 8.8%

As expected, the segmentation using the last purchase only (row three) outperforms any other combination of elder purchases with respect to the resulting observed CR (rows four to six). Considering the last purchase together with elder purchases further improves the CR, but the difference to last purchase declines successively when adding purchases further in the past (rows one and two). Hence, the predictive value of a purchasing sequence exhibits a temporal structure: the combination of all three purchases is important, but the more recent the purchase, the higher the individual predictive value of a purchase. The goal of the mechanisms we propose is to focus on recent purchases while still capturing the predictive value of longer purchasing sequences.

3.1 Sequence-Set Model

Before we develop the Sequence-Set-Model (SSM), we will first introduce some notation. Let $\{P_0, \{P_i\}\}$ be a set of I offered products or services $i \in \{1, \ldots, I\}$ and an artificial product or service P_0, later serving as a dummy product in case no product has been purchased. Let further $S_{cL} = \langle s_{c1}, \ldots, s_{cl}, \ldots, s_{cL} \rangle$ denote a sequence S of products s_{cl} purchased by a customer c of length L, with $c \in Z^+$, $L \in Z^+$, and $l \in Z^+ \leq L$, where the first element s_{c1} contains the most recent product purchased by customer c. The last element in the sequence vector, s_{cL}, indicates the L-th to last product purchased, i.e., the most ancient product purchase stored in the vector.

Considering the dummy-product P_0, for each customer we can assign a purchasing sequence S_{cL} of length L by



(a) Formal Definition **(b)** Interpretation **(c)** Example

Fig. 1 Definition of *SSM* with different parameterizations (*left-hand side*), interpretation of models parameter combinations (*center*), and example instantiations of the models with specific parameterizations (*right-hand side*). $SSM_{1,1}$ groups sequences according to the product purchased most recently. $SSM_{5,3}$ groups sequences with identical vectors up to the third purchase, and the same portfolio of two products purchased earlier. $SSM_{5,5}$ indicates a grouping where the last five product purchases are identical

filling-up the empty fields in the vectors from right to left with P_0 if less then L products have been purchased. For instance, assuming $L = 5$ and a customer c purchased the products P_5 and P_1 only, with P_5 being his last and second to last product purchase, than his respective purchasing sequence would be $\langle P_5, P_5, P_1, P_0, P_0 \rangle$. If a sequence has more than L elements, prior purchases are ignored.

An element in a sequence can also contain an unordered itemset $n\text{−set} = (\{\mathrm{P_i}\})_n, \quad j \in \{0, \ldots, I\}, \quad \text{with} \quad n \in Z_0^+$, where n indicates the number of products to be selected from the set, potentially including multiples of the same product. As an example, consider the definition of a sequence type $\langle P_5, (P_1, P_2)_2 \rangle$. The following sequences belong to this type: $\langle P_5, P_1, P_2 \rangle$, or $\langle P_5, P_2, P_1 \rangle$, while $\langle P_5, P_2, P_3 \rangle$ is not of the defined sequence type.

With the two parameters L, indicating the total length of a sequence and a threshold parameter $t \in Z_0^+$, $t \leq L$, we now define SSM as shown in Eq. (1).

$$\begin{aligned} SSM_{L,t} &= \langle S_{c,t}, n\text{−set} \rangle \\ n &= L - t \end{aligned} \tag{1}$$

Parameter t indicates the number of the most recent purchases where the models demand a mandatory order, while for the remaining $L - t$ product purchases no order is required, but purchases stem from the n−set (portfolio). The relationship among the different types of models is illustrated in Fig. 1. The parameter space is formed by all integers in the gray-shaded area.

As aforementioned, the length of the purchasing sequence considered by the model increases with L. By increasing t, the strict sequential order of the t most recent purchases is considered in the model. Hence, SSM relaxes the sequential order constraint of elder purchases to obtain a less restrictive grouping of sequences while still capturing sequence information to a predefined degree by setting parameters L and t. By setting $t = 0$ the purchasing history is considered as a portfolio ignoring the order of purchases. Setting $t = L$ the model specifies an exact sequence of the last L purchases. As an example, a model with $L = t = 2$ would group together the sequences $\langle P_1, P_2, P_4 \rangle$ and $\langle P_1, P_2, P_6 \rangle$, while a model with $L = t = 3$ would derive two different segments from the sequences.

As appropriate parameters for L and t depend highly on the data at hand, Sect. 3.3 proposes a statistical framework of bias–variance trade-off, which gives a general recommendation on search of the parameters based on training data. *SSM* can be used for segmentation as well as for the prediction of conversion rates. For prediction, the segments and the corresponding CR are estimated from the training data, e.g., for $SSM_{1,1}$ as the last purchase. For the test data, a customer with the respective last purchase is assigned the corresponding CR. The customer segments are then sorted in the descending order to be selected up to a certain campaign size. If the segment did not appear in the training data, the segment is assigned the baseline value.

3.2 Weighted-Productspace Clustering

The second approach we propose is a technique we will refer to as Weighted-Productspace Clustering (WPC). With WPC, for each customer we assign geometrically descending weights to the elements in the corresponding purchasing sequence vector S_{cL}, thereby weighting more ancient product purchases less than more recent ones.[3] With parameter $\lambda \in R_0^+ \leq 1$ indicating the discount per purchase in a vector, the weight of a product at position l in S_{cL} is calculated using Eq. (2).

$$w_l = \lambda^{l-1} \tag{2}$$

Total weight of a product P_i is then determined as the sum of weights associated with the positions of P_i-occurrences in the purchasing sequence. With b_{icl} being a binary variable indicating whether P_i is at the l-th position in S_{cL} of c,

[3] Geometrically descending weights are widely-used techniques to model a discounted importance of observations, such as in time series forecasting (Brown 2004).

we compute the total weight W_{ic} of P_i in S_{cL} as shown in Eq. (3).

$$W_{ic} = \sum_{l=1}^{L} b_{icl} w_l, b_{icl} = \begin{cases} 1 & \text{if } s_{cl} = P_i \\ 0 & \text{else} \end{cases} \quad (3)$$

The resulting vector of product weights $D_{cL} = (W_{1c}, \ldots, W_{ic}, \ldots, W_{Ic})$ for customer c is then positioned into product space– the I-dimensional space spanned by the offered products– where for each vector W_{ic} serves as coordinate along the dimension associated with product i. Thereby, a discrete purchasing sequence is projected onto a quasi-continuous product-space.

Figure 2 illustrates how WPC works. Consider a product portfolio of only three products P_1, P_2, and P_3, and the three product purchasing sequences $S_{1,3} = \langle P_2, P_2, P_2 \rangle$, $S_{2,3} = \langle P_3, P_2, P_1 \rangle$, and $S_{3,3} = \langle P_3, P_2, P_2 \rangle$. With $\lambda = 0.5$ and $L = 3$, these sequences are represented as points in product space as shown in the figure. In addition, the plot shows all potential locations of data points (the smaller points) with $\lambda = 0.5$ and $L = 3$.

WPC reflects the sequential order of purchases implicitly, as the coordinate along the i-th dimension is higher if P_i is the product purchased more recently, and decreases with the order in a sequence. If a customer has not bought P_i in his L last purchases, the coordinate for the corresponding product is zero. The decreasing importance of products purchased further in the past depends on λ, taking values in $[0, 1]$. $\lambda = 1$ sets the coordinates of all purchased products to one; hence, it does not project sequential but only portfolio information. With decreasing λ, coordinates increase less for older purchases so that sequential information is captured in the respective data point. As λ approaches zero, the coordinates for elder purchases also converge to zero.

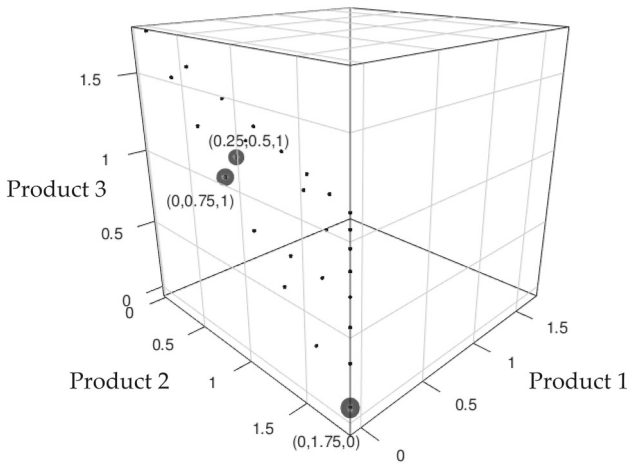


Fig. 2 All possible sequence projections consisting of three products with $L = 3$ and $\lambda = 0.5$. *Labeled points* represent the sequence $S_{1,3} = \langle P_2, P_2, P_2 \rangle$ with $D_{1,3} = (0, 1.75, 0)$, $S_{2,3} = \langle P_3, P_2, P_1 \rangle$ corresponding to $D_{2,3} = (0.25, 0.5, 1)$ as well as $S_{3,3} = \langle P_3, P_2, P_2 \rangle$ corresponding to $D_{1,3} = (0, 0.75, 1)$

In WPC, distance-based clustering is then applied to segment data points D_{cL} in product space. Hence, clusters represent similar sequences with a diminishing importance of older product purchases, adjustable by λ. In this work WPC is implemented with k-means clustering, a widely used distance-based clustering method (MacQueen et al. 1967), as the temporal information is reflected in the distance between data points. In addition, k-means has also the advantage that segments will follow the empirical density of data as it will tend to locate cluster centers where many observations are located, therefore enforcing high support of many segments. Finally, the marketing division of a company can set k, the number of segments, to a reasonable and manageable number. We denote WPC with its parameters as $WPC_{\lambda,k}$.

Cluster centers are strongly related to purchasing behavior; the higher the coordinate the more recently a product has typically been purchased in a cluster. Using $\lambda < 0.5$ assures that weights definitely point to products that have been purchased more recently. Assuming $\lambda = 0.5$, a cluster centroid with coordinates (0.53, 1.02, 0) for products P_1, P_2, and P_3 aggregates customers who purchased P_2 most recently after they purchased P_1, as well as customers that additionally purchased P_1 and/or P_2 earlier, leading to a coordinate exceeding one for P_1 and a coordinate exceeding 0.5 for P_2.

Similar to *SSM*, WPC can also be used for CR prediction and customer targeting. Therefore, the clusters are learned on training data (including their observed CR) and customers in the test set are then assigned the CR of the cluster with the nearest centroid. However, in contrast to *SSM*, all customers are assigned to one of the clusters learned. Finally, we order the determined segments in descending order by conversion rate and target the customers in the segments until reaching the campaign size, e.g., 10% of the customer base.

3.3 Estimation of Loss in Lift

As aforementioned, our prescriptive task is to maximize the out-of-sample lift with a targeting model, where lift is formally defined as the quotient of the CR within a defined percentile of customers selected by a model divided by the average CR over all customers $\overline{CR}$ (Piatetsky-Shapiro and Masand 1999).

We reformulate this task as minimization of the portion of lift that has not been achieved by a targeting model. We refer to this quantity as *loss in lift*. In this regard, the problem at hand is a customer sorting rather than an accuracy maximization problem, where the estimated CR of customers relative to the estimated CR of other customers decides whether a customer is targeted, and errors

result from wrong customer orderings, i.e., wrong relative CR estimations.[4]

We now turn to the question whether we can anticipate the loss in lift with different aggregation levels and parameters of SSM and WPC in order to find near-optimal configurations of the methods with low loss. In statistical learning, this is usually done by decomposing the resulting loss when applying a model into a systematic loss component due to generalization (*bias*), a component due to fluctuations in a data set (*variance*), and a *random loss* (or error), component that cannot be influenced or reduced (James et al. 2013).

A unified decomposition for any type of predictive task was defined by Domingos (2000). We will now briefly describe our specific proposition for loss of lift based on the unified decomposition. For a loss function $L(t, y)$ the expected value of the loss in lift $E_{D,t}$ is decomposed into random loss $N(x)$, bias $B(x)$ and variance $V(x)$, where t signifies the true value, y stands for the prediction of the respective training instance x within a training set D. The task is now to define the three components constituting loss in lift, our criterion to be minimized. We will start with the random component, that refers to the loss that cannot be reduced by any model and is therefore not a part of a detailed investigation.

Random Loss Consider an optimal targeting model, i.e., one that correctly predicts the CR per (non-aggregated) sequence and targets the customers in the highest-ranked segments (by CR) up to campaign size. Further consider as an example a segment with CR of 90%. Even with this optimal model, 10% of the targeted customers would not buy the product of interest later on. This percentage of expected "wrong" classifications is referred to as the Bayes error rate. It can be estimated on the training data by ranking the non-aggregated sequence observations by CR and determining the difference of the mathematically maximal achievable lift if segments were "pure" and consisted only of buyers or non-buyers, determined by $\frac{1}{CR}$ until the expected number of buyers is already targeted, and the lift achieved with in-sample optimal model (non-aggregated sequences).

Bias The bias component is the systematic loss in lift with a targeting (aggregation) model compared to the optimal in-sample model, described in the previous paragraph, that achieves the highest in-sample lift using full sequential information. The intuition is that any kind of aggregation will result in further loss in lift (referred to as bias component) compared to the in-sample optimal model, and we will experience this loss also on the test data later on. With $lift_{max,C}$ denoting the in-sample maximum achievable lift for a campaign size C, and $lift_{a,C}$ denoting the lift for C with aggregation model a, the bias is shown in Eq. (4). Therein, bias is defined as relative in-sample loss of the aggregation model due to generalization compared to optimal in-sample lift.

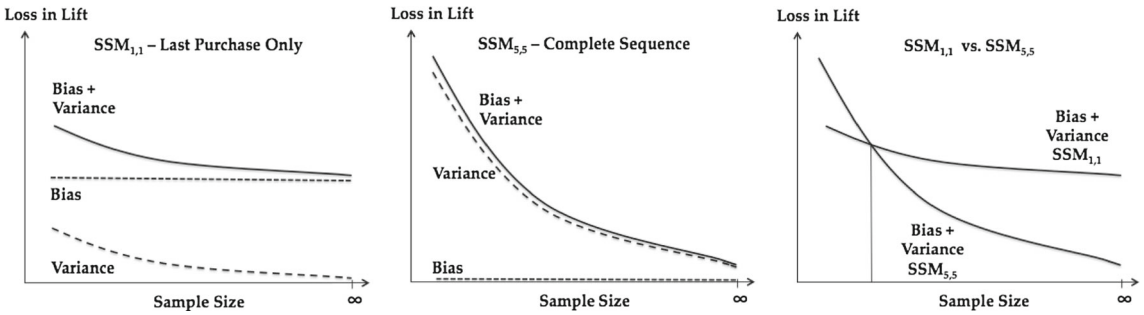
$$Bias = \frac{lift_{max,C} - lift_{a,C}}{lift_{max,C}} = 1 - \frac{lift_{a,C}}{lift_{max,C}} = 1 - lift_{norm,C} \tag{4}$$

The bias is closely related to the value of the (in-sample) normalized lift $lift_{norm,C}$ that quantifies the amount of information captured by the model a as naturally both sum up to 1. Obviously, no aggregation means zero bias as it corresponds to the optimal model in-sample. The higher the aggregation-level the higher the bias will be, i.e., the larger the segments the more they will tend towards the baseline CR. Increasing support per cluster increases the robustness of the inference as aggregation stabilizes the predicted CR and reduces the probability of wrongly selecting customers in segments with high CR in the training data. On the other hand, aggregation introduces the loss in lift on test data that can be drawn from in-sample observed CR, as a part of information from the data was lost because of (too strong) aggregation.

Variance The variance component quantifies the loss in lift due to randomness in the training data that is captured by a model. Given small segments, training CR will vary strongly depending on the particular training sample observed (see example in the introduction). Targeting based on the CR observed with small segments will tend to pick the segments (i.e. customers) with (then often randomly) higher CR and therefore overestimate the lift achievable out-of-sample. The result is a sub-optimal customer targeting based on such CR estimates. The variance component of loss in lift, therefore, results from fluctuations of customers between the samples of data (or folds within cross-validation).

To model the variance component, we compute the variance of the in-sample predicted CR on customer level for all customers, which would be addressed in at least one sample of training data. The aggregation over the customers is done by computing the mean of the variance estimates. The computation is shown formally in Eq. (5). For each customer c who entered the C top customers at least once among all training folds $f \in 1, \ldots, F$ of cross-validation and therefore belonging to the set labeled X_C we estimate variance of observed conversion rates $PCR_{c,f}$ around the mean predicted conversion rate m_c over all F training sets.

[4] This is very different from tasks such as class prediction, where a classifier is typically assessed by the total accuracy or its (potentially weighted) confusion matrix computed over all test data instances. The discrepancy of the business-oriented objective of lift and the traditional accuracy measures as well as its implications are extensively discussed in Baumann et al. (2015).



(a) Bias and variance for $SSM_{1,1}$

(b) Bias and variance for $SSM_{5,5}$

(c) Comparison of both

Fig. 3 Bias and variance components of the loss in lift for $SSM_{1,1}$ and $SSM_{5,5}$. Which of both models will lead to higher asymptotic loss depends on training sample size. At the intersection of both loss curves both models achieve the same asymptotic loss

The variance component can be reduced by stronger aggregation, which however, would lead to a higher bias on test data because of CR estimates closer to the baseline. Therefore, training CR (bias) and cluster support (variance) are in conflict, typically referred to as bias–variance trade-off, which needs to be understood and solved in order to minimize the loss in lift.

$$Variance = \frac{1}{|X_C|}\sum_{c=1}^{|X_C|}\frac{1}{F}\sum_{f=1}^{F}(PCR_{c,f} - m_c)^2 \quad (5)$$

The asymptotic behavior of bias and variance over increasing training size is illustrated in Fig. 3. The figure depicts the loss in lift for two sequential models sketching two extremes for our data with respect to aggregation level, namely $SSM_{1,1}$ and $SSM_{5,5}$. $SSM_{1,1}$, depicted in the left-hand graph, is the model resulting in the highest aggregation levels and can be expected to show the highest bias combined with low variance even for small training samples. In contrast, as depicted in the center, the complete sequence per definition exposes no bias on the training data, but variance should strongly increase with decreasing sample size, as small changes of the training sample can strongly impact the predicted CR of segments. The right-hand side figure visualizes the expected size of the overall reducible loss in lift with both models. With decreasing sample size – and decreasing support per segment - a point will be approached where $SSM_{1,1}$ will lead to lower loss than $SSM_{5,5}$.

Overall, we expect a bias–variance behavior as described above, making appropriate sequence aggregation a mandatory step. This will be analyzed in our empirical study that will now be described.

4 Empirical Data and Evaluation Design

We now investigate the out-of-sample lift (henceforth: lift) achieved with WPC and SSM on empirical data and compare the results to those obtained with benchmark approaches for customer targeting as they are broadly used today. Subsequently, we will discuss the discounting of weights towards more recent purchases and its impact on lift. Finally, we will analyze how well the estimates of the bias–variance behavior predict the loss in lift and helps us to find beneficial parameterizations of our models. In the next section, we will briefly describe the data used in our evaluation.

4.1 Available Data

The data set available comprises purchasing sequences of 236,327 customers, with purchasing sequences over a period of ten years. An item in a sequence indicates from which of the ten available product categories the purchased product originates from (henceforth we will omit the term product category as we do not differentiate between products within a category). The products are certain hosting services or products starting with Web domains as a basic service over different hosting offers to business solutions like online-shops. For reasons of confidentiality, product categories are labeled $P_i, i = 1, \ldots, I$.

The raw sequence data (categorical encoding of the sequence per sequence position) is used to generate ten different datasets using the following preprocessing procedure. We select one out of the ten products as target product P^d for which we analyze purchasing probabilities depending on customers' prior purchasing histories. If P^d is contained in a customer's purchasing history, we store the last L product purchases prior the P^d purchase and label the sequence S^d_{cl}, to indicate that this sequence resulted in a P^d purchase. If a customer purchased P^d more then once, we take the first purchase of P^d and derive the purchasing sequence up to that purchase. If P^d is the first purchase of the customer we omit the customer from the sub-dataset. We set $L = 5$ as sequence length as less than 10% of the

Table 3 Data statistics by target product

P^d	Baseline CR	# of customers	# of sequence types
P1	0.098	102,226	1710
P2	0.085	168,841	3580
P3	0.017	220,753	3863
P4	0.111	106,119	2563
P5	0.009	230,103	4000
P6	0.011	227,568	4209
P7	0.009	232,179	4203
P8	0.003	234,637	4817
P9	0.012	230,748	4814
P10	0.005	235,222	4829

The table shows the corresponding baseline CR within a specific data subset, the number of customers and the number of unique sequences (sequence types)

sequences in our data exceed a number of five product purchases. For a customer that does not contain P^d in the purchasing sequence at all, we generate the respective purchasing sequence S_{cl} from the last L products purchased (including duplicates). As P^d is one of the ten products offered by the company, sequences can consist of nine possible previous products plus P_0.

Table 3 provides descriptive statistics of the ten sub-datasets. For each P^d, the table shows the corresponding baseline CR within a specific data subset, the number of customers and the number of unique sequences (sequence types).

4.2 Evaluation Design

The treatment structure used in our evaluation is shown in Table 4.

The first treatment considers the model. If not stated otherwise, we will refer to a model with a particular parameterization also simply as model. With a maximum sequence length of five, for SSM we apply nine models: five models with $t = L$ from one to five with mandatory purchasing sequences of the most recent t purchases and four SSM-models with $L = 5$ and $t \in \{0, 1, 2, 3\}$, representing sequence types with a mandatory order for the most recent t purchases, and portfolios of product purchases further in the past. For WPC, we analyze 16 models with $\lambda \in \{0.1, 0.5, 0.9, 1.0\}$ combined with different numbers of clusters $k \in \{10, 50, 200, 500\}$. For both models we estimate CR for a resulting segment on training data (*training CR*). Customers in the test data are targeted by the training CR of their associated segments, in descending order by CR until a defined campaign size C is approached. Evaluation criterion is the lift achieved for targeted customers. Estimates for bias and variance are also computed on training data.

As benchmarks for two-step approaches, segmentation and subsequent prediction, we use the Levenshtein–Ward (LW) clustering that is broadly used in the marketing literature, with $k \in \{10, 50, 200, 500\}$. As with WPC, clusters are then sorted and test customers are assigned the CR or their segment to determine targets. In addition, we benchmark against widely used model-based clustering and classification approaches, namely Association Rules (AR), Hidden Markov Model (HMM) and Self-Organizing Maps (SOM). The support parameter of AR encodes the minimum support for a rule, resulting in approximately 20 ($support = 10^{-5}$) or 200 ($support = 10^{-4}$) customers as support threshold.[5] For HMM we employ $k \in \{10, 50\}$ states. Most studies use HMM with less than ten states (see Schweidel et al. 2011 or Netzer et al. 2008), which did, however, perform poorly in our setting and we included a variant with 50 states that performed better. The prediction is derived using previously learned emission probabilities given the next predicted latent state.

SOM allows for clustering using a predefined grid, so that the resulting assignment of an observation to a grid cell can be mapped to the corresponding in-sample CR. For SOM we employ grids with $k \in \{10, 50, 200, 500\}$ cells. Data preparation for SOM allows for implicit dummy coding, so that the complete sequential order is preserved in the input data. A binary encoded vector per customer indicates whether a customer purchased a certain product category as his last, previous to last, etc. purchase.[6]

As further benchmarks we apply Logistic Regression (LR) and Singular Value Decomposition (SVD) as it is a standard matrix factorization tool in recommender systems (Dunlavy et al. 2011). As with SOM, LR is applied to the complete sequential information, encoded in 48 dummies. SVD operates on the same weighted vectors as WPC, but is represented as a customer–product matrix, where a row corresponds to a customer and a column to a product dimension. As with WPC, we use $\lambda = \{0.1, 0.5, 0.9, 1.0\}$. As WPC operates on the same data representation, the difference of outcomes can only be driven by the clustering used in WPC instead of using SVD. These two approaches in contrast to the previous ones do not build segments.

The second treatment is the target product, where we analyze the models on the ten target product-specific data sets. The sample size of the training data, as the third treatment, is reduced to 1/2, 1/4, 1/8, 1/16, 1/32 of the

[5] Evaluations with higher and lower parameter values delivered clearly worse results and are not further considered in this article.

[6] This results in 48 dimensional binary vector encoding 9 + 9 potential products for the first two purchases, and 10 + 10 + 10 potential products when including P_0 for the three prior purchases.

Table 4 The treatments considered are (1) the aggregation and targeting model with its parameterization, (2) the target product-specific data subset, and (3) the degree to which the target product-specific subsets have been down-sampled

For each model, we indicate (in brackets) whether it performs segmentation (S) and predictions are then derived by the CR of segments, and/or performs prediction (P) directly

Model type	Cardinality (no. of distinct parameterizations)
Sequence set model (S/P)	5 models $SSM_{i,i}$ with $i \in [1,5]$ and 4 models $SSM_{5,j}$ with $j \in [0,3]$
Weighted-productspace clustering (S/P)	16 models $WPC_{n,\lambda}$ with $n \in \{10,50,200,500\}$ and $\lambda \in \{0.1,0.5,0.9,1.0\}$
Association rules (S/P)	2 models $AR_{support}$ with $support \in \{10^{-4},10^{-5}\}$
Hidden Markov model (S/P)	2 models HMM_n with $n \in \{10,50\}$
Levenshtein–Ward (S/P)	4 models LW_n with $n \in \{10,50,200,500\}$
Logistic regression (P)	1 model based on 45 dummies (complete sequence)
Self-organizing maps (S/P)	4 models SOM_n with $n \in \{10,50,200,500\}$
Temporal SVD (P)	4 models SVD_λ with $\lambda \in \{0.1,0.5,0.9,1.0\}$
Data subset	10 Product-specific data sets
Sample size	6 training sample sizes (1, 1/2, 1/4, 1/8 ,1/16, 1/32)

available sequences using random subsampling to study the model behavior when the overall support declines. Unlike the training size, the test data is the same for all six training sample sizes.

The evaluation criterion is lift when selecting a campaign size of 1 up to 10% of customers that we analyze for ten target product-specific sets and six down-sample levels using ten-fold cross-validation, resulting in 6000 model comparisons. If not stated otherwise, we will use the out-of-sample *normalized lift* (or transformations) as criterion to allow for comparisons across products.

5 Empirical Results

For our evaluation we regress the Box-Cox-transformed normalized lift $lift^{BC}_{norm}$ on our treatments. We follow a multiple regression-based evaluation approach as this allows to control for influences such as the dependent product category in a straightforward and concise fashion using dummies (as done for example by Hsu et al. 2016). The aim is to avoid an overload of the evaluation section by many case separations, pairwise testing, and various results on filtered datasets related to particular treatment combination. Furthermore, lift develops exponentially with campaign size, which can be directly considered in a multiple regression settings using a prior Box–Cox transform.[7] While estimates are not directly interpretable due to the transformation of the lift, the transformation does not change the order of the estimates and the significance of their differences. With C denoting campaign size, P^d the target product of interest, $P^d \cdot C$ the interaction of product dummy and campaign size controlling for different slope per product (how fast lift decreases with percentile) and $a \in 1,\ldots,A$ as a dummy for the particular model (or model family), the regression formula is shown in Eq. (6).

$$lift^{BC}_{norm} = \beta_0 + \beta_1 \cdot C + \sum_{d=1}^{I-1} \beta_{d+1} \cdot P^d + \sum_{d=1}^{I-1} \beta_{I+d} \cdot P^d \cdot C + \sum_{j=1}^{A-1} \beta_{2 \cdot I+j-1} \cdot a_j + \epsilon \quad (6)$$

Table 5 shows the regression estimates.

Table 5 Aggregated results

	(1) Estimate Best C1-10		(2) Estimate All C1-10
Campaign size C	−0.019**		−0.013***
$AR_{10^{-4}}$	0.515***	AR	0.331***
HMM_{50}	0.202***	HMM	0.097***
LW_{500}	0.975***	LW	0.832***
$SSM_{5,0}$	1.111***	SSM	1.053***
SOM_{500}	1.136***	SOM	0.902***
$SVD_{0.9}$	0.770***	SVD	0.655***
$WPC_{0.9,200}$	1.159***	WPC	1.047***
R^2	0.716		0.804
R^2_{adj}	0.715		0.804
F-statistic	771.665		6641.716
DF	7973		41973

The columns depict the regression coefficients of the different models considered. Column (1) (*Estimate Best*) shows the coefficients with models considering only the overall best parameterization of a model type over all campaign sizes. Column (2) (*Estimate All*) shows the coefficients when encoding all parameterizations of a model type with one model dummy, thereby reflecting the performance for all instances per model type. Baseline level for the method dummy is LR. The three best performing methods for all regressions are WPC, SOM and SSM

Significance codes: *** 0.001; ** 0.01; * 0.05; . 0.1

[7] We apply $\lambda_{Box-Cox} = 0.26$, as in our dataset we observe approximately white noise error structures with this value.

Column (1) (*Estimate Best*) depicts the coefficients considering C and only the overall best parameterization of a model type. Column (2) (*Estimate All*) shows the coefficients when one model type is encoded with one dummy independently of the parameters used. The latter regression does not control for different cardinalities of model instances per type. Hence, the coefficients are shown to allow for a general assessment of the sensitivity of outcomes with a model to improper parameter choices. On the one hand, due to the fact that the best parameterization is not known upfront, this comparison provides a more generic view of model types. On the other hand, extreme parameterizations such as $SSM_{5,5}$ lead to strongly overfitted results and poor performance, which negatively impacts the estimate of the model family. Therefore, both results are presented and discussed.

As expected, campaign size has a negative association with $lift^{BC}_{norm}$; the higher C, the lower the lift. As to the models, model estimates express the increase in average lift compared to the baseline model *LR*. As an example, the best *AR* model significantly increased average lift compared to the best *LR* model by 0.515.

The table further shows that the best performing individual models are WPC, SOM and SSM models (Column 1). These also dominate the other models when ignoring the specific parameterization and cumulating results over all parameterizations of a model (Column 2). This indicates that the ranking of the models is somewhat robust against suboptimal parameter choices.

As to the best individual model, $WPC_{0.9,200}$ achieved the highest lift with a coefficient of 1.159. The second best individual model, SOM_{500}, has a respective coefficient of 1.136, followed by $SSM_{5,0}$ with a coefficient of 1.111. As Table 5 determines significant lift differences between the models and *LR*, we conducted further statistical test for the best WPC against SOM – the strongest benchmark model on our data. Using two-sided paired Wilcoxon test[8] on the original lift values over all products and C up to 10%, $WPC_{200,0.9}$ achieves a significantly higher lift than SOM_{500} (p-value = 7e-06).

Aggregated results of all model independent of their parameterization (Column 2) show SSM with the highest estimate of 1.053, followed by WPC with an estimate of 1.047, and SOM with a coefficient of 0.902. The slightly inferior performance of WPC in Column (2) is due to very poor performing parameterizations with $k = 10$ (see Fig. 5 below) for very sparse data as no temporal discounting ($\lambda = 1$) is employed. When removing the WPC variants with $k = 10$, the lift over all remaining WPC parameterizations would then increase to a value of 1.113(***). Overall, these results show the importance of a proper parameterization.

Interestingly, SVD – although operating on the same data representation as WPC, where temporal weights are stored in the customer-to-product matrix instead of a vector per customer – achieves a much lower lift. This means that in our case a k-means clustering algorithm on the weighted product space data is superior to the application of SVD to the same data. HMM leads to only slight improvements over the baseline [with estimates of 0.202 (0.097)].

The superior performance of WPC compared to AR and LW is of particular interest. While all three approaches work with clustering based on sequence-similarity, only WPC considers where the (dis)similarity of sequences stems from; hence, the recency of purchases. In this respect, discounting more ancient purchases seems to be beneficial and will now be further discussed.

Figure 4 shows the empirical function of average lift over WPC discount (y-axis). A discount factor of 1.0 indicates no discounting, and the discounting increases when approaching zero. On the left-hand side, the curve is shown for WPC with ten clusters ($k = 10$). The right-hand side plot shows the curve with $k = 50$. The curves are computed for a campaign size of 10%. We see an inverse U-shaped relation. Initially, the lift increases with the introduction of a discount ($\lambda < 1$), indicating the higher importance of more recent purchases. With an increasing discount, at some point the lift decreases again, indicating that also older purchases are important, which are then discounted strongly for lower values of λ. The highest impact of a temporal discount is observed for a small number of clusters ($k = 10$), where a stronger discount leads to a significantly higher performance. For 50 clusters WPC shows the best result for $\lambda = 0.9$, which can be

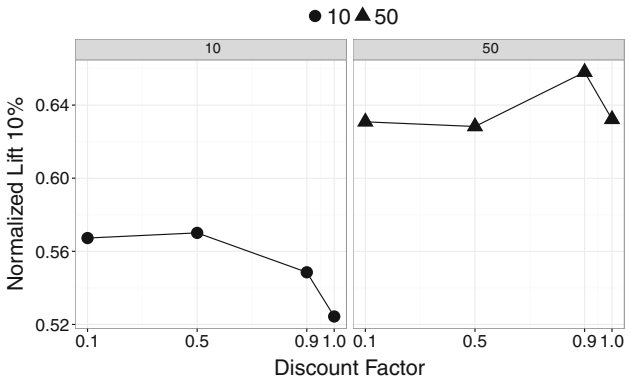


Fig. 4 The graphs show the average normalized lift with WPC over the discount factor (1.0 indicates no discounting) for C = 10%. The *left-hand* (*right-hand*) plot shows the curve with k = 10 (50). We see an inverse U-shaped relation

[8] We used Wilcoxon test as a more conservative approach but a t-test has also been conducted on Box–Cox transformed values, also confirming the significance in lift difference.

Table 6 Results with different sample sizes

	Estimate (1/1)	Estimate (1/2)	Estimate (1/4)	Estimate (1/8)	Estimate (1/16)	Estimate (1/32)
Campaign Size	−0.046***	−0.044***	−0.040***	−0.030***	−0.033***	−0.036***
LR	−0.868***	−0.817***	−0.806***	−0.825***	−0.954***	−0.895***
SSM	0.141***	0.138***	0.113***	0.090***	0.090***	0.071***
WPC	0.136***	0.144***	0.136***	0.127***	0.126***	0.105***
R^2	0.866	0.877	0.875	0.872	0.839	0.802
R^2_{adj}	0.866	0.877	0.874	0.872	0.839	0.802
F-statistic	8801.697	9755.849	9502.107	9253.521	7122.936	5516.304
DF	29,977	29,977	29,977	29,977	29,977	29,977

The baseline method is SOM. The order of model types by achieved lift is stable over decreasing sample size, simultaneously, the advantages of WPC over SSM increase with decreasing sample size

Significance codes: *** 0.001; ** 0.01; * 0.05; . 0.1

interpreted as a high amount of sequential information in the data.

First, these findings provide empirical support for the theoretical considerations that the original, high-dimensional data should not be utilized without prior (unsupervised) aggregating of the sequence data. Second, results underpin that the explicit consideration of temporal structures in purchase vectors and the diminishing importance over time provide additional predictive value.

5.1 Sensitivity to Sample Size

We will now study the impact of down-sampling the available training data on the models' performances. For reasons of brevity, we consider the models that have shown the best results on the complete datasets, namely SOM with four, SSM with nine and WPC with 16 model parameterizations as well as LR as the weakest benchmark on the full data set.

In contrast to Table 5 we use SOM as baseline model, as this was the strongest external competitor, so the estimates and p-value are provided with respect to this model type. Again, the model dummy indicates a particular model type independent of the parameters selected. Table 6 presents the average normalized transformed lift over all treatments of the respective model type and all campaign sizes between 1 and 10%.

In line with previous results, WPC reveals the highest estimates starting from 1/2 even with suboptimal parameterizations, and estimates are negative with LR for all down-sampled training subsets. However, the difference between the estimates increases with decreasing sample size, and we observe an increasing advantage with WPC as the samples get smaller. In order to determine an appropriate parameterization of WPC or SSM in practical settings and draw a more general recommendation, we will now discuss the empirical bias–variance behavior of the models and parameterizations.

5.2 Empirical Bias–Variance Behavior and Sensitivity to Sample Size

As discussed, models producing more segments should perform better with increasing sample size relative to models with fewer (but larger) segments. For instance, models considering more structure in sequences (e.g., by considering the full sequence) should improve relative to models considering only the most recent purchase because of the decreasing variance given a sufficient amount of training data (combined with the low bias component). This is studied in Fig. 5 for WPC and SSM and $C = 10\%$. The plots depict the average normalized lift (y-axis) over the average number of customers per cluster for sample sizes from 1 and 1/32.

With decreasing sample size the best WPC model employs a stronger discount: the top performing models for 1/32 sample size operate on $\lambda = 0.1$ or $\lambda = 0.5$. In the same spirit, the SSM-portfolio model, $SSM_{5,0}$, that strongly aggregates sequences, typically performs well with small sample sizes. Furthermore, the number of segments used by the best models decreases with decreasing sample size as the variance component increases. Also, the optimal average size of clusters can be derived from the plots: around 200 with complete sample, about 50 for 1/32.

Interestingly, the average cluster sizes with the best WPC parameterizations systematically exceed the ones used with SSM, meaning that WPC achieves comparable or better results with fewer clusters. The impact of the discounting factor on segment size is best seen for $WPC_{0.5,10}$ as it is able to accomplish both, higher lift value and larger segments as compared to $WPC_{1.0,10}$ or $SSM_{1,1}$ with comparable numbers of segments.

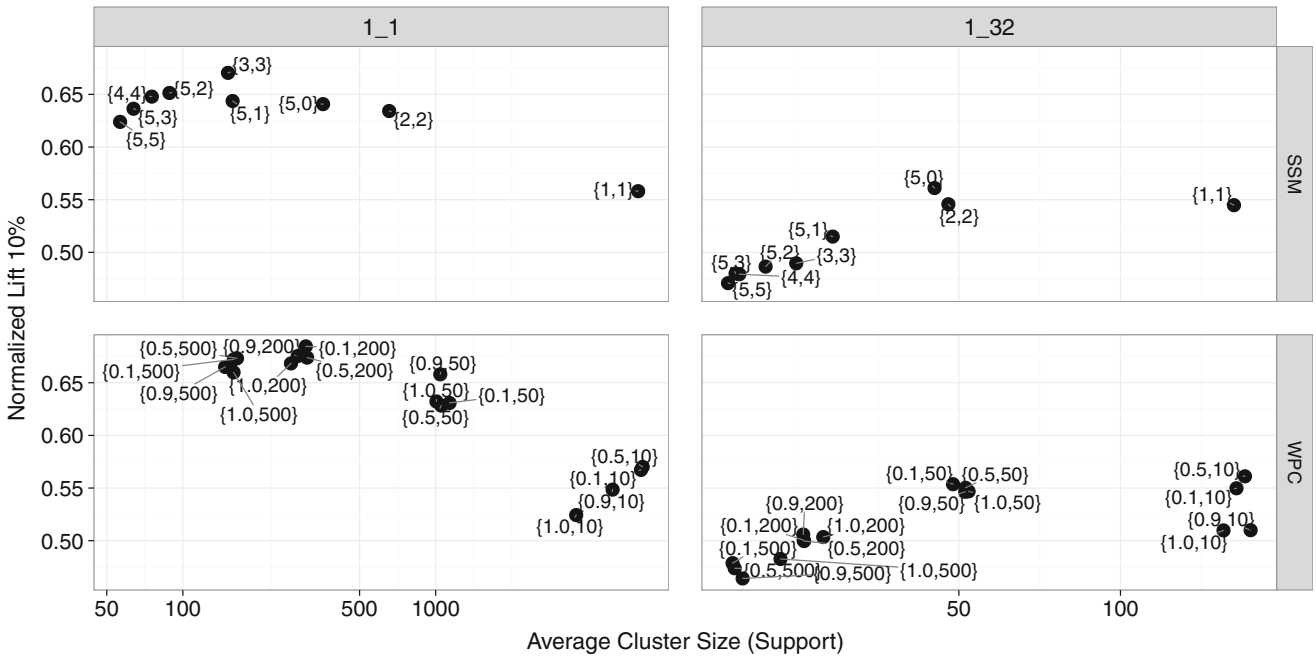


Fig. 5 Normalized lift for the best WPC and SSM parameterizations and their average cluster sizes. While the optimal average cluster size increases with sample size, WPC needs fewer clusters – that are consequently higher supported than SSM clusters

The bias–variance behavior of both model types will now be discussed in more detail. Figure 6 (left) displays the mean bias observed with selected models for subsamples of decreasing size.

As expected, the highest bias is obtained with $SSM_{1,1}$ as it is the simplest model and does differentiate by the last purchase only. On the other hand, $SSM_{5,5}$ per definition exhibits no bias, while model $SSM_{5,0}$ (the portfolio model) results in a bias in between both extremes. The WPC model with only 10 clusters has higher bias than the portfolio model, while it shows a performance similar to portfolio with 50 clusters, and clearly has a much lower bias when applied with k = 200 clusters.

Figure 6 (right) displays the estimated variance component of the loss in lift. Herein, the $SSM_{1,1}$, differentiating by the last purchase only, has a very low variance over all sample sizes. In contrast, $SSM_{5,5}$, the complete sequence, has the highest variance. The variance of $WPC_{0.5,10}$ is very close to the one of $SSM_{1,1}$. As with bias, we also observe similar (low) variance estimates for $SSM_{5,0}$ and $WPC_{0.9,50}$. $WPC_{0.1,200}$ has relatively high variance in the 1/32 sample, which can be explained by a number of clusters too large for this small training data set, leading to insufficient support and variance increase.

Figure 7 provides another view to the bias–variance trade-off: it shows the average normalized lift for 10% of customers with respect to training data size. The labels represent the average number of segments included in the respective customer selection and intersections show where the trade-off is better solved with another model. In summary, the best SSM and WPC models address the bias–variance trade-off by adjusting the parameterization in order to obtain average cluster sizes in beneficial ranges, while WPC better solves the trade-off and, therefore, achieves higher lift with lower numbers of clusters except for the 1/32 sample, best captured also by the portfolio $SSM_{5,0}$. $SSM_{1,1}$ shows low but stable performance when sample sizes decrease.

5.3 Estimating Loss in Lift on Training Data

Overall, we see empirical bias–variance behavior very well in line with our hypothesized developments. We will now study how well loss in lift[9] can be predicted by regressing it on the bias and variance estimates derived on the training data for the complete data set. Outcomes are shown in Table 7. In the linear regression, both bias and our variance estimate were normalized to the interval [0, 1]. Furthermore, we used a set of variables for the target product as this variable explains about 50% of the variance and is omitted for reasons of brevity.

With an R^2_{adj} of 87% the model has high explanatory power on the complete data set. As expected, bias and variance are positively associated with loss in lift, and the trade–off between both is well reflected by the interaction

[9] Note, that loss in lift and normalized out-of-sample lift sum up to 1.

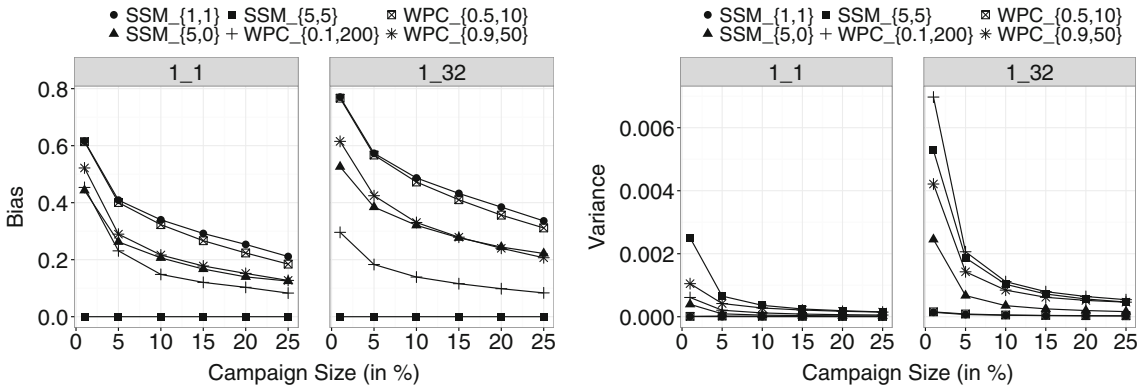


Fig. 6 *Left* Bias obtained by a specific model compared to the maximal sequential information of $SSM_{5,5}$ (on the training data). With higher aggregation, the bias increases, as a simpler model is not able to incorporate more information. *Right* Variance for selected models as the mean variance of customer-specific prediction for all customers up to campaign size. $SSM_{1,1}$ has the lowest absolute value and minimal increase as the sample decreases, while $SSM_{5,5}$ typically has the highest variance

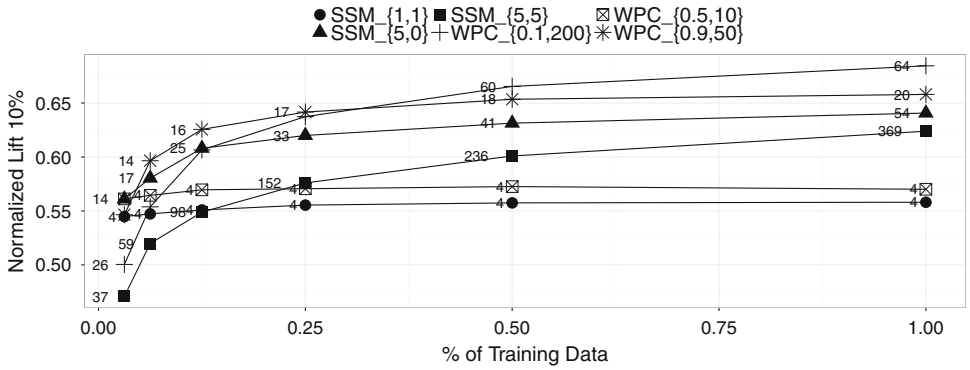


Fig. 7 Influence of the training sample size on average normalized lift for C = 10%. The numbers represent the number of segments within the campaign size. WPC shows the best results when selecting a high λ and 50 clusters for small samples of up to 25% of the original sample size, and low λ and a higher number of clusters of around 200 for larger samples

Table 7 Regression of loss in lift for 10% of customers over all products and the complete data set on bias and variance estimates

	Estimate
Bias	0.221***
Variance	0.108***
Bias × variance	−0.465***
R^2	0.880
R^2_{adj}	0.874
F-statistic	145.214
DF	237

Controls for target products are included but not shown. Bias and variance lead to significantly higher loss. The interaction of bias and variance is negative, indicating the trade-off between both

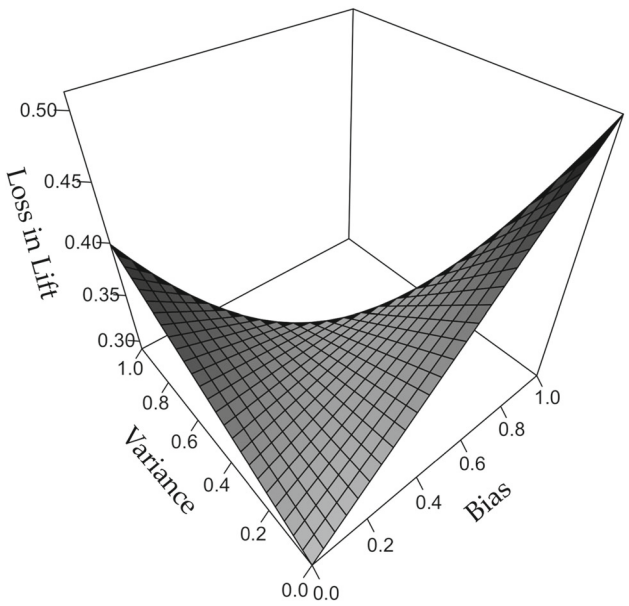


Fig. 8 Illustration of the bias–variance trade-off as a regression interaction surface with bias component (*x-axis*), estimated variance component (*y-axis*), and the loss in lift (*z-axis*)

of both with the negative coefficient. Figure 8 visualizes their interaction effect. Thereby, the x-axis represents the value of bias, the y-axis the magnitude of the estimated variance component, and the z-axis the loss. Although bias is the major driver of loss due to relatively high amount of training data given by the complete data sample, the

variance has a comparable influence. The U-shape clearly reveals the bias–variance trade-off that needs to be addressed appropriately with a segmentation model.

6 Discussion and Conclusion

Motivated by the increasing availability of sequential data in corporate information systems, e.g., traces of customer purchases, we introduced two novel mechanisms to aggregate similar sequences to generalized types. We argued both, from a statistical perspective and empirical results, that a prior unsupervised aggregation of sequences is mandatory to increase support and operationalize this information as a feature in predictive analytical models, where a segment can be used as a categorical feature in combination with additional data in a standard prediction model. Also, the resulting segments can be used for descriptive analysis directly by marketing practitioners allowing the derivation of marketing strategies based on discovered purchasing patterns. We have provided an explanation and have shown empirically that supervised aggregation can easily lead to unreliable predictions and, therefore, poor prescriptive performance, for instance with logistic regression.

The two similarity-based aggregation mechanisms we propose allow for adjustable, order-dependent weighting with higher weights given to more recent purchases, while still capturing the predictive value of older purchases and longer sequences. We further recommend learning weighting schemes of items depending on their order (here, recency) instead of truncating sequences. In our empirical evaluation, the two mechanisms achieved superior results compared to existing approaches that do not consider or learn weighting schemes, and did also lead to higher lift values than state-of-the-art class predictors. As the benchmark methods used the raw data without any preprocessing, more sophisticated supervised approaches or additional preprocessing steps might be evaluated in future.

We also discussed the bias–variance trade-off when applying the models to target customers in CRM campaigns with respect to the lift criterion. To our knowledge, for customer targeting problems this trade-off has not been formally discussed so far. As in any predictive task, we believe that a sound understanding of how to solve the bias–variance trade-off is key to learning appropriate models and deriving sound prescriptive models. For our use case and the empirical data set at hand, we have shown that the expected bias–variance behavior well predicts the lift achieved with the mechanisms, thereby providing guidance on how to parameterize the models.

Our data stems from one telecommunications provider and the models are evaluated on the specific task of customer targeting in cross-selling campaigns. Therefore, the question arises to which extent the results and recommendations are generalizable to other domains, business functions, and prescriptive tasks. While we are confident that the findings generally carry over to customer targeting problems of other larger telecommunications providers. The question whether to aggregate sequences in a supervised or unsupervised fashion, and whether to apply recency-dependent weighting when determining similarity does not clearly depend on the environment, the data, and the particular problem to be solved. Here, pseudo-out-of-sample results, e.g., using the bootstrap or cross-validation and proposed estimates for bias and variance, should already indicate which weighting scheme might be appropriate. Furthermore, depending on domain knowledge, any other weighting scheme than a geometric discount function might be used for WPC.

A limitation of the approaches we introduced is the number of products that can be considered. For the approaches to work well, the number of observations is required to increase superlinearly with the number of different products considered. Even on our dataset, both SSM and WPC did not perform well at product level (although on average still outperforming the benchmark methods), and we needed to conduct the analysis on product category level, which than achieved even higher lift values compared to applying the models on the full set of around 50 products. However, this also shows that a product palette, depending on product topology, can often be reduced to a number of categories, which can be better handled by the proposed algorithms.

A further issue is the small number of categories in our data, which on the one side reduces the complexity and a possibility of data artifacts, such as lock-in effects or tariff changes, on the other hand brings the disadvantages of the clustering with rising dimensionality. As to the general lock-in effect, a purchase of a domain technically does not demand a purchase of other components from the same supplier.

Finally, the consideration and explicit modeling of simultaneous product purchases might be of interest, as this might hint to different customer purchasing types. Also, it would be of interest from a statistical perspective whether and when the additional representation complexity (variance) can be compensated by the corresponding reduction of the bias component.

References

Back B, Holmbom A, Eklund T (2011) Customer portfolio analysis using the SOM. Int J Bus Inf Syst 8(4):396–412

Baumann A, Lessmann S, Coussement K, De Bock KW (2015) Maximize what matters: predicting customer churn with decision-centric ensemble selection. In: ECIS 2015 completed research papers. http://aisel.aisnet.org/ecis2015_cr/15/. Accessed 25 June 2017

Bicego M, Murino V, Figueiredo MA (2003) Similarity-based clustering of sequences using hidden Markov models. Machine learning and data mining in pattern recognition. Springer, Heidelberg, pp 86–95

Bose I, Chen X (2009) Quantitative models for direct marketing: a review from systems perspective. Eur J Oper Res 195(1):1–16

Brown RG (2004) Smoothing, forecasting and prediction of discrete time series. Courier Dover Publications, Mineola, NY

Chan CCH (2008) Intelligent value-based customer segmentation method for campaign management: a case study of automobile retailer. Expert Syst Appl 34(4):2754–2762

Cho YB, Cho YH, Kim SH (2005) Mining changes in customer buying behavior for collaborative recommendations. Expert Syst Appl 28(2):359–369

Daoud RA, Amine A, Bouikhalene B, Lbibb R (2015) Combining RFM model and clustering techniques for customer value analysis of a company selling online. In: Computer systems and applications (AICCSA), 2015 IEEE/ACS 12th international conference, IEEE, pp 1–6

Domingos P (2000) A unified bias-variance decomposition. In: Proceedings of 17th international conference on machine learning. Morgan Kaufmann, Stanford, CA, pp 231–238

Dunlavy DM, Kolda TG, Acar E (2011) Temporal link prediction using matrix and tensor factorizations. ACM Trans Knowl Discov Data TKDD 5(2):10

Han SH, Lu SX, Leung SC (2012) Segmentation of telecom customers based on customer value by decision tree model. Expert Syst Appl 39(4):3964–3973

Hsu MW, Lessmann S, Sung MC, Ma T, Johnson JE (2016) Bridging the divide in financial market forecasting: machine learners vs. financial economists. Expert Syst Appl 61:215–234

James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning, vol 6. Springer, Heidelberg

Joh CH, Timmermans HJ, Popkowski-Leszczyc PT (2003) Identifying purchase-history sensitive shopper segments using scanner panel data and sequence alignment methods. J Retail Consum Serv 10(3):135–144

Kaski S, Nikkilä J, Kohonen T (1998) Methods for interpreting a self-organized map in data analysis. In: In Proc. 6th European Symposium on Artificial Neural Networks (ESANN98). D-Facto, Brugfes, Citeseer

Khajvand M, Tarokh MJ (2011) Estimating customer future value of different customer segments based on adapted RFM model in retail banking context. Proced Comput Sci 3:1327–1332

Kohonen T (2001) Self-organizing maps. Springer, Heidelberg

Kruskal JB (1983) An overview of sequence comparison: time warps, string edits, and macromolecules. SIAM Rev 25(2):201–237

Levenshtein VI (1966) Binary codes capable of correcting deletions, insertions and reversals. Cybern Control Theory 10:845–848

Li S, Sun B, Wilcox RT (2005) Cross-selling sequentially ordered products: an application to consumer banking services. J Mark Res 42(2):233–239

MacQueen J et al (1967) Some methods for classification and analysis of multivariate observations, vol 1. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, California, pp 281–297

Miguéis V, Van den Poel D, Camanho A, Cunha J (2012) Predicting partial customer churn using Markov for discrimination for modeling first purchase sequences. Adv Data Anal Classif 6(4):337–353

Moeyersoms J, Martens D (2015) Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector. Decis Support Syst 72:72–81

Moon S, Russell GJ (2008) Predicting product purchase from inferred customer similarity: an autologistic model approach. Manag Sci 54(1):71–82

Mooney CH, Roddick JF (2013) Sequential pattern mining—approaches and algorithms. ACM Comput Surv 45(2):19:1–19:39

Netzer O, Lattin JM, Srinivasan V (2008) A hidden Markov model of customer relationship dynamics. Mark Sci 27(2):185–204

Ngai E, Xiu L, Chau D (2009) Application of data mining techniques in customer relationship management: a literature review and classification. Expert Syst Appl 36(2):2592–2602

Park DH, Kim HK, Choi IY, Kim JK (2012) A literature review and classification of recommender systems research. Expert Syst Appl 39(11):10,059–10,072

Piatetsky-Shapiro G, Masand B (1999) Estimating campaign benefits and modeling lift. In: Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining, ACM, New York, KDD '99, pp 185–193. doi:10.1145/312129.312225

Prinzie A, Van den Poel D (2007) Predicting home-appliance acquisition sequences: Markov/Markov for discrimination and survival analysis for modeling sequential information in NPTB models. Decis Support Syst 44(1):28–45

Sahoo N, Singh PV, Mukhopadhyay T (2012) A hidden Markov model for collaborative filtering. MIS Q 36(4):1329–1356

Schweidel DA, Bradlow ET, Fader PS (2011) Portfolio dynamics for customers of a multiservice provider. Manag Sci 57(3):471–486

Shirley KE, Small DS, Lynch KG, Maisto SA, Oslin DW (2010) Hidden Markov models for alcoholism treatment trial data. Ann Appl Stat 4:366–395

Steinmann S, Silberer G (2010) Clustering customer contact sequences—results of a customer survey in retailing. European Retail Research. Gabler, Wiesbaden, pp 97–120

Van den Poel D, Buckinx W (2005) Predicting online-purchasing behaviour. Eur J Oper Res 166(2):557–575

Wong KW, Zhou S, Yang Q, Yeung JMS (2005) Mining customer value: from association rules to direct marketing. Data Min Knowl Discov 11(1):57–79