Purchasing History Representations for Prediction and Targeting in Customer Relationship Management Campaigns

Zur Erlangung des akademischen Grades eines
Doktors der Ingenieurwissenschaften

(Dr.-Ing.)

von der Fakultät für
Wirtschaftswissenschaften
am Karlsruher Institut für Technologie (KIT)

genehmigte

DISSERTATION

von

M. Sc. Wi.-Ing. Katerina Shapoval

Tag der mündlichen Prüfung:	15.03.2017
Referent:	Prof. Dr. Thomas Setzer
Korreferent:	Prof. Dr. Hansjörg Fromm
Prüfer:	Prof. Dr. Christof Weinhardt

2017 Karlsruhe

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Motivation

In this chapter a brief introduction to Customer Relationship Management (CRM) is given. Thereby, the general tasks of predictive analytics are presented within a typical customer lifecycle. Next, the importance and crucial challenges of purchasing histories within predictive models for CRM are outlined. Based on the identified challenges the research questions are derived. The chapter concludes with the structure of the thesis.

1.1 Predictive Analytics in Customer Relationship Management (CRM)

In the context of CRM the main task translates into using available customer data for understanding and developing long-term relationships with the customer base. Predictive analytics, a tool of analytical CRM and as well as within information systems (Shmueli and Koppius, 2010), aims at providing explanatory and predictive models. One of the possible application fields is choosing the most receptive customers for a certain marketing campaign, e.g. a product offer, using all available customer information in a company's information systems. For example, given a specific target product, campaign details, and the number of customers to be contacted, a typical predictive task is to identify the customers most likely to purchase this target product, i.e., if the target product is a probable next purchase.

As the customer bases nowadays include millions of customers, for any data-driven CRM activity the complexity of the heterogeneous customer information needs to be reduced to relevant and interpretable generalized customer types, e.g. by segmentation of customers. Therefore, predictive analytics uses statistical prediction models, thereby generating added value by employing customer

data, available in enterprise IS. In the context of customer's lifecycle, there are three types of predictive models with respect to their goal (Lo, 2002).

- An acquisition model has the goal of identifying the non-customers most likely to become new customers.
- A development model searches for the most receptive customers to an offer of an additional product (cross-selling).
- A retention or a prevention model aims at identifying customers who are likely to be pursuable to not canceling the contract (prevention) or to taking back the cancellation (retention).

The focus of this work is explicitly the modeling based on available customer data within an enterprise, i.e., for individuals who are already customers; acquisition models are out of scope of this work. Therefore, the investigation is dedicated to prediction of cross-selling probabilities in the context of customer development and modeling of cancellation probabilities for targeting in prevention campaigns.

1.2 Purchasing History in Prediction Models

As large amount of information is available in enterprise databases, the challenge nowadays is rather the identification and aggregation of the relevant one, than a lack of available information. Although different kinds of information, like demographics or revenue history, are available to a company, the focus of this work is only the purchasing history consisting of data on product portfolio and purchasing sequence of customers. Information about customer's actions available to a company is crucial for its marketing efforts. One of the vital tasks for service companies is predicting a probability of a certain action, e.g. for a next purchase or for a contract cancellation. As the preferences of a customer are not known in advance, these have to be derived based on available information and known actions of similar customers.

One challenge in predictive modeling is the generation or selection of features from available customer data that allow targeting the most affine customers (Ngai et al., 2009). The purchasing history has been shown to have a strong predictive power in the literature. Several approaches successfully used additional features such as the customer lifetime value or particular products purchased in the past, e.g. published by Chan (2008) and Khajvand and Tarokh (2011),

where in particular purchasing history is assigned high significance. Van den Poel and Buckinx (2005) along with Bose and Chen (2009) have shown that a more detailed consideration of customers' past purchasing behavior can provide additional predictive value over models solely based on customers' socio-demographics and RFM features considering *R*ecency, *F*requency, and *M*onetary value of customers' purchases. As an example for RFM features, a customer would be characterized by the number of days since last purchase, total number of purchases and the total value of these.

Beyond the pure value of the purchasing history, there is evidence that the temporal relation between purchases provides additional predictive value, if a logical order of purchases exists. Especially, the purchasing sequence is of a particular interest. Given a temporal structure in the data, it can be used to derive the next offer for a customer and constitute a useful feature for prediction of contract cancellation.

There are several studies in different domains, like financial services (Li et al., 2005) or electrical devices retail (Prinzie and Van den Poel, 2007), investigating the existence of such "logical" order using different methods, such as different Markov models, sequence pattern analysis and Cox regression. The goal of this explicit consideration is to investigate the influence of the sequence information and therefore of the presence of such typical sequences – logical order of purchases – in the examined area of information and communications technology. A simple example for such sequences is the purchase of a domain, followed by a web space acquisition, and finally by a purchase of a server solution.

As to the application of purchasing history data relevant for this work, there are two main cases, namely cross-selling and prevention in the context of direct marketing. For the case of cross-selling, a detailed survey on the process and data used in direct marketing is presented in Bose and Chen (2009). This work investigates the ability of aggregated purchasing history to predict the probability of the next purchase.

The second area of application is prevention. Thereby, prevention aims at predicting a cancellation announcement for a contract in order to make a preventive offer. In contrast, the case of retention has the goal of predicting the success of a reacquisition offer after cancellation announcement. Both cases can be summarized to the term churn prediction, which is widely used in the literature. For a general review see Kamalraj and Malathi (2013). As to the importance of the portfolio information, several studies have shown predictive power of features derived from the purchasing history in general context of churn (e.g. Miguéis et al. (2012)). Particularly, in telecommunications bundling, which is in some

way a forced portfolio decision, a positive impact on churn reduction has been demonstrated, as shown for example by Prince and Greenstein (2014). This work investigates the influence of a chosen product portfolio on churn predictability in the context of prevention, where not only the portfolio combination but also the timing of contract expiration is evaluated.

1.3 Challenges of Sequential Data and Evaluation Metrics in CRM

Unfortunately, a purchasing sequence easily becomes a high-cardinality feature, a non-quantitative feature with many categories, but only few observations (support) per category (Moeyersoms and Martens, 2015). 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 purchases of a customer (Van den Poel and Buckinx, 2005), possession of a product in certain categories (Back et al., 2011) or customer value (Han et al., 2012). However, even for moderate numbers of product categories combined with low number of past purchases taken into account, combination of which constitutes exponentially growing representation complexity of purchasing history, raw purchasing vectors cannot be operationalized within a prediction model in case of only few observations per sequence due to unreliable inference and poor predictive value.

The following two challenges are related to the evaluation metric. The first issue in this regard in predictive models for CRM is the problem of imbalanced data sets with quite low baselines (Ling and Li, 1998). For such settings, accuracy is not an appropriate metric. Imagine, a baseline purchasing probability in a data set is 1 %. One would gain an accuracy of 99 % just by assigning the negative class to all observations. As the goal is to identify positive cases, e.g. likely buyers, false negatives are a more appropriate metric.

The second issue is not only the unequal importance of false negatives and false positives, but that the nature of a classification problem in CRM setting is different. Typical classifiers assign instances to two classes, positive and negative, but the goal in marketing is different, namely choosing best-suited customers for a marketing campaign. As the size of a marketing campaign is restricted, a more granular statement than division in two classes is needed. Therefore, direct marketing requires algorithms, which allow for ranking of targets, e.g. by probability. Additionally, the respective measures for the quality of such ranking are necessary.

State-of-the-art methodology for predictive modeling, and more broadly class prediction, includes a palette of class prediction techniques such as Logistic Regression, Hidden Markov Models, or Association Rules (an extensive survey is given by (Ngai et al., 2009)). The goal of these algorithms is to assign a probability for certain reactions like a product purchase to each customer. However, the problem of performance assessment of such algorithms also changes due to the described switch from a classification to an ordering problem, so that accuracy or a confusion matrix, with false negatives being a part of it, is not appropriate as an evaluation metric.

As a solution to both problems above a specific performance criterion, the so called lift, was proposed by Piatetsky-Shapiro and Masand (1999), which considers the sorting of customers and therefore is crucial for customer selection in marketing, where only the top portion of customers is addressed by an offer due to a constrained campaign size. It is defined as a quotient of the relative success rate for the selected percentage of customers and the average success rate among all customers. Hence, not only it is important whether a prediction (purchase vs. no purchase) is correct, i.e., accuracy related measures, but the position of each customer (ranked by purchasing probability) in the population is crucial, as it determines if this customer will be included into a marketing activity. Lift is the most applied metric in the context of direct marketing, whereas the so-called top-decile lift is the lift value (lift obtained for 10 % of best customers with respect to the evaluated model) is the common evaluation metric (Baumann et al., 2015). Therefore, lift serves as an evaluation metric for the practical application in CRM tasks as well as for the empirical evaluation in this work.

1.4 Sequence Aggregation in the Context of Bias–Variance Trade-Off

While it has been shown that customers' purchasing history is a promising feature to better anticipate the behavior of customers, the operationalization of explicit purchasing sequence is challenging. The difficulty of such high-cardinality features, namely categorical features with high number of categories and little support per unique sequence, is that any technique making inference on such data would not deliver a robust prediction based on such low case numbers. Hence, sequential data must be grouped prior to applying predictive techniques to increase the statistical support and therefore the robustness of the prediction, thereby optimizing the bias–variance trade-off discussed next.

A typical bias–variance trade-off has the goal of balancing the systematical error due to generalization (*bias*) against the error due to fluctuations in the training sample (*variance*), especially in cases with low support the latter has a higher importance. Applying a stronger generalization (or aggregation) would reduce the variance component due to higher support per segment but would simultaneously increase the bias, and vice versa. The application of this principle in purchasing sequence segmentation can be achieved by changing the number vs. granularity of the resulting segments. Higher granularity allows better selection of segments on training data reducing the bias, but bears the risk of generalization loss as the results are instable due to high variance within segments with small support. Hence, the challenge is to find the "right" aggregation mechanism using an appropriate similarity criterion for grouping combined with the "right" parameterization, so that the bias–variance trade-off is solved optimally. The goal is thereby to determine a small number of segments with high discriminatory power and high predictive performance.

As the purchasing history is the focus of this thesis, an example of the conversion rate (CR) – empirical probability of a purchase – for a specific product based on the sequence of previous purchases is shown in Table 1.1 to motivate the research questions at hand. The table shows an excerpt of empirical purchasing sequences observed in the customer database of the sample company, a leading multinational telecommunications company. 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 palette of the telecommunications company can be differentiated into two branches. The first comprises the products related to the information and communications technology (*hosting* products), ranging from basic starter-products like Internet domains, up to various professional server solutions for large-scale businesses. The second branch comprises access products, including mobile telephony, mobile Internet access (e.g. for tablets) as well as digital subscriber lines (*access* products).

Table 1.1 contains purchasing sequences of hosting products. Each row of the table corresponds to one purchasing sequence (the last five purchases) prior to a purchase of a particular product P_2 – intended to be advertised in a cross-selling campaign – after observing a sequence from the ten products offered by the company ($P_1, \ldots . . P_{10}$). For instance, the first row displays the sequence $< P_1, P_1, P_3, P_1, P_4 >$ (ID 168) with a purchase of P_1 as the most recent one. Rows are sorted by the observed (historic) purchasing ratio, i.e., conversion rate. *ID* displays the relative position in the set, *Support* quantifies number of observa-

Table 1.1: Examples of purchasing histories from the sample company's data. Rows are sorted by CR per sequence. ID displays the relative position in the set. The baseline purchasing frequency (CR when randomly selecting customers) is 8.8 %. Sequences with high observed CR tend to have little support, while frequent sequences are close to the baseline CR.

ID	Last	2ndL	3rdL	4thL	5thL	Support	CR
…	…	…	…	…	…	…	…
168	P1	P1	P3	P1	P4	4	0.75
176	P1	P1	P5	P1	P3	2	0.50
…	…	…	…	…	…	…	…
380	P9	P4	P4	P4	P4	23	0.22
436	P9	P4	P1			56	0.18
…	…	…	…	…	…	…	…
656	P4	P4	P1			2956	0.09
…	…	…	…	…	…	…	…

tions per unique sequence. The baseline purchasing probability over all customers independent of previous purchases is 8.8 %.

Sequences with high CR appear to be the most promising targets. Unfortunately, CR estimates with top-ranked sequences are not only unstable because of low support, but CR will be systematically overestimated and the targeting will be biased. That is because even a small change in the distribution of purchases within the segment will randomly increase or decrease a segment's rank. For instance, only four customers exhibit a purchasing history with ID 168, displayed in the first row, while three of them purchased the target product (CR of 0.75). A single buyer less among the four customers would lead to a change of the CR estimate to 0.5. Unfortunately, sequences with higher support (for instance, sequence with ID 656) approach the baseline CR of 8.8 % and are of no value as targeting this segment would not improve over a random selection. Overall, the result is a strong overestimation of the top-segments' CR estimates. Therefore, a targeting using such segments will also be biased.

Obviously, a supervised aggregation (considering the information if a target purchase was done) at the level of sequence instances with high but individually overestimated CR values would not be an optimal solution as this would not stabilize the CR estimates per segment. Hence, unsupervised aggregation techniques, which do not consider target purchase information, are required prior to any supervised approach to customer targeting. However, it is far from obvious how sequences should be grouped to (i) maximize the discriminatory power of the resulting customer clusters with respect to a next purchase and exploiting the temporal structure of purchases, while (ii) keeping the number of clusters (segments) small and (iii) allowing an intuitive interpretation of the resulting

groups as generalized purchasing behavior for marketing experts.

The motivation for the first goal is of practical value as the clusters having no difference in conversion rates are not useful, as they provide no additional information. Keeping the number of clusters small reduces the complexity and gives an effective aggregation mechanism simultaneously allowing for maximal information preservation. As the application of such grouping is also interesting from the managerial perspective, the resulting segmentation should also be interpretable for the marketing decision makers.

1.5 Research Questions and Structure of the Thesis

In summary, a challenge of predictive modeling is to identify the important features or characteristics for customers most prone to specific marketing campaigns. Although it has been shown in some domains that both, portfolio and purchasing history can provide predictive value, these are often reduced to segment information, number of contracts or not used at all. One reason is the high complexity of the sequence representation, which is not properly manageable due to the exponential increase of possible item sets and, therefore, a very high number of attributes required for such representation. The second important issue is the optimal level of aggregation with respect to the bias–variance trade-off for the performance criterion of the lift.

Based on the introduced context and specifics of the setting, this work concentrates on analytical aspects in CRM, in particular for two application fields within the customer lifecycle, namely cross-selling and churn prediction for prevention activities. The focus from the data perspective lies on definition and evaluation of features based on the purchasing history as well as on aggregation and incorporation of these into predictive models in CRM within these two particular fields. Additionally, the proposed methods are also useful for segmentation and target group analysis as they deliver interpretable segments based on purchasing histories. The domain-specific facet of this work lies in telecommunications industry, as the available real world data stems from it.

Dealing with purchasing history, the incorporation of purchasing sequences is motivated by the presence of temporal patterns, which then are used for segmentation and prediction. This idea summarizes the first group of research questions and considers hosting products of the sample company, which in the contrast to the access products, where a purchase of a mobile access does not necessarily lead to an interest in a digital subscriber line, may constitute a logical purchas-

ing order. Starting with the hypothesis that a certain "logical" purchasing order exists for hosting products, the first step is to provide evidence of the existence of such pattern. In this work this is done by employing a Hidden Markov Model, which provides the possibility of reducing the sequence variety to probabilistic segments, the so-called latent states, which statistically summarize the "reason" for the observed sequence. If any temporal pattern exists, the model will identify the corresponding grouping of sequences to the segments (latent states).

RQ 1.a **Presence of Temporal Patterns**
Does a logical order of purchases exist for products related to information and communications technology (hosting products)?

If such patterns exist, the first goal is to explicate and to interpret these at the explorative level in order to provide business understanding of the patterns, which can be summarized to the task of segmentation of purchasing sequences. The main hypothesis is that the products reflect a technological maturity of a customer as these can be built "on top" of each other and a next purchase would probably constitute an upgrade compared to an existing technology. The operationalization of this question is done by interpretation of the resulting model parameters, sequences belonging to the segments, additionally combined with the external available data (e.g. revenue per customer).

RQ 1.b **Interpretation of Temporal Patterns**
Do the identified segments correspond to a technological level or a need of the customer?

The next step is then an estimation of the predictive value for the identified segments, so an anticipated gain can be estimated prior to further activities. This is done by assigning the corresponding CR of the target product to each segment (latent state of the model) and then evaluating the predictive performance of segments via cross-validation.

RQ 1.c **Predictive Value of Temporal Patterns**
Which predictive performance can be achieved using aggregated sequential data with respect to lift?

After the presence and predictive performance of sequential data is shown, as the next step two aggregation models for purchasing sequences are introduced and evaluated. The intuition of both is a temporal discount of purchases done further in the past. *Sequence Set Model (SSM)* truncates the sequences or uses a combination of a certain sequence length and generalizes older purchases to a portfolio, i.e., neglecting the order purchases and considering only if a cer-

tain product was purchased further in the past. *Weighted-Productspace Clustering (WPC)* projects the sequences using geometrically descending weights into continuous space, where a distance-based clustering is done. The second method is more flexible due to a possibility of continuous parameterization, whereas the first relies on predefined categories, which do not necessarily have enough support. The second advantage of WPC lies in application of clustering, which is able to follow the data distribution in the space. Therefore, the hypothesis is that WPC provides higher predictive performance.

Both models are empirically evaluated and analyzed for prediction of the next purchase. As several established methods for sequence segmentation and prediction of the next purchase exist, a question is also how the proposed models perform compared to the state-of-the-art methods, leading to the next research question.

RQ 2.a **Comparison with Benchmark Models**
Which of the proposed aggregation models achieves higher predictive performance using the idea of temporal discount compared to other sequence aggregation methods in the context of cross-selling?

After introduction and evaluation of the models, two theoretical issues arise. The first is whether the predictive error can be estimated on the training data. This is a common issue and is usually investigated using the framework of the bias–variance trade-off. In this work, a respective definition of the estimates is proposed and evaluated for lift criterion, which has not been done yet in the literature. Thereby, bias refers to the error due to generalization of the data and variance refers to the error component resulting from fluctuations in the training data. The latter also refers to the challenge described in Table 1.1, where a single customer can influence the CR estimate of one particular sequence instance drastically, causing a non-optimal customer selection. As mentioned, both components constitute a trade-off, meaning an increase of the bias through stronger aggregation reduces variance, and vice versa. The proposed estimates are evaluated with regard to the question, if these are able to predict the resulting lift achieved on test data. Given such prediction is possible, the second issue is solved simultaneously, namely the one of the optimal parameterization, as the optimal aggregation with respect to the bias–variance trade-off can be estimated for any sequential data set.

RQ 2.b **Estimates for Bias–Variance Trade-Off**
To which extent do the proposed estimates of bias and variance on training data predict the resulting predictive performance for lift?

In the context of CRM it can be important to model additional information, for example in order to analyze bundles (products purchased simultaneously). Another issue, which was motivated by the empirical data, is the presence of repetitive purchases of the same product. The goal of this investigation would be the evaluation of additional predictive value of such serial purchases or if these can be skipped in the sequence. Furthermore, the length of the purchasing history can be interesting for further differentiation of purchasing histories, for instance, in order to identify a more detailed view on purchased products. Respectively, three extensions allowing for modeling of additional features like bundling, length of a purchasing sequence, and handling of repetitive purchases are presented and evaluated for WPC. Additionally, metrics indicating a possibility of the application for each extension based on the data are proposed.

RQ 2.c **Performance of WPC Extensions**
Which additional predictive performance is achieved by proposed WPC extensions?

A similar question about the resulting test performance through the estimation of the bias–variance trade-off arises with respect to more straightforward data metrics, for example the number of available data instances or average support per sequence. The idea is to estimate the potentially achievable lift already given the raw data, based on which simple metrics are computed. The evaluation is then conducted in a similar manner as for the last research question, i.e., by evaluating the ability of the proposed data-based metrics to predict the lift resulting from a prediction.

RQ 2.d **Data-Based Performance Prediction**
To which extent do the proposed data-based metrics predict the resulting predictive performance?

An additional application context is introduced with prediction of churn announcements for the access products. Typically, in the field of telecommunications a certain term of contract is given, which is the motivation for the hypothesis that contract cancellations are unequally distributed over contract duration. As a consequence, the time within this period is assumed to be a good predictor for the cancellation probability, probably with an increase of the probability

towards the end of the contract term. Additionally, the portfolio variety of a customer (number of product categories in customer's possession) as an estimate for the loyalty to the firm is included into the model. Overall, the contribution of features based on contractual information for contract cancellation prediction is evaluated compared to other available information.

RQ 3 **Features for Churn Prediction**
Which additional predictive performance is achieved using proposed features based on the contractual information for access products?

This work is structured along the formulated research questions. Chapter 2 outlines the related work with respect to the state-of-the-art methods in CRM, especially sequence aggregation methods and churn prediction, as well as summarizes the contribution of the presented approach compared to the existing literature. Chapter 3 investigates means for temporal pattern identification and the measurement of the resulting predictive value for hosting products. Chapter 4 lies out the formal definition for both models, additional extensions for certain data specifics as well as the discussion on bias–variance trade-off. Therein, also the results of an empirical application on a set of telecommunication data are presented. Then, the features proposed for aggregation of the contractual information are presented and evaluated in the context of churn prediction for access products in Chapter 5. The contribution, limitations and future research directions are finally discussed in Chapter 6.

This work partly contains the insights, evaluations and textual paragraphs close to the source from published and working papers. The presence of temporal patterns in Chapter 3 is discussed in Shapoval et al. (2015). Chapter 4 is based on Shapoval and Setzer (2017b) and Shapoval and Setzer (2017a), the evaluation in Chapter 5 additionally uses Shapoval and Setzer (2015). Chapter 2 partly contains paragraphs from all the listed publications.

Chapter 2

Aggregation and Application of Purchasing Histories in CRM

This chapter outlines the general context of CRM as an intersection between information systems as its technical aspect and direct marketing as one of the application fields. As a next step, typical tasks within a customer's lifecycle and the respective modeling approaches are introduced in detail. Then a methodical side of sequence aggregation mechanisms is specified, where existing types of approaches are listed, including the shortcomings of these. As an additional aspect, the application of purchasing history for churn prediction is dedicated an own section, as it constitutes a broad research field. The chapter also outlines the contribution with respect to each relevant research field.

2.1 CRM in Context of Information Systems and Marketing

CRM includes a broad palette of levels and tasks. Due to Farquad et al. (2014) there are three main levels of CRM: (1) strategic CRM concerned with the corresponding aspects like business culture or values; (2) collaborative CRM having cooperation of different departments in its focus; (3) analytical CRM with its goal of using available customer data and data mining methods for different tasks like cross-selling or churn prediction. The latter applications are the focus of this investigation.

From the global perspective CRM summarizes all processes used for supporting the maximization of customer value with a company along his complete lifecycle, whereas these processes or activities include customer acquisition, cross-selling and retention (Ngai et al., 2009). Therefore, CRM is an integral approach for companies to conduct interactions along the customer's lifecycles, from cus-

tomer acquisition, through cross-selling campaigns, to prevention and retention activities. Such definition accentuates the lifecycle of the customer within a CRM view thereby embracing the activities related to it. In the following, two aspects of CRM relevant for this work will be explicated based on the definition provided in the literature: the positioning of CRM within the customer lifecycle and the interconnection of IT and marketing by the area of predictive modeling.

As to the marketing side of CRM, with the rise of individual communication channels, such as telephone and e-mail, a shift from mass marketing, where an undifferentiated offer is brought to a wide range of customers, to direct marketing was enabled and also fulfilled. Direct marketing implies using knowledge of the customer for more targeted offers, which should result in higher response rates compared to the mass marketing. So overall, direct marketing aims at identifying likely responders at individual level and addressing these using promotion activities like email of telephone campaigns (Ling and Li, 1998).

On the other hand, CRM emerged in the context of information systems in mid 1990s and summarizes also a variety of technology-based solutions (Payne and Frow, 2005), so the second ingredient of CRM consists of information systems, constituting the prerequisites of operational implementation. Therefore, another definition points out that CRM is unifying marketing and IT resources for creation of profitable long-term relationships with the customer base (Payne and Frow, 2005), so the both areas, namely marketing and IT, are intertwined for a common goal. As a result, CRM processes are strongly supported by database and information systems, which provide infrastructure and services for more effective customer interaction.

Overall, this work contributes to the body of literature on analytical CRM with a particular application in direct marketing. The goal of the latter is the optimal selection of customers for marketing campaigns based on the available data in enterprise information systems.

2.2 Supervised and Unsupervised Tasks in CRM

A dimension for differentiating modeling tasks in CRM is the division into unsupervised and supervised techniques (Tsiptsis and Chorianopoulos, 2011). A supervised technique refers in general to a task, where a label or class for each data instance has to be predicted based on a trained statistical model, therefore constituting a typical predictive task. In contrast, an unsupervised technique refers to an explorative task over all available data without any predefined classes and

Table 2.1: CRM Tasks along a typical customer's lifecycle (based on Ngai et al. (2009)).

Lifecycle CRM Phase	Analytical CRM Task	Method Type	Description
Customer Acquisition	Target Customer Analysis	Supervised	Group profiling with available variables
	Segmentation	Unsupervised	Explorative modeling of customer base
Customer Development	Up/Cross-Selling	(Un)Supervised	Prediction of next-best-offer
Customer Prevention and Retention	Churn Prediction	(Un)Supervised	Prediction of a contract cancellation and prediction of a winback success

is consequently a typical explorative task. For this work both tasks as well as corresponding approaches are relevant and a summary of these is given in Table 2.1 based on Ngai et al. (2009), where the goals and corresponding tools for each specific phase of the CRM lifecycle listed in Section 1.1 are presented.

The first phase comprises a customer's acquisition and includes activities related to analyzing the target population using both types of methods. First, target customer analysis uses supervised methods, whereby a specific profitable target group is profiled with respect to available characteristics. Second, customer segmentation is listed, which uses the opposite direction by taking the entire database and dividing it into groups using available characteristics without a predefined grouping but using explorative data-driven techniques, which constitutes a typical unsupervised task.

The second phase is customer development. The main task of direct marketing in this phase is the selection of customers most receptive to a marketing activity, which is mostly operationalized by predicting of the next purchase or the probability of purchasing a certain product. This phase includes unsupervised methods like market basket analysis or sequence mining, as well as supervised methods, such as logistic regression or decision trees.

The last phase includes prevention and retention activities. The former are aimed at customers who have not churned yet but are assumed to. On the other hand, retention activities are directed at customers with a stated cancellation and aim at convincing the customer to reverse his decision and to stay with the company. Within this phase, mostly supervised techniques are applied in order to predict which customers are most likely to cancel their contracts for prevention campaigns and which would be most prone to revise the churn decision.

Although the acquisition is not in the scope of this work as usually no data of non-customers is available, the tasks of customer segmentation and target customer analysis are also relevant for customer development. This is due to the fact that these activities can also serve the goal of cross-selling. Segmentation supports the understanding of the customer base and thereby of potential target groups for a certain product. Given a certain product, an ex-post analysis

of a target group is also a very useful tool for understanding customers, who purchased a certain product, as it will support further planning of cross-selling campaigns. Therefore, this work contributes to supervised and unsupervised aspects of CRM in phases of development and prevention.

2.3 Sequence Aggregation of Purchasing Histories

Observed customer behavior is often a strong predictor of future behavior (Bose and Chen, 2009). Moreover, in certain industries such as financial services (Li et al., 2005) or home-appliances retail (Prinzie and Van den Poel, 2007) a certain logical order of purchases has been found, which can be used to anticipate a customer's need for a next product, and can therefore be operationalized in cross-selling campaigns. In contrast to the established RFM approach (recency, frequency, monetary value), which divides customers into groups (segments) based on the corresponding values of purchasing recency, frequency and resulting monetary value (McCarty and Hastak, 2007), this work conducts a behavioral aggregation of purchasing histories based on customer's purchasing behavior, namely the order of products purchased.

Several streams of research investigate how to efficiently search and identify frequent sequence patterns (Mooney and Roddick, 2013). The goal of these approaches is to identify frequent (sub)patterns. However, these approaches only identify the most frequent patterns but do not group sequences together, and thus do not address the problem of sequence generalization, especially with regard to prediction methods.

The research with predictive focus in the literature as well as the practical side mostly uses one of two approaches. First group considers the purchasing history only and applies methods belonging to the field of recommender systems (for a review see e.g. 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). In the field of recommender systems, Sahoo et al. (2012) employ purchasing sequence-aware models to propose articles for a reader, which is done using Hidden Markov Models. As a disadvantage, these incorporate only the purchasing history, but no additional covariates like age or tenure of a customer.

The second group contains data mining methods, such as logistic regression (McCarty and Hastak, 2007), incorporating a set of predictors including purchasing history as a part of the variables and therefore overcoming the disadvantage

of purely purchasing-history based models. But the challenge in this group is that in this case, the complete sequential information or even binary encoding of portfolio information leads to a so-called high-cardinality feature. Typical examples are postal codes or, as in this case, purchasing sequences. Given the exponential growth of the number of potential purchasing sequences, utilizing "sequence" as a feature in predictive modeling requires preprocessing or aggregation steps to mitigate the high cardinality with potentially millions of categories, as already 100 categories are stated as impracticable within a prediction model (Moeyersoms and Martens, 2015). Thereby, a particular challenge is the grouping of observed "similar" customer sequences, where individual sequences often have a support (number of representatives per sequence type) too low to allow for a reliable prediction of the probability of a next purchase based on observing such as sequence.

Approaches for grouping related sequences to overcome the problem of low support have also been proposed. The most important streams of research can be categorized into proximity-based, feature-based and model-based approaches (Bicego et al., 2003).

Model-based approaches employ a set of models to best describe the clusters, so that each model corresponds to a segment. In the marketing literature, Self-Organizing Maps (Kohonen, 1990), a neural network-based dimensionality reduction method, are often used for the segmentation of product portfolio information (Cho et al. (2005), D'Urso and De Giovanni (2008), Back et al. (2011)). Other established approaches are Hidden Markov Models (HMM), that have been successfully applied for the prediction of sequential information – amongst others – by Netzer et al. (2008) and Sahoo et al. (2012). A shortcoming of this and related methods is, however, the lack of interpretability of the results, as the characteristics of sequences leading to higher conversion rates are not directly revealed.

As this work also uses HMM for detection of temporal patterns, in the following a short discussion of this methodology in marketing research is presented. In the CRM context, the methodology of HMM has been applied to customer purchasing data by Netzer et al. (2008), amongst others, where the latent states (as underlying segments derived from the data as "statistical cause" of the observed sequence) represent the degree of customer satisfaction. Sahoo et al. (2012) employ HMM for the derivation and prediction of the reading preferences in blogs, where the latent states represent the affinity of readers to certain topics.

Schweidel et al. (2011) conducted a dynamic portfolio analysis using HMM. The authors encode complete portfolios as observation states and the latent

states were interpreted as a product affinity. For the purpose of this work, this approach would be impractical due to the vast number of possible portfolios. Although the authors also study the telecommunication industry, the type of products differs from the ones analyzed in this work. Schweidel et al. (2011) analyzed access products such as Internet access, in contrast to hosting related products or services in this work. The important difference between these categories is the fact that access products are independent but hosting products can constitute a logical purchasing order, as they can be built "on top" of each other.

Overall, HMM was shown to be useful for descriptive and predictive analysis of customer behavior in several contexts. However, neither the presence of typical purchasing patterns nor the predictive power of these was investigated for the telecommunications sector.

Feature-based methods for sequence clustering are aimed at finding and constructing features from sequence information to represent sequences more concisely. Approaches have been proposed to transform high-cardinality features into continuous attributes. One of approaches is *weight of evidence*, which assigns a success percentage (in this case CR) to a category and uses this as a similarity measure for aggregation (Moeyersoms and Martens, 2015). In the setting at hand, however, this approach would neither allow for a meaningful interpretation of resulting segments as very different purchasing histories would be mapped to the same value, nor would address the selection bias inherent with supervised techniques.

Some approaches include the number of purchased categories or the last purchasing category as a binary feature (see e.g. Li et al. (2005)). Research by Miguéis et al. (2012) has successfully used dummy-coded aggregation of sequences, but neither a systematic framework nor a theoretical reasoning for parameter selection was provided. Furthermore, Moon and Russell (2008) extracted principal components from portfolio representation for a logistic regression so that the typical product combinations can be retrieved and used for prediction. For instance, Moon and Russell (2008) proposed encoding customer product portfolios. Although a major part of portfolio information is considered with their approach, the temporal order of purchases is not incorporated. Wang and Wang (2007) encoded the sequential information of all customers so that typical sequences can be retrieved and used for prediction. However, the patterns had to be specified upfront by marketing specialists, especially with respect to the timing of purchases.

Finally, proximity-based approaches use a distance measure for sequences, the so-called group of Sequence Alignment Methods (SAM) (Kruskal, 1983), and

have been successfully applied in many analytical tasks (see e.g. Joh et al. (2003) or Tsai et al. (2011)). In the context of marketing, the Levenshtein distance (Levenshtein, 1966) is often used to determine sequence dissimilarity. The measure computes the number of operations (deletion, insertion, substitution) required to transform one sequence into another with subsequent clustering of the resulting distance matrix. 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, which represent the customers by their behavioral pattern (Steinmann and Silberer, 2010). Several studies have shown that the application of clustering based on the Levenshtein distance metric leads to a superior performance compared to other distance measures such as Euclidean distance, in particular when using Ward clustering afterwards (Joh et al. (2003), Murtagh (1983)). The Levenshtein distance, however, does not consider the temporal information in purchasing sequences, i.e., if sequences differ in most recent purchases.

In summary, several techniques exist to cluster similar portfolios and sequences, followed by using the resulting clusters for purchase prediction. The approaches, that will be introduced in Chapter 4, differ from existing approaches in two aspects. First, a higher weight is assigned to more recent purchases by explicitly modeling a decreasing importance of subsequences further in the past. The motivation stems from several studies in various industries, e.g. in the financial industry (Li et al., 2005), or for home appliances (Prinzie and Van den Poel, 2007), that have shown that not only a certain logical order of purchases exist that typically leads to the purchase of a target product, but that the categories of the last purchases have the strongest predictive value. Weighting more recent observation higher is also an established procedure in the context of time series forecasting (e.g. exponential smoothing in Brown (2004)) and regression analysis (Goodwin, 1997). Second, a major drawback of many of the existing techniques is the interpretability of the resulting clusters. Clusters with proposed models can be easily visualized and intuitively linked to specific types purchasing behavior that leads to lower or higher next-purchase probability.

2.4 Churn Prediction in Telecommunications

As the literature review up to this point mostly regarded the area of cross-selling, in the following a short summary of the research directions on churn prediction with a focus on the field of telecommunications is provided.

With annual churn rate estimated at up to 30 % (Tamaddoni Jahromi et al., 2010) and low marginal costs per customer, customer churn prevention and retention is of crucial importance in telecommunications (Kim and Yoon, 2004), and several articles study the determinants of churn risk (Kim and Yoon, 2004; Ahn et al., 2006; Keramati and Ardabili, 2011; Lu, 2002; Zhang et al., 2012). Therefore, preventing customer churn is an important task in CRM, in which the identification of customers with an intention to terminate one or more contracts plays a pivotal role. Identified customers at high churn risk are then subject to CRM activities aimed at avoiding churn.

An extensive survey on churn prediction as well as the closest research to the presented work so far is given by Gerpott et al. (2015), where case studies and the relevant variables of the studies are given. This survey has identified four main direction of research on customer churn: (i) comparative study of predictive performance for different techniques without consideration of variable interpretation, (ii) in-depth analysis of drivers of churn concentrating on interpretation of statistical models, (iii) identification of churn reasons by means of questionnaires, (iv) success of different winback activities. Furthermore, four variable groups are identified for churn prediction, namely (i) contract characteristics, (ii) socio-demographics, (iii) usage behavior and (iv) perception of service or its quality. This work contributes to the research of the first two directions by comparing a performance of different models as well as interpreting the results with respect to the importance of employed variables from the group of contract characteristics.

The study by Gerpott et al. (2015) considers only subscribers of mobile packages bundled with Internet access, whereas this work has both – a broader portfolio of products as well as an extensive feature generation and evaluation based on contract features. In addition, this work studies the impact of product variety in a customer's portfolio on the churn probability, as there is evidence from both, theory and practical experience in other industries, e.g. in the financial sector (Van den Poel and Lariviere, 2004) or the retail industry (Miguéis et al., 2012)), that product variety can be related to loyalty.

One of the primary churn prediction methods is survival analysis (see, for instance, Lu (2002) and Van den Poel and Lariviere (2004)). The palette of the churn prediction methods also includes techniques such as binomial logistic regression (Ahn et al., 2006), random forests (Xie et al., 2009), and neuronal networks (Tsai and Lu, 2009). These classification-oriented approaches do not allow for covariate analysis over time. For this reason, survival analysis is typically preferred in practice for this task. However, survival analysis assumes a proportional, time-

invariant influence of covariates. In telecommunications, for instance, these assumptions are questionable because of existing fixed-term contracts and term of notice clauses. These can be expected to result in non-monotonous cancellation probabilities over time, with increased frequencies of cancellation in time periods before minimum subscription periods end.

Therefore, this work focuses on predicting a customer's churn risk, considering the specific contractual settings in the telecommunications sector, i.e., a presence of a binding contract term. With respect to the contractual setting, a feature generation procedure for modeling each customer's individual contract information is presented. These features are then included in churn prediction model. Consequently, among the research directions for churn management proposed by the survey in Hadden et al. (2007), this work additionally contributes to the literature by analyzing and modeling novel features for churn prediction.

Chapter 3

Modeling of Temporal Patterns Using Hidden Markov Models (HMM)

This chapter investigates the presence of temporal patterns in purchasing sequences for the available data set for products related to information and communications technology. The main hypothesis is that, given such temporal structures in the data, a clear pattern of typical paths or segments should be identified using a statistical model. First, stochastic modeling with Hidden Markov models (HMM) is used for the extraction of temporal patterns in purchasing histories. Second, the correspondence of the resulting patterns to a technological level or need of a customer is investigated. Third, the predictive performance of the resulting segments is evaluated in a product-specific setting.

As introduced before, HMM is a powerful tool in predictive analytics, as it helps to gain insights into behavior of customers by analyzing empirical purchasing sequences. It is possible through identification of typical underlying customer types, which are assumed to represent the "cause" of the resulting behavior. This is done by implicitly assigning latent states to customers' purchasing sequences, which serve as segments and will be profiled for better understanding afterwards. The resulting segments are evaluated for descriptive purposes (segmentation using purchasing sequences) as well as for predictive purposes using out-of-sample evaluation. Within the segmentation the resulting groupings are examined for the correspondence of latent states to a technological level or need of a customer through the analysis of products purchased in a certain state.

3.1 Methodological Background on HMM

For the purposes of this work HMM will be operationalized as a probabilistic approach to determine underlying patterns in observed sequences. In contrast to the so-called visible Markov models, the HMM also includes an unobserved (latent) process based on the observed sequence of events. Formally, a HMM consists of two stochastic processes (Ibe, 2013), one of which is not observable, and is overall defined as a tuple $(A, \omega, X, \Phi, \pi)$:

- $A = a_1, \ldots, a_N$ is a finite set of N latent states,
- $\omega = o_1, \ldots, o_I$ is a finite set of possible emitted symbols I,
- $X = x_{mn}$ is the transition matrix of state-transition probabilities, where each element assigns a probability for the transition of the system from latent state m to latent state n,
- $\Phi = \phi_n(o_i)$ is the emission matrix, where $\phi_n(o_i)$ is the emission probability of the system being in the latent state ϕ_n for the observable sequence symbol o_i,
- $\pi = \pi_n$ is the initial distribution of the states, where π_n is the probability that the system starts in the latent state π_n.

The basic elements of HMM are presented in Figure 3.1. The horizontal line represents the border of the observable world. In the following the mapping of the presented elements of the model to the setting at hand will be introduced. The set of latent states is assumed to be the cause of customers' behavior and therefore to represent purchasing type or related level of technological needs. The probability of the possible switch between latent states, i.e., changes of purchasing behavior or technological needs, are described by the transition matrix. The next element relates to the observable world and is the set of possible observable sequence symbols, in this case the purchased product. The emission matrix indicates the probability to observe a certain sequence symbol given a latent state in terms and is within this context the probability distribution of the purchased products.

There are three fundamental estimation problems related to HMM (Ibe, 2013):

- Evaluation problem: determine how likely it is that the observed sequence $O = o_1, \ldots, o_L$ of length L was generated by a given model (X, Φ, π),

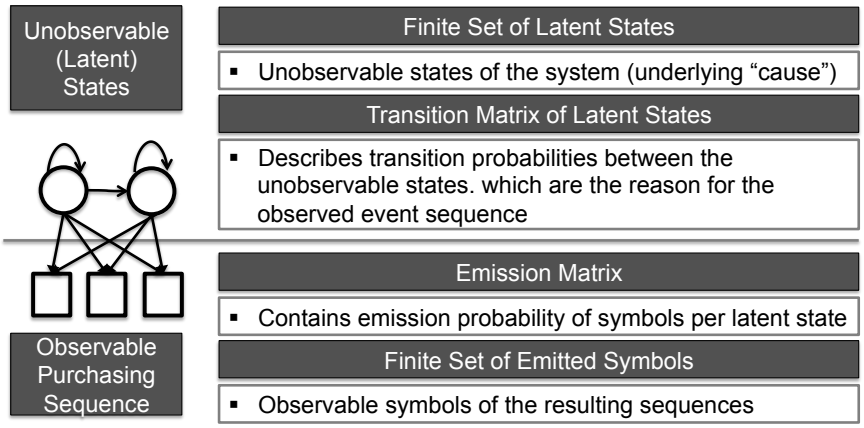


Figure 3.1: Elements of a Hidden Markov Model. The horizontal line represents the border of the observable world. The set of latent states is assumed to be the cause of the order of observable sequence symbols. The transition matrix describes the probability of the possible switch between latent states. The emission matrix indicates the probability to observe a certain sequence symbol given a latent state.

- Decoding problem: determine the most likely sequence of latent states given observed sequence $O = o_1, \ldots, o_L$ of length L and a model (X, Φ, π),
- Learning problem: determine the most likely model given a set of observed sequences.

In this research, the model parameters are estimated first (learning problem) and then the most probable latent state sequence per observed sequence are decoded using the Viterbi algorithm as an established method for decoding problem to derive the underlying sequence of latent states (Forney, 1973). The input parameters for the learning problem are the number of latent states, the initial distribution of latent states, the initial values for the emission and transition matrices (also called initial topologies) as well as the set of possible observable symbols. The number of latent states is often based on prior knowledge. If none is available, as in this case, the number of states is chosen by estimating several models and choosing the one with the best fit. The initial distribution of latent states is assumed to be equally distributed probabilities and the initial values of the emission matrix are set to the product distribution in the training sample.

The most interesting parameter in the case at hand is the initial transition matrix (or topology), as it incorporates additional knowledge about the process by defining the assumed transition probabilities. For instance, that a linear switch between the states is possible, so that all latent states constitute a chain. This

work uses several typical topologies in order to test, which is most appropriate in order to describe the pattern in the data. The four common topologies including the description and interpretation in this setting are depicted in Figure 3.2 and included in the empirical study.

HMM fit is usually assessed using the Bayesian information criterion (BIC) and involves both, measuring the fit to the data by means of log-likelihood and penalizing term for model complexity based on the number of model parameters (Schwarz, 1978). The latter is important as the number of parameters differs significantly for different topologies, while the transition matrix of an ergodic topology requires n^2 parameters to be estimated and the linear topology starts with only $2n - 2$ (Fink, 2014).

First, a descriptive study is conducted, which aims at analyzing customers' behavior on all available data in order to find potential temporal patterns. All models are computed for increasing numbers of latent states $n = 2,\ldots,25$ for all the topologies shown in Table 3.2. The models are evaluated using the BIC. The fitted parameters from the best model are then analyzed and profiled using additional customer information. The goal of the descriptive analysis is the identification of typical development paths and the interpretation of the resulting latent states. Predictive analyses are conducted by means of cross-validation for product specific data sets described later in Section 3.2. Using the product-

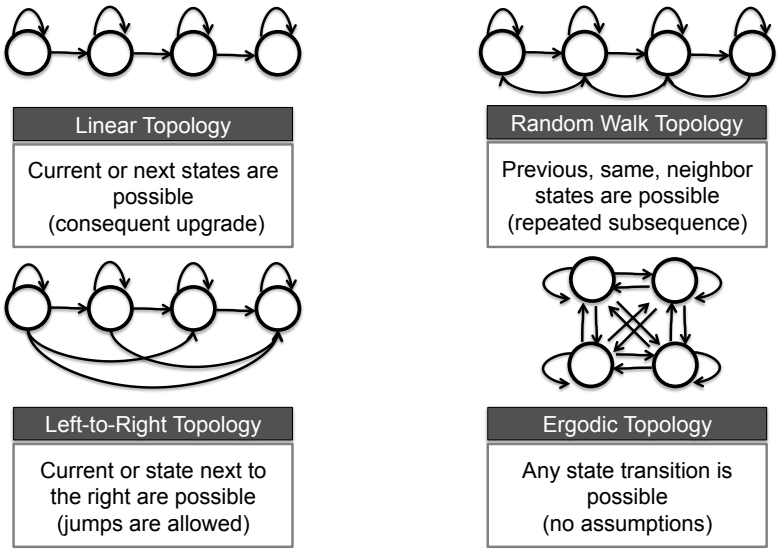


Figure 3.2: Types of latent state transitions (topologies). The figure illustrates the most common topologies for HMM, gives a short description of possible state switches and maps the topologies an interpretation in the context of this study (denoted in brackets).

specific model fitted on the corresponding training data including the target product, the sequences of the latent states are estimated for each test sequence (decoding problem). Then, the test sequences, not containing the information if the target product was purchased, are assigned a purchasing probability for the target product using transition and emission matrices. Overall, a model is trained for each combination of product, topology and number of states. The R-package *hmm.discnp* (Turner and Liu, 2009) was used for both parts of the empirical study. The resulting prediction accuracy is measured by means of lift (Piatetsky-Shapiro and Masand, 1999).

3.2 Available Data

The sample contains customers' purchasing histories, which belong to a particular regional market in order to ensure the same product portfolio as well as temporal dynamics, as the company operates on several international markets. For these customers the contract history for a multi-year period up to August 2016 is available, including information on the product category of contract and purchase dates. The products (or services) purchased belong to 10 different categories in the realm of hosting products, which are labeled Product 1 to Product 10 due to reasons of confidentiality with 96.885 sample customers having an average sequence length of 3.7.

The basic hypothesis is that latent states relate to a certain technological level or need, therefore a qualitative description of products is presented in Figure 3.3, where these are profiled in two dimensions: price and technological level of the products. These products are separated into several groups with rising technological level and also with the rising price of the product.

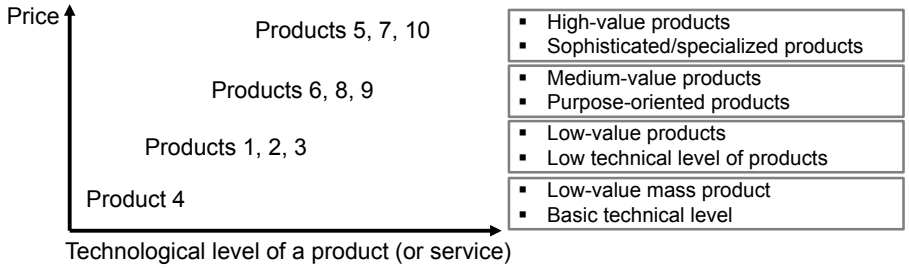


Figure 3.3: Product hierarchy - qualitative description of products. The products can be separated in four groups with respect to the technological level and the price starting from low-value mass product to the group of specialized high-priced products.

Specific data sets for predictive evaluation are prepared by splitting the purchasing sequence at the first purchase of a specific product and treating the product of the respective last purchase as target product. This is done as the goal is to predict the first purchase of the respective product. An example of the product-specific data set derivation is presented in Figure 3.4, which shows how a particular purchasing sequence is prepared for different data sets, where P1 stands for Product 1 etc. As Product 1 is the one purchased first, the corresponding sequence is not used for the data set of Product 1. The example sequence is not considered in the data set of Product 3 as it would lead to a preceding sequence length of one product, which does not constitute a purchasing path. Also the example sequence is included in the data sets for all remaining products not contained in it as one belonging to the class of "non-purchasers".

3.3 Descriptive Results

The BIC resulting from each model fitting based on the complete sample is presented in Figure 3.5. The Random Walk topology is outperformed for all state numbers and will therefore not be further considered. Other models seem to be comparable considering BIC. The curve converges at approximately 20 states. However, cross-validation for predictive results showed that using at most 10 latent states is optimal (see Table 3.4), meaning that more states lead to overfitting. Therefore, a model having ten or less states based on relative BIC gain per additional state and separation quality of latent states with respect to emitted

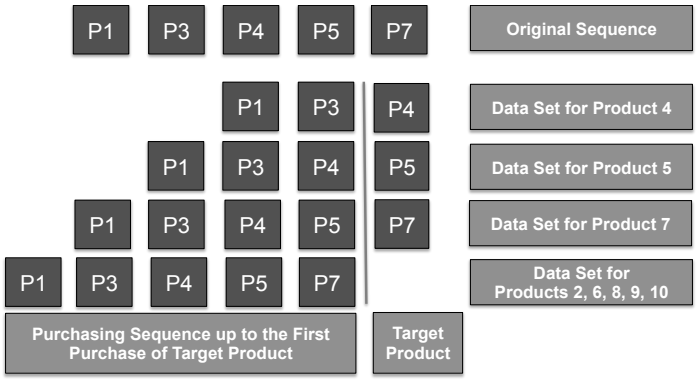


Figure 3.4: Example of sequence preparation for product-specific data sets. As an example for a single sequence is shown how it is transformed for each product-specific set. As Product 1 (P1) is one purchased first, the corresponding sequence is not used for the data set of Product 1.

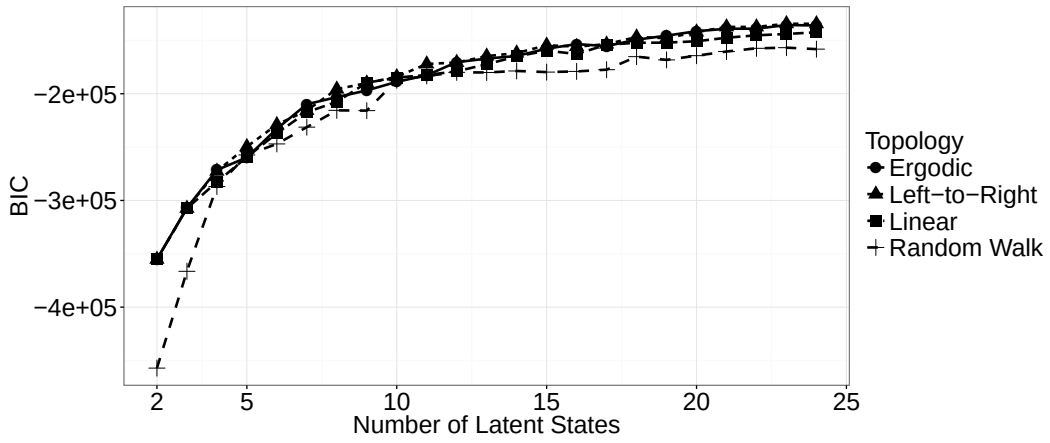


Figure 3.5: BIC-based descriptive model selection. The Random Walk topology is outperformed for all state numbers and will therefore not be further considered. Other models show comparable results considering BIC.

products was chosen. The final choice was an ergodic model with $n = 6$ as the ergodic topology performed best for most products. A higher number of states led to redundant states or states emitting almost all products, therefore BIC only was not sufficient for the choice of the descriptive model in this case but required a detailed analysis of all models.

First, the transition matrix of the fitted model is analyzed. Transitions with probability of at least 1 % are displayed in Figure 3.6, ordered by the state stability (percentage of the customers residing in the same state). Interestingly, there are two "pseudo-sink" states that are rarely left, with State 1 as an extreme example. On the contrary, State 2 seems to be left quite quickly. States 2, 4, 5, 1 seem to constitute a typical path, as they were ordered regarding the most probable next state. Another possible path is the one taken from State 6 to State 4, which 43 % of customers take.

The next element of the fitted model is the emission matrix shown in Table 3.1. The table contains the purchasing probability of a product given a latent state. All products having an emission probability of at least 5 % are printed bold. Furthermore, a potential interpretation of the latent state is provided based on the products purchased in the corresponding latent state.

Although State 1 and State 2 have the same exclusive emission probability for Product 4, the difference can be seen in Figure 3.6. Here, State 2 commences a customer development, in contrast to State 1, which is left in less than 1 % of cases and is considered a low-value end state. Remarkably, State 3 leads to purchases of high-value and technologically sophisticated products or services, as it

Table 3.1: Emission matrix for an ergodic topology and six latent states with a potential interpretation. A clear division of emitted products into categories corresponding to Figure 3.3 can be seen. For instance, State 3 is the only one emitting a considerable number of high-value products.

Latent State Description	Stable State of Low-Value Purchases	Promising Low-Value Purchases	High-Value Service Solutions	Middle-Value Service Solutions	Customers with Typical Bundle Solution	Customers with Middle-Value Solution
Product	State 1	State 2	State 3	State 4	State 5	State 6
Product 1	0.01	0.00	**0.20**	0.00	**0.53**	**0.99**
Product 2	0.00	0.00	0.01	**0.72**	0.00	0.00
Product 3	0.00	0.00	0.02	**0.14**	0.00	0.00
Product 4	**0.99**	**1.00**	**0.11**	0.00	**0.47**	0.00
Product 5	0.00	0.00	**0.34**	0.00	0.00	0.00
Product 6	0.00	0.00	0.01	**0.07**	0.00	0.00
Product 7	0.00	0.00	**0.24**	0.00	0.00	0.00
Product 8	0.00	0.00	0.02	0.00	0.00	0.01
Product 9	0.00	0.00	0.00	**0.06**	0.00	0.00
Product 10	0.00	0.00	**0.05**	0.00	0.00	0.00

includes Products 5, 7, 10, which are moreover purchased in a substantial number of cases only in this state. This seem to be the state, supporting the hypothesis of the relationship between the latent state and the technological level (or need) of a customer as no other state contains a considerable purchasing probability for this type of a product. In a similar manner, State 4 is responsible for all purchases of middle-value services and products. State 6 is similar to State 2 as it is relatively unstable, emits a rather low-value product and leads to further more valuable latent states. State 5 constitutes a typical bundling combination of products, which is relatively stable and has low to medium value.

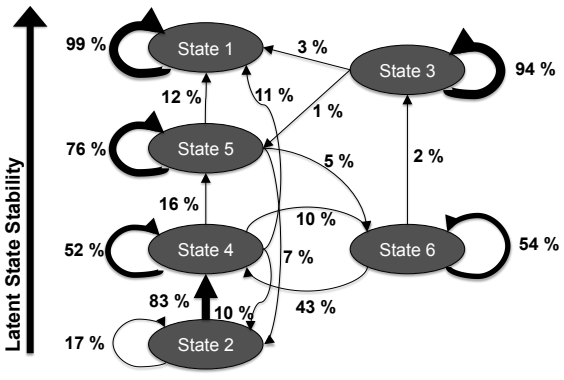


Figure 3.6: Transition matrix for the ergodic topology with six latent states. A clear presence of paths is shown in the data. For instance, States 2, 4, 5, 1 seem to constitute a typical path.

Table 3.2: Product distribution over complete sequence given final latent state in percent. At the end of a customer's sequence, no customer is in State 2, therefore no distribution for this state is shown. For each state the three products with the highest percentage are marked bold.

Product	State 1	State 3	State 4	State 5	State 6
Product 1	**6.03**	**25.90**	**22.01**	**63.05**	**72.33**
Product 2	**3.41**	1.63	**43.24**	**0.97**	**11.79**
Product 3	1.25	2.18	6.11	0.83	5.8
Product 4	**88.01**	13.07	**21.38**	**34.66**	**6.14**
Product 5	0.21	**29.38**	0.03	0.06	0.16
Product 6	0.64	0.85	3.22	0.30	2.57
Product 7	0.20	**19.85**	0.04	0.04	0.05
Product 8	0.05	2.63	0.20	0.05	0.68
Product 9	0.15	0.22	3.75	0.03	0.49
Product 10	0.04	4.28	0.02	0.00	0.00

For the next evaluation the last estimated latent state (final latent state) per sequence is shown in Table 3.2. Rows of the table represent the product and the columns denote the corresponding relative percentage of purchased products per latent state. The difference to the emission matrix in Table 3.1 is that Table 3.2 presents the distribution reflecting all purchases done by the customer from all latent states the customer has been in, including the final one. For each state the three products with the highest percentage are marked bold. State 2 is not contained in the list of the final states. As seen in Figure 3.6, it is the unstable state, which is rather a beginning of a customer's path and is left by all customers at the end of the observation period. This fact can be interpreted as the initial state, which is left quite quickly after a customer has joined the company.

Table 3.3 shows another result of the descriptive evaluation with additional customer data, e.g. the overall revenue with the company, which was not used for the estimation of the model. The purpose is to characterize the resulting latent states by external information in order to get a better interpretation of the states and additionally to illustrate the differences between latent states with respect to the hypothesis of the technological level correspondence of latent states. Due to reasons of confidentiality the absolute values are not shown. Instead, the numbers in the table show the relative ratio per variable, i.e., mean for the customers of a specific final latent state from Table 3.2 divided by the mean computed over all customers (with 1 corresponding to the average level over the all customers). Highest and lowest values per row are highlighted in bold. Additionally, the percentage of customers per final latent state is shown.

The customers in State 3 are typically those with the highest tenures and by far highest revenue; however, this is the case only for 5 % of the sample. Also, the

Table 3.3: Profiling of the segments by external data given the final state. The numbers in the table show the relative ratio per variable with 1 corresponding to the average level over the all customers. Highest and lowest values per row are highlighted in bold. State 3 contains customers with the highest revenue, required support level and number of active domains, which can be interpreted as professional Internet presence.

Category	Variable Name	State 1	State 3	State 4	State 5	State 6
General	Percentage of Customers	35 %	5 %	30 %	24 %	6 %
	Tenure (in months)	0.90	**1.21**	**0.89**	**1.21**	1.12
	Average Revenue (12 months)	**0.58**	**5.03**	0.89	0.81	1.19
Support	Number of Support Cases	0.88	**1.82**	1.03	**0.85**	1.11
	Total Duration of Support Calls	**0.76**	**1.70**	1.23	0.77	1.06
Contracts	Active Contracts	**1.69**	0.69	**0.57**	0.65	0.73
	Number of Upgrades (12 months)	0.89	**1.40**	1.26	**0.72**	1.23
	Number of Downgrades (12 months)	**1.69**	0.71	1.55	**0.36**	0.42
	Active On-Features	1.26	**2.09**	**0.57**	0.88	0.92
Domains	Number of Active Domains	1.10	**3.02**	**0.55**	1.03	1.16
	Number of Redirection Domains	1.16	**2.02**	**0.68**	1.03	0.92
	Average Number of Words	**0.78**	1.26	1.03	1.14	**1.36**
	Average Number of Subsites	**0.65**	1.10	1.15	1.23	**1.27**

support level increases with the technological level of the products. Regarding the number of contracts, State 1 has the highest level, which is mostly driven by multiple purchase of Product 4. Furthermore, State 3 seems to be very prone to upgrading to a contract with higher value within the same product or service and a downgrading rate below the average compared to States 1 and 4, which downgrade more often. The highest relative number of additional in-product (on-features) purchases such as additional space for the same product or service is also the highest for State 3. The last two profiling blocks contain information about crawled domains from the customer sample. Customers of State 6 seem to have mostly content-intensive sites, e.g. blogs, and customers of State 3 the highest number of related domains suggesting intensive Internet presence, e.g. a company with several re-direction links.

Overall, the model estimates, the sequences corresponding to the latent states as well as external data have shown the presence of typical purchasing patterns. The interpretation of the latent states has shown a correspondence of the products purchased therein to the potential technological level of a customer. The resulting segmentation can be used to identify the types of customers and to derive a marketing strategy for each segment.

3.4 Predictive Results

So far, the behavior of customers based on the complete sample with the goal of identification of customer types was investigated. In the following, the results of a 10-fold cross-validation aimed at predicting the next purchases are presented. Simultaneously, gaining predictive performance would confirm the initial hypothesis regarding the presence of certain purchasing patterns among the customers, which can be used for next-best-offer marketing campaigns. First, the best model per product is displayed in Table 3.4, sorted by the level of products with respect to both, price level and technological complexity. In most cases using an ergodic topology shows the best results. Interestingly, for Product 7 and 9 as target product, assuming a stepwise upgrade to the next higher product, linear topology shows better results. In contrast, the path leading to the purchase of Product 1 rather follows a random walk pattern, which can be interpreted such that this product or service is purchased in different orders and does not logically follow a particular prior purchasing sequence.

A further evaluation regarding prediction quality of the resulting latent states by means of the lift is shown in Figure 3.7. The curve depicts average performance over ten folds for a certain percentage of customers chosen by sorted model outcome (campaign size).

The product order in Table 3.4 corresponds to the decreasing performance in the lift chart. High-level products allow for better prediction. This can be explained by longer sequences preceding the purchase of these products and is evidence for the presence of purchasing patterns allowing for such prediction. With Product 3 as an exception the trend of decreasing predictive performance continues into the category of low-level products or services. For low-level products, especially Product 4, there was little to almost no predictive power of the best model. The reason is exactly contrary to high-level products – Product 4 is mostly purchased in the beginning of a sequence and further purchases do not

Table 3.4: Best model per product determined via 10-fold cross-validation with respect to topology and number of latent states. The products are grouped corresponding to their technological level.

High Level			Middle Level			Low Level		
Product 10	Ergodic	n = 10	Product 8	Ergodic	n = 7	Product 3	Ergodic	n = 7
Product 5	Ergodic	n = 8	Product 6	Ergodic	n = 7	Product 1	R. Walk	n = 6
Product 7	Linear	n = 8	Product 9	Linear	n = 8	Product 2	Ergodic	n = 6
						Product 4	Ergodic	n = 10

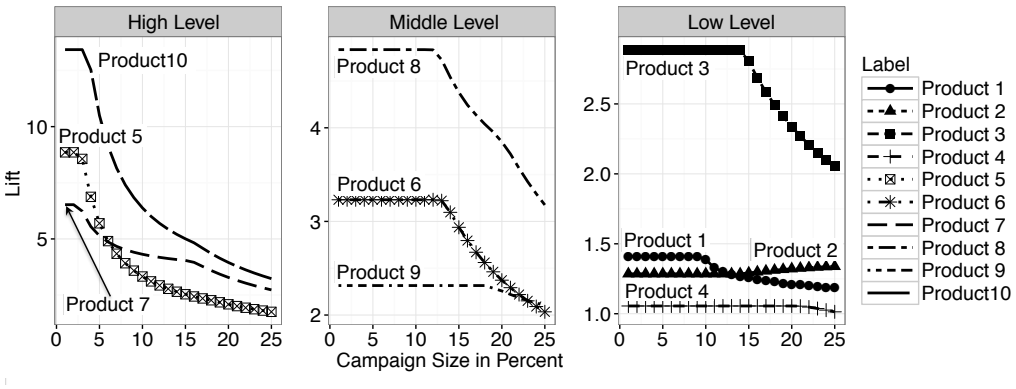


Figure 3.7: Lift chart shows the mean performance over ten folds by product and its category. For high-level products a higher lift, whereas for low-level products a prediction is rather difficult as the lift value is close to random selection, i.e., to the value of 1.

seem to follow a regular pattern. A later increase for Product 2 demonstrates a non-optimal selection of customers, as the segments in the higher campaign size range expose higher lift values, which should not be the case.

3.5 Summary

In this chapter the application of HMM to purchasing sequences of customers in the telecommunications sector was shown. The results allow for a representation and an interpretation of customer behavior. Thereby, common purchasing patterns were identified, which can be related to technological development paths. Using the purchasing sequences, transition and emission matrices, it was possible to interpret the resulting latent states, which is often considered a drawback and major challenge when using HMM. In addition, the results of the descriptive model analysis are confirmed and extended by profiling using additional customer data. Although ergodic topology as an initial parameterization was beneficial for most products, the most suitable model topology and the number of states varies across products.

With respect to the research question, the presence of typical purchasing paths was shown and interpreted. The correspondence of the latent states to the technological level or need of a customer was shown by using the purchasing sequences, model parameters and external profiling data. Furthermore, out-of-sample evaluations revealed the presence of the predictive value of the determined segments, in particular for high-value products. Considering predictive performance, segmentation with a low number of states based only on purchas-

ing history is not sufficient for detailed prediction as it builds relatively large segments. However, in this case even such segmentation was sufficient and delivered predictive performance for some products.

Regarding the managerial impact, the introduced methodology can be used to analyze customer segments based on purchasing data and for building strategies regarding campaign assignment, i.e., targeting. For instance, customers in a latent state with low value and low probability of switching to more profitable states should be excluded from marketing activities and in contrast typical development paths can be used to select customers cross-selling campaigns.

Chapter 4

Methods for Aggregation of Purchasing Histories

This chapter provides four insights. After a short motivational section, where the value of temporal information is once again demonstrated for the available data from industry of information and communications technology, two aggregation mechanisms for sequences are presented and evaluated in the context of the next purchase prediction. The respective bias–variance trade-off for lift is explicated next. The third aspect considers marketing-relevant extensions for the second mechanism, e.g. for incorporation of bundling information or sequence length. The last section regards an ex-ante estimation of the sequential information in order to anticipate the potential predictive performance of sequential information based on raw data, before an extensive modeling is conducted.

4.1 Preliminary Evaluation of Temporal Information

The problem of the sequence or portfolio representation is that with rising number of products, the number of possible portfolios increases exponentially. As a result, the groups become too small for robust prediction from a statistical perspective and too small to form relevant segments for marketing activities from a practical perspective. Therefore, there is a need for an aggregation mechanism.

In line with findings in other industries, tests on the available telecommunications data set indicate a decreasing predictive value of older purchases, which will be shown in the following. As demonstrated in Chapter 3, a temporal pattern is present within the sequences. However, the results of HMM display the presence and the paths, but not the exact value of each position in the sequence.

Table 4.1 shows the in-sample CR when considering a combination of the last three purchases for predicting the next one of an arbitrarily selected target prod-

uct. As expected, the segmentation using the last purchase outperforms any other combination of past purchases with respect to the resulting CR.

Further consideration of the older purchases in the past has additional predictive value, but the difference from considering another purchase declines successively. Therefore, the purchasing sequence shows a clear temporal structure. Capturing this structure is the goal of the two approaches for sequence segmentation and prediction introduced in this chapter.

As shown in Table 4.1, the temporal order has a clear but decreasing information value. This chapter proposes two unsupervised mechanisms to group customers to generalized purchasing types. Both techniques project and aggregate raw sequence data-based on the intuition that (i) there is predictive value in the whole order of purchases but (ii) the predictive value of a product purchase increases with its recency. Discounting the weights of past purchases addresses the temporal trade-off between only considering the most recent purchases and the whole sequential information. The overall goal is to aggregate sequences most similar in more recent purchases. The results of such aggregation are then shown in an evaluation of predictive performance in terms of lift for the context of cross-selling.

As a first approach, the Sequence Set Model, is introduced that consecutively relaxes the sequential order of purchases by hierarchically grouping samples that differ primarily for older purchases. The second approach, Weighted-Productspace Clustering, assigns descending weights to purchased products according to their sequential order, projects the resulting weighted vectors into the product space (a continuous space spanned by the product dimensions), and then applies distance-based clustering. Segments are then targeted according to the model prediction up to a certain campaign size. The bias–variance trade-off

Table 4.1: Conversion rate for the best 1 % and 10 % of customers due to annotated segment based on a combination of previous purchases. Rows are sorted by conversion rate (CR) per segmentation type. The numbers in brackets state the difference to the next best approach. Considering the last purchase only (3rd row) outperforms any other combination of past purchases. However, considering of older purchases delivers an added value. The baseline CR when randomly selecting customers is 8.8 %.

Segmentation Feature	CR Best 1 %	ΔCR Best 1 %	CR Best 10 %	ΔCR Best 10 %
Combination of Three Last Purchases	28.4 %		14.6 %	
Combination of Two Last Purchases	24.3 %	(4.1 %)	13.8 %	(0.8 %)
Only the Last Purchase	18.4 %	(5.9 %)	12.9 %	(0.9 %)
Second and Third Last Purchase	15.7 %	(2.7 %)	11.1 %	(2.0 %)
Only the Second Last Purchase	13.6 %	(2.1 %)	10.6 %	(0.5 %)
Only the Third Last Purchase	12.8 %	(0.8 %)	9.6 %	(1.0 %)

is discussed for application of the proposed grouping mechanisms from a statistical perspective. The empirical evaluation for cross-selling campaigns is conducted on over 230 000 purchasing sequences available in the sample company's database and products.

4.2 Formal Definition of Aggregation Models

This section contains a detailed motivation for each aggregation model, the respective formal definition and a short example for the application for both models. Additionally, after an introduction of the both methods is given, these are compared with respect to the expected strengths and weaknesses.

4.2.1 Sequence Set Modeling (SSM)

The first proposed model unifies two perspectives of sequence aggregation. The first perspective uses a truncated sequence: the last purchase, two last purchases, etc. as a separate category, for which a relative occurrence of the target purchase is computed and evaluated. A similar idea was used by Miguéis et al. (2012). The distinctive feature of the model proposed in this work is the possibility to vary the proportion of sequential information as it allows for a variable combination of sequential information with the second perspective of portfolio. Thereby, 1 is assigned to a product category if a certain product was purchased by a customer, and 0 if it was not. The portfolio is a binary representation of product categories, as used by Boztuğ and Reutterer (2008).

These two perspectives, truncated sequence and portfolio, are unified within the proposed model. Overall, each customer is assigned a combination of both views encoding the respective possession of products. Sequential information is in this way added successively to the portfolio by creating a category, e.g. for last purchase, and considering the portfolio of older purchases. The number of purchases into the past can be varied, until the complete sequence up to a certain length is considered, which represents complete sequential information for this sequence length. This idea is formalized in the following.

Formal Definition

Before the Sequence-Set-Model is developed, the notation will be introduced. 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 prior to the target product.

Let further $S_{cL} =< s_{c1}, \ldots, s_{cl}, \ldots, s_{cL} >$ denote a sequence S of products s_{cl} of length L purchased by a customer c, 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 oldest product purchase stored in the vector.

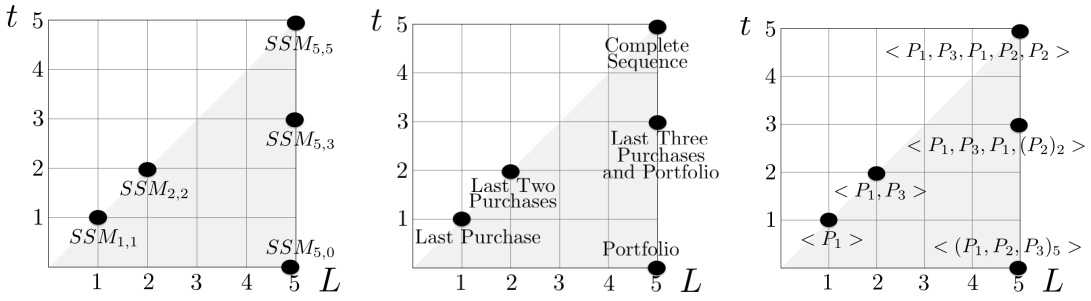
Considering the dummy product P_0, for each customer a purchasing sequence S_{cL} of length L can be assigned by filling-up the empty fields in the vectors from right to left with P_0 if less than L products have been purchased. For instance, assuming $L = 5$ and a customer c purchased the products P_5 and P_1 only, with P_5 being his last and second to last product purchase, then his resulting purchasing sequence would be $< P_5, P_5, P_1, P_0, P_0 >$. If a sequence has more than L elements, prior purchases are ignored.

An element in a sequence can also contain an unordered item set $n\text{-}set = (\{P_i\})_n$, $i \in \{0, \ldots, I\}$, with $n \in Z_0^+$, 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 $< P_5, (P_1, P_2)_2 >$. The following sequences are instances of the type: $< P_5, P_1, P_2 >$ or $< P_5, P_2, P_1 >$, while $< P_5, P_2, P_3 >$ is not of the defined sequence type.

With the two parameters L, indicating the total length of a sequence and a threshold parameter $t \in Z_0^+$, $t \leq L$, SSM is then defined as a product purchasing sequence type as shown in (4.1). Thereby, parameter t indicates the number of the most recent purchases where the models demands a mandatory order, while for the remaining $L - t$ product purchases no order is required.

$$\begin{aligned} SSM_{L,t} &= < S_{ct}, n\text{-}set > \\ n &= L - t \end{aligned} \tag{4.1}$$

The relationship among the different types of models is illustrated in Figure 4.1. 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 older purchases to obtain a less restrictive grouping of sequences while still capturing sequence information to a predefinable 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



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

Figure 4.1: Definition of SSM with different parameterizations (left-hand side), interpretation of models parameter combinations (center), and example instantiations of the model 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.

an exact sequence of the last L purchases. As an example, a model with $L = t = 2$ would group together the sequences $< P_1, P_2, P_4 >$ and $< P_1, P_2, P_6 >$, while a model with $L = t = 3$ would derive two different segments from the sequences.

Example of Segmentation with SSM

In summary, the two parameters for aggregating the purchasing history are the completeness (length of the sequence considered as parameter L) and the degree of the temporal information contained (parameter t). The completeness is reflected by the number of products aligned along the timeline of the purchases that are examined, for example the last one, the last two purchases and so on. The second dimension regards the temporal order of the purchases. If no temporal order is considered, but only the possession of certain products, then the representation is a portfolio. The other extreme is using all sequence information by creating a category per sequence instance, i.e., the complete sequence constitutes a category.

The complete sequence contains all available sequence information, whereas the last purchase captures the least. However, the trade-off is always that the sequence is more specific, but it comes at the cost of a smaller target group and higher variance. The goal of evaluating the discrete model is to assess the amount of information captured in each additional information step. Different cases are demonstrated in Table 4.2, where aggregation levels outperform others due to generalization or specialization of sequence information. To show the

impact for the sequence of three or more products, the example of preceding sequences $< P_9, (P_i)_n >$ for the target purchase of Product 2 is chosen. Each of the patterns can also have also further purchase prior to the denoted ones, which are not shown but are implied in respective aggregation.

For a sequence $< P_9, P_1, (P_i)_n >$ the generalization by means of "Last" ($SSM_{1,1}$) is better than a more detailed sequence, as it delivers one of the highest conversion rates as well as the highest number of customers in the segment. A more detailed view of sequence information has a decreasing impact on the CR (for example $SSM_{2,2}$).

As the next example, the combination of Products 9 and 4 preceding the target purchase of Product 2 is investigated. Considering only the last purchase, Product 9 is a strong predictor for the purchase of the considered product with CR of 18.56%, as the CR is about twice as high as the baseline. The preceding purchase of Product 1 before Product 9 reduces the probability of a subsequent purchase.

On the contrary, a recent acquisition of Products 4 and 9, with Product 9 being the last increases the target purchase probability about four times above the conversion rate. An additional increase in CR is brought by the aggregation to portfolio. The portfolio contains all combinations of Products 4 and 9 within a sequence without any other products, whereas $SSM_{1,1}$ comprises also sequences with other products in the elder part of it. The interpretation is then that the pure sequence consisting of Products 4 and 9 is better than one containing any others. In this example also a trade-off between the increase in conversion rate and decreasing aggregation possibility of the groups is shown: $SSM_{1,1}$ has the most customers but a lower CR, corresponding to a high bias of the model.

Overall, the results presented in this section show the suitability of the proposed SSM approach to: (i) draw first conclusions about the importance of pre-

Table 4.2: Example of SSM application to available data. The resulting segmentation using the denoted model can be used for both, interpretation and selection of customers for product offers based on the purchasing history. The second column describes one segment taken from the model to illustrate clear differences in CR for different aggregation levels. With a segment $SSM_{2,2} =< P_9, P_4 >$ the highest CR is achieved.

Notation	Model Segment Description	Number of Customers	Conversion Rate
$SSM_{1,1}$	Product 9 as Last	3 125	18.56%
$SSM_{2,2}$	9-1 as Last	1 956	13.39%
$SSM_{5,0}$	9-1 Portfolio	1 980	13.28%
$SSM_{2,2}$	9-1 as Sequence	1 708	13.76%
$SSM_{2,2}$	9-4 as Last	611	35.51%
$SSM_{5,0}$	9-4 Portfolio	549	39.16%
$SSM_{2,2}$	9-4 as Sequence	442	43.21%

ceding product purchases, (ii) evaluate a potential impact of sequence information by different levels of segmentation, (iii) systematically analyze the best aggregation level by means of the lift.

4.2.2 Weighted-Productspace Clustering (WPC)

The introduced discrete approach SSM allows segmenting of sequences on any discrete level of portfolio or sequence information. In the following an approach to translate the presented idea into a continuous product space with the possibility of analogous aggregation but with an additional ability to control the degree of the importance of purchases further in the past. Introducing an exponential discounting factor does this. Geometrically descending weights are a widely-used technique to model a discounted importance of observations, for instance in exponential smoothing methods (Brown, 2004) or in time series forecasting (Goodwin, 1997). Exponential smoothing was also shown to be successful in the context of recommender systems (Dunlavy et al., 2011).

The second method consists of two steps: (i) projection of customers' purchasing sequences into the so-called product space with diminishing weights towards older purchases and (ii) consecutive application of a clustering algorithm on the transformed data in order to derive the segments.

Formal Definition

The second approach proposed is referred to as Weighted-Productspace Clustering (WPC). With WPC, geometrically descending weights are assigned to the elements in each purchasing sequence vector S_{cL}, thereby weighting more ancient product purchases less than more recent ones. With parameter $\lambda \in R_0^+ \leq 1$ as discount per purchase in a vector, the weight of a product at position l in S_{cL} is calculated using (4.2).

$$w_l = \lambda^{l-1} \tag{4.2}$$

The 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 customer c, the total weight W_{ic} of P_i in S_{cL} is computed as shown in (4.3).

$$W_{ic} = \sum_{l=1}^{L} b_{icl} w_l, b_{icl} = \begin{cases} 1 & \text{if } s_{cl} = P_i \\ 0 & \text{else} \end{cases} \tag{4.3}$$

The resulting vector of product weights $D_{cL} = (W_{1c}, \ldots, W_{ic}, \ldots, W_{Ic})$ for customer c is the resulting projection into product space – the I-dimensional space spanned by the offered products – where each vector W_{ic} serves as coordinate along the dimension associated with product i.

Figure 4.2 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 $S_{1,3} =< P_2, P_2, P_2 >$, $S_{2,3} =< P_3, P_2, P_1 >$, and $S_{3,3} =< P_3, P_2, P_2 >$. 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 highest if P_i is the product purchased most recently, and decreases with the position of a product purchase 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. $\lambda = 1$ sets the coordinates of all purchased products to 1; hence, it does not project sequential but only portfolio information. With decreasing λ, coordinates increase less for older purchases so that that sequential information is captured in the respective data point. For example, with $\lambda = 0.99$, the difference in weighting of the last and the fifth to last purchase is only $1 - 0.96 = 0.04$, while the coordinate change for the older prod-

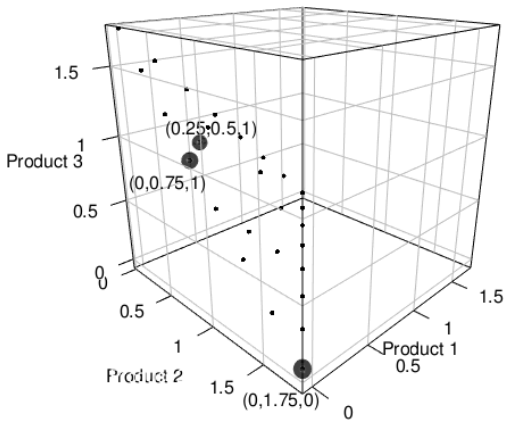


Figure 4.2: Illustration of all possible sequence projections consisting of three products using $L = 3$ and $\lambda = 0.5$. Labeled points represent the sequence $S_{1,3} =< P_2, P_2, P_2 >$ with $D_{1,3} = (0, 1.75, 0)$, $S_{2,3} =< P_3, P_2, P_1 >$ corresponding to $D_{2,3} = (0.25, 0.5, 1)$ as well as $S_{3,3} =< P_3, P_2, P_2 >$ corresponding to $D_{1,3} = (0, 0.75, 1)$.

uct purchases approaches to 1 with λ approaching zero.

In WPC, distance-based clustering is then applied to segment the D_{cl} data points in product space. Hence, clusters represent similar sequences considering a diminishing importance of older product purchases, adjustable by λ. Although WPC is not restricted to a specific clustering technique, in this work WPC is implemented with k-means clustering, a widely used distance-based clustering method (MacQueen et al., 1967). This is motivated by the fact that the temporal information is reflected in the distance between data points, in line with the spirit of the sequence projection approach. In addition, k-means has the advantage that, in particular for smaller values of k, segments will follow the empirical density of data points in the product space as it will tend to locate cluster centers where many observations are located, therefore enforcing high average support per segment. Finally, the marketing division of a company can set k, the number of segments, to a reasonable and manageable number. WPC is denoted with its parameters as $WPC_{\lambda,k}$.

It is worthwhile to notice that – in contrast to SSM – a centroid derived by WPC does not necessarily represent one or more empirical sequences, but rather a common characteristic of purchasing sequences. Cluster centers are strongly related to purchasing behavior; the higher the coordinate the more recently a product has typically been purchased in a cluster. 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. Finally, the determined segments are ordered by conversion rate in descending order and the segments are targeted until reaching the campaign size, e.g., 10% of the customer base.

Example of Segmentation with WPC

In the following an example for WPC is presented in Table 4.3. The result of the application for the same product as in Table 4.2 to provide comparability. The available data was projected with $\lambda = 0.5$ and clustered with $k = 10$, the table displays the resulting centroids.

For instance, cluster two represents a sequence that solely consists of P_6 as last purchase without previous purchases. This cluster can be directly associated with a real purchasing sequence. For instance, cluster with ID 3 and the highest lift of 18.6 %, has a centroid with coordinates above zero for $P_9 = 1.02$,

Table 4.3: Centroids resulting from a WPC computation using $k = 10$ and $\lambda = 0.5$. The best performing cluster with ID 3 is similar to the one obtained by SSM as it contains Product 9 as the last purchase and Products 1 and 4 as previous to last purchases. The numbers equal to or over 0.1 are marked bold.

ID	Count	P1	P3	P4	P5	P6	P7	P8	P9	P10	CR
1	328 431	**1.00**	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	9.07 %
2	1 811	**0.14**	0.01	0.04	0.09	0.00	0.07	0.00	0.00	**1.10**	4.31 %
3	3 127	**0.35**	0.01	**0.11**	0.00	0.02	0.00	0.00	**1.02**	0.00	18.58 %
4	1 420	**0.34**	0.05	0.05	0.02	0.01	0.01	**1.05**	0.00	0.00	8.94 %
5	9 413	**0.16**	0.01	0.08	0.00	**1.03**	0.00	0.00	0.00	0.00	14.87 %
6	27 166	0.07	**1.04**	0.05	0.00	0.00	0.00	0.00	0.00	0.00	10.42 %
7	7 138	**0.12**	0.01	0.04	**1.22**	0.00	0.04	0.00	0.00	0.02	2.54 %
8	4 695	**0.20**	0.01	0.07	0.04	0.01	**1.11**	0.00	0.00	0.01	2.90 %
9	317 512	0.04	0.00	**1.13**	0.00	0.00	0.00	0.00	0.00	0.00	8.32%
10	34 660	**1.57**	0.01	0.03	0.00	0.00	0.00	0.00	0.00	0.00	9,12%

$P_1 = 0.35$, $P_4 = 0.11$. Reconsidering that $\lambda = 0.5$ means that product purchases are weighted descending with weight 1 for the first, 0.5 for the second, 0.25 for the third, 0.125 for the fourth, and 0.0625 for the last entry, sequences grouped together for cluster with ID 3 therefore have the following characteristics. P_9 is the most recent purchase. Furthermore, sequences in cluster three typically have P_1 at position two or three and P_4 between position three and four. The table also shows that for each cluster, centroids typically have between one and three coordinates exceeding 0.1, while the remaining coordinates are often zero or close to zero.

Reconsidering SSM results presented in Table 4.2, the clusters found by WPC correspond to the best performing sequences obtained by SSM but as a result of an automatic procedure and not an extensive search through different aggregations.

4.2.3 Comparison of Proposed Aggregation Methods

The presented WPC approach allows for differentiation of the customers given their previous behavior, similar to the discrete one, but with following advantages:

- The complete sequence is encoded and retrievable (except for $\lambda = 1$) and the reduction to segments is conducted afterwards, allowing to select a predefined number of clusters in contrast to SSM, where the number of segments is already given by the number of categories in the data.
- The same idea of discounting older purchases is used as for the discrete

approach, but the variation of the continuous discount parameter allows seamless control of the importance of the older purchases and a reduction of the data sparsity in the multi-dimensional space.

- Due to distance-based clustering in WPC the relevant regions are captured based on the data distribution on a flexible scale of discount compared to categories in SSM. In this way, WPC is able to capture dense data regions by placing centroids there.
- Interpretation of the resulting groups with respect to their characteristics based on the cluster centroids is still possible, but additionally a visual interpretation of the sequences is possible as shown above.

Overall, WPC proposes a more generic aggregation mechanism as it is more flexible with respect to parameterization and follows the data distribution. As to the optimal number of segments with respect to predictive performance, the next section is dedicated to the topic of parameter selection.

4.3 Bias–Variance Trade-Off for Lift

The question in this section is how to anticipate the lift achievable with different aggregation levels of SSM and WPC and therefore choosing the optimal one by modeling the error behavior on training data. Usually, this is done by understanding a bias–variance trade-off, originally proposed for a linear regression, by decomposing the resulting prediction error (or loss) into a systematic error component due to generalization (*bias*) and an error based on fluctuations in a data set (*variance*) (James et al., 2013).

Also, a unified decomposition for any type of predictive tasks was defined by Domingos (2000), which will be used for the bias and variance estimates proposed in this work. Therein, for a general error (loss) function $L(t,y)$ the expected loss $E_{D,t}$ is defined in Equation (4.4), where t signifies the true value and y stands for the prediction of the respective instance. Then the expected value of the loss $E_{D,t}$ is decomposed into noise or random error $N(x)$, bias $B(x)$ and variance $V(x)$, whereas x is a training instance within a training set D and c_1, c_2 are multiplicative values depending on the loss function.[1]

$$E_{D,t} = c_1 N(x) + B(x) + c_2 V(x) \tag{4.4}$$

[1] In case of the linear regression loss function is the squared error with $c_1 = 1$ and $c_2 = 1$.

Herein, the loss of lift due to aggregation (generalization) of preceding sequences is investigated. What makes the assessment of predictive models in the marketing context different from classification settings is the performance criterion, that changes the nature of the usual error decomposition. While a classifier is typically assessed by total accuracy or its (potentially weighted) confusion matrix, the criterion in direct marketing is the lift of CR over the baseline among the targeted customers. The discrepancy of the business-oriented objective of lift and the traditional accuracy measure as well as its implications are discussed, e.g. by Baumann et al. (2015).

As the task is to maximize the (out-of-sample) lift with, e.g. 10% of customers targeted (campaign size), the problem at hand is a customer sorting rather than an accuracy maximization problem. In this setting, 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. A decomposition of the prediction model error in a random, a bias, and a variance (selection) related component is done in order to understand and improve the outcome on the test data.

Random error: Within the unified decomposition, the random error is defined as $N(x) = E_{D,t}(L(t, y*))$, where $y*$ is the optimal prediction. Usually, random error is estimated by the Bayes error rate observed in-sample with non-aggregated observations. This concept is transferred to the problem setting by calculating the difference between the maximum lift achievable if all sequences would be unique (i.e., assuming all sequences having a support of one, and, hence, maximum purity), and the in-sample lift achievable with a model considering the full sequential information constituting the optimal prediction $y*$ (non-aggregated sequences). As random error is the unavoidable component of the loss as it is irreducible by the optimal model $y*$, the random error is excluded from further consideration, and bias and variance are studied.

Bias: In general, bias is the systematic loss resulting from a model application due to generalization. In order to measure this component the so-called main prediction y_a is defined as the prediction, which differs least from all possible predictions over the training sets. Then the bias is defined as $B(x) = L(y*, y_a)$, so it quantifies the difference between the optimal result $y*$ compared to the main prediction y_a. The bias of a model a is then estimated by taking the difference between the mean value of in-sample achieved lift compared to the lift $y*$ achieved with full sequential information, i.e., complete sequence.

With $y* = lift_{max,C}$ denoting the maximum achievable lift for a campaign size C, and $lift_{a,C}$ denoting the lift for C with aggregation model a, the bias is deter-

mined as shown in (4.5). Hence, bias measures the share of loss in lift because of non-captured information due to sequence aggregation in the training data. The bias is closely related to the value of the (in-sample) normalized lift $lift_{norm,C}$ that quantifies the amount of captured information as naturally both sum up to 1. Obviously, no aggregation means zero bias, and the higher the aggregation level (the larger the segments) the more they will tend towards the baseline CR, and the higher the bias will be.

$$Bias = \frac{lift_{max,C} - lift_{a,C}}{lift_{max,C}} = 1 - \frac{lift_{a,C}}{lift_{max,C}} = 1 - lift_{norm,C} \tag{4.5}$$

Variance: Variance expresses the fluctuation around central tendency incurred by differences in the training set. The variance is then $E_{D,t}(L(y_a, y))$, meaning the expected loss of all predictions with respect to the main prediction. The variance is estimated using the training CR for a particular customer, as for WPC the result can be incomparable with respect to strongly differing and therefore unmatchable centroids, so a cluster-based estimate of variance is not possible. Furthermore, lift-based estimation did not show systematical patterns as given one relatively large segment, the result for a particular campaign size can be quite robust for training data but do not carry over to the test data.

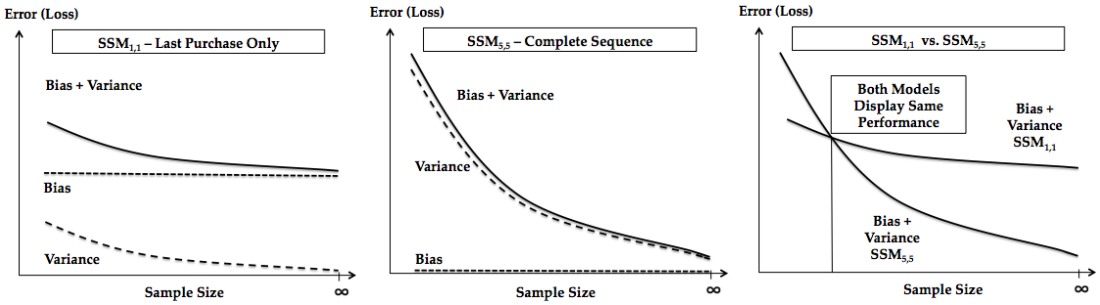
As to the estimation of variance on customer level, given small clusters, training CR will vary strongly depending on the training data, targeting based on such training CR will tend to pick the segments (i.e., customers) with higher CR and therefore clearly overestimate the lift achievable out-of-sample. Increasing support per cluster due to aggregation reduces and stabilizes the predicted CR and therefore, wrongly selected segments based on the in-sample top CR clusters. Therefore, cluster support and training CR are in conflict, typically referred to as bias–variance trade-off, which needs to be understood and solved.

Overall, the variance is estimated as follows. For each customer c who entered the C top customers, corresponding to the campaign size, at least once among all training folds $f \in 1, \ldots, F$ of cross-validation and therefore belongs to the set labeled X_C, the variance is estimated as variance of observed conversion rates $PCR_{c,f,a}$ given an aggregation model a around the mean predicted conversion rate $m_{c,a}$ over all F training sets. Then, an aggregation is conducted over all customers in X_C by computing the mean of the variance for CR estimate as shown in (4.6).

$$Variance = \frac{1}{|X_C|} \sum_{c=1}^{|X_C|} \frac{1}{F} \sum_{f=1}^{F} (PCR_{c,f,a} - m_{c,a})^2 \tag{4.6}$$

The expected asymptotic behavior of bias and variance over increasing training sample size is illustrated in Figure 4.3. The figure depicts the lift for two sequential models sketching two extremes for available 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 error with both models. With decreasing sample size, and therefore decreasing support per segment, a point will be approached where $SSM_{1,1}$ will lead to lower error than $SSM_{5,5}$.

Overall, a bias–variance behavior as shown in Figure 4.3 is expected, making sequence aggregation a mandatory step. Furthermore, it allows for an anticipation of the optimal aggregation level given the data. This will be analyzed in the following empirical study that will now be described. Furthermore, a study whether this trade-off is better solved with models in-between both extremes and the shifting of weights towards more recent purchases is conducted.



(a) Bias and variance for $SSM_{1,1}$ (b) Bias and variance for $SSM_{5,5}$ (c) Comparison of both

Figure 4.3: Bias and variance components of the lift error for $SSM_{1,1}$ and $SSM_{5,5}$. Which of both models will lead to higher asymptotic lift depends on training sample size. At the intersection of both error curves both models achieve the same asymptotic lift.

4.4 Empirical Evaluation of SSM and WPC

This section concentrates on the evaluation of the proposed methods in the context of next purchase prediction. After a short description of the available data the evaluation design is presented in detail for both proposed methods. The evaluation design includes a comparison to established benchmarks, a product-specific evaluation of the performance, a sensitivity analysis for the sample size as well as a detailed empirical discussion of the bias–variance trade-off.

4.4.1 Available Data

The data set available comprises purchasing sequences of 236 327 customers, allowing observation of customers' purchasing behavior over a period of 10 years. An item in a sequence indicates to which of the ten available product categories the purchased product belongs (henceforth the term product category is omitted as there is no differentiation 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 and cloud servers. For reasons of confidentiality, product categories are labeled P_1, P_i, ... P_{10}. On average, a customer purchased about 4 products (median=2).

The raw sequence data is used to generate ten different datasets using the following preprocessing procedure. One of the ten products is selected as target product P^d for which purchasing probabilities are analyzed depending on customers' prior purchasing histories. The process is as shown in Figure 3.4.

If P^d is contained in a customer's purchasing history, the last L product purchases prior the P^d purchase are indicated and the sequence S^d_{cl} is labeled to mark that this sequence resulted in a P^d purchase. If a customer purchased P^d more than once, the first purchase of P^d is taken and the purchasing sequence up to that purchase is derived. If P^d is the first purchase of the customer the customer is omitted from the data sets. $L = 5$ is set as sequence length as only 10% of the sequences in the data exceed a number of five product purchases.

For a customer that does not have P^d in the purchasing sequence at all, the respective purchasing sequence S_{cl} is generated from the last L products she purchased (including duplicates). These sequences are required later to determine the impurity of the clusters and the lift, respectively. As P^d is one of the ten products offered by the company, sequences can consist of nine possible previous products plus P_0, resulting in 3 628 800 potential sequences.

Table 4.4 provides descriptive statistics of the ten sub-datasets. For each P^d,

the number of customers that bought P^d at least once is displayed (after the preprocessing described above) together with the baseline (conversion rate), i.e., the ratio of S_{cl}^d and S_{cl} for each individual P^d. Additionally, the table shows the number of customers contained in each subset (i.e., the number of sequences considered), and the number of unique purchasing sequences (*Sequence Types*).

Table 4.4: Data statistics by target product. The table shows the corresponding baseline within a specific data subset, the number of customers, and the number of unique sequence types.

P^d	Baseline	# Customer	# 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

4.4.2 Evaluation Design

In the following the out-of-sample lift (henceforth: lift) achieved with different aggregation and targeting methods is investigated. Also the empirical bias–variance behavior for lift and how well the respective estimates of the expected error predict the resulting lift are analyzed. The treatment-structure is shown in Table 4.5.

The first treatment considers the respective model and its parameterization. Henceforth, a model with a particular parameterization is referred to simply as model. Given that a maximum sequence length of five is considered, for SSM nine models are applied: five models with $t = L$ ranging from one to five, demanding a mandatory purchasing sequences of the most recent t purchases. Furthermore, four SSM models with $L = 5$ and $t = \{0,1,2,3\}$ are applied, representing sequence types with a mandatory order for the most recent t purchases, and identical portfolios of product purchases further in the past. The latter four models allow a comparison with models using only "hard" sequences up to a certain length for clustering.

For WPC, 16 models with $\lambda \in \{0.1,0.5,0.9,1.0\}$ combined with different numbers of clusters $k \in \{10,50,200,500\}$ are analyzed. As benchmarks for the distance-based segmentation and prediction approach, Levenshtein-Ward (LW)

Table 4.5: The treatments considered are (i) the aggregation and targeting model with its parameterization, (ii) the target product-specific data subset, and (iii) the degree to which the target product-specific subsets have been downsampled. For each model, the letters in brackets indicate whether it performs segmentation (S) and predictions are then derived by the CR of segments, or performs prediction (P) directly.

Model Type	Cardinality (Number 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)

clustering technique with $k \in \{10,50,200,500\}$ is used as one of the most common approaches used in the marketing literature. In addition, the proposed approaches are benchmarked against widely used model-based clustering approaches, namely Association Rules (AR), Hidden Markov Model (HMM) and Self-Organizing Maps (SOM). The support parameter of AR encodes the minimum support for which a rule should apply, resulting in approximately 20 ($support = 10^{-5}$) or 200 customers ($support = 10^{-4}$) as threshold support. With AR, obviously only rules exposing an in-sample lift value exceeding 1 are considered as potential target segments. HMM has the number of latent states as a parameter, which in this case can be interpreted as implicit customer types. As most studies employing HMM use about 10 latent states (see Schweidel et al. (2011) or Netzer et al. (2008)), only a lower range of cluster numbers with $k \in \{10,50\}$ is employed to provide comparability. The prediction for test sequence is derived using previously learned emission probabilities given the next predicted latent state. SOM allows clustering using a predefined grid, so the resulting mapping of an observation to a grid cell can be mapped to the corresponding in-sample CR. Four SOM grids with $k \in \{10,50,200,500\}$ cells are employed. Data preparation for SOM allows for implicit dummy coding of the sequence structure, so that the 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. This encoding represents (10 products - 1) times 5 possible purchases as a 48 dimensional

binary vector. As only sequences with at least two purchases are considered, there are 9+9 potential different encoding for the first two purchases, and 10 (including P_0) for the three purchases beforehand.

As further benchmarks the following prediction methods are used: Logistic Regression (LR) that is also widely applied in the field of predictive modeling and a singular value decomposition (SVD) based approach as it is a standard tool in recommender systems (Dunlavy et al., 2011). SVD operates on the same weighted vectors as WPC, but is represented as customer-to-product matrix. As for WPC, $\lambda \in \{0.1, 0.5, 0.9, 1.0\}$ is employed. The potential difference between WPC and SVD is driven by the clustering in product space used in *WPC* instead of a Singular Value Decomposition on a matrix of values. LR is applied to the complete sequential information, encoded in 48 dummies, as also SOM does. However, the two last benchmark methods do not allow a segmentation of purchasing sequences.

The second dimension considers products, which result in 10 product specific data sets depending on the target product. The sample size, as a third dimension, is systematically reduced to 1/2, 1/4, 1/8, 1/16, 1/32 of the available sequences to study the behavior of the different models when the overall support decreases. Thereby, only the size of the training data is downsampled. The test data is the same for all 6 training sample sizes in order to provide comparability of prediction results.

The evaluation criterion is lift when selecting up to 10% of customers (so called top-decile lift (Baumann et al., 2015)), that are analyzed for all ten product specific sets and all downsample-levels in a ten-fold cross-validation. As there are different product-specific ranges of the lift, *normalized lift* is used for some evaluations in order to conduct comparisons across product categories. Its value is computed by dividing the lift value of a model by the corresponding value of the $SSM_{5,5}$ model, which is the full sequential information contained in the data. This way the proportion of sequential information captured by a given model is quantified.

Overall, SSM, WPC and selected benchmark models are evaluated by their results in the cross-validation per target product, sample size, training sample and lift percentile up to 10 %, resulting in 6 000 data points (10 product specific sets x 6 sizes x 10 cross-validations x 10 lift percentiles) per model. In addition, the number and interpretability of segments required to target 10 % of the company's customers is studied. The smaller the number of segments, the better marketing specialists can work with these.

First, aggregated results in terms of a comparison of the average normalized

lift achieved when targeting 10 % of customers with the different models over all data samples are presented. Second, the drill-down to product level is conducted and the best performing model per product type for 1 % or 10 % of customers is analyzed. Third, the results for different downsample ratios of the training data within the cross-validation and results of the effect of different parameters on lift are analyzed in the context of the expected trade-off between bias and variance entailed by the parameterization.

4.4.3 Aggregated Results

For the evaluation the Box-Cox-transformed normalized lift $lift^{BC}_{norm}$ is regressed on the treatments (Box and Cox, 1964). As lift develops in a non-linear way (close to a decreasing exponential trend) with campaign size, which can be directly considered in a multiple regression setting using a prior Box-Cox transform. $\lambda_{Box-Cox} = 0.26$ was applied as the value for the data set, where the observed errors then expose approximately white noise error structures. 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.

A multiple regression-based evaluation approach is used as this allows to control for influences such as the dependent product category in a straightforward and concise fashion using dummies (as done e.g. 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.

With C denoting campaign size, P^d denoting the target product of interest, interaction of product dummy and campaign size $P^d \cdot C$ controlling for different slope of products (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 (4.7).

$$lift^{BC}_{norm} = \beta_0 + \beta_1 \cdot C + \sum_{d=1}^{I-1} \beta_{d+1} \cdot P^d + \sum_{d=1}^{I-1} \beta_{I+1} \cdot P^d \cdot C + \sum_{j=1}^{A-1} \beta_{2 \cdot I+1} \cdot a_j + \epsilon \quad (4.7)$$

Table 4.6 shows the regression estimates for campaign sizes 1 % to 10 % (denoted as C1-10). 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 in-

Table 4.6: Aggregated evaluation results for SSM and WPC. 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, SSM.

	(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

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

dependently 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 considered.

As expected, campaign size has a negative association with $lift^{BC}_{norm}$ (henceforth: lift); 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 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 dominated 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 robust against suboptimal parameter choices.

As to the best individual model WPC achieved the highest lift. The application of $WPC_{0.9,200}$ on average improves lift achieved with LR by 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 4.6 determines significant lift differences between the models and LR, a further statistical test for the best WPC against SOM was conducted – the strongest benchmark model on the data. Using two-sided paired Wilcoxon test[2] 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 models, 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 Figure 4.5 below) due to clustering with a very low number of clusters for very sparse data as no temporal discounting 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 clearly show the importance of proper parameterization on the one hand.

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 this 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 coefficient 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 sequence-similarity-based clustering, only WPC considers where the (dis)similarity of sequences stems from; hence, the recency of purchases. In this respect, the discounted consideration of more ancient purchases seems to be beneficial and will now be discussed.

Figure 4.4 shows the empirical function of the 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 %. The plot shows an inverse U-shaped relation. Initially, the lift increases with the introduction of a discount (a value below 1.0), indicating the higher importance of more recent

[2]Wilcoxon test was used as a more conservative approach but a t-test was also significant.

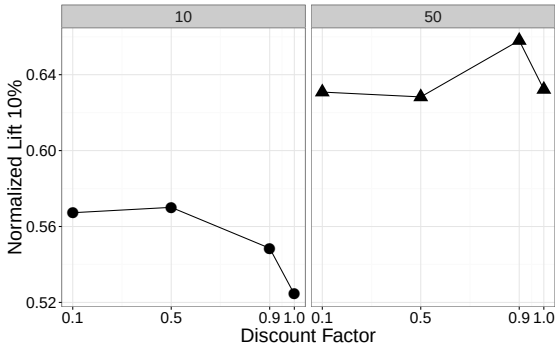


Figure 4.4: Influence of discounting parameters for 10 and 50 clusters on out-of-sample lift. The graphs show the average normalized lift with WPC over the discount factor (1.0 indicates no discounting) for C =10 %. The left-hand (right-hand) plot shows the curve with k=10 (50). Both display an inverse U-shaped relation.

purchases. With an increasing discount, at some point the lift decreases again, indicating a lower predictive value of older purchases, i.e., of complete sequences. The highest impact of a temporal discount is observed when learning a small number of clusters ($k = 10$), where a stronger discount ($\lambda = 0.5$ or $\lambda = 0.1$) leads to a significantly higher performance. Interestingly, for 50 clusters a discounting factor $\lambda = 0.9$ shows the best result, which can be interpreted as a high amount of sequential information captured from the data given more clusters.

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

4.4.4 Results at Product Level

Table 4.7 summarizes the top performing models per product for 1% and 10% of the customer base in terms of (untransformed) lift. Overall, WPC dominates the field of models for all products, as the application of WPC with a discounting factor $\lambda < 1$ always appeared among top-performing models.

The second most frequent among top 3 models is SSM, however, with different parameterizations, like portfolio ($SSM_{5,0}$) or a sequence representation, e.g. ($SSM_{2,2}$). Interestingly, $SSM_{5,0}$ appears together with WPC using higher values of λ, e.g. for Product 10, so that no incorporation of sequential information seems

Table 4.7: List of the three best models per product for 1 % or 10 % (C1 and C10 correspondingly) of selected customers considering mean performance over the folds for the complete training data, the numbers in brackets display the standard deviation. WPC is with one exception in all products/lift combinations.

C1 Product 1	Product 2	Product 3	Product 4	Product 5
$WPC_{0.9,500}$ 3.20 (0.55)	$WPC_{0.5,50}$ 2.76 (0.44)	$WPC_{0.5,50}$ 4.87 (0.94)	$WPC_{0.9,200}$ 2.35 (0.56)	$WPC_{0.1,200}$ 18.44 (3.06)
$SSM_{5,1}$ 3.18 (0.50)	$WPC_{0.9,50}$ 2.76 (0.27)	$WPC_{0.1,200}$ 4.86 (1.09)	$WPC_{1.0,200}$ 2.24 (0.34)	$SSM_{2,2}$ 18.11 (3.07)
$WPC_{0.1,200}$ 3.15 (0.46)	$SSM_{2,2}$ 2.71 (0.39)	$WPC_{0.9,50}$ 4.82 (0.84)	$WPC_{0.1,200}$ 2.22 (0.36)	$WPC_{0.5,200}$ 17.98 (2.47)
C10 Product 1	Product 2	Product 3	Product 4	Product 5
$WPC_{0.9,500}$ 1.94 (0.11)	$WPC_{0.1,200}$ 1.62 (0.11)	$WPC_{0.9,200}$ 2.54 (0.21)	$WPC_{0.1,200}$ 1.96 (0.11)	$WPC_{0.1,200}$ 5.42 (0.41)
$WPC_{0.1,200}$ 1.93 (0.11)	$WPC_{0.5,200}$ 1.62 (0.10)	$WPC_{1.0,500}$ 2.52 (0.18)	$WPC_{0.1,500}$ 1.96 (0.10)	$WPC_{0.5,200}$ 5.39 (0.48)
$WPC_{0.9,200}$ 1.93 (0.11)	$WPC_{0.1,500}$ 1.60 (0.10)	$WPC_{0.9,500}$ 2.44 (0.21)	$SSM_{3,3}$ 1.96 (0.10)	$WPC_{0.9,50}$ 5.33 (0.43)
C1 Product 6	Product 7	Product 8	Product 9	Product 10
$WPC_{0.1,50}$ 3.36 (1.10)	$WPC_{1.0,500}$ 11.77 (1.71)	$WPC_{0.1,50}$ 9.39 (4.33)	$AR_{10^{-4}}$ 5.09 (2.27)	$SSM_{5,0}$ 20.40 (2.42)
$SSM_{2,2}$ 3.20 (1.09)	$WPC_{0.9,50}$ 11.60 (2.09)	$SSM_{2,2}$ 9.33 (4.01)	$SSM_{5,0}$ 4.73 (0.61)	$WPC_{1.0,500}$ 20.09 (2.83)
$WPC_{0.5,50}$ 3.05 (0.96)	$WPC_{1.0,200}$ 11.41 (1.79)	$SSM_{5,0}$ 8.92 (2.85)	$SSM_{5,2}$ 4.60 (0.85)	$WPC_{0.9,500}$ 19.43 (1.89)
C10 Product 6	Product 7	Product 8	Product 9	Product 10
$WPC_{0.1,200}$ 2.09 (0.23)	$WPC_{1.0,200}$ 4.56 (0.33)	$WPC_{0.9,200}$ 4.85 (0.48)	$WPC_{0.1,200}$ 3.13 (0.25)	$WPC_{0.1,200}$ 7.35 (0.49)
$WPC_{0.5,500}$ 2.07 (0.20)	$WPC_{0.1,200}$ 4.54 (0.31)	$WPC_{0.9,500}$ 4.80 (0.51)	$WPC_{0.5,500}$ 3.10 (0.26)	$WPC_{0.9,200}$ 7.20 (0.54)
$WPC_{1.0,200}$ 2.05 (0.28)	$WPC_{0.9,50}$ 4.50 (0.30)	$SSM_{5,0}$ 4.79 (0.70)	$WPC_{0.1,500}$ 3.10 (0.25)	$WPC_{0.5,200}$ 7.20 (0.60)

to be beneficial for these products. On the other side, in cases where sequential *SSM* parameterizations appear, a higher discount is preferable, so for Product 2, 4, 5, 6. This can be seen as an indication of sequential order of products because preserving a high amount of sequential information is beneficial. For Product 1, a stronger discount with $\lambda = 0.1$ leads to a lower number of required segments while keeping comparable performance, so that a strong discount is beneficial for this product, which can be sign of high data sparsity or a very high importance of last purchases. As an exception, Association Rules have shown the best result for Product 9 and 1 % of customers.

4.4.5 Sensitivity to Sample Size

This section studies the impact of downsampling the available training data on the models' predictive performance. For reasons of brevity, only the models showing the best results on the complete dataset are considered, namely SOM with 4, SSM with 9 and WPC with 16 model parameterizations as well as LR as the weakest benchmark. In contrast to Table 4.6, SOM is the baseline model, as this was the strongest external competitor, so the estimates and p-value are provided with respect to this model. Again, the model dummy indicates a par-

Table 4.8: Results of the benchmark models with different sample sizes. The baseline method is SOM. The order of model types by achieved lift is stable over decreasing sample size, with advantages of WPC over SSM decreasing with sample size.

	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	29977	29977	29977	29977	29977	29977

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

ticular model type independent of the parameters selected. Table 4.8 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 has the highest estimates starting from 1/2 even with suboptimal parameterizations, and estimates are negative with LR for all downsampled training subsets. However, the difference between the estimates increases with decreasing sample size, and an increasing advantage with WPC as the samples get smaller is shown. In order to determine an appropriate parameterization of WPC or SSM in practical settings and draw a more general recommendation the empirical bias–variance behavior of the models and parameterizations are discussed in the following.

The top performing models for WPC and SSM per sample size and for $C =$ 10 % of customers as well as the respective average number per cluster (x-axis) are shown in Figure 4.5. The plots depict the average normalized lift (y-axis) over the average number of customers per cluster for sample sizes 1 and 1/32.

As discussed, models producing more segments should perform better with increasing sample size. 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).

As the training data gets smaller, the number of segments used by the best models also decreases. Additionally, the temporal discounting of WPC is stronger: the top-performing models for 1/32 sample size use $\lambda = 0.1$ or $\lambda = 0.5$. Portfolio ($SSM_{5,0}$) is often a well performing model for small sample sizes. The

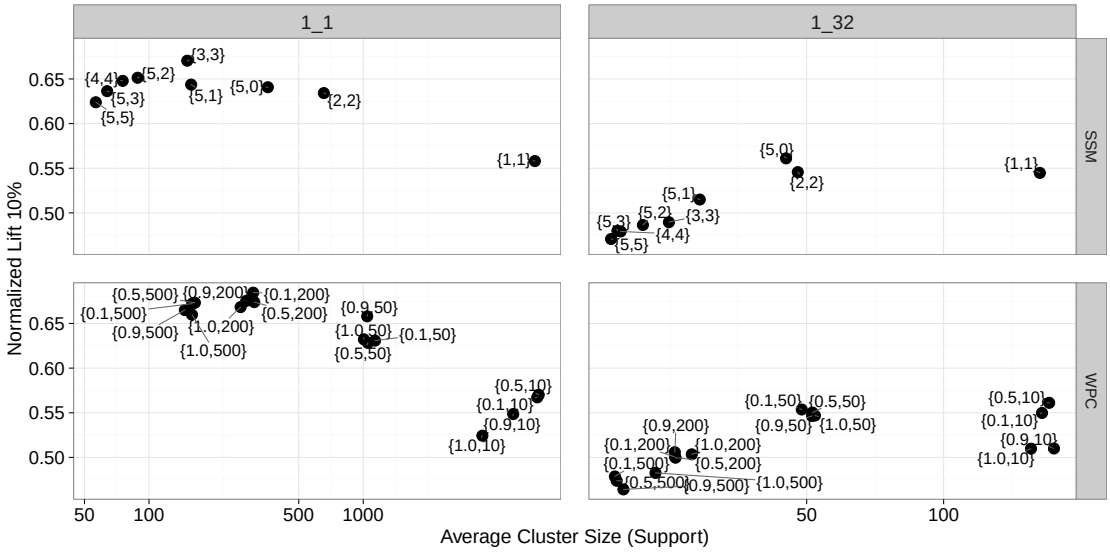


Figure 4.5: Normalized lift for the best WPC and SSM parameterizations and their respective average cluster sizes. While the optimal average cluster size increases with sample size, WPC need fewer clusters – that have consequently higher support than SSM clusters.

impact of discounting factor on the resulting segments size can be seen best for 10 clusters, where $WPC_{0.5,10}$ is able to accomplish both, higher lift value and larger segments in comparison to $WPC_{1.0,10}$. Considering the related $SSM_{1,1}$, although the average segment size is comparable, $WPC_{0.5,10}$ seems to form more beneficial segments. With decreasing sample size, the best WPC model employs a stronger discount: the top performing models for 1/32 sample size uses $\lambda = 0.1$ or $\lambda = 0.5$. In the same spirit, portfolio $SSM_{5,0}$, 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. 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. Also, the optimal average size of clusters can be derived from the plots: around 200 with complete sample and about 50 for 1/32.

4.4.6 Evaluation of Bias–Variance Trade-Off

The previous section has shown several tendencies with respect to the sample size. For instance, given a smaller sample sizes an application of models with a

lower number of segments was beneficial. The bias–variance behavior of both model types will now be discussed in more detail.

Figure 4.6 (left) displays the mean bias observed with selected models for complete sample on the left and for sample size 1/32 on the right. The x-axis of each subgraph depicts the campaign size. As expected, the highest bias is obtained with $SSM_{1,1}$ as it is the simplest model and does only differentiate the last purchase. On the other hand, $SSM_{5,5}$ per definition exhibits no bias, while model $SSM_{5,0}$ (the portfolio model) results in a bias between both extremes. The WPC model with only 10 clusters has a higher bias than the portfolio model, while it shows a performance similar to the portfolio with 50 clusters, and clearly has a much lower bias for 200 clusters.

Figure 4.6 (right) displays the estimated variance component of the test error. Here, the last purchase only ($SSM_{1,1}$) has a very low variance over all sample sizes. In contrast, the complete sequence ($SSM_{5,5}$) has the highest variance. The variance of $WPC_{0.5,10}$ is very close to the one of $SSM_{1,1}$. As with bias, similar (low) variance estimates for $SSM_{5,0}$ and $WPC_{0.9,50}$ can be observed. $WPC_{0.1,200}$ has a relatively high variance in the 1/32 sample, which can be explained by a number of clusters too high for this small training data set, leading to insufficient support and variance increase.

Figure 4.7 provides another view on the bias–variance trade-off: it shows the average normalized lift for 10 % of customers with respect to sample size. The

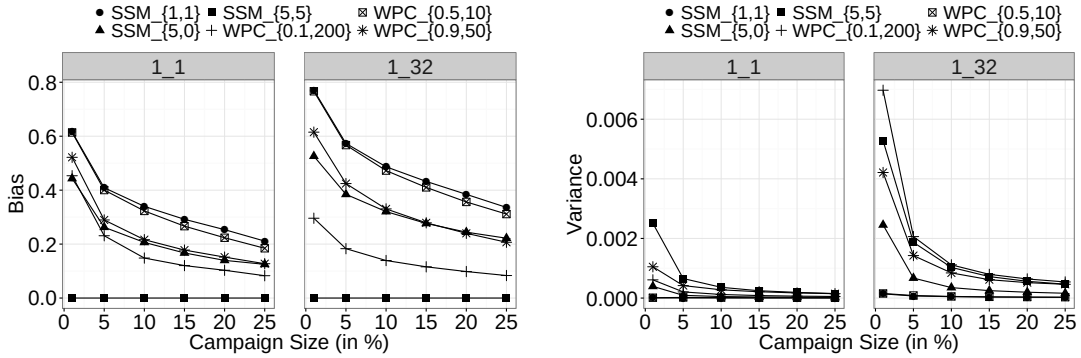


Figure 4.6: Empirical evaluation of bias and variance estimates. *Left:* Bias obtained by a specific model compared to the maximal sequential information of $SSM_{5,5}$ (on the training data). With higher aggregation, the bias increases as a simpler model is not able to incorporate more information. *Right:* Variance for selected models as the mean variance of customer-specific prediction for customers up to campaign size. $SSM_{1,1}$ has the lowest absolute value and minimal increase as the sample size decreases, while $SSM_{5,5}$ typically has the highest variance (for 1/32 it is WPC with a high number of clusters).

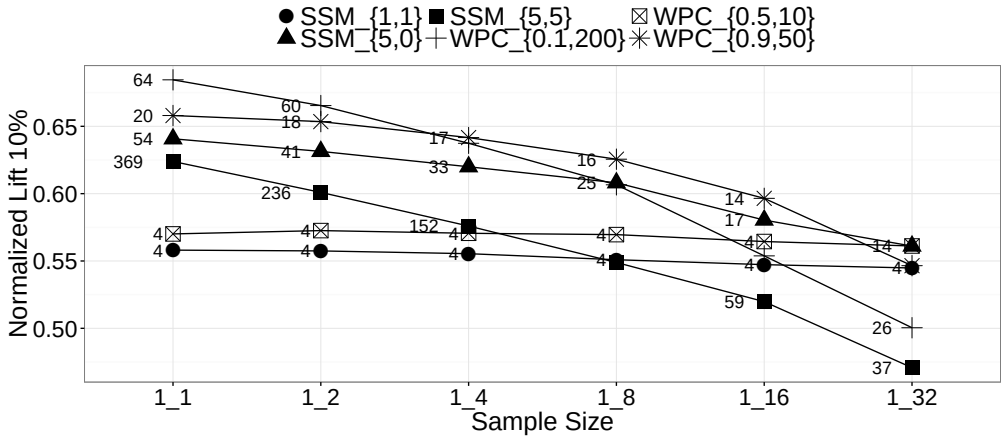


Figure 4.7: Influence of training sample size on average normalized lift for 10 % of customers. The numbers represent the number of segments within the campaign size. WPC shows the best results when selecting a low discount (high λ) and 50 clusters for small samples of up to 25 % of the original sample size, and high discount (low λ) and a higher number of clusters of around 200 for larger samples.

labels represent the average number of segments included in the respective customer selection. Intersections show that trade-off would be better solved with an alternative model.

$SSM_{1,1}$ shows low but stable performance when sample sizes decrease. The performance of $SSM_{5,5}$ drops strongly as the sample size gets lower. 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 which is best captured by the portfolio ($SSM_{5,0}$) or $WPC_{0.5,10}$.

4.4.7 Estimating Test Lift on Training Data

The estimates for bias and variance components of the test error overall show a behavior in line with general theory on statistical learning as shown in the previous section. The next question is, whether test error can be estimated by both values. This is studied by test error, by regressing *Test Error*, which in combination with the normalized lift sums up to 1. Outcomes for the complete sample size are shown in Table 4.9.

In the linear regression, both bias and variance estimates were normalized to the interval $[0,1]$. Furthermore, a set of variables for the target product was

Table 4.9: Regression of test error for 10 % of customers over all products and the complete data set on bias and variance estimates. Controls for target products are included but not shown. Bias and variance lead to significantly higher test error. The interaction of bias and variance is negative, indicating the trade-off between both.

	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

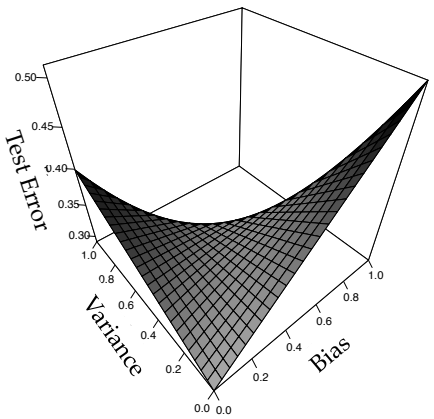


Figure 4.8: Illustration of the bias–variance trade-off as a regression interaction surface with bias component (x-axis), estimated variance component (y-axis), and the test error (z-axis).

used in the regression as this variable explains about 50 % of the variance in the regression, but is omitted in the table for reasons of brevity.

With an R^2_{adj} of 87 %, the model has high a explanatory power on the complete data set. As expected both, bias and variance, are positively correlated with the test error, and the trade–off between both is well reflected by the interaction of both with negative coefficient. Figure 4.8 visualizes the interaction effect from the regression above. The x-axis represents the value of the bias component, the y-axis the magnitude of the estimated variance component, and the z-axis the test error. Although the bias is the main source of error, the variance shows a comparable influence level. The U-shape for the interaction of both clearly reveals the bias–variance trade-off that needs to be addressed appropriately with a segmentation model.

4.4.8 Summary

This section introduced and empirically tested two unsupervised mechanisms for segmenting purchasing sequences and using the segments for next-purchase next purchase prediction, namely Sequence-Set-Model and Weighted Productspace Clustering. Two key issues both methods are using sum up to the diminishing importance of the older purchases and the robustness of the prediction based on data-driven aggregation of sequences. Both models were empirically evaluated and demonstrated to have a competitive predictive performance

in an out-of-sample evaluation.

Comparison to established segmentation and classification methods in the literature (e.g. Logistic Regression) has shown that supervised classification on full data dimensionality leads to poor performance, whereas an unsupervised aggregation is beneficial in the presented case. Additional evaluation with respect to the available training sample size has shown the necessity of an appropriate parameterization prior aggregation, especially for small training samples.

Between the two proposed methods, WPC has mostly shown better results, even using a smaller number of interpretable segments. The discounting factor for older purchases was also shown to be an important parameter for the predictive performance. The closest candidate from SSM group was the portfolio representation for small data samples. These results are important for marketing activities as they do not only allow for increasing lift, but also deliver interpretable segments as they concisely represent general purchasing types.

The proposed estimates for bias and variance have shown a behavior in line with the statistical theory, as well as predictive ability for achieved out-of-sample performance. The latter allows for finding optimal parameterizations for proposed models.

4.5 Extensions of WPC with Respect to Time Aspects and Data Specifics

This section is dedicated to different extensions of the proposed WPC method, which are motivated by general marketing tasks like analysis of bundle purchases and also inspired by different data issues, which were specific to the available data set, for instance repetitive (serial) purchases from the same category. However, the proposed extensions are a general approach and can be applied to other sequential data in context of CRM. These methods extend the aggregation mechanism of WPC and allow additional modeling capabilities.

Due to the bias–variance trade-off shown above a higher modeling granularity leads to a higher variance component. This is due to the fact that the data is then dispersed even more granularly over possible categories, which leads to a higher sensitivity to the training data. Therefore, the reduction of the bias by an additional modeling complexity has to compensate the growing variance component in order to provide an overall performance achievement, which means that the additional modeling complexity should be applied exactly at the source of the bias.

Among the proposed extensions are (i) additional modeling of bundling purchases if such occur, (ii) additional modeling of sequence length, (iii) reduction of representation complexity by filtering of serial purchases. This section also contains descriptive metrics, which illustrate each case within the data and can be used as an indication for the potential application of each extension. Afterwards, this section contains a comparative study with the results of the originally proposed WPC method.

4.5.1 Consideration of Specific Sequence Characteristics

In the previous evaluation, only the data consisting of at least two purchases prior to the target purchase was used. In this study sequences with only one purchase are also used, so that the length distribution is even more skewed. The data set containing all sequences will be referred to as a *complete set* and the data set containing at least two previous purchases is then referred to as *reduced set*. The latter is the one used in Section 4.4.

Sequence Length Distribution

The first data characteristic under study is the sequence length distribution. Considering the complete data set, on average across all 10 products, 73.56 % (SD=2.97 %) of customers had only one previous purchase. On average of 16.19 % (SD=1.54 %) had two previous purchases of sequences and 4.78 % (SD=0.06 %) of sequences contain exactly three purchases. The reduced data set contains 64.03 % (SD=3.63 %) of sequences with length of two purchases, 19.09 % (SD=0.86 %) with exactly three purchases.

Therefore, the distribution of sequence length is clearly skewed, so there is a need for additional separation. In the context of the proposed WPC approach, such distributions can lead to a distortion, as the employed distance-based clustering algorithm is also influenced by the distribution of the instances. The single purchase sequences dominate the data and therefore the centroids of the clusters. As a result, the final clustering would converge to the one of the last purchase, namely $SSM_{1,1}$. The remaining sequential information would thus be lost to a large proportion.

The proposed solution is introducing of an additional dimension, the so-called Product 0 (P0) dimension to the product space. Such explicit modeling of "Non-Purchases" might be adequate to separate longer sequences from shorter ones and thereby restoring the sequential information from longer sequences. This

feature allows to model "empty" purchases - and therefore considers sequence length. The distance between sequences with similar order of recent purchases but differing length increases, allowing the clustering mechanism to build new clusters considering sequence length. The remaining product dimensions are computed as presented before.

In the context of the bias–variance trade-off, such an increased representation complexity should reduce the bias error component (as more information is incorporated, so the representation is "closer" to the sequence) and simultaneously increase the variance component. Overall, if the information in longer sequences is more valuable, such an extension should improve the predictive performance, but only if the gain in the reduction of the bias error component outweighs the additional variance.

Presence of the Same-Product Sequences

Due to the specifics of the offered telecommunications products (or services), it is often the case for some of the products are bought in multiple quantities, sometimes within a short period of time. This fact motivates the following investigation. The second characteristic explored here is a sequential purchase of the same product. Table 4.10 shows the relative share of "same-product" or "serial purchases" in the data, i.e., sequences with subsequent purchases of the very same product (category) prior to a purchase of a specific target product. Products with comparably high percentage of consecutive same-product purchases are Product 1 and Product 2, so these are potential candidates for added (or reduced) value of serial purchases.

Given a strong discount, WPC weights multiple purchases such that an earlier purchase is heavily discounted. However, depending on the target product, the information of four serial purchases may distort the performance due to building segments being biased towards such serial purchases, so that product preceding

Table 4.10: Summary statistics of the available data set without single-purchase sequences. Percentage of consecutive purchases from the same category by target product is shown Product 4 has the lowest amount of such purchases.

Distribution of Serial Purchases			
Product 1	9.54 %	Product 6	6.68 %
Product 2	8.01 %	Product 7	6.59 %
Product 3	6.63 %	Product 8	6.59 %
Product 4	2.84 %	Product 9	6.67 %
Product 5	6.57 %	Product 10	6.59 %

these has no significant influence. So given no informational value of such purchases, consideration of one instead of for example four such purchases would be a solution. On the other hand, if such serial purchases have an additional predictive value, e.g. once a customer bought at least four domains, this might be an indicator for a build-up of a new web page and a purchase of web space is very likely. Having no prior knowledge about the importance of serial purchases, both (filtering and no filtering) should be evaluated.

The problem with too many, unnecessary clusters is decreasing support per segment, resulting in lower predictive performance. Therefore, this work proposes pre-processing of purchasing sequences data by aggregating consecutive purchases of the same product to only one single purchase. If these consecutive purchases have predictive value, than a performance decrease will be observed, marking an explicit informational value of these. With respect to the bias-variance trade-off, this extension would lead to an increase of the bias component, as a part of the information from the data is lost, instead the variance component would decrease as several sequences containing such serial purchases would be aggregated, increasing support per group.

Modeling of Bundling Purchases

The third sequence characteristic investigates the share of simultaneous purchases within a sequence (so-called bundles). Table 4.11 displays such share of bundles per product-specific data set. A bundle purchase occurred if purchases were conducted within one day. The table demonstrates that Product 1, Product 2 and Product 4 have strongly deviating shares of bundles within sequences compared to the baseline of other purchases.

Within the basic definition of WPC the order of purchases is mandatory. In

Table 4.11: Summary statistics of the available data set without single-purchase sequences. Percentage of sequences having bundles (multiple purchases within one day from the same category) depending on target product is shown. Product 1 has the highest and Product 4 the lowest share of bundle purchases in the sequence prior to the target purchase.

Distribution of Bundle Purchases			
Product 1	19.83 %	Product 6	15.84 %
Product 2	18.30 %	Product 7	15.92 %
Product 3	15.68 %	Product 8	15.90 %
Product 4	7.69 %	Product 9	15.92 %
Product 5	15.89 %	Product 10	15.90 %

this case no bundle information is stored, therefore, this work proposes an extension, where several purchases at the same time are modeled explicitly. For this purpose, if two purchases are done within the same day the corresponding index i is given to the respective categories. In comparison to the previous model, within a two product space, not only (1,0.5) and (0.5,1) but also (1,1) are possible sequence representations.

Using this idea, the representation complexity and consequently the variance component of the error grow. Once again, a sufficient amount of bias captured by this modeling extension must compensate the variance increase for the overall predictive performance to increase. So, given a high number of bundles, the WPC method is likely to allocate additional clusters for bundles if these have a high share within the data. If these additionally have predictive value, a performance increase might be gained. Otherwise, the increased representation complexity reduces bias at costs of higher variance leading to a less robust prediction.

An illustration of all three approaches is given in Figure 4.9, where the mechanisms are presented using two exemplary purchasing sequences. The first extension additionally weights sequence length thereby including it as additional dimension for determination of sequence similarity. The second extension considers serial purchases from the same category and WPC model. This method is particularly useful in case where an additional purchase of the same category does not have much predictive value (example: ordering 210 or 211 domains).

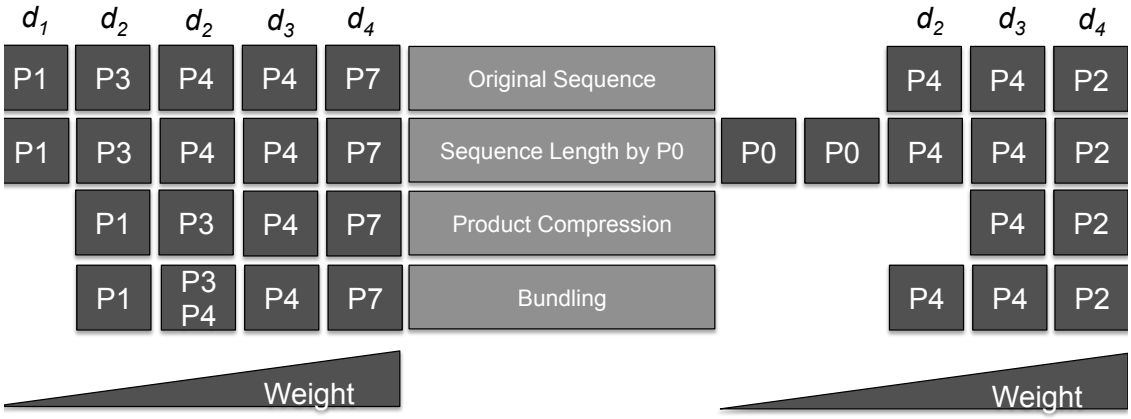


Figure 4.9: Different methods for representation of two example sequences (left-hand and right-hand columns). The rows represent the corresponding alternative representation method and the result of its application. Products are encoded as P_i and time of purchase as d_t.

The third extension regards frequent simultaneous co-purchase of several products (bundles), as a bundle can be a strong predictor for a next purchase as it might imply a different usage intention compared to two separate purchases.

4.5.2 Empirical Evaluation of Proposed Extensions

After the introduction of the additional extensions, this section presents an empirical evaluation of these compared to the performance of basic WPC in a 10-fold cross-validation. As before, the evaluation metric is lift. The aggregated results are evaluated by means of a linear regression of the logarithmized lift value controlling for the campaign size (*Campaign Size*), product category (*Product*) as well as for the extension type (*Extension*), both as a dummy variables. As the lift curve is shaped exponentially with respect to the *Campaign Size* the lift value was logarithmized. The column of the Table 4.12 shows the results of the regression up to a specific campaign size in order to illustrate different performance among the range of campaign sizes. For reasons of brevity the regression results are divided into several parts, although these stem from the same regression. First, Table 4.12 shows outcomes of a regression involving dummies for *Product* and a continuous variable for *Campaign Size* as well as their interaction.

The goal of this regression is to demonstrate general effects regarding products. Product 4 is used as baseline as the one with the lowest total lift value; the values are not normalized over different products, as these are important for the further interpretation. All product dummies have significant effects compared to the baseline. The practical explanation of the low lift value for Product 4 is the fact that this is a starter product with a low price and low technical sophistication. Overall, the coefficients can be divided into several groups with respect to the performance measured, which are consistent with Figure 3.3.

The first group is the one of best predictable products, such as Product 5, Product 7 and Product 10, which belong to more sophisticated and expensive product categories. This can be explained by the fact of longer preceding sequences, as the technological level is built up successively. Other products can be placed between these two extremes regarding both, value and technological complexity. These product-specific findings are in line with those presented in Chapter 3.

The estimate for *Campaign Size* is negative (for 10 % and 20 % significant). This fact illustrates the naturally decreasing slope of the lift function and is in line with expectations. The next group of coefficients estimated in this regression quantifies the interactions between target product-percentile and lift. This interaction illustrates the slopes of the product-specific lift curves. Product 4 shows

Table 4.12: Regressing lift on target product, percentile and their interaction. The baseline is Product 4 as the one having the lowest lift value. Overall, strong differences in both, dummy coefficients as well as for the interactions with campaign sizes can be seen.

Independent/Dependent Variable	Log (Lift 5 %)	Log (Lift 10 %)	Log (Lift 20 %)
Intercept	1.042***	1.051***	1.061***
Product 1	0.271***	0.195***	0.124***
Product 2	0.159***	0.096***	0.029***
Product 3	0.623***	0.467***	0.329***
Product 5	1.968***	1.805***	1.515***
Product 6	0.229***	0.180***	0.131***
Product 7	1.484***	1.378***	1.173***
Product 8	0.972***	0.970***	0.919***
Product 9	0.497***	0.515***	0.468***
Product 10	2.016***	2.044***	1.856***
Campaign Size	-0.002	-0.006***	-0.007***
Product 1 x Campaign Size	-0.048***	-0.025***	-0.008***
Product 2 x Campaign Size	-0.042***	-0.022***	-0.009***
Product 3 x Campaign Size	-0.098***	-0.043***	-0.016***
Product 5 x Campaign Size	-0.172***	-0.117***	-0.061***
Product 6 x Campaign Size	-0.034***	-0.015***	-0.006***
Product 7 x Campaign Size	-0.119***	-0.084***	-0.045***
Product 8 x Campaign Size	-0.033***	-0.033***	-0.025***
Product 9 x Campaign Size	-0.015***	-0.022***	-0.013***
Product 10 x Campaign Size	-0.093***	-0.106***	-0.071***
R^2_{adj}	0.933	0.936	0.924

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

the lowest absolute lift value, but simultaneously exhibits the lowest decrease in performance per percentile. So the performance is robust over the percentiles; at the same time there seems to be no top group of customers to target with this product using the preceding purchasing sequence. On the contrary, for the products with high absolute lift values the performance drop is steepest, e.g. for Product 5 or Product 7. This relationship can be explained as follows. First, the lift value strongly depends on the baseline of the sample. Given a very low baseline within the sample, a shift of one customer has a high influence on the lift value. Second, the general predictability of the product has an additional influence on the shape of the curve. For example, Product 1 and Product 2 have a comparable purchase probability baseline. Therefore, the difference in favor of Product 1 with respect to the dummy variable as well as to the interaction coefficients is driven by a higher predictability of a product purchase for given purchasing history.

As a second part of the evaluation, Table 4.13 shows the regression estimates for the application of different extensions combined with the parameterizations. Here, the performance of the extensions is compared to the baseline of the WPC method on the same data. The dummy variables are decoded as follows: adding

Table 4.13: Extension-specific results when regressing the logarithm of the lift up to a certain percentile on (i) filtering for serial purchases (FSPR), (ii) bundling (BDL) and (iii) explicit sequence length modeling (P0). Baseline levels for factors is the original WPC method with $\lambda = 0.1$.

Independent/Dependent Variable	Log (Lift 5 %)	Log (Lift 10 %)	Log (Lift 20 %)
$\lambda = 0.5$	0.00551	0.00181	0.00132
$\lambda = 0.9$	0.00864.	0.00587.	0.00196
$\lambda = 1$	-0.01871***	-0.01716***	-0.02203***
Number of Segments k	0.00020***	0.00021***	0.00019***
Extension FSPR	0.00859	0.00077	-0.00378
Extension BDL	-0.07920***	-0.08181***	-0.07936***
Extension P0	0.00669	0.00628	0.00916*
Extension FSPR x $\lambda = 0.5$	-0.00253	-0.00122	0.00003
Extension FSPR x $\lambda = 0.9$	0.00847	0.00475	0.00261
Extension FSPR x $\lambda = 1$	0.01293.	0.00950*	0.00805**
Extension BDL x $\lambda = 0.5$	0.03107***	0.05041***	0.05938***
Extension BDL x $\lambda = 0.9$	0.07789***	0.08461***	0.08314***
Extension BDL x $\lambda = 1$	0.07437***	0.08364***	0.08189***
Extension P0 x $\lambda = 0.5$	-0.00382	-0.00166	-0.00031
Extension P0 x $\lambda = 0.9$	-0.01083	-0.00618	-0.00450
Extension P0 x $\lambda = 1$	-0.03198***	-0.02779***	-0.01882***
Extension FSPR x Number of Segments	-0.00004**	-0.00002**	-0.00002**
Extension BDL x Number of Segments	0.00008***	0.00007***	0.00006***
Extension P0 x Number of Segments	0.00003*	0.00003***	0.00001*

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

a clustering dimension for the length (P0), compressing consecutive purchases of the same type (FSPR), explicit consideration of bundles (BDL).

Additional controls summarize input parameters for all methods, namely the number of segments and the discount factor. The discount factor has the same levels $\lambda \in 0.1, 0.5, 0.9, 1$. The number of segments k was set to $k \in 10, 50, 200, 500$ and was included as a continuous variable. Also, interactions of the parameters are included in order to illustrate extension-specific differences in predictive performance. The baseline for the discount parameter is $\lambda = 0.9$. As a general effect, no temporal information with $\lambda = 1$ and therefore the highest sparsity of the data segments has shown a strong negative effect on the resulting lift value compared to the baseline level $\lambda = 0.1$. Interestingly, using $\lambda = 0.9$ and therefore preserving more temporal information than $\lambda = 0.1$ demonstrates the benefit of higher levels of sequential information. As to the number of segments k, the general effect per additional segment is positive.

Considering the different methods, bundling (BDL) shows a significant negative effect. Simultaneously, $\lambda = 0.5$ to $\lambda = 1$ have shown strongly significant positive effects with respect to predictive performance. This indicates that discounting of older purchases seems to be non-beneficial for this method. The interaction with the number of segments k has the highest positive effect among

this group of interactions. This fact supports the hypothesis that bundling needs a higher number of segments in order to outperform the competitors due to more complex representation and higher data sparsity, explained by a strong variance increase.

The extension of filtering for serial purchases (FSPR) is overall not significantly different from the WPC method except for two interactions. First, the interaction with $\lambda = 1$ is positive. This might be explained by the fact that the data is pre-aggregated by FSPR and therefore the sparsity is already reduced. Contrary to all other interactions with the number of segments, the interaction for this extension is negative, which can be explained by the previous aggregation, so that additional segments rather lead to overfitting and weaker performance.

Explicit modeling of the sequence by P0 has shown a weak positive effect for campaign sizes of up to 20 %. In contrast to all other methods, $\lambda = 1$ is strongly negative within the interaction. This shows the trade-off when considering an additional dimension without a preceding sparsity reduction through a discount factor leads to higher variance and therefore lower predictive robustness. The interaction with the number of segments is positive, indicating that a higher number of segments required. Having a more distorted distribution of sequence lengths, the positive effect for P0 was also significant on the method level for 20 % of customers (0.00804*) and more product interactions became significant across all percentiles.

Table 4.14 contains the interaction effects of products and different extensions. Only estimates significant for at least one campaign size are shown for reasons of clarity. Interactions of FSPR with Product 1 and Product 8 are significant. As shown in Table 4.10, Product 1 has a high share of serial purchases of the same number. The positive interaction with this product suggests that the number of same purchases is important for the prediction. Consequently, important information was lost by filtering serial purchases. For Product 2, also having a high share of purchases, such compression did not have an effect, suggesting no further predictive relevance of the same category for the purchase prediction of this product. For Product 8, the data reduction by FSPR was beneficial, meaning serial purchases are not important for predictive value.

More product-specific interaction effects were significant for BDL, which is an indicator for a strong influence of bundling modeling. As before, results for Product 1 and Product 2 were negatively influenced by this method, which suggests that the additional data complexity is not beneficial for these products, as well as for Product 5, Product 8, Product 9 and Product 10. Therefore, for these products there is clear advice to use the original WPC model, which solves the

Table 4.14: Interaction between product-specific and extension-specific estimates when regressing the logarithm of the lift value up to a certain percentile with (i) filtering for serial purchases (FSPR), (ii) bundling (BDL) and (iii) explicit sequence length modeling (P0). Baseline levels for the regression are Product 4 (having the lowest lift value) and the original WPC method.

Independent/Dependent Variable	Log (Lift 5 %)	Log (Lift 10 %)	Log (Lift 20 %)
Product 1 x Extension FSPR	-0.039***	-0.033***	-0.026***
Product 8 x Extension FSPR	0.021.	0.022**	0.019***
Product 1 x Extension BDL	-0.060***	-0.055***	-0.043***
Product 2 x Extension BDL	-0.022.	-0.021**	-0.016***
Product 5 x Extension BDL	0.005	-0.004	-0.012**
Product 6 x Extension BDL	0.025*	0.005	-0.004
Product 8 x Extension BDL	-0.071***	-0.072***	-0.063***
Product 9 x Extension BDL	-0.024*	-0.031***	-0.039***
Product 10 x Extension BDL	0.002	-0.004	-0.008.
Product 1 x Extension P0	-0.006	-0.008	-0.009*
Product 2 x Extension P0	-0.013	-0.009	-0.009.
Product 3 x Extension P0	-0.019	-0.017*	-0.016***
Product 5 x Extension P0	-0.012	-0.011	-0.009.
Product 6 x Extension P0	-0.009	-0.008	-0.010*
Product 8 x Extension P0	-0.036**	-0.028***	-0.022***
Product 9 x Extension P0	-0.017	-0.016*	-0.016***
Product 10 x Extension P0	-0.007	-0.011	-0.013**

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

bias–variance trade-off of sequences better with lower representation complexity. Slight improvements were achieved for Product 6. This is evidence that the previous purchase of bundles is an important indication for prediction of the purchase of the respective target product.

Considering the extension P0, the data set of the presented regression excluded sequences of length 1. Thereby, only the interaction with Product 8 was negatively significant meaning that additional dimension lead to a performance decrease. An additional analysis for the set including one-purchase sequences was conducted, as including the sequences of 1 leads to an even more skewed distribution of purchases. In large parts, the results were comparable. However, on product level, interactions of the extension P0 were significantly positive for several products, such as Product 1, 3, 5, 8, 9, 10. This leads to the conclusion that a certain level of "skewness" of sequence length distribution is necessary so that the additional dimension, resulting in representation complexity and variance increase, can become beneficial. For the available data, this was beneficial only if also taking one-purchase sequences into account.

4.5.3 Summary

The proposed extensions were motivated by specific characteristics of the data set of the sample corporation. However, these address frequent modeling issues in customer purchasing behavior, e.g. bundling or serial purchases.

The extension for sequence length modeling had a positive impact; a proposed question for the future research is the threshold of the "minimal skewness" for a distribution required for the extension to work. The argument is that an application of such a model extension might be interesting also for all domains having rather short event histories. This might be the case, for instance, in the area of medical histories, because most patients interact rather rarely with their physician. Providing an interaction length contains therefore crucial information about the patient history providing a more meaningful segmentation of these.

The second extension aggregates serial purchases of the same product, which has shown mixed results depending on the target product-specific dataset. This extension was not beneficial in cases where the number of subsequent purchases presumably had additional information, so that predictive performance decreased. For one product the application was beneficial. Besides the direct effect of performance improvement in the latter case, application of this model extension also contributed to the understanding of the target group of the respective product, namely that certain serial purchase behavior leads to a better prediction of the consecutive purchase. The value of understanding the customer is given also for the cases with performance decrease. For example, the serial consecutive purchases of domains and their number were important for anticipation of further purchasing intentions.

The extension for modeling bundles within a purchasing sequence was non-beneficial for all but one product. This can be explained by the fact that bundling requires higher representation complexity and therefore leads to a higher sparsity of data. This extension seems to be more beneficial if both, share and informative value of bundles are high, which does not seem to be the case for the available data.

As to the question of anticipation when a certain extension might be applied, this section proposed extension-specific metrics for the data set. Originally, these were used as a motivation, however, in some cases they also constituted an indication for a potential performance effect. These are the sequence length distribution, share of sequences with serial purchases and the share of sequences with bundle purchases. The dominance of short sequences in a data set was shown to be of importance, as the set with a more skewed set towards short sequences has

shown better performance when using the proposed extension of the sequence length dimension. The share of serial purchases was only a quantitative indicator to show the potential influence but not if the application would lead to a performance increase.

Finally, the proportion of bundle purchases was analyzed. Though the final application did not always increase predictive performance, the interpretative side was nevertheless valuable, as it has shown, that bundling modeling does not increase the predictive performance, therefore, bundles do not have sufficient predictive value for compensating the representation complexity.

Overall, the proposed metrics as well as the proposed modeling extensions for WPC method have shown to be useful for predictive performance as well as for business understanding, thereby contributing to research on sequence aggregation, especially in the context of the next purchase modeling and prediction.

4.6 Metrics for Temporal Information in Sequences

In times of large amounts of available data in information systems, it is of interest to estimate the value of the information contained in the data upfront, in order to anticipate the utility of further data collection and processing. This section is dedicated to metrics, which are aiming at determination of the sequential information value in a particular data set with respect to lift based on raw purchasing sequences and prior to any out-of-sample evaluation. Second, it links the proposed metrics to the aggregation level for sequential information, so that the aggregation level can be better estimated upfront. The proposed metrics are evaluated on the previously introduced data set. The results show the ability of the proposed metrics to capture the added value of sequential information and to provide guidance for data preprocessing and aggregation in analytical and predictive tasks.

4.6.1 General Framework for Sequence Metrics

Herein, a general differentiation into several tasks in the context of sequence aggregation is given, which is then used for categorization of the proposed metrics. The categorization is conducted along two dimensions. The first dimension regards the (un)supervision of a certain task (see Section 2.2).

In the context of the information value estimation an unsupervised metric can be specified as estimation of the sequential information without the considera-

tion of a target class. So in its essence the unsupervised metric quantifies the information over all sequences in the data set without a specific predictive goal (i.e., a target variable), so therefore making a general statement about the data set as a whole.

As the next step, one could proceed with an evaluation of the information with respect to the target class, thereby undertaking a supervised task. So the question here would be, if the sequential information contains any additional value with respect to the class to predict. Correspondingly, the supervised metrics are used for the estimation of the predictive performance.

The second dimension differentiates the proposed metrics into model-dependent and model-independent. In this context a model refers to an aggregation model for sequential data. Starting with the latter, the model-independent category has as the goal to estimate the value of the non-aggregated data set without any model applied to the data, so that the overall information contained in the non-aggregated data set is evaluated. In contrast, the model-dependent metric regards a certain aggregation and has the goal to quantify the value of sequential data after the aggregation. So in total, the goal of the evaluation using a model-dependent metric is to estimate the remaining information after an aggregation was done, also in comparison to the non-aggregated data.

Overall, a differentiation is undertaken between unsupervised (UNSUP) metrics that do not consider the label of the target category and supervised (SUP) metrics that aim at evaluating the information with respect to the usefulness for a predictive model given a target class. For supervised approaches, there is a further distinction between metrics computed over to the non-aggregated complete sequential data and those that work on a certain pre-aggregation of the data done by an aggregation model. The metrics for the non-aggregated data will be referred to as aggregation model-independent (MINDEP), the metrics computed on pre-aggregated data are referred to as aggregation model-dependent (MDEP). The metrics will be introduced next.

4.6.2 Specification of Sequence Metrics

The first group of eight metrics describes the data independently from target class and aggregation model (unsupervised). First, the average sequence length is measured for the purchasing sequence. The basic assumption is that the longer the sequences the higher the potential information value.

Second, the number of sequence instances refers to the number of unique sequence combinations (sequence variety), so a higher number of sequence in-

stances is suspected to be an indicator for higher sequential information.

The number of data instances refers to a total number of observed sequences (including multiple observations per instance). A higher number of observations leads to a more robust prediction.

The next metric, the average number of data instances per sequence, measures how many customers are available per sequence instance, so this metric is directly related to the variance in the data as higher number here would make the prediction more robust to small data perturbations and therefore reduce the variance.

The next group of metrics summarizes three measures, which are computed per sequence position, meaning that the ordered sequence is divided into last purchase, previous to last purchase and so on and the metrics are computed with regard to this position. As the data at hand is categorical data, if shorter sequences are in the set, these are filled up with the category "No Purchase".

The first metric in this group is the Shannon entropy of the position (Lin, 1991). The entropy is the classical measure of the informational value: the higher its value the lower is the uncertainty reduction given the segments, or in this case preceding purchases.

An idea for measuring the "skewness" of the distribution over preceding purchases is reflected in the next metric, namely the predominance of a symbol Shenkin et al. (1991). It measures the proportion of the most frequent class in the data, thereby quantifying how skewed a distribution is towards the most frequent class. The higher this metric the lower is the informational value of such information. This fact can simply be visualized by the extreme example of a position for which only one class exists, therefore no additional information is provided by this position.

The third metric also relates to the distribution of the products in the position and is closely related to the entropy, namely the Kullback-Leibler divergence (Kullback and Leibler, 1951). Using the discrete variant of this metric quantifies the difference between any reference distribution and the given one. The distance to the uniform distribution was chosen as it reflects the maximal entropy. Therefore, a higher distance over all classes would mean a higher potential information value.

As the number of positions in a sequence can be quite high, the proposition is to use principal component analysis (PCA) for dimensionality reduction for this as well as for all position-dependent metrics.

The supervised metric used herein is the baseline, i.e., the probability of the target class in terms of the proportion of positive cases. The intention of using

the measure here is to quantify how difficult the separation of the positive class would be and how high the maximally achievable lift is (per definition it is the reciprocal of the baseline).

The last metric of this block is the average number of instances per model segment and is closely related to the one for the complete data, but is applied to the segments instead of sequence instances. It is used in order to estimate the sparsity of the data, therefore, also the expected variance component of the aggregation model.

As to the third and fourth metric of the last block, it has the same idea but as average number of instances, it is applicable to both, complete data (model-independent) as well as for a model-dependent evaluation. The idea of this metric is to compute the distance between the baseline of the target distribution and the baseline of the sequence instance (or segment). In order to generalize over the data set, the value of the respective segment is weighted with the number of observations per sequence and summarized. Overall, this measure computes the possible supervised information in the data set, therefore a higher value indicates a higher potential value of the sequences.

An overview over the proposed metrics is given in Table 4.15. The table categorizes the metrics into respective categories unsupervised (UNSUP), supervised (SUP), model-dependent (MDEP), model-independent (MINDEP). All unsupervised metrics can be clearly used also for supervised tasks. However, as the distinguishing dimension is with respect to the task (and implicitly the data not including a target class), these are assigned to the unsupervised group.

Table 4.15: Considered metrics for sequential information and their types: unsupervised (UNSUP) vs. supervised (SUP), model-dependent (MDEP) vs. model-independent (MINDEP).

Proposed Metric for Sequential Information	UNSUP	SUP	MINDEP	MDEP
Average sequence length	x		x	
Number of sequence instances	x		x	
Number of data instances	x		x	
Average number of data instances per sequence	x		x	
Position dependent entropy	x		x	
Distribution of sequence symbols using	x		x	
Kullback-Leibler divergence (compared to uniform distribution)	x		x	
Predominance of a symbol	x		x	
Baseline of consequent target purchase		x	x	
Average number of instances per model segment	x			x
Squared distance to baseline per sequence (weighted)		x	x	
Avg. sq. distance to baseline per model segment (weighted)		x		x

4.6.3 Empirical Evaluation of Proposed Metrics

The evaluation data set is the same, which was used throughout this chapter. Overall, the training data from 10-fold cross-validation is used for the estimation of the proposed metrics. For this evaluation two lift values are interesting. The first is the training (in-sample) lift, which is achieved using training data and evaluation lift in-sample, i.e., only for segmentation and not for prediction. This is done in the way that for a certain sequence (or sequence segment) in the training data, the respective conversion rate (CR), i.e., percentage of the buyers within a sequence (segment), is computed. As lift is the measure of the value for sequential information later, the sequences are sorted with respect to a descending conversion rate and select the best 10 % of customers on the training data. The lift on the test data is then used as a measure for the information value of the out-of-sample test.

First, the presence of a (linear) relationship of the proposed metrics and both lift values was evaluated using Person correlation coefficient. Second, in order to address the value of sequential information for prediction, the lift on the test data is used as the dependent variable in a linear regression model using the metrics computed on the training data as dependent variables. The percentage of the variance explained is evaluated in this way. Additionally, the same regression is used for training data lift performance as a measure of how the in-sample performance is captured by the metrics.

Furthermore, six sample sizes from the complete training data set per product, namely 1/2, 1/4, 1/8, 1/16 and 1/32 were employed in order to investigate the effects of downsampling.

The aggregation levels used are the non-aggregated data referred to as complete sequence ($SSM_{5,5}$), the last purchase ($SSM_{1,1}$) and portfolio ($SSM_{5,0}$). $SSM_{1,1}$ states the strongest sequence aggregation level with respect to the sequence order, as it has the minimal number of levels and therefore the highest bias and the lowest variance. In contrast, $SSM_{5,5}$ has the highest variance as the model has the highest granularity, where a sequence can have only one observation, and therefore be maximally unstable towards data perturbations, simultaneously having the lowest bias. Aggregation level $SSM_{5,0}$ is used to evaluate if there is an additional value over a non-sequential representation of the purchasing history. This level contains no sequential information and is used in order to anticipate the value of sequential information over a more simple aggregation model without temporal information.

In the following, the empirical results of the evaluation are presented. First, a

simple Pearson correlation coefficient of the training and test lift with all available metrics is conducted in Table 4.16. This evaluation is done on the raw sequential data set, therefore only the model-independent metrics are involved.

A higher sequence length leads to a higher performance, however, the strength of the relationship decreases slightly from training to test lift, probably due to overfitting on training data and less robust inference on test data.

On the other hand, a higher number of sequence instances, i.e., sequence variety in the data set, increases test lift. This relationship seems contra-intuitive, as a higher number of sequences should identify a higher data sparsity. However, this might be due to the fact that the number of sequences is positively correlated with the sample size, which on the other hand increases the robustness of the prediction. The latter is shown in the next row of the table and constitutes a well known relationship of the sample size (number of data instances) and prediction quality. However, the importance declines from training to the test data.

A clear illustration of the variance aspect is given with the number of data per sequence. While it is not beneficial to have large segments in-sample, the sign of the correlation switches from training to test data. So a higher support (number of observations per sequence) is clearly an important variance-reducing factor.

For the next three metrics a principal component analysis (PCA) is conducted for the five position-dependent results corresponding to the five sequence positions. In all cases a strong correlation with lift values can be noted.

As to the baseline, the reciprocal relationship of the baseline is well established

Table 4.16: Pearson correlation coefficients for proposed metrics with training and test lift for 10 % of customers. All coefficients are significant; except for correlation of the number of sequence instances with training lift. The second column contains the respective abbreviation used in later analyses.

Person Correlation Coefficient	Abbreviation	Training Lift	Test Lift
Average sequence length	ASL	0.155***	0.108***
Number of sequence instances	NSI	-0.020	0.312***
Number of data instances	NDI	0.699***	0.534***
Average number of data instances per sequence	ANPS	-0.209***	0.169***
Position dependent entropy (PCA 1)	PDA1	0.689***	0.550***
Position dependent entropy (PCA 2)	PDA2	0.340***	0.297***
Kullback-Leibler divergence (PCA 1)	KLD1	0.451***	0.426***
Kullback-Leibler divergence (PCA 2)	KLD2	-0.213***	-0.256***
Predominance of a symbol (PCA 1)	PDS1	0.653***	0.492***
Predominance of a symbol (PCA 2)	PDS2	0.173***	0.142***
Baseline of consequent target purchase	BL	-0.749***	-0.581***
Squared distance to baseline per sequence (weighted)	SDW	-0.626***	-0.415***

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

and only serves as a validity check. Interestingly, the relationship is weaker for the test data, indicating a smaller importance.

The average distance to the baseline, which was expected to be a strong indicator of predictive performance, also has a reciprocal relationship. At first sight, the sign seems contra-intuitive, as the higher distance should be beneficial. However, this measure seems to capture the selection bias of the data. Considering low baselines and, especially for small sample sizes, a very low support for all sequences, the sum of distances gets higher as the segments tend to contain buyers and non-buyers only.

The next step is to assess how well the metrics explain the achieved lift in combination. Therefore, the variables were checked for collinearity, where the variables average number of data instances per sequence (ANPS), first and second PCA component of Kullback-Leibler divergence (KLD1 and KLD2) and the baseline (BL) did not show a mutual correlation over an absolute value of 0.8. After a Box-Cox transformation of training and test lift with a coefficient of $\lambda_{Box-Cox} = -0.828$ in order to fulfill normality requirements of errors in linear regression, model-independent regression outcomes are displayed in Table 4.17.

While the absolute regression coefficients are not directly interpretable due to a preceding transformation, the direction and ranking still remains the same. As in the correlation analysis, the average number of data instances changes its sign. The Kullback-Leibler divergence is difficult to interpret due to a preceding PCA, however both components are significant. The baseline has a significant negative coefficient, therefore being plausible in the context of the regression. The variables overall show significant effects and explain 88 % of variance for training and 63 % for test data.

In a next step, the model-dependent performance evaluation is presented in

Table 4.17: Results of regressing the Box-Cox transformed lift value for training and test data. All considered variables show significant effects and explain 88 % of the variance for training and 63 % for test data.

Independent variable	Training Lift	Test Lift
(Intercept)	1.014***	0.650***
Average number of data instances per sequence	-0.003***	0.003***
Kullback-Leibler divergence (PCA 1)	-0.014.	0.050*
Kullback-Leibler divergence (PCA 2)	-0.277***	-0.514***
Baseline of consequent target purchase	-3.289***	-3.453***
R^2	0.883	0.640
R^2_{adj}	0.882	0.637
Degrees of freedom	595	595

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

Table 4.18. All correlations of the metric variables are under 0.35 in their absolute value. However, as in the previous regression, a Box-Cox transformation for the data is used. The additional variables in this regression are dummies for the aggregation models. The strongest aggregation of sequences is used, namely the category of the last purchase ($SSM_{1,1}$). Further categories are portfolio ($SSM_{5,0}$) without the sequential aspect and the complete sequence ($SSM_{5,5}$).

The coefficient for the baseline has again a strong influence with a plausible direction. The portfolio dummy plays a week role on training data, switching to test data the performance is not significantly better compared to the last purchase. Consequently, it is better suited for segmentation but not for prediction. On average, the complete sequence performs worst, which is not surprising, especially on test data, as it tends to overfit strongly.

A higher weighted distance per segment (SDW) has a positive influence on the performance. As the baseline level is the last purchase, the corresponding level of this metric is very low. During this measure in the context of the correlation analysis on the non-aggregated data in Table 4.16 was a measure of the selection bias, the result here is rather an informational gain. The interaction of this variable with model dummies signifies the benefit of the additional potential information for the model. Thereby, the complete sequence profits stronger as it is able to incorporate more information due to a higher number of levels than the portfolio.

Although the direction of the next variable, the average number of data instances per sequence segment, seems non-intuitive, for the baseline model of the last purchase it is plausible. As a minimal number of segments possible with

Table 4.18: Results of regressing Box-Cox transformed lift value on training and test data with model-specific interaction effects.

Independent variable	Training Lift	Test Lift
(Intercept)	0.878***	0.7474***
Baseline of consequent target purchase	-6.104***	-4.736***
Portfolio Dummy $SSM_{5,0}$	0.027.	0.020
Complete Sequence $SSM_{5,5}$	-0.220***	-0.170***
Avg. sq. weighted distance to baseline per model segment (SDW)	0.015***	0.011***
Average number of data instances per sequence (ANPS)	-0.00002***	-0.00001***
Portfolio x SDW	0.198***	0.158***
Complete Sequence x SDW	1.993***	1.562***
Portfolio x ANPS	-0.0001*	-0.0001*
Complete Sequence x ANPS	0.003***	0.002***
R^2	0.526	0.535
R^2_{adj}	0.523	0.533
Degrees of freedom	1790	1790

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

respect to purchases is used, larger segments lead to a worse performance in this case. The direction changes strongly for the counterpart, complete sequence. These coefficients can be interpreted as a hint for a possible direction of the information aggregation. For this case, some aggregation methods in-between of portfolio and complete sequence would be a more proper choice, as the positive coefficient for the complete sequence can be interpreted as a lack of support. However, the portfolio seems not to be in the need of further support.

Overall, the explanatory power of the models in Table 4.18 declined strongly compared to a generally estimated result for non-aggregated data in Table 4.17. This can be interpreted as a potential for an investigation of further metrics or different relationship (e.g. quadratic), which are responsible for the variance of the lift performance.

4.6.4 Summary

This section presented a framework for estimating the value of purchasing sequences for the lift achieved in-sample and out-of-sample. Thereby, the categorization of such metrics with respect supervision (unsupervised/supervised) and with respect to a potentially used aggregation model (model-independent/model-dependent) was conducted.

Using the same data set of a telecommunication provider, it was shown that the proposed metrics are strongly related to the lift performance for training and test data. Combination of proposed metrics, such as for the baseline, number of instances and Kullback-Leibler divergence, were able to explain about 88 % of training and 64 % of the test variance of lift. So overall, the metrics are capable to anticipate the predictive value of sequential information to a high degree.

Also, the regression analysis provided a guideline for a direction of aggregation in the model-dependent analysis. However, the explanatory power of a model-dependent regression was lower. Therefore, this investigation is a potential starting point for an extensive study of possible metrics, as only a selection of the most common and intuitive ones was provided, so the gap in explanatory power can be filled.

Chapter 5

Churn Prediction Using Purchasing Histories

In this chapter, customer-specific contract duration dates are considered within established methods of survival analysis in order to predict churn probabilities for access products. A novel feature generation procedure for contract term is introduced. In addition, a study of the impact of product variety in a customer's portfolio on his churn probability is conducted, as there is evidence from both, theory and practical experience in other industries that product variety can be related to loyalty. In the empirical part of the chapter, proposed extended model is evaluated using data of the telecommunication company. Results show that generated features significantly increase churn prediction performance in out-of-sample tests.

5.1 Methodological Background on Survival Analysis

In the context of churn there is a difference between prevention and retention campaigns, whereby the first precedes the churn notification and the second aims at winning the customer back after a cancellation notification. This chapter uses the established tool of survival analysis for the prediction of a cancellation notification, which can be used for selection of customers for prevention campaigns.

Survival analysis models customers' time-dependent churn probabilities. For this purpose, the time to a churn event as well as relevant covariates are recorded for every customer, starting at a certain point in time (*origin of time*). Customers without churn events in the time period considered are *censored* observations. For a censored observation information is available only starting from a certain

or up to a certain point of time. The latter is the case in this setting, as at every point of time it only known if a customer have churned so far. Therefore, the question when the customer will churn remains open. Based on this censored observations, multivariate models are trained to estimate the parameters for several covariates.

The most common techniques for survival analysis are accelerated failure time modeling and the Cox proportional hazards model. Both techniques are briefly introduced in the following.

The *Accelerated Failure Time Model (AFT)* is a parametric model discussed in Lawless (1982) and shown in (5.1).

$$log(T_c) = Z_c\beta + \sigma\epsilon, \tag{5.1}$$

In this model $log(T_c)$ is the logarithm of a customer c's churn time, Z_c is the vector containing c's covariate values, and β is a vector of global regression parameters for all customers on the covariates. σ is a scaling parameter for the model's overall error distribution ϵ. AFT assumes that the time to churn event T_c is exponentially distributed and that every covariate either accelerates or decelerates the underlying process (Wei, 1992).

The *Cox Proportional Hazards Model* is a semi-parametric model shown in (5.2).

$$\lambda_c(t) = \lambda_0(t)exp(Z_c\beta), \tag{5.2}$$

The hazard function $\lambda_c(t)$ quantifies the risk that customer c churns in an interval later than t, i.e., c's conditional churn probability given that c already survived t time periods. As in AFT, Z_c is a vector of covariate values for customer c, and β is a vector of regression parameters. In order to allow for a straightforward comparison of both models, Equation (5.2) is logarithmized. The result is shown in (5.3). Comparing (5.3) and (5.1), the key difference between AFT and the Cox regression model is the time-dependent (but individual-independent) baseline function $\lambda_0(t)$. $\lambda_0(t)$ does not have to be specified explicitly and can have any shape, which classifies this model as semi-parametric. Similar to the AFT model, the Cox model assumes a constant influence of the covariates over (logarithmized) time and customers.

$$log(\lambda_c(t)) = Z_c\beta + log(\lambda_0(t)) \tag{5.3}$$

In summary, both models assume a monotonous, additive influence of covariates on the logarithm of the time to a churn event, T_c in case of AFT, and churn

probability in case of Cox regression. In a contractual setting, for instance in the telecommunication industry, these assumptions are questionable due to contract duration and term of notice clauses, where certain time intervals can be expected to have higher churn likelihood than others, such as the time shortly before the minimum subscription time ends, potentially with a lead time corresponding to the term to notice. However, these influences are customer-specific. As a consequence, models are required that consider each individual customer's contract dates and durations.

5.2 Proposed Features for Churn Prediction

In this section the description of *Contractual Information* is given, which can be used to extend a customer's portfolio information at a particular point in time. From the data, three different representations of product portfolio information are generated. The representations are calculated as follows. Let $i = 1,\ldots,I$ be all possible product categories, from which a customer can obtain a product. The logical variable $x_{c,i}$ indicates c's possession of at least one contract within category i. First, the feature $s_c = \sum_{i=1}^{I} x_{c,i}$ is defined, indicating the number of product categories within c's portfolio. This feature represents the highest aggregation level of portfolio information for a customer that is considered here. The intuition of this feature is that a higher number of product categories might lead to a longer tenure with the company, as shown in Kamakura et al. (2003) for the financial industry.

Second, the binary variable $x_{c,i}$ itself is used as a feature, resulting in three features (as three categories are considered) per customer. These indicate possession of at least one contract within i. Using $x_{c,i}$, the goal is to investigate whether differences in a customer's tenure depend on a particular product category, as has also been observed in the financial industry (Van den Poel and Lariviere, 2004). Third, the feature $X_c = \{x_{c,1},\ldots,x_{c,3}\}$ contains the complete (binary) representation of a customer's portfolio which corresponds to $SSM_{L,0}$ from Chapter 4 for the three main categories of access products for the available data set. For instance, the customer $c = 1$ with $X_1 = \{1,0,0\}$ owns at least one contract in category $i = 1$, and no product in any other category.

In this work, churn is considered as the complete cancellation of all contracts corresponding to a resulting portfolio $X_{churn} = \{0,0,0\}$. In terms of regression analysis, this portfolio representation corresponds to an interaction effect of the category-specific $x_{c,i}$ values. Using this feature this section investigates whether

certain product combinations in customers' portfolios are correlated with significantly different duration of customer tenure.

Besides portfolio-related features, the influence of *Last Contract*, the time elapsed since the last contract activation, is studied.[1] The general idea is that a complete churn is only possible when the minimal contract duration of the last activated contract is reached (as complete churns are considered as churns in the context of this work). This feature will be used to model the influence of minimum contract duration and term of notice lead times in a survival model. Therefore, the last contract activated serves as an anchor for the customer and determines the duration a customer will at least stay with the company.[2]

In the following a description of how the features developed from *Last Contract* data are used in the prediction model. The intuition is illustrated in Figure 5.1.

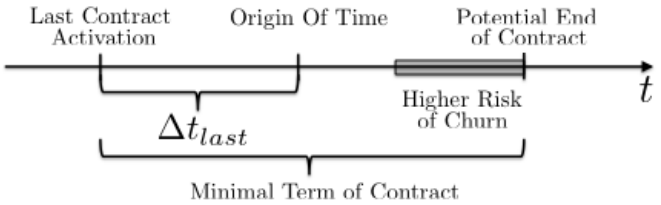


Figure 5.1: Schematic illustration of the relationship between Δt_{last} and contract duration. A customer's last contract activation determines a customer's shortest potential time to churn. Higher churn announcement (cancellation) probabilities are expected when approaching the end of the contract duration (shown as gray area).

Given the origin of time as the beginning of the observation within the study, Δt_{last} is defined as the time since the last contract activation. Therefore, Δt_{last} corresponds to the number of months a contract already existed at the beginning of the observation period. Although a cancellation notification can occur at any time during the contract duration, e.g. as a result of an interaction with the company, it is assumed that the cancellation probability is not uniformly distributed over the contract duration. In general, a customer might be more sensitive to – and aware of – offers by other companies when approaching the end of the contract duration. Therefore, the hypothesis is an accumulation of contract cancellation notifications shortly before the end of contract durations.

[1] Please note that a snapshot of the portfolio information at the origin of time is used.

[2] As alternative measure one might use the time (from now) to the latest subscription deadline of a customer, i.e., when the last contract will expire. Unfortunately, this information is not available in the data. However, almost all contracts have the same duration, and therefore the proposed predictor is expected to capture similar information and predictive value.

Based on the assumption of non-uniformly distributed cancellation probabilities over the contract duration, a transformation function $p_{canc}(\Delta t_{last})$ is proposed. The goal is to map the number of months since the last contract activation Δt_{last} to an empirical cancellation frequency from historical data, taken as a proxy for cancelation probability p_{canc}. By doing so, a better predictive performance is expected in comparison to using Δt_{last} as a feature. This is because, if the churn risk indeed has a non-linear behavior around the end of contract term, the additive (linear) character of the model is not able to capture this development solely based on Δt_{last}.

The probability distribution is then estimated by means of a kernel function, which smoothed possible peaks. Another possibility is to use the unsmoothed conditional probability $p_{canc}|\Delta t_{last}$. Additionally, $log(\Delta t_{last})$ is tested, as it scales the feature to $log(T_c)$, the dependent variable.

The values of $p_{canc}(\Delta t_{last})$ are obtained by using the contract duration of historical values for churn frequencies given the number of months of the contract existence Δt_{last}. The proposed features are listed in Table 5.1, together with their notation.

Table 5.1: Generated predictors by group. The first group comprises portfolio representations at different aggregation levels. The second group is related to transformations of data on a customer's most recent contract activation.

Predictor Group	Predictor	Notation
Portfolio Information	Number of Categories	Model s_c
	Contract Category Possession	Model $x_{c,i}$
	Portfolio Representation	Model X_c
Last Contract	Probability Distribution	Model $p_{canc}(\Delta t_{last})$
	Conditional Probability	Model $p_{canc}\|\Delta t_{last}$
	Logarithm of Δt_{last}	Model $log(\Delta t_{last})$

5.3 Empirical Evaluation of Proposed Features

This section describes the available data, outlines the evaluation design and presents the results. First, the predictive performance for cancellation notifications is evaluated using lift. Thereby, the models without the proposed features are compared to the ones with the proposed features. In the descriptive analysis the estimated coefficients are interpreted. Furthermore, an ex-post analysis of the distributions for the proposed features is conducted.

5.3.1 Available Data

The available data comprises contract data in three different product categories: Internet access for households, mobile telephone services (incl. data volume) as well as mobile Internet services for (tablet) PCs. Data are available for 374 942 customers from January 2002 to January 2015, including the purchasing history with timestamps of contract activation, cancellation notification and product category.

The covariates of the study span several categories and are presented in Table 5.2. The first category, *Customer Information*, summarizes general customer attributes like customer type (business/private), tenure with the company, and several more attributes. The data also includes historical average values for some covariates. These averaged values are prepared for the last 3, 6, 12, and 24 months into the past from the origin of time (that can be set to any tome in the past). However, these values are highly correlated, and this collinearity naturally provides a problem for any regression method. The procedure dealing with collinearity will be presented later in Section 5.3.

The next category, *Product Usage*, includes usage information such as the average duration of telephone conversations and average data transfer volume (per month). Revenue data include the regular average monthly revenue (monthly fee) as well as on-top revenue for services not included into monthly fees (e.g., roaming). *Contact history* data includes the number of support cases initiated by a customer. Additionally, information on the number of contacts within the firm's customer base is known, e.g. through product recommendations, reflecting how many persons within the circle of acquaintances are also firm's customers.

The last category, *Portfolio Information*, comprises data related to a customer's

Table 5.2: Available customer data attributes by group, which are used as predictor in a basic model serving as a baseline for all additional models in the empirical study.

Predictor Group	Predictor
Customer Information	Customer Type
	Tenure (month)
Product Usage	Average Telephony Duration
	Average Data Volume
Revenue Data	Average Regular Revenue
	Average On-Top Revenue
Contact History	Number Support Cases
	Number of Contacts
	Number of Interactions Online
Portfolio Information	Number of Terminated Contracts
	Last Contract Activation

contracts, such as the number of terminated contracts, each customer's product portfolio at a given point in time, the dates of contract activations and cancellation notifications (when churns have been announced), and much more information.

5.3.2 Evaluation Design

In the presented analysis, the predictive accuracy of the two introduced models is studied, AFT and the Cox proportional hazards model. The predictive accuracy of the models is compared, considering the proposed additional features, with the outcomes of the models when solely the features in Table 5.2 are considered.

The benchmark criterion is lift, as it is the standard evaluation metric in marketing not only for cross-selling but also for churn prediction. The previously used benchmark of the top-decile lift is also used in this evaluation (Neslin et al., 2006). The lift values are determined for the prediction horizons of 1, 3, 6, 9, and 12 months (where the observed churn events take place) as done by Lu (2002).

The origin of time is set to the 6th of January 2014 and churns are then observed over a period of one year. A prior total churn rate of around 6.5 % was measured over the complete period of 12 months. As mentioned before, churn is defined as the complete cancellation of all contracts corresponding to the resulting portfolio $\{0,0,0\}$. Hence, churn is defined for a complete product portfolio. The time of churn event T is defined as the cancellation notification related to the last contract of the portfolio.

Initially, the data is splitted to 60 % for model training and 40 % as validation set. As a preliminary evaluation showed a random downsampling of non-churners to a proportion of 1:2 (churner to non-churner) results in better results, a total of 42 414 customers was used for model building. The test set includes 154 566 observations with an overall churn rate of 6.4 % within 12 months. All models are built and evaluated on the same data samples. To check the hypothesis that a transformation of Δt_{last} into empirical cancellation probability increases predictive accuracy a model using Δt_{last} is included in the analysis.

The estimation of the survival models is done in SAS using the LIFEREG procedure for AFT and the PHREG procedure for the Cox proportional hazards model. In order to estimate the transformation function $p_{canc}(\Delta t_{last})$, the UNIVARIATE procedure is used in combination with a normal kernel density function for multimodal distributions. For AFT, different distribution functions were tested for the error term, such as the Weibull, lognormal, log-logistic and expo-

nential distribution. The log-logistic distribution has shown best results and is chosen for the analysis.

As for some predictors the historical average values for 3, 6, 12 and 24 month into the past are also known, e.g. for *Average Telephony Duration*, this additional features are also incorporated into the models. However, these values are highly correlated per covariate (an observed Pearson correlation coefficient above $r = 0.80$ for several covariates). To decorrelate, a principal component analysis (PCA) is applied and project a covariate's values onto the first principal component, which, in this case, captures over 80 % of the variance within a predictor group.

5.3.3 Predictive Results

Empirical results with the different models studied are shown in Table 5.3. Each row represents lift values per model. The table is divided into four parts; in each part, models are sorted by lift (1 month) in descending order. The first part (at the top of the table) contains models with only the basic predictors (*Basic* model); the next two parts show the outcomes with models with one additional feature, grouped by models considering portfolio or last contract information. The part at the bottom of the table shows the outcomes with the best AFT and Cox models, considering baseline predictors, *Portfolio Data* and *Last Contract* data.

Table 5.3 shows that including portfolio features clearly improves the performance of *Basic* models for all portfolio representations. Amongst the portfolio-aware models, s_c, that includes the number of categories in a portfolio, adds the lowest value to the *Basic* model. Model $x_{c,i}$, which includes the logical variable of contract category possession, further improves predictive accuracy. Therefore, as stated in the hypothesis, it is not only important from how many categories a customer has products but in which particular category a customer has a contract. The best performance, among the models including portfolio information, is achieved with the categorical representation of the portfolio (model X_c). Hence, model X_c, that explicitly considers the different combinations of product categories, leads to the best predictive value of the portfolio-aware models.

In the following the performance of the models, when considering the last contract activation time, is studied. Model Δt_{last} improves the performance over *Basic* for both model types. The proposed transformation of months since the last contract activation into the corresponding churn probability further improves lift values (model $p_{canc}(\Delta t_{last})$); for long-term prediction over 12 months, even by 33 % (for the Cox regression model). The feature $p_{canc}|\Delta t_{last}$, which corre-

Table 5.3: Aggregated empirical results over all models and time horizons. Each row corresponds to the outcomes with one model, sorted (descending) by the top-decile lift. The lift values are shown for different observation period durations, from one to twelve months. The table shows that including both features groups, portfolio information as well as information on a customer's last contract activation time, improves lift. The two "Best" models at the bottom of the table combine the basic predictor set with the best-performing feature representation of both groups.

Model Type	1mth	3mth	6mth	9mth	12mth
AFT Basic Model	1.806	1.674	1.563	1.483	1.440
Cox Basic Model	1.763	1.637	1.520	1.447	1.410
Cox Model X_i	2.493	2.305	2.117	2.026	1.951
AFT Model X_i	2.449	2.268	2.097	2.019	1.951
Cox Model $x_{i,c}$	2.436	2.249	2.062	1.987	1.920
AFT Model $x_{i,c}$	2.402	2.235	2.069	2.008	1.936
Cox Model s_i	1.897	1.768	1.634	1.508	1.469
AFT Model s_i	1.875	1.742	1.619	1.497	1.452
Cox Model $p_{canc}\|\Delta t_{last}$	2.225	2.014	1.810	1.720	1.656
AFT Model $p_{canc}\|\Delta t_{last}$	2.212	2.005	1.810	1.715	1.651
Cox Model $log(\Delta t_{last})$	2.091	2.030	1.950	1.880	1.816
Cox Model $p_{canc}(\Delta t_{last})$	2.065	2.123	2.029	1.980	1.924
AFT Model $p_{canc}(\Delta t_{last})$	2.056	2.128	2.040	1.995	1.927
AFT Model $log(\Delta t_{last})$	2.022	2.037	1.957	1.873	1.808
AFT Model Δt_{last}	1.892	1.812	1.727	1.634	1.591
Cox Model Δt_{last}	1.845	1.637	1.679	1.589	1.547
Cox Model *Best*	2.583	2.395	2.175	2.071	1.980
AFT Model *Best*	2.570	2.393	2.172	2.074	1.981

sponds to the conditional churn probability, performs better for short-time prediction of one month only. Usage of $log(\Delta t_{last})$ leads to a performance between $p_{canc}(\Delta t_{last})$ and $p_{canc}|\Delta t_{last}$ for all time horizons. In summary, all three transformations of Δt_{last} outperform the *Basic* model as well as model Δt_{last}.

Finally, the combined models at the two rows at the bottom of the table are discussed. The models contain the portfolio predictor X_c and $p_{canc}|\Delta t_{last}$ as best performing combination of the two feature groups and allow achieving a better performance over all time horizons compared to any other model.

Overall, a prediction for a longer time interval seems to be more difficult for all models, although the number of available churners increases over time. Using the same data and predictors also allows a comparison of both model types, AFT and Cox, indicating that Cox regression provides slightly better results in this setting. In general AFT model leads to better results only for the *Basic* model and model Δt_{last} combined with durations not exceeding three months.

Table 5.4: Hazard ratio estimates for Cox regression models based on the standard-features (*Basic*), and the additional features proposed in this work. The regression baseline category (BL) for interpretation of categorical variables is given in brackets.

Feature	*Basic*	s_c	$x_{c,i}$	X_c	Δt_{last}	$log(\Delta t_{last})$	$p_{canc}(\Delta t_{last})$	$p_{canc}\|\Delta t_{last}$	*Best*
Business Customer (private)	0.869***	0.888**	0.930.	0.932.	0.853***	0.838***	0.875**	0.871***	0.937
Tenure months	0.996***	0.996***	0.996***	0.996***	0.993***	0.992***	0.997***	0.997***	0.997***
Average Telephony Duration (PCA)	0.918***	0.928***	0.928***	0.927***	0.919***	0.913***	0.917***	0.918***	0.926***
Average Data Volume (PCA)	0.997	0.984	0.939***	0.941***	0.999	0.997	0.994	0.995	0.940***
Average On-Top Revenue (PCA)	0.993	0.986	0.969**	0.965 **	0.992	0.996	0.997	0.994	0.966**
Number of Interactions Online	0.983***	0.991.	0.995	0.996	0.992.	0.999	0.983***	0.982***	0.995
Number Support Cases (PCA)	1.001.	1.001	1.001	1.001.	1.001	1.001	1.001.	1.001.	1.001
Number of Contacts (PCA)	0.836***	0.847***	0.877***	0.875***	0.846***	0.845***	0.829***	0.832***	0.873***
Average Regular Revenue (PCA)	1.030***	1.066***	1.129***	1.138***	1.039***	1.021**	1.014.	1.023	1.132***
Number of Terminated Contracts (PCA)	1.033***	1.039***	1.03***	1.029***	1.054***	1.054***	1.027**	1.030***	1.025***
$s_c = 1$ (BL $s_c = 3$)		1.486***							
$s_c = 2$ (BL $s_c = 3$)		1.079							
$x_{c,1} = 0$ (BL $x_{c,1} = 1$)			0.935**						
$x_{c,2} = 0$ (BL $x_{c,2} = 1$)			1.766***						
$x_{c,3} = 0$ (BL $x_{c,3} = 1$)			1.509***						
$X_c = \{0,0,1\}$				1.654***					1.714***
$X_c = \{0,1,0\}$				1.404***					1.432***
$X_c = \{0,1,1\}$				0.988					0.992
$X_c = \{1,0,0\}$				2.896***					2.957***
$X_c = \{1,0,1\}$				1.748***					1.752***
$X_i c = \{1,1,0\}$				0.903					0.927
Δt_{last}					1.006***				
$log(\Delta t_{last})$						1.249***			
$p_{canc}(\Delta t_{last})$							1.176***		
$p_{canc}\|\Delta t_{last}$								1.090***	1.111***

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

5.3.4 Interpretation of the Parameter Estimates

In contrast to more complex methods like neural networks, AFT and Cox regression allow a straightforward coefficient interpretation. The coefficients of the Cox proportional hazards model are now presented and discussed, as it showed slightly better predictive results. Table 5.4 displays the estimates for the models studied, including the models with the best combination of the basic predictor set, X_c and $p_{canc}|\Delta t_{last}$.

First, the basic estimates are interpreted (*Basic* Model). The Hazard ratio is an exponential parameter estimate and can be directly interpreted. For instance, one additional online interaction with the customer within the last 3 months reduces the hazard rate by 1.7 %. An additional year of tenure with the company reduces churn risk – on average – by $12 \cdot 0.4 = 4.8\%$. As the other covariates are obtained using principal component analysis (marked with *PCA*), these do not allow a direct quantification of their impacts but only the direction of the influence. The direction of each covariate remained the same. For instance, higher values for the historical values of revenue results in higher values for the principal component, also resulting in strong positive correlation with each original covariate. In general, higher numbers of contacts within the firm, higher average telephony usage, and higher regular revenue have a significant positive effect on tenure, while the number of terminated contracts has a significant negative influence on the tenure. Interestingly, private customers show a longer tenure with the company.

As the first level of portfolio information, the portfolio variety s_c is used. The regression baseline category is $s_c = 3$, corresponding to the maximal portfolio variety (meaning a customer possesses at least one contract in each category). According to the Cox regression results, it does not significantly influence the tenure whether a customer has contracts in two or three categories, as the effect of $s_c = 2$ is not significant in comparison to the baseline category of one purchase. However, customers with only one contract category show a significantly higher churn risk of 48.6 % compared to the baseline category $s_c = 3$.

Next, the results for the binary representation of contract possession are presented. The estimates are measured to the baseline of each category $x_{c,i} = 1$ and meaning that the hazard ratio is interpreted for the non-possession of the contract category. Interestingly, the possession of a product of the category $x_{c,1} = 1$ is connected with a higher churn risk.

The last portfolio category summarizes the effects of the product combination obtained by the customer, where the baseline is the complete portfolio, also cor-

responding to $s_c = 3$. As previously shown by model s_c, the number of products influences the tenure of a customer. All variations of $s_c = 1$ show significantly higher hazard ratio. But there are differences depending on the product combination for $s_c = 2$, namely only portfolio $X_c = \{1,0,1\}$ is significantly worse for the tenure compared to other combinations of products. This makes $x_{c,2}$, i.e., the second product category (mobile telephony) a rather critical product within two-category combinations with respect to long customer relationships. Interestingly, some variables are not significant as the portfolio information is added. For example, the dummy for private customer is not significant anymore. One potential explanation is that portfolio information provides a better differentiation between customer groups.

As to the features related to the *Last Contract* data, all of these show a significant influence. As expected, higher Δt_{last} and $log(\Delta t_{last})$ values lead to higher churn probabilities. In other words, churn (announcement) probability increases when approaching contract end term. The value of $log(\Delta t_{last})$ has a higher hazard ratio due to log-transformation. Also $p_{canc}(\Delta t_{last})$ and $p_{canc}|\Delta t_{last}$ are significant. Considering the influence of the other covariates, the covariate *Regular Revenue* gets insignificant. However, no obvious explanation can be found here. For the model *Best*, the influence of the additional features, although combined, remains robust with respect to both direction and magnitude.

5.3.5 Descriptive Analysis of Feature Distributions

The results of the empirical evaluation confirmed the predictive value of both groups of features introduced in this study. This section provides data statistics that support the empirical results, namely that both features have predictive value.

Table 5.5 shows the churn rate depending on the portfolio combination for this customer sample. The results illustrate the dependence of the churn rate on the number of products: the higher the product diversity, the lower the churn rate tends to be. This confirms the results obtained with model s_c. However, values for $s_c = 2$ and $s_c = 3$ are closer than the ones for $s_c = 1$, which explains the absence of a significant effect for $s_c = 2$ in Table 5.4. This finding is intuitive as of stronger customer ties and barriers due to contract terms. Second, the churn rate varies depending on the product category within the portfolio; possessing a product in category 1 exhibits higher churn rates. Third, portfolio combinations display different churn behavior over time, e.g. $X_c = \{1,1,1\}$ almost doubles from the third month to the sixth month, whereas $X_c\{1,0,0\}$ grows only at a rate

Table 5.5: Descriptive analysis of cumulative churn rate depending on the portfolio at the origin of time. The churn is evaluated after the given number of months. Due to confidentiality all values are normalized relative to the minimal value.

Feature	1 mth	3 mth	6 mth	9 mth	12 mth
$X_c = \{1,1,1\}, s_c = 3$	1.000	2.158	4.057	5.674	6.532
$s_c = 2$	1.422	2.868	4.753	6.590	7.505
$s_c = 1$	3.202	5.806	9.089	11.273	12.247
$x_{c,1}$=1	3.733	6.590	10.120	12.775	13.897
$x_{c,2}$=1	1.886	3.497	5.554	7.184	7.951
$x_{c,3}$=1	2.258	4.349	7.106	9.187	10.192
$X_c = \{0,0,1\}$	2.849	5.364	8.633	10.862	11.918
$X_c = \{0,1,0\}$	2.928	5.146	7.897	9.604	10.406
$X_c = \{1,0,0\}$	7.933	12.806	18.053	21.390	22.145
$X_c = \{0,1,1\}$	1.225	2.457	4.036	5.637	6.361
$X_c = \{1,0,1\}$	2.269	5.014	8.947	12.389	14.470
$X_c = \{1,1,0\}$	2.456	3.937	5.185	6.313	6.989

of approximately 50 % in the same time. Overall, these facts are in line with the empirical results and motivate the incorporating of contractual information into churn modeling.

Second, the churn rate depending on the elapsed contract duration is explored. Figure 5.2 shows the relative percentage of contract cancellation (corresponds to $p_{canc}|\Delta t_{last}$) depending on the overall contract duration. Due to confidentiality reasons the exact scaling of the axes cannot be provided. The representation is based on the contracts owned by the customers at the origin of time (January 2014), which are then cancelled until January 2015. The line is the kernel density estimate for $p_{canc}(\Delta t_{last})$. The distribution is multimodal, with peaks at approximately the typical contract duration and its multiplicities, which corresponds to multiple standard contract durations. The distribution explains that the churn probability depends on the contract term, which is captured by Δt_{last} or $log(\Delta t_{last})$

The higher hazard ratios for Δt_{last} or $log(\Delta t_{last})$ in Table 5.4 are supported by the ascending slope of the distribution function from the first month to the first peak.

5.4 Summary

In this chapter, two groups of features for churn prediction models in the telecommunications sector (for access products) are proposed and analyzed. First, customer-specific contract duration dates are considered within established methods of survival analysis. This group of features is motivated by the

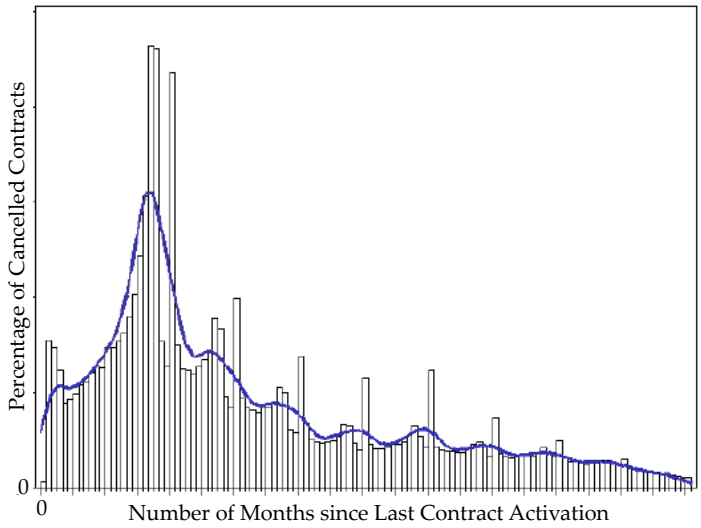


Figure 5.2: Distribution of cancellation notifications for all contracts at the origin of time. The x-axis depicts the number of months from contract activation to contract cancellation. The y-axis represents the total percentage of all contracts. Due to confidentiality reasons numerical scales are not displayed. However, the peaks correspond to the multiples of typical contract duration. The line depicts a kernel density estimation function.

important role of fixed-term contracts in telecommunications, as it can be expected that this results in non-monotonous cancellation probabilities over time, with increased frequencies of cancellation in time periods before minimum subscription periods end. However, this violates basic assumptions of survival analysis models and thus required non-standard feature generation and modeling. Second, the impact of product variety in a customer's portfolio on his churn probability is studied, as it can be related to customer loyalty. This information had positive impact on prediction of churn.

In the empirical evaluation of the proposed extended model is conducted using data provided by one of the largest telecommunication companies in Europe. Results show that both model extensions significantly increase churn prediction performance in out-of-sample tests. The results also show improved churn prediction when using the historical distribution of cancellation notifications instead of contract duration information. The impact of both feature groups was robust for both model types, and a detailed discussion of the impacts of the basic features typically included in survival analysis and the introduced features was provided.

From a managerial point of view, the introduced features allow a better target-

ing of customers at risk to churn, at least in contractual settings, and, therefore, more cost efficient prevention campaigns. In the context of general research on churn modeling, as suggested by Burez and Van den Poel (2009), downsampling was beneficial in this setting. Considering the importance of portfolio information, the result is different from the one described by Van den Poel and Lariviere (2004), where the authors found a low importance of product portfolio information in the financial industry.

Chapter 6

Conclusion and Future Research

This work contributed to the literature on analytical CRM as well as predictive analytics by means of the systematical evaluation of purchasing histories in the context of the next purchase and churn prediction. The main hypothesis of the work is that the purchasing history of products related to information and communications technology contains a temporal pattern, which can be used for segmentation and interpretation of generalized purchasing patterns and as well as prediction of the next customer action. Such action in the context of this work is a probability of a certain product purchase or a probability of a contract cancellation (churn).

The presence and predictive performance of temporal patterns within the sequential data of purchasing histories was shown for the context of telecommunications industry, namely for hosting-related product categories. In order to use the existing patterns in predictive models, two sequence aggregation methods for grouping the purchasing history to segments were proposed and studied in comparison with state-of-the-art models. The key feature of the proposed methods is the ability to group sequences with respect to the similarity in the most recent purchase by employing a weighting scheme assigning decreasing weights to the purchases in the past. Also marketing-relevant extensions for the aggregation method were proposed and evaluated with respect to the resulting predictive performance gain. The extensions include modeling of bundle purchases (products purchased simultaneously), sequence length and handling of repetitive purchases from the same product category.

The main challenge for integration of features based on purchasing histories is the problem of low number of observations per sequence type. This fact makes previous aggregation of purchasing sequences mandatory, as no reliable inference can be done based on such data. As the next challenge, the question of the optimal aggregation level arises, as a trade-off between the aggregation level increasing the generalization ability of the model (reducing variance of the training

data per sequence) and systematic information loss (bias due to generalization) exist, typically referred to as bias–variance trade-off. This work proposed the respective estimates for bias and variance for the lift criterion, typically used in CRM. Analyzing and understanding of these estimates allows for the choice of the optimal aggregation level. Furthermore, a framework of metrics for estimation of the predictive performance based on raw sequential data were provided and analyzed.

Overall, the proposed methods and metrics as well as presented empirical evaluations are contributing to both, theoretical understanding of the sequence aggregation as well as to the practical application of features based on sequential data within predictive models.

6.1 Contribution and Limitations

This section summarizes the results for research questions formulated in Section 1.5 and based on these outlines the contribution as well as limitations of each result for research and practice.

Presence of Temporal Patterns: Dealing with purchasing history, the incorporation of purchasing sequences is motivated by the presence of temporal patterns. This first question regards the hypothesis that a certain "logical" purchasing order exists for products related to information and communications technology. The evidence for existence of such patterns was provided using Hidden Markov Models. The presence of such pattern was shown by the clear structure of the resulting model. Using the proposed approach an analysis for a data set of another company or another context with potential temporal pattern can be conducted.

Interpretation of Temporal Patterns: The second question regards the relation of the identified segments (latent states), statistically derived from the data by the Hidden Markov Model, to a certain customer type. The hypothesis is that the products reflect a technological maturity of a customer, which should be seen in the resulting segmentation. The parallels between the technological level of products bought by customers and their respective segment were clearly shown using resulting model parameters, sequences belonging to the segments as well as external information, which were used for an ex-post evaluation. The proposed approach overcame one of the main critiques on Hidden Markov Models, namely, a difficult interpretation of latent states. Therefore, the proposed approach can be transferred to other similar contexts as a framework for interpre-

tation of latent states in order to provide business understanding of the identified patterns.

Predictive Value of Temporal Patterns: As one possibility of sequence segmentation, the resulting latent states can be used for the estimation of purchasing probability of a target product. This investigation serves as an indicator for potential incorporation of the identified segments in predictive models. The results of this investigation have shown a presence of the predictive value, which depended strongly on the type of the product. However, in the benchmark comparison later on Hidden Markov Models were a rather weak competitor in the task of the prediction, so other segmentation techniques were a better choice for predictive purposes. Nevertheless, the proposed methodology can be used as a general predictive performance assessment for segments based on sequential data, especially as using the proposed framework for profiling latent states, it can also be used for analyzing and interpreting purchasing patterns.

Comparison with Benchmark Models: After the presence and predictive value of sequential data was shown, as the next step two aggregation models for purchasing sequences were introduced and evaluated. One of the key contributions is constituted by proposed and formally defined mechanisms for sequence aggregation, which can be used for both, descriptive and predictive tasks within the analytical CRM context. The intuition of both is a temporal discount of purchases done further in the past.

Sequence Set Model (SSM) truncates the sequences or uses a combination of a certain sequence length and generalizes elder purchases to a portfolio, i.e., neglecting the order purchases and considering only if a certain product was purchased further in the past. *Weighted-Productspace Clustering (WPC)* projects the sequences using geometrically descending weights into continuous space, where a distance-based clustering is done. The second method is more flexible as due to parameterization with a discount factor and number of segments for the clustering, it is able to follow the data distribution in the space, whereas SSM relies on predefined categories, which do not necessarily have enough support for certain categories.

Both models were empirically evaluated with respect to their predictive performance in the context of the next purchase prediction of hosting products. The empirical evaluation has shown a significant dominance over existing approaches established in CRM, namely Logistic Regression, Hidden Markov Models, Association Rules, Levenshtein-Ward, temporal Singular Value Decomposition and Self-Organizing Maps.

The first two approaches have shown the weakest predictive performance.

Logistic Regression does not aggregate sequences upfront and therefore suffers from a low number of observations per sequence type. In contrast, the proposed methods profit from the aggregation done prior to learning the corresponding purchasing probability. Hidden Markov Models aggregate the sequences by means of a probabilistic estimation of underlying statistical patterns, but the problem with these is the absence of temporal relation further that one step ago due to the Markov property of the latent states, so the proposed methods clearly profit from incorporation of older purchases in a discounted fashion.

Association Rules, Levenshtein-Ward and Singular Value Decomposition deliver an average performance compared to competitors. Association Rules are able to generate rules based on the different purchases but do not provide a grouping of similar sequences, therefore not ensuring a necessary support for a reliable prediction. Levenshtein-Ward does not capture the similarity of most recent purchases therefore not able to capture temporal aspect of the differences between sequences. Although temporal Singular Value Decomposition operates on a similar data as WPC, the clustering is the differentiating feature of both methods, which a more useful approach compared to Singular Value Decomposition.

Among the three best models are the two proposed methods as well as Self-Organizing Maps. The latter are able to use implicitly non-linear weighting, which is, however, quite difficult to grasp due to both, parameterization of the temporal discount, as this is done automatically within the algorithm, as well as the resulting weighting scheme learned from the data, which is quite difficult to interpret for a non-statistician.

As to comparison of both proposed models, WPC outperformed SSM except for very small sample sizes, as the amount of data was insufficient for incorporation of the sequential information. The two main differences between the proposed methods are the exponential temporal discounting of older purchases resulting in a more advantageous aggregation of sequential information as well as the application of distance-based clustering leading to a better allocation of the available segments. So overall, based on the data in the evaluation, WPC is a dominating approach for sequence aggregation. Using this method and its flexible parameterization any other sequential data set can be segmented delivering interpretable segments with respect to underlying temporal sequential patterns given the presence of these in the data.

Estimates for Bias–Variance Trade-Off: The question of optimal parameterization of the proposed methods is discussed in the context of the bias–variance trade-off for lift, which constitutes the second key contribution of the work as it

also addresses the issue of estimation of the test result given the aggregation on training data. This work proposed and evaluated the first estimates for bias and variance for lift criterion. The empirical study has shown that the behavior of estimates is in line with the required statistical properties. Also, an ability of these estimates to predict the test error (loss of lift within prediction) was empirically shown. As a consequence, the proposed estimates can be used for both, optimization of the aggregation level for sequential data as well as for the prediction of resulting performance for any categorical sequential data.

Performance of WPC Extensions: In order to incorporate further marketing-relevant information from the purchasing history, additional extensions for WPC were proposed and evaluated. These include modeling of bundles, modeling of sequence length and handling of repetitive purchases. Bundling information was rarely more beneficial. This can be explained by additional representation complexity compared to basic WPC. Hence, the problem of low observation number is stronger, leading to higher variance. The increase in variance was not compensated by the bias reduction, which lead to a poorer performance. A further explanation is that the additionally modeled segments did not have a strong discriminative power compared to the baseline probability.

The additional modeling of the sequence length was beneficial with respect to predictive performance in cases where the distribution was strongly skewed towards very short sequences. The rationale is similar to the modeling of bundles, namely, the additional representation complexity has to be compensated by the predictive value of new segments, which was not the case. Similar trade-off is also given when considering repetitive purchases. However, in contrast to the other cases excluding the information on the second, third etc. purchase from the same category had a negative impact on the predictive performance. Here, the increase of representative complexity was beneficial as the number of repetitive purchases was crucial information for prediction of the next purchase.

Overall, the proposed extensions delivered mixed results depending on the data. Nevertheless, the general approach addresses common modeling issues in marketing context. Furthermore, the metrics for an estimation of the potential application of each metric were provided. These, however, do not estimate the test error but solely provide indication, if any of the proposed extensions is to evaluate.

Data-Based Performance Prediction: A step further with respect to the prediction of the test results is done by proposing data-based metrics for an ex-ante estimation of lift achieved using the raw sequential data for prediction, such as the number of available data set instances or average support per sequence. The

respective metrics were defined and evaluated. Whereas, the performance of different aggregation levels on training data was estimated quite well, a decrease of the explanatory power was observed as one moves from training to test performance prediction. Nevertheless, the proposed metrics were able to explain 64 % of the test performance measured in lift. As an additional insight, the direction of the aggregation (stronger or weaker) can be anticipated using the proposed methodology. Furthermore, a generic framework of metrics for the estimation of sequential information was proposed, which can serve as a ground for further proposition and evaluation of metrics with related goal of data-based prediction of resulting performance.

Features for Churn Prediction: Definition and application of features based on contractual information for access products (like mobile Internet access) was the final investigation of this work. As no logical pattern is expected for this group of products, the portfolio of a customer was evaluated and has shown to be a strong predictor of churn. Not only the number but also the combination of the products possessed by a customer (i.e., customer's product portfolio) was in important indicator for churn probability. The transformation of the time-related information of contract term delivered additional predictive value with respect to lift as it was able to provide a scaling of a feature, which satisfied the assumptions of survival models better than untransformed values.

The key limitation of the empirical study is the fact that if was done using only data stemming from one telecommunications company with a specific product palette, which raises the question of generalization ability of the results. However, the evaluation results are surprisingly robust throughout the ten data samples with different target products, and the results will most probably carry over to other corporations in particular in the telecommunications sector. Without a doubt, the performance of WPC relies strongly on the presence of temporal patterns, whereas the events further are less important. This is also the crucial point for the transfer to other application contexts. Still, such contributions as a definition of two aggregation models or the theoretical framework for bias–variance trade-off are otherwise independent of data set or application context.

Furthermore, the average sequence length was quite short, keeping the representation complexity and computation effort tractable. With Hidden Markov Models being a computationally expensive methodology, using this type of models might not be appropriate for very large and complex data sets. A further issue is the small number of categories in the data, which on the one side reduces the complexity and a possibility of data artifacts, such as tariff changes, on the other hand brings the disadvantages of the clustering, if the dimension-

ality of the problem rises. Even given a high number of products, any product palette can be aggregated to a number of categories or only a certain category can be subdivided into smaller ones. The proposed algorithms easily handle the lower number of derived categories.

Overall, as to managerial implications, the proposed methodology allows for the estimation and monetization of purchasing histories within the CRM context. An integrated approach supporting segmentation and prediction activities based on purchasing histories of customers was proposed and evaluated. The results contribute to general understanding of the purchasing behavior along the customer's lifecycle with the company as well to the operationalization of sequential information within predictive models from the statistical point of view.

6.2 Future Research

After the value of purchasing histories for segmentation and prediction was discussed in the context of CRM, this section is intended to motivate further research in the outlined directions. Based on the results the three main research paths are proposed. First, integration of the identified segments with further data. The focus of this work was dedicated to purchasing histories, as these constitute a challenging information source with issues related to the low number of observations per sequence and its temporal pattern. However, further important information is available about the customer, which can be in the next step integrated with customers' demographics or product usage information in a similar way as done in the context of churn in Chapter 5.

Second, the prediction of the performance using the aggregated sequential segments based on the raw data is a topic of a high relevance. Hypothetically, having a powerful tool for a reliable ex-ante estimation of informational value, big amounts of data can be quantified with respect to their predictive value prior to costly storage and processing activities. This work proposed a first approach beyond which further metrics (like Gini-coefficient or related measures) can be tested in this context. An even more interesting question is the one of functional dependencies between several metrics for a more detailed analysis. For instance, given a certain sequence variety (number of unique sequence types) which number of observations is necessary in order to provide a prediction of a certain quality. In this context also metrics estimating the bias–variance trade-off on a raw data would make the estimation of the required aggregation level even more simple that the estimates proposed in this work. However, search for such

generalizable functional dependencies is quite difficult. An investigation on a simulated data set with controllable effects might be useful for a generalizable and a more theoretical discussion of the proposed metrics.

Third, a rather practical issue, is an additional investigation of the product purchase characteristics like the channel of purchase, if the product was a part of a specific marketing advertisement with different pricing or simply several different countries. Thereby, the goal is to compare the patterns among different treatments. The available data might be filtered for the respective events. Then, generalized patterns can be derived using the proposed methods and compared among different groups.

As to general application context, predictive maintenance and health monitoring are proposed as possible candidates for a successful application of proposed models. Defects in predictive maintenance can be assumed to show temporal patterns as often a common underlying problem is causing the defects, which are then observed as several sequential events with a "logical" order (including repair events). Therefore, the identification of typical patterns might lead to a better understanding and prediction of the subsequent events. In a similar way, also medical events might constitute an interesting field of sequential data for application of the proposed methods.

References

Ahn, J.-H., S.-P. Han, and Y.-S. Lee (2006). Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. *Telecommunications Policy 30*(10), 552–568.

Back, B., A. Holmbom, and T. Eklund (2011). Customer portfolio analysis using the SOM. *International Journal of Business Information Systems 8*(4), 396–412.

Baumann, A., S. Lessmann, K. Coussement, and K. W. De Bock (2015). Maximize what matters: Predicting customer churn with decision-centric ensemble selection. In *ECIS 2015 Completed Research Papers*.

Bicego, M., V. Murino, and M. A. Figueiredo (2003). Similarity-based clustering of sequences using Hidden Markov Models. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pp. 86–95. Springer.

Bose, I. and X. Chen (2009). Quantitative models for direct marketing: A review from systems perspective. *European Journal of Operational Research 195*(1), 1–16.

Box, G. E. and D. R. Cox (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological) 26*(2), 211–252.

Boztuğ, Y. and T. Reutterer (2008). A combined approach for segment-specific market basket analysis. *European Journal of Operational Research 187*(1), 294–312.

Brown, R. G. (2004). *Smoothing, forecasting and prediction of discrete time series*. Courier Dover Publications.

Burez, J. and D. Van den Poel (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications 36*(3), 4626–4636.

Chan, C. C. H. (2008). Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer. *Expert Systems with Applications 34*(4), 2754–2762.

Cho, Y. B., Y. H. Cho, and S. H. Kim (2005). Mining changes in customer buying behavior for collaborative recommendations. *Expert Systems with Applications 28*(2), 359 – 369.

Domingos, P. (2000). A unified bias–variance decomposition. In *Proceedings of 17th International Conference on Machine Learning*, pp. 231–238.

Dunlavy, D. M., T. G. Kolda, and E. Acar (2011). Temporal link prediction using matrix and tensor factorizations. *ACM Transactions on Knowledge Discovery from Data (TKDD) 5*(2), 10.

D'Urso, P. and L. De Giovanni (2008). Temporal self-organizing maps for telecommunications market segmentation. *Neurocomputing 71*(13), 2880–2892.

Farquad, M., V. Ravi, and S. B. Raju (2014). Churn prediction using comprehensible support vector machine: An analytical CRM application. *Applied Soft Computing 19*, 31 – 40.

Fink, G. A. (2014). *Markov models for pattern recognition: from theory to applications*. Springer Science & Business Media.

Forney, G. D. (1973). The Viterbi algorithm. *Proceedings of the IEEE 61*(3), 268–278.

Gerpott, T. J., N. Ahmadi, and D. Weimar (2015). Who is (not) convinced to withdraw a contract termination announcement? – A discriminant analysis of mobile communications customers in Germany. *Telecommunications Policy 39*(1), 38–52.

Goodwin, P. (1997). Adjusting judgemental extrapolations using Theil's method and discounted weighted regression. *Journal of Forecasting 16*, 37 – 46.

Hadden, J., A. Tiwari, R. Roy, and D. Ruta (2007). Computer assisted customer churn management: State-of-the-art and future trends. *Computers & Operations Research 34*(10), 2902 – 2917.

Han, S. H., S. X. Lu, and S. C. Leung (2012). Segmentation of Telecom customers based on customer value by decision tree model. *Expert Systems with Applications 39*(4), 3964–3973.

Hsu, M.-W., S. Lessmann, M.-C. Sung, T. Ma, and J. E. Johnson (2016). Bridging the divide in financial market forecasting: machine learners vs. financial economists. *Expert Systems with Applications 61*, 215–234.

Ibe, O. (2013). *Markov processes for stochastic modeling*. Newnes.

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

Joh, C.-H., H. J. Timmermans, and P. T. Popkowski-Leszczyc (2003). Identifying purchase-history sensitive shopper segments using scanner panel data and sequence alignment methods. *Journal of Retailing and Consumer Services 10*(3), 135 – 144.

Kamakura, W. A., M. Wedel, F. De Rosa, and J. A. Mazzon (2003). Cross-selling through database marketing: A mixed data factor analyzer for data augmentation and prediction. *International Journal of Research in marketing 20*(1), 45–65.

Kamalraj, N. and A. Malathi (2013). A survey on churn prediction techniques in communication sector. *International Journal of Computer Applications 64*(5), 39–42.

Keramati, A. and S. M. Ardabili (2011). Churn analysis for an Iranian mobile operator. *Telecommunications Policy 35*(4), 344–356.

Khajvand, M. and M. J. Tarokh (2011). Estimating customer future value of different customer segments based on adapted RFM model in retail banking context. *Procedia Computer Science 3*, 1327–1332.

Kim, H.-S. and C.-H. Yoon (2004). Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. *Telecommunications Policy 28*(9), 751–765.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE 78*(9), 1464–1480.

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

Kullback, S. and R. A. Leibler (1951). On information and sufficiency. *The Annals of Mathematical Statistics 22*(1), 79–86.

Lawless, J. (1982). *Statistical methods and model for lifetime data*. Wiley&Sons, New York.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet physics doklady 10*(8), 707–710.

Li, S., B. Sun, and R. T. Wilcox (2005). Cross-selling sequentially ordered products: An application to consumer banking services. *Journal of Marketing Research 42*(2), 233–239.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory 37*(1), 145–151.

Ling, C. X. and C. Li (1998). Data mining for direct marketing: problems and solutions. In *KDD*, Volume 98, pp. 73–79.

Lo, V. S. (2002). The true lift model: A novel data mining approach to response modeling in database marketing. *ACM SIGKDD Explorations Newsletter 4*(2), 78–86.

Lu, J. (2002). Predicting customer churn in the telecommunications industry – an application of survival analysis modeling using SAS. *SAS User Group International (SUGI27) Online Proceedings*, 114–27.

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

McCarty, J. A. and M. Hastak (2007). Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. *Journal of business research 60*(6), 656–662.

Miguéis, V., D. Van den Poel, A. Camanho, and J. Cunha (2012). Predicting partial customer churn using Markov for discrimination for modeling first purchase sequences. *Advances in Data Analysis and Classification 6*(4), 337–353.

Miguéis, V. L., D. Van den Poel, A. S. Camanho, and J. F. e Cunha (2012). Modeling partial customer churn: On the value of first product-category purchase sequences. *Expert Systems with Applications 39*(12), 11250–11256.

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

Moon, S. and G. J. Russell (2008). Predicting product purchase from inferred customer similarity: An autologistic model approach. *Management Science 54*(1), 71–82.

Mooney, C. H. and J. F. Roddick (2013, March). Sequential pattern mining – approaches and algorithms. *ACM Comput. Surv. 45*(2), 19:1–19:39.

Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal 26*(4), 354–359.

Neslin, S. A., S. Gupta, W. Kamakura, J. Lu, and C. H. Mason (2006). Defection detection: measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research 43*(2), 204–211.

Netzer, O., J. M. Lattin, and V. Srinivasan (2008). A Hidden Markov Model of customer relationship dynamics. *Marketing Science 27*(2), 185–204.

Ngai, E. W., L. Xiu, and D. C. Chau (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications 36*(2), 2592–2602.

Park, D. H., H. K. Kim, I. Y. Choi, and J. K. Kim (2012). A literature review and classification of recommender systems research. *Expert Systems with Applications 39*(11), 10059–10072.

Payne, A. and P. Frow (2005). A strategic framework for customer relationship management. *Journal of Marketing 69*(4), 167–176.

Piatetsky-Shapiro, G. and B. Masand (1999). Estimating campaign benefits and modeling lift. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, pp. 185–193. ACM.

Prince, J. and S. Greenstein (2014). Does service bundling reduce churn? *Journal of Economics & Management Strategy 23*(4), 839–875.

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

Sahoo, N., P. V. Singh, and T. Mukhopadhyay (2012). A Hidden Markov Model for collaborative filtering. *MIS Quarterly 36*(4), 1329–1356.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*, 461–464.

Schweidel, D. A., E. T. Bradlow, and P. S. Fader (2011). Portfolio dynamics for customers of a multiservice provider. *Management Science 57*(3), 471–486.

Shapoval, K., M. Reisser, and J. Baldinger (2015). Term of contract and portfolio aware churn modeling in telecommunication campaigns. In *Multikonferenz Wirtschaftsinformatik (MKWI) 2016*, Volume 1, pp. 26–32. IEEE.

Shapoval, K. and T. Setzer (2015). Term of contract and portfolio aware churn modeling in telecommunication campaigns. In *2015 IEEE 17th Conference on Business Informatics*, Volume 1, pp. 26–32. IEEE.

Shapoval, K. and T. Setzer (2017a). Customers purchasing sequence representation for response modeling. *Working Paper in Preparation for European Conference on Information Systems*.

Shapoval, K. and T. Setzer (2017b). Next purchase prediction using projections of discounted purchasing sequences. *Accepted in Business & Information Systems Engineering*.

Shenkin, P. S., B. Erman, and L. D. Mastrandrea (1991). Information-theoretical entropy as a measure of sequence variability. *Proteins: Structure, Function, and Bioinformatics 11*(4), 297–313.

Shmueli, G. and O. Koppius (2010). Predictive analytics in information systems research. *Robert H. Smith School Research Paper No. RHS 35*(3), 06–138.

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

Tamaddoni Jahromi, A., M. M. Sepehri, B. Teimourpour, and S. Choobdar (2010). Modeling customer churn in a non-contractual setting: the case of telecommunications service providers. *Journal of Strategic Marketing 18*(7), 587–598.

Tsai, C.-F. and Y.-H. Lu (2009). Customer churn prediction by hybrid neural networks. *Expert Systems with Applications 36*(10), 12547–12553.

Tsai, C.-Y., C.-C. Lo, and C.-W. Lin (2011). A time-interval sequential pattern change detection method. *International Journal of Information Technology and Decision Making 10*(01), 83–108.

Tsiptsis, K. K. and A. Chorianopoulos (2011). *Data mining techniques in CRM: inside customer segmentation*. John Wiley & Sons.

Turner, R. and L. Liu (2009). hmm. discnp: Hidden Markov Models with discrete non-parametric observation distributions. *R package version 0.1–1. URL http://CRAN. R-project. org/package= hmm.discnp*.

Van den Poel, D. and W. Buckinx (2005). Predicting online-purchasing behaviour. *European Journal of Operational Research 166*(2), 557–575.

Van den Poel, D. and B. Lariviere (2004). Customer attrition analysis for financial services using proportional hazard models. *European Journal of Operational Research 157*(1), 196–217.

Wang, H. and S. Wang (2007). Mining purchasing sequence data for online customer segmentation. *International Journal of Services, Operations and Informatics 2*(4), 382–390.

Wei, L. (1992). The accelerated failure time model: A useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine 11*(14-15), 1871–1879.

Wong, K. W., S. Zhou, Q. Yang, and J. M. S. Yeung (2005). Mining customer value: From association rules to direct marketing. *Data Mining and Knowledge Discovery 11*(1), 57–79.

Xie, Y., X. Li, E. Ngai, and W. Ying (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications 36*(3), 5445–5449.

Zhang, X., J. Zhu, S. Xu, and Y. Wan (2012). Predicting customer churn through interpersonal influence. *Knowledge-Based Systems 28*, 97–104.