“New Approaches to Customer Relationship Management in Fashion Retail Online”

Dissertation
zur Erlangung des Grades eines Doktors der Wirtschaftswissenschaft
der Rechts- und Wirtschaftswissenschaftlichen Fakultät
der Universität Bayreuth

vorgelegt von
Björn Stöcker
aus
Kulmbach

Dekan:	Herr Prof. Dr. Jörg Schlüchtermann
Erstberichterstatter:	Herr Prof. Dr. Daniel Baier
Zweitberichterstatter:	Herr Prof. Dr. Claas Christian Germelmann
Tag der mündlichen Prüfung:	13. April 2021

Danksagung

„Wenn ich weiter gesehen habe als andere, so deshalb, weil ich auf den Schultern von Riesen stehe.“

Bernhard von Chartres um 1120

Aus meiner Sicht gibt es kein treffenderes Gleichnis, welches die Entstehung meiner Dissertation beschreibt. Ich durfte in meiner Arbeit meinen wissenschaftlichen Horizont erweitern und habe dabei in vielerlei Hinsicht auf den Schultern von Riesen gestanden, denen ich zutiefst zu Dank verpflichtet bin. An dieser Stelle möchte ich den wichtigsten Personen danken, die meine Dissertation möglich gemacht haben.

Mein besonderer Dank gilt meinem Doktorvater, Prof. Dr. Daniel Baier. Lieber Daniel, ich kann mich noch lebhaft an den „schicksalshaften“ Kaffee in der Cafeteria in der Uni erinnern, bei dem wir uns über die Methode des Uplift-Modelings unterhalten haben. Nach einigem Diskutieren waren wir uns einig, dass dieses Thema so spannend ist, dass man es im Rahmen einer Doktorarbeit untersuchen könnte, was der Startschuss für mein Dissertationsprojekt war. Während der letzten zwei Jahre warst Du mir mit Rat und Tat zur Seite gestanden und hattest immer ein offenes Ohr für meine Anliegen. Eine berufsbegleitende Dissertation – mit den zusätzlichen Auflagen, die man als Fachhochschulabsolvent erfüllen muss – ist immer eine besondere Herausforderung, auch für den Betreuer. Auch in dieser Hinsicht möchte ich mich bei Dir noch einmal ganz herzlich für meine Betreuung bedanken und für die geduldige Beantwortung der vielen organisatorischen Rückfragen! Weiterer Dank gilt Prof. Dr. Claas Christian Germelmann, der sich bereit erklärt hat, das Zweitgutachten für diese Arbeit zu übernehmen.

Zusätzlich möchte ich mich sehr bei den Professoren Dr. Daniel Baier, Dr. Claas Christian Germelmann und Dr. Herbert Woratschek und den Mitgliedern des Arbeitsbereichs Marketing und Services

bedanken. Ich durfte während meiner Zeit als Doktorand an zwei Doktorandenseminaren teilnehmen, aus denen ich sehr viele Impulse für die Ausgestaltung dieser Arbeit mitnehmen konnte. Gerade für externe Doktoranden ist dies eine wunderbare Einrichtung, sich zum einen besser mit den anderen Doktoranden des Arbeitsbereichs zu vernetzen. Zum anderen war es mir möglich, wertvolle, konstruktive Rückmeldung zum aktuellen Stand der Dissertation zu erhalten. Ich würde mich freuen, wenn diese Einrichtung noch sehr lange Bestand hat!

Außerdem möchte ich mich bei meinem Arbeitgeber, der Firma BAUR, bedanken: einerseits für die schon lange gepflegte Beziehung mit der Universität Bayreuth, die nicht unwesentlich zum Zustandekommen dieser Arbeit beigetragen hat. Die Ergebnisse dieser Zusammenarbeit zeigen, wie wertvoll diese für Unternehmen und Studenten sein kann. Andererseits möchte ich mich für die unkomplizierte und vorbehaltlose Bereitstellung der Daten bedanken, welche die hohe Qualität meiner Forschung erst ermöglicht hat. Dabei danke ich allen Kollegen, die mich auf meinem beruflichen Weg begleitet und mir in vielen Situationen mit gutem Rat beigestanden haben.

Auch der Dank an meine Familie soll nicht zu kurz kommen, wenn die Zeit in den letzten Monaten auch oft rar war. Ein herzlicher Dank gilt meinen Eltern, die mir den schulischen Weg geebnet haben, und ohne deren Engagement ich heute nicht diese Zeilen schreiben könnte. Eine berufsbegleitende Dissertation ist auch für die Familie eine Herausforderung, wenn man nach der Arbeit noch einmal an die Arbeit geht. Ein ganz besonderer Dank gilt dir, liebe Franzi, die du mich ermuntert hast, diese Chance anzunehmen, und mir in allen Situationen den Rücken gestärkt hast. Eine Dissertation ist ja nicht nur das Schriftstück, sondern ein Prozess mit vielen Höhen und Tiefen. Danke, dass du für mich in allen Situationen da warst, ganz besonders in diesem ereignisreichen Jahr 2020! Und zu guter Letzt möchte ich mich bei meinen Töchtern bedanken. Ihr habt mich mit eurer Fröhlichkeit und Wissbegierde immer wieder dran erinnert, wie schön es ist, Neues zu entdecken. Euch beiden widme ich diese Arbeit!

Kronach, November 2020 Björn Stöcker

Abstract

This thesis focuses on new approaches to customer relationship management (CRM) in online fashion retail. For this purpose, this thesis turns to two essential and current CRMs and further develops existing methodical approaches.

The first part of this thesis presents two papers examining the customer-company interface. How the penalty reward contrast analysis (PRCA) behaved with skewed response distributions was investigated. This circumstance partially led to misinterpretation of the results, and a cubic regression was applied to avoid. Another paper examining new approaches in returns management from the customer perspective is presented. For the first time in the literature, this paper presents the examination of the entire customer journey to derive valuable insights. It is shown, among other things, that the expectations of current mail-order customers continue to be ineffective, different dynamics exist, and future differentiation potentials can be identified.

The second part turns to the optimization of direct marketing campaigns. For the third paper presented, previously established uplift modeling applications are discussed, and shortcomings pointed out. These shortcomings were transferred to a profit perspective by applying three statistical methods (Heckman sample selection model, zero-inflated negative binomial regression model, and random forest-based regression) and adapting the previous procedure. This third paper makes an essential contribution, as its research applied the modeling of continuous values to a real-world dataset, and the paper calls for a stronger focus on continuous variables, profit, and return on investment (ROI). The fourth paper considered the influence of different marketing campaign costs on uplift modeling's validity, respectively, to estimate the causal effect. Current research has excessively focused on cost per contact, which has led to a failure of the method in the presence of, for example, respond-variable costs. The optimal approach based on typical cost constellations was

examined. These insights led to far-reaching adjustments of the established method and a generalization of currently applied performance measurements.

This thesis addresses current and essential CRM methods by presenting four research papers, discussing their weaknesses, and presenting improvements for the methods used. It thus makes an essential contribution to research and further provides practitioners with essential insights and improvements.

Table of Contents

Chapter 1

Introduction

1. Motivation

The consistent orientation of a company towards its customers is nowadays an established part of corporate management due to the advent of relationship marketing in the 1990s, which led to customer relationship management (CRM) in the 2000s (e.g., Payne and Frow 2005; Bruhn 2016). Despite this lengthy time, CRM continues to be dynamically developed; therefore, particularly in recent years, there has been much movement in this area. However, what are the most relevant topics in CRM today, particularly for fashion retail online? A recent study has shown that "80% of customers say the experience a company provides is as important as its products and services", and "fifty-seven percent of customers have stopped buying from a company because a competitor provided a better experience" (Salesforce 2018, p. 8). Furthermore, in another study, Forrester (2020) has stated that "customer experience (CX) is still king in 2020", and CRM is the core technology for customer engagement. Although customer orientation and CRM have been established concepts for years, they are increasingly important. Particularly in a buyer's market with a high purchase frequency and comparatively low switching costs, such as with online retail fashion, customers are no longer willing to accept inadequate or ordinary service. In addition to the general need to ensure customer satisfaction, the management of returns is of paramount importance. Particularly in the online retail fashion sector, a 50% return rate is common. This high rate means that half of the ordered clothing is returned to the distributor. This circumstance is a great challenge for the customer, the retailer, and the environment. Enhancement in this field, therefore, would lead to an improvement for all three involved parties.

Significant challenges can be further identified in another field of CRM. Two real-world data sets from a fashion retailer in Germany have shown that the marketing efficiency of direct marketing measures such as discount mailings is only approximately 70%. In other words, nearly one-third of marketing campaign costs have no measurable effect on sales. A look at the data shows that about one-third of mailing customers' sales are additionally made in the control group. This baseline sale is not caused by the marketing campaign itself but results from other preceding measures. A more ideal prediction of customer behavior, more precisely whether a customer is likely to buy, including without a campaign, represents the second part of this thesis. Based on the available literature, it can be assumed that this problem has been found with other retailers and industries as well.

Following from these priorities, in this thesis, four full research papers are presented that have addressed two essential domains of CRM; therefore, this thesis is divided into two parts:

Part A focuses on the optimal allocation of resources regarding their impact on customer satisfaction. The first paper has dealt with the multi-factor theory in customer satisfaction measurement. It can be seen how the established penalty reward contrast analysis (PRCA) method fails when a majority of customers are satisfied or delighted with a service. This first paper developed an alternative method and proved the robustness regarding skewed distributions in a direct method comparison. The second paper has focused on return management as an integral part of the customer experience in online retail fashion. This second paper has extended current theory through a comprehensive view of return management, the presentation of new technological approaches for return averting and avoidance, and showing the measures' diffusion, with the so-called segmented Kano perspective (Baier et al. 2018). This paper included current technological possibilities and explored how much differentiation potential existed in different measures.

Part B of this thesis is dedicated to the optimal selection of direct marketing campaigns. The two papers have critically examined a particular method known as uplift modeling or causal effect modeling. This discipline enhances response-modeling approaches by predicting the response to treatment and the reaction to the absence of treatment. In recent decades, this problem has been nearly exclusively labeled as a classification problem. Examining the two papers, it can be seen that this restricted view is insufficient and has, further, led to the method's failure. These papers have emphasized the prediction of discrete values, and for the first time in the literature, this method has been embedded in cost and activity accounting. This holistic assessment has resulted in far-reaching adjustments to the previous approach, starting with new measurement methods and ending with different modeling strategies.

The postulation and embedding of the four research questions into the context of the various CRM fields follow in Section 3.

2. Theoretical Background

The role of marketing has been subject to different perspectives in prior decades (Meffert et al. 2008; Bruhn 2009): In the 1950s and 1960s, marketing concentrated on products, and, in the 1970s and 1980s, marketing concentrated on markets and competitors, respectively. From the 1990s onwards, the reference point changed from a company or market view to a customer view (e.g., Bruhn 2016). The customer and its needs have increasingly become the starting point for business decisions. During this time, customers learned to emancipate themselves and make entirely different demands towards companies and their products and services. Customers can no longer be categorized into homogeneous groups, referred to as hybrid consumption (e.g., Ehrnrooth and Gronroos 2013). Customers increasingly have different expectations for different products. For example, it is not unusual for a customer to set the highest value on quality and taste in food but choose low-cost clothing. This heterogeneity has become a significant challenge for companies.

A preliminary stage of CRM is relationship marketing (approximately 1990). In the early days, this involved establishing and maintaining a customer relationship with the company (Morgan and Hunt 1994; Webster Jr 1992). With a new institutional approach, economic theory was introduced to explain the development and break-up of customer relationships, for example, transaction cost theory (Rindfleisch and Heide 1997). The basic assumption of relationship marketing is that managing the customer relationship is beneficial for the firm (Reichheld et al. 1996; Reichheld et al. 2000). However, it was later discovered that the customer relationship (loyalty) does not necessarily lead to more profit (Reinartz and Kumar 2000; Reinartz and Kumar 2002). This insight led to CRM's development in the 2000s and included further success criteria for a customer relationship. Recent approaches have focused on the customer lifetime value (CLV), that is, all the value of a customer throughout the customer relationship, alternatively called the customer equity (value of the entire customer base) (e.g., Gupta et al. 2006). Successful companies have increasingly distinguished themselves by identifying customer-related insights more quickly and deriving and implementing appropriate actions.

The term "CRM" is defined differently by many sides; software providers, in particular, tend only to consider the application. For this thesis, however, CRM is defined in the broader sense, namely, as the company's task of gaining an advantage from customer understanding. Kumar and Reinartz (2018, p. 5) summarize the task of CRM (similar to Zeithaml et al. 2001) as follows:

> *"CRM is the strategic process of selecting customers that a firm can most profitably serve and shaping interactions between a company and these customers. The ultimate goal is to optimize the current and future value of customers for the company."*

CRM's task is to identify different customer needs and customer segments then develop concepts tailored to these needs. All these considerations are based on an essential concept: the customer's value to the company. In addition to campaign effectiveness and efficiency, customer value provides a key indicator for customer segmentation. In this context, customer value is primarily considered from a monetary perspective.

3. Research Agenda

In order to outline the broad field of CRM, Reinartz et al. (2004; Kumar and Reinartz 2018) suggest the covering of three primary areas of application:

Customer-facing level or frontstage

The customer-facing level describes the collection and processing of customer information to make appropriate deductions regarding tactics or strategies. The concept of the customer's single view is the focus here. Only from this holistic perspective can customers be systematically managed. The data is collected at the boundary between the company and the customers, and the derived measures can be re-experienced at this boundary. This field additionally encompasses measuring customer satisfaction and the subsequent derivation of the optimal deployment of resources. Due to its close association with service marketing, this level can be described as the "frontstage."

Functional level or backstage

The functional level includes all the processes that have customer-related tasks. This level has a clear focus on processing the customer relationship. These are applications within the company that are intended to support sales personnel or marketing managers in their work. These tasks can be electronic forms for capturing leads from exhibitions or technologically sophisticated algorithms for selecting marketing campaigns. Providers of CRM software frequently only focus on these applications; consequently, CRM is generally used synonymously with the functional level at this point. Here, no direct interaction with the customer occurs, but the gathering, processing, and providing information related to the customer occurs. Therefore, this area can additionally be called the "backstage."

Strategic level

The strategic level considers the implementation and use of customer knowledge throughout the organization and establishes customer-centricity. It addresses business management issues that are consistently free of technological dependencies. Alternatively, Stauffer (2001) describes this level as follows: “[…]. It’s letting customers determine how you organize” (Galbraith 2005, p. 6).

This thesis deals with practical questions from the frontstage (customer-facing) and backstage (functional) levels and develops the existing methods that had been applied in the literature thus far. The strategic level is not examined for this thesis.

Following the insights of Salesforce (2018) and Forrester (2020), these research questions were addressed:

Part A—Frontstage

RQ1—“How do skewed distributions in the response behavior of service satisfaction surveys influence the validity of the penalty reward contrast analysis (PRCA), and how can this be avoided?”. This has been assessed by **Research Paper #1**.

RQ2—“Which measures in return management currently show the highest potential in terms of increasing customer satisfaction and differentiation, and what further insights can be gained when the return process is viewed holistically?”. This has been assessed by **Research Paper #2**.

Part B—Backstage

RQ3—“Which statistical method is best suited to optimize profit in direct marketing campaigns using uplift modeling regarding the prediction of continuous values?”. This has been assessed by **Research Paper #3**.

RQ4—“How can direct marketing campaigns be optimized in terms of return on investment (ROI), taking into account different marketing campaign cost structures?”. This has been assessed by **Research Paper #4**.

This thesis includes four full research papers that were published or are under review in renowned international journals. Each paper has investigated one specific research question and has contributed to the research mentioned above (Fig. 1).

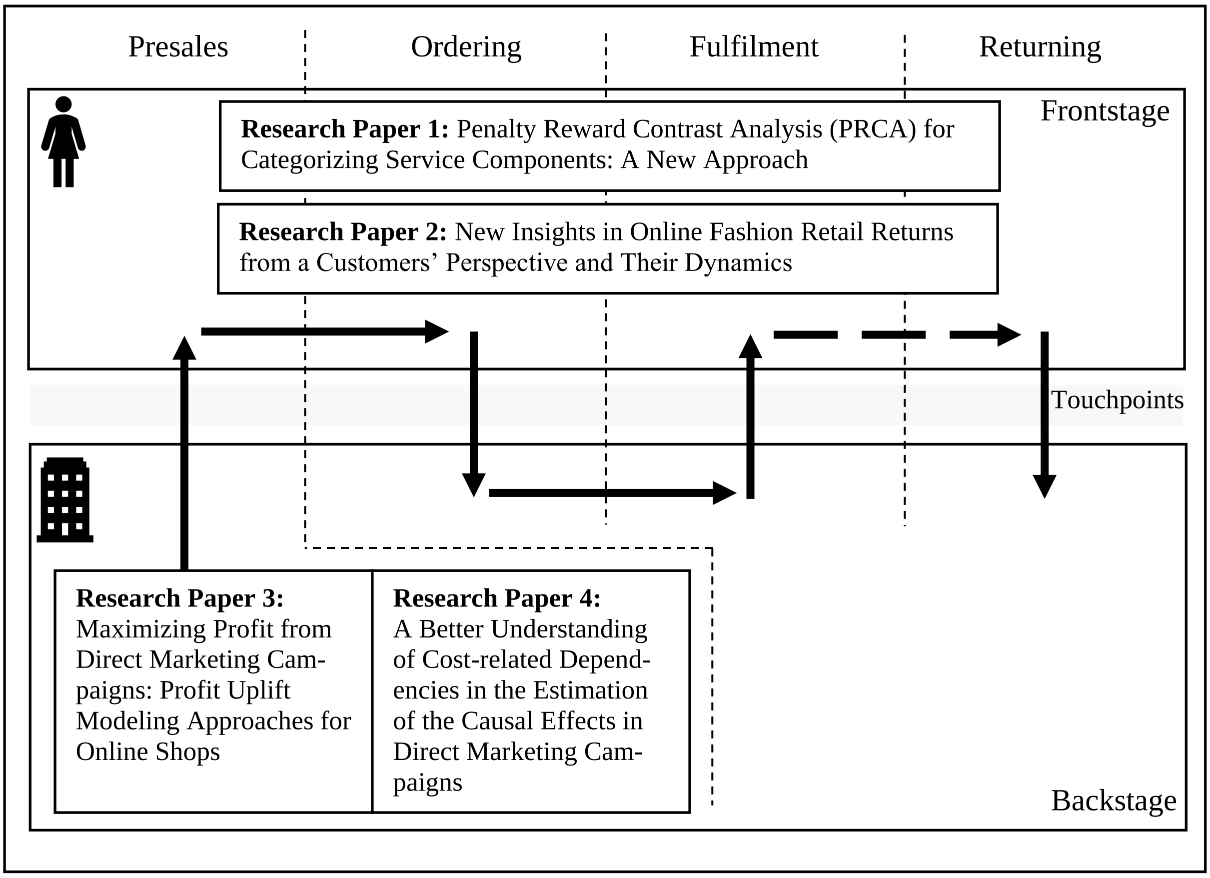


Fig. 1 CRM area of application along the customer journey and positioning of the research papers

As shown in Figure 1, the first research paper has covered the entire customer journey, as an improvement in the method of measuring customer satisfaction can be applied to all touchpoints. Research Paper #2 has further extended across the entire purchasing process since a significant contribution of this paper is that it has dealt with the emergence of returns in the presales phase through to the actual return at the end of the purchase process, thus enabling the direct comparability of the measures. Research Papers #3 and #4 have focused on the presales phase. The optimization of customer selection for marketing campaigns is carried out without the customers' direct involvement and is designed to initiate a purchase efficiently.

3.1 Part A—Frontstage

The task of a CRM, here, is to obtain a holistic or singular view of the customer (single point of truth) to consistently control marketing channels and messages, leading to the appropriate deployment of resources. The emphasis here is on the holistic customer experience. In the final step, the marketing intelligence thus gained must be distributed to all customer-facing functions for the feedback loop to restart once more. New measures result in new data and new insights.

In order to conceptualize CRM in terms of the frontstage or customer-facing level, several references can be found in the literature. The primary task is to establish and maintain a customer relationship (Morgan and Hunt 1994; Webster Jr 1992); the company and the customer enter into a conscious association. The next assumption is that those customer relationships are not static independent events but develop dynamically (Dwyer et al. 1987). Therefore, CRM further contains a time series in which events build upon and influence each other, and companies should proactively manage customer interactions at each stage of the lifecycle in a different customer-individual way (Srivastava et al. 1998). Fourthly, the value arising for a company from the customer relationship is not equally distributed throughout the entire customer base (Mulhern 1999; Niraj et al. 2001). There must be a product-related dimension and a customer-related dimension (customer value) in cost and activity accounting.

In this environment, Paper #1 has investigated the relationship between satisfaction with a service or touchpoint and a company's overall satisfaction, which is frequently non-linear. Kano (1984) has illustrated this relationship and pointed to the effects on the strategy to be adopted. Based on this understanding, the measurement of customer satisfaction has shifted to the determination of this non-linearity. At this point, services or touchpoints are no longer in the concept stage. Mikulić and Prebežac (2011) have described that, in this context, the PRCA (Brandt 1987) is suitable. The application of the PRCA is non-standardized; consequently, Tahir Albayrak and Meltem Caber (2013) have listed three different calculation methods.

Following this, the question of whether the central implicit assumption that satisfaction with the service and overall satisfaction scatters around the mean value is commonly fulfilled could be raised. In surveys on service quality, loyal customers participate; thus, answer patterns are primarily skewed. Based on this point, Paper #2 has proposed a new method: the application of a cubic regression. By comparing the different calculation methods of PRCA and the cubic regression, one can see that the calculation method and the skewed response patterns can occasionally turn statements into the opposite. There have been two standardized service surveys of a German mail-order company (Study 1 in 2011 with n = 480 and Study 2 in 2013 with n = 500). Paper #2 has classified the service components in four ways: three variations of the PRCA and once with cubic regression. In Paper #2, it can be seen that the new application of cubic regression is less sensitive to skewed data and should be superior to PRCA. Furthermore, an interpretation of the beta coefficients concerning Kano's categories could be provided.

Paper #2 has further addressed customer satisfaction with a broader focus on the design phase. One of the most critical service processes in the mail-order business, which is at once distinctive and unique to it—return management—is examined. Due to the information gap online, many returns are not intentionally caused but are affected due to the system. However, returns are annoying for the company or the customer and constitute a significant environmental burden due to the return transport and the possible necessary reconditioning. Paper #2's objective was to determine the most promising strategies for averting and avoiding returns.

In the literature, only isolated purchase process steps have been examined. For the first time, as presented in this paper, the entire purchase process was examined to make measures directly comparable to the different phases. In the measures examined, current technological developments, several of which had not yet been widely used (e.g., self-measurement via webcam), were examined. To investigate the conception phase of which measures should be followed up in particular, Kano's model and method were used. More than 8,000 customers of a mail-order company responded via an online survey. Utilizing the unique Kano questioning technique, all measures could be described in terms of their effect on customer satisfaction. However, many measures were only classified as indifferent or attractive, which, according to Kano (2001), indicates an early phase in the product life cycle. One of the significant findings was that monetary incentives currently show the strongest influences in the entire process, followed by improved presentations.

From other studies such as that of Nilsson-Witell and Fundin (2005), it was previously known that it was possible to identify sub-segments with a more mature life cycle stage in these categories. For this purpose, the segmented Kano perspective (Baier and Rese 2018; Rese et al. 2019) was applied as a second method. In this perspective, explicit dynamics could be identified and named.

3.2 Part B—Backstage

Through the skillful combination of data and technology, an improved understanding of customers, products, and markets is expressed in improved management decisions (Kelly 2000). This process generally follows three steps (Barton and Court 2012): capturing customer information, transferring it, and discovering the knowledge. Companies have increasingly begun to segment their customers according to customer behavior and value contribution during this time. In this environment, the actual modeling and selection of customers for marketing campaigns is located as well. Thematically, this task is closely linked to database marketing, and there is considerable overlap with the frontstage as well. Database marketing additionally refers to obtaining knowledge from data and transferring this knowledge in strategies to maintain a long-lasting (profitable) customer relationship (Hughes 1996; Blattberg et al. 2001). The research papers presented in this thesis focused on eCommerce retailing, which overlaps considerably with direct marketing. The prerequisites for maintaining one-to-one contact with the customer (Blattberg and Deighton 1991) are excessively provided in eCommerce retailing, for example, via newsletter, mailings, app-push, retargeting, and others. As presented in the following, the methodical approaches of database marketing were focused on for the papers presented in this part.

In database marketing, insights are gained from the analysis of previously completed or ongoing marketing campaigns. Typical questions include the following: “Was the campaign successful?”, “How many customers participated in the campaign?”, or “What return on investment (ROI) was achieved?”. However, there is another strength of database marketing: predicting a customer’s development or so-called predictive modeling. Nevertheless, this modeling is not about a perfect prediction of the future. Instead, imperfect targetability describes that profit and sustainable competitive advantages can be created by applying more effective models than using chance alone (Chen et al. 2001; Chen and Iyer 2002).

A fundamental requirement in database marketing is statistically utilizable trials with a test and control group. From the customer characteristics with statistic methods, the drivers for a different development can be isolated and used as a predictive model for future similar marketing campaigns. Neslin et al. (2006) have described the procedure as follows: define the problem, prepare the data, estimate the model, and select the targets. Various statistical methods can be used for modeling at this point (Blattberg et al. 2008): linear regression, logistic regression, the Tobit-model, decision trees, neural networks, and machine learning algorithms, to name a few. Nevertheless, here as well, the development is just at the initial stage. Thus, the application of artificial intelligence (AI) has become increasingly crucial for CRM (Salesforce 2018). The growing role of AI is reflected in personalized offers as well as automation. AI-powered CRM systems now guide sellers through decision-making processes to continuously improve their products or campaigns. Furthermore, robotic process automation is no longer only a pipe dream but has already achieved a clear ROI in individual companies. Additionally, the use of AI in personalization has continuously advanced, known as hyper-personalization. Here, as well, AI can set new impulses, for example, by adding unstructured interaction data to purchasing histories and similar strategies (Salesforce 2018). Therefore, to tackle this field, the papers presented in this part chose methods based on AI as well.

This part of the thesis contributes to the understanding of a particular form of predictive models and the determining of the causal effect (Holland 1986) or uplift (Radcliffe 2007). The causal effect describes the marketing campaign's effectiveness and is the difference between the treatment and the otherwise expected baseline. A customer who buys because of a campaign might have made a purchase without the campaign. The causal effect estimation hints to customers who can be excluded from marketing campaigns to improve the ROI.

Papers #3 and #4 have dealt with the selection of customers for specific promotional offers. The optimal allocation of the advertising budget results in an investment in customer satisfaction and sales promotions. While response predictions and the recency-frequency-monetary value approach have been long-established methods for optimizing (direct) marketing campaigns, a new approach has emerged since the beginning of the 1990s: modeling causal effects. While response models can only predict the behavioral change in the case of a sales promotion, the causal effect (or uplift or net lift (Devriendt et al. 2018)) considers the behavioral change that can be traced back to a treatment. In concrete terms, a customer would react favorably to a treatment, but, simultaneously, there would be a high probability of buying if the treatment was not applied. This surplus reflects the treatment's actual effectiveness and can determine the efficiency concerning the costs incurred. The marketing campaigns' optimization consists of identifying the customers with the highest causal effect in order to prioritize contacting them. In contrast, customers with a pessimistic prediction should more ideally not be addressed. Work previously existing to the research papers presented in these papers have primarily focused on the case of a dichotomous response, such as a purchase. Additionally, no research has considered different campaign cost structures.

Research Paper #3 has focused on predicting and optimizing continuous values, such as sales or profit. It is not untypical that the response rate to marketing campaigns is in the single-digit percentage range in practice; thus, dealing with an excess of zeros has been necessary. For this purpose, the Heckman sample selection model, the zero-inflated negative binomial regression model, and random forest-based regression were compared. Paper #3 used the freely available Hillstrom dataset (Hillstrom 2008) with 64,000 records (conversion rate 0.9%) and a new dataset from a large German eCommerce retailer with 155,388 datasets (conversion rate 9.25%), each with an equally distributed test and control group. With a predetermined transformation from observed sales to profit, the effectiveness of the method was proven. In a direct model comparison, all three methods applied were convincing with nearly equally strong results.

Research Paper #4 inherited the work on the excesses of zeros and the prediction of profit from Paper #3. At this point, the literature has only dealt with the effectiveness of causal effect estimation in terms of cost per contact. Paper #4 has examined the cost side for the first time. Based on the results from Paper #3, the results could be examined from an ROI perspective. Three ideal cost structures were simulated on a new dataset with 295,040 records with an equally distributed test and control group from a large German eCommerce retailer: fixed costs per contact (e.g., an advertising brochure), response-fixed (e.g., a voucher for purchase), and response-variable (e.g., a discount for a purchase). Theoretically, it could be deduced that there was a significant difference between discounts and other cost structures. Instead of showing the causal effect as a difference, a quotient yielded a significant performance improvement.

Additionally, how the causal effect manifested, either by more purchases or by higher shopping baskets, led to a hardly noticed shortcoming: In the latter case, no uplift was predicted in the previously widespread approach, which had only focused on purchases. This form of causal effect only became visible when the shopping basket value was predicted. The role of redemption behavior in response-dependent marketing costs was examined as well. Existing model evaluation methods were still overly specific for this new approach; thus, the previously used metrics in the literature had to be generalized. The results were validated in a Monte Carlo setting, with 100 randomly selected training and validation splits. These clearly showed that the cost structure had a massive impact on the results.

The findings confirmed the conventional approach's former results as long as only costs per contact were incurred. However, particularly in the case of the other cost structures that had not yet been considered, these failed. Thus, it could be practically proven that a representation of the causal effect as a quotient for response-variable costs such as discounts or rebates was the more ideal choice.

Table 1 summarizes all the research papers included in Chapters 2–5. Chapter 6 provides the conclusion of this thesis.

Table 1 Summary of research papers included in the thesis

	Title	*Authors*	*Content of the research paper*	*Methodological approach*	*Database*	*Status*
			Part A: Frontstage / customer-facing level			
Research Paper #1 Chapter 2	Penalty Reward Contrast Analysis (PRCA) for Categorizing Service Components: A New Approach	Stöcker, B., Nasseri, A.	- Discussion of penalty reward contrast Analysis for measuring the non-linearity concerning service portfolio management - Development of a new approach with cubic regression - Application and comparison of the results	Penalty reward contrast analysis (PRCA), cubic regression	980 (two quantitative CATI-surveys)	Published *Archives of Data Science*
Research Paper #2 Chapter 3	New Insights in Online Fashion Retail Returns from a Customers' Perspective and Their Dynamics	Stöcker, B., Baier, D., Brandt, B	- Investigation of potential measures suitable to avert or avoid returns in the pre-purchase, purchase, and post-purchase phases - Development of a three-stage return management approach - Application of the segmented Kano perspective to reveal Kanos' dynamics	Theory of attractive quality, segmented Kano perspective	8,393 (three quantitative online surveys)	First revision completed *Journal of Business Economics*
			Part B: Backstage / functional level			
Research Paper #3 Chapter 4	Maximizing Profit from Direct Marketing Campaigns: Profit Uplift Modeling Approaches for Online Shops	Baier, D., Stöcker, B.	- Discussion of current uplift approaches, primarily covering binary outcomes - Development of an adaption to a new profit-based uplift modeling - Application and comparison of three different modeling techniques	Heckman sample selection, zero-inflated negative binomial regression, and random forest-based regression	155,388 64,000 (CRM datasets from controlled A/B Test)	First revision completed *Journal of Business Economics*
Research Paper #4 Chapter 5	A Better Understanding of Cost-related Dependencies in the Estimation of the Causal Effects in Direct Marketing Campaigns	Stöcker, B., Baier, D.	- Discussion of the actual shortcomings regarding the lack of campaign cost perspective, sources of uplift, and the influence of redemption - Development of a new approach by integrating cost and activity accounting framework and generalized metrics - Validation of the new approach through 100 random data splits (Monte Carlo method)	Zero-inflated negative binomial regression, cost and activity accounting, Rubin's causal model	295,404 (CRM datasets from controlled A/B Test)	Under revision *International Journal of Research in Marketing*

References

Baier, Daniel; Rese, Alexandra; Roeglinger, Maximilian (2018): Conversational User Interfaces for Online Shops: A Segmented Kano Perspective. In *Proceedings on the 39th International Conference on Information Systems (ICIS*), San Francisco, December 13-16, 2018.

Barton, Dominic; Court, David (2012): Making advanced analytics work for you. In *Harvard Business Review* 90 (10), pp. 78–83.

Blattberg, Robert C; Getz, Gary; Thomas, Jacquelyn S (2001): Customer Equity: Building and Managing Relationships as Valuable Assets. In *Harvard Business School Press*.

Blattberg, Robert C.; Deighton, John (1991): Interactive marketing: exploiting the age of addressability. In *Sloan management review* 33 (1), pp. 5–15.

Blattberg, Robert C.; Kim, Byung-Do; Neslin, Scott A. (2008): The Predictive Modeling Process. In Robert C. Blattberg, Byung-Do Kim, Scott A. Neslin (Eds.): Database Marketing, vol. 18. New York, NY: Springer New York (International Series in Quantitative Marketing), pp. 245–287.

Brandt, Randall D. (1987): A procedure for identifying value-enhancing service components using customer satisfaction survey data. In *Add Value to Your Service, Chicago: American Marketing Association*, pp. 61–65.

Bruhn, Manfred (2009): Das Konzept der kundenorientierten Unternehmensführung. In Hans H. Hinterhuber, Kurt Matzler (Eds.): Kundenorientierte Unternehmensführung. Kundenorientierung - Kundenzufriedenheit - Kundenbindung. 6., überarbeitete Auflage. Wiesbaden: Gabler Verlag / GWV Fachverlage GmbH Wiesbaden, pp. 33–68.

Bruhn, Manfred (2016): Kundenorientierung. Bausteine für ein exzellentes Customer Relationship Management (CRM). München: C.H. Beck (dtv Beck-Wirtschaftsberater, 50950).

Chen, Yuxin; Iyer, Ganesh (2002): Research note consumer addressability and customized pricing. In *Marketing Science* 21 (2), pp. 197–208.

Chen, Yuxin; Narasimhan, Chakravarthi; Zhang, Z. John (2001): Individual marketing with imperfect targetability. In *Marketing Science* 20 (1), pp. 23–41.

Devriendt, Floris; Moldovan, Darie; Verbeke, Wouter (2018): A Literature Survey and Experimental Evaluation of the State-of-the-Art in Uplift Modeling: A Stepping Stone Toward the Development of Prescriptive Analytics. In *Big Data* 6 (1), pp. 13–41. DOI: 10.1089/big.2017.0104.

Dwyer, F. Robert; Schurr, Paul H.; Oh, Sejo (1987): Developing buyer-seller relationships. In *Journal of Marketing* 51 (2), pp. 11–27.

Ehrnrooth, Hanna; Gronroos, Christian (2013): The hybrid consumer: exploring hybrid consumption behaviour. In *Management Decision* 51 (9), pp. 1793–1820. DOI: 10.1108/MD-12-2012-0867.

Forrester (2020): Top CRM Trends For 2020. With assistance of Kate Leggett. Available online at https://go.forrester.com/blogs/top-crm-trends-for-2020/, checked on October 27, 2020.

Galbraith, Jay R. (2005): Designing the customer-centric organization. A guide to strategy, structure, and process. 1. ed. San Francisco, Calif.: Jossey-Bass. Available online at http://www.loc.gov/catdir/enhancements/fy0621/2005001675-b.html.

Gupta, Sunil; Hanssens, Dominique; Hardie, Bruce; Kahn, Wiliam; Kumar, V.; Lin, Nathaniel et al. (2006): Modeling customer lifetime value. In *Journal of Service Research* 9 (2), pp. 139–155.

Hillstrom, Kevin (2008): The MineThatData e-mail analytics and data mining challenge. In *MineThatData blog*. Available online at https://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.html, checked on October 27, 2020.

Holland, Paul W. (1986): Statistics and Causal Inference. In *Journal of the American Statistical Association* 81 (396), p. 945. DOI: 10.2307/2289064.

Hughes, Arthur M. (1996): Complete database marketer. Second-generation strategies and techniques for tapping the power of your customer database. Rev. ed. New York: McGraw-Hill.

Kano, Noriaki (1984): Attractive quality and must-be quality. In *Hinshitsu (Quality, The Journal of Japanese Society for Quality Control)* 14, pp. 39–48.

Kano, Noriaki (2001): Life cycle and creation of attractive quality. In *Proceedings of the 4th QMOD Conference*. Linkoping, Sweden, September 12-14, 2001, pp. 12–14.

Kelly, Sean (2000): Analytical CRM: the fusion of data and intelligence. In *Interactive Marketing* 1 (3), pp. 262–267.

Kumar, V.; Reinartz, Werner (2018): Customer Relationship Management. Berlin, Heidelberg: Springer Berlin Heidelberg.

Meffert, Heribert; Burmann, Christoph; Kirchgeorg, Manfred (2008): Marketing. Grundlagen marktorientierter Unternehmensführung; Konzepte - Instrumente - Praxisbeispiele. 10., vollst. überarb. und erw. Aufl. Wiesbaden: Gabler (Meffert-Marketing-Edition).

Mikulić, Josip; Prebežac, Darko (2011): A critical review of techniques for classifying quality attributes in the Kano model. In *Managing Service Quality: An International Journal*.

Morgan, Robert M.; Hunt, Shelby D. (1994): The commitment-trust theory of relationship marketing. In *Journal of Marketing* 58 (3), pp. 20–38.

Mulhern, Francis J. (1999): Customer profitability analysis: Measurement, concentration, and research directions. In *Journal of Interactive Marketing* 13 (1), pp. 25–40.

Neslin, Scott A.; Gupta, Sunil; Kamakura, Wagner; Lu, Junxiang; Mason, Charlotte H. (2006): Defection detection: Measuring and understanding the predictive accuracy of customer churn models. In *Journal of Marketing Research* 43 (2), pp. 204–211.

Nilsson-Witell, Lars; Fundin, Anders (2005): Dynamics of service attributes: a test of Kano's theory of attractive quality. In *International Journal of Service Industry Management* 16 (2), pp. 152–168. DOI: 10.1108/09564230510592289.

Niraj, Rakesh; Gupta, Mahendra; Narasimhan, Chakravarthi (2001): Customer profitability in a supply chain. In *Journal of Marketing* 65 (3), pp. 1–16.

Payne, Adrian; Frow, Pennie (2005): A strategic framework for customer relationship management. In *Journal of Marketing* 69 (4), pp. 167–176.

Radcliffe, Nicholas J. (2007): Using control groups to target on predicted lift: Building and assessing uplift models. In *Direct Marketing Analytics Journal* 1, p. 1421.

Reichheld, Frederick F.; Markey Jr, Robert G.; Hopton, Christopher (2000): E-customer loyalty-applying the traditional rules of business for online success. In *European Business Journal* 12 (4), p. 173.

Reichheld, Frederick F.; Teal, Thomas; Smith, Douglas K. (1996): The loyalty effect. In *Harvard Business School Press*.

Reinartz, Werner; Krafft, Manfred; Hoyer, Wayne D. (2004): The customer relationship management process: Its measurement and impact on performance. In *Journal of Marketing Research* 41 (3), pp. 293–305.

Reinartz, Werner; Kumar, V. (2002): The mismanagement of customer loyalty. In *Harvard Business Review* 80 (7), 86-94, 125.

Reinartz, Werner; Kumar, V. (2000): On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing. In *Journal of Marketing* 64 (4), pp. 17–35.

Rese, Alexandra; Schlee, Tobias; Baier, Daniel (2019): The need for services and technologies in physical fast fashion stores: Generation Y’s opinion. In *Journal of Marketing Management* 35 (15-16), pp. 1437–1459.

Rindfleisch, Aric; Heide, Jan B. (1997): Transaction cost analysis: Past, present, and future applications. In *Journal of Marketing* 61 (4), pp. 30–54.

Salesforce (2018): State of the Connected Customer. Insights from 6,700+ consumers and business buyers on the intersection of experience, technology, and trust (Second Edition). Available online at https://c1.sfdcstatic.com/content/dam/web/en_us/www/documents/e-books/state-of-the-connected-customer-report-second-edition2018.pdf, updated on October 27, 2020.

Srivastava, Rajendra K.; Shervani, Tasadduq A.; Fahey, Liam (1998): Market-based assets and shareholder value: A framework for analysis. In *Journal of Marketing* 62 (1), pp. 2–18.

Stauffer, David (2001): What Customer-Centric Really Means: Seven Key Insights. In *Harvard Management Update* 6 (8), pp. 1–3.

Tahir Albayrak; Meltem Caber (2013): Penalty–Reward-Contrast Analysis: a review of its application in customer satisfaction research. In *Total Quality Management & Business Excellence* 24 (11-12), pp. 1288–1300.

Webster Jr, Frederick E. (1992): The changing role of marketing in the corporation. In *Journal of Marketing* 56 (4), pp. 1–17.

Zeithaml, Valarie A.; Rust, Roland T.; Lemon, Katherine N. (2001): The customer pyramid: creating and serving profitable customers. In *California management review* 43 (4), pp. 118–142.

Chapter 2

Penalty Reward Contrast Analysis (PRCA) for Categorizing Service Components: A New Approach

Björn Stöcker and Aydin Nasseri

Abstract

Ever since Noriaki Kano's research, we have known that the relationship between performance and customer satisfaction is not just linear. Depending on the performance, different customer requirements exist, which are visualized in the Kano Model with three curves. In this article, we would like to present a new method that uses Kano's model to characterize different service components using a cubic term. We then compare the results of the Penalty Reward Contrast Analysis (PRCA) and the cubic terms and recommend how the cubic terms can be interpreted, based on two surveys of an online retailer collected via CATI (study 1 in 2011 with n=480 and study 2 in 2013 with n=500). This paper makes three contributions: 1) we compare three different and popular applications of the PRCA on real customer data, then 2) contrast the results with our new approach of using cubic terms and 3) give hints towards causal relations of different service components to the overall customer satisfaction in the fashion online business.

This chapter has been published in:

Stöcker, Björn; Nasseri, Aydin (2020): Penalty Reward Contrast Analysis (PRCA) for Categorizing Service Components: A New Approach. In *Archives of Data Science, Series A (Online First)* 6 (2).

1. Introduction

Predatory competition in the retail sector has been taking place for years. The market environment is characterized by an overcapacity of goods and services. In this highly competitive environment (buyer's market), it is existential to know where investments can be used most profitably. There are many studies concerning the effect of investments in service quality on repeat purchase (Szymanski and Henard 2001), retention (Bolton 1998), loyalty (Anderson and Sullivan 1993), retail sales performance (Gomez et al. 2004), and profitability (Anderson et al. 1994; Bernhardt et al. 2000). A lot of research has been done in recent decades. The original assumption that the relationship between experienced (service) quality and overall satisfaction is simply linear is outdated.

The studies show how important it is to understand the impact relationship on each individual service component. Kano's example: A Must-be factor must maintain a certain performance level in order not to have a negative effect on satisfaction; an investment beyond this level has no economic benefit. One dimensional factors, on the other hand, are always perceived by the customer; poor or good performance influence customer satisfaction and thus indirectly the success of the company. On the other hand, if Attractive factors are not expected, and if they are not present, they do not lead to dissatisfaction, but these can lead to a differentiation in the market.

PRCA is often used to determine the current service performance of a company and the effect of the individual components on overall satisfaction. However, this method is associated with many limitations.

In this paper, we would like to show that in this environment, Kano's model can also be determined using cubic terms and that this sometimes leads to diametrically different findings than PRCA. We start by laying the theoretical foundations for the emergence and correlations of service quality, the different applications of PRCA, and why it is so difficult to get answers from dissatisfied customers in section 2. In section 3, we introduce the survey data, apply different PRCA strategies, and the new approach with cubic terms, and compare the results. In section 4, we draw our conclusions, talk about the limitations of the new approach, and give a short outlook.

2. Theoretical Background

The initial assumption of a linear correlation between (service) performance and (service) satisfaction has been challenged (Mittal et al. 1998; Anderson and Mittal 2000). Non-linear relationships can be found in the prospect theory (Daniel Kahneman and Amos Tversky 1979), in regression analysis and cross-sectional survey data in health care and automobile settings (Mittal et al. 1998), in hypermarkets (Ting and Chen 2002), the automotive industry (Matzler et al. 2004), educational program e-portal (Cheung and Lee 2005, 2009) and e-services (Finn 2011). Kano (1984) described two different non-linear response functions and classified them as Attractive or Must-be. Oliver et al. (1997) later described the same response functions as monovalent satisfier and monovalent dissatisfier. Herzberg et al. (1959) also described this asymmetry: Hygiene factors which, if positive, prevent the development of dissatisfaction but do not contribute to satisfaction and Motivators, thus change satisfaction, but their absence does not necessarily lead to dissatisfaction. Brandt (1987), in the PRCA, identifies two characters called penalty and reward factors. The PRCA also gives hints towards the best service design by calculating the driver's strength.

2.1 Service Component Categories

Kano was the first person to describe two non-linear relations. In his work, he supplemented the linear relationship, which was the initial assumption towards the drivers for customer satisfaction in the early days (Fig. 1(a)) and argued that the degree to which customer requirements are met depending on the importance of the product or service component has different effects on customer satisfaction. Quality components whose poor fulfillment leads to great dissatisfaction, but when done well, not to satisfaction are classified as Basic factor. Secondly, the Attractive factors describe those components which contribute to a high degree of customer satisfaction when done well but have no negative effect when poorly fulfilled. The One-dimensional factors show a proportional correlation between the degree of fulfillment and satisfaction (Kano 1984, 1968, 1987, 1995; Berger 1993; Sauerwein 2000; Löfgren and Witell 2005; Mark C. Lee and John F. Newcomb 1997; Högström 2011). In addition to its use to categorize services, the Kano model is also represented in other areas, such as conversational user interfaces (Baier et al. 2018) or digitalization cases for e-commerce retailers (Baier et al. 2019).

Kano (2001, p.°1) and Fundin (2005, p.°18) found that the classification of the components is not static but changes over time to follow an attribute lifecycle.

Non-linear response functions are claimed in different shapes. Components showing an asymmetry towards satisfaction are linked to customer delight (Oliver et al. 1997). Furthermore, customer delight mostly arises from unexpected positive customer experience (Rust and Oliver 2000). An explanation for the asymmetry towards dissatisfaction can be found in the prospect theory (Daniel Kahneman and Amos Tversky 1979). Customer satisfaction or dissatisfaction arises from the difference between expected and experienced individual's performance standards. People tend to weigh losses greater than gains (loss aversion), shown in a steeper slope. Mittal found in his work regarding services and products that: "overall satisfaction displays diminishing sensitivity to attribute level performance" (Mittal et al. 1998, p.°33), later also called "satisfaction maintaining attributes" (Anderson and Mittal 2000, Fig. 2, Panel 2). The graph seems to represent a cube root (Fig. 1(c)). And Woodruff proposed to modify the confirmation/disconfirmation paradigm by, amongst other things adding a "zone of indifference" (Fig. 1(d)). "For all practical purposes, perceived performance within some interval around a performance norm is likely to be considered equivalent to the norm." (Woodruff et al. 1983, p.°299). Here the graph represents a monotonically increasing cubic.

In addition to his model, Kano has also developed a method to classify the components. He proposes using two questions on a 5 point Likert scale to categorize a specific component. The one question asked is functional ("What would you say if the product has . . ."), and the other is dysfunctional ("What would you say if the product has not …"). The two answers are used to categorize the component via the two-dimensional evaluation chart (Kano 1984, p.°173). For example, high values in the functional question (“I like it if [component] is fulfilled”) and mean values in the dysfunctional question (“I’m indifferent when [component] is not fulfilled”) means that the component is categorized as "Attractive").

Another way of determining the character of a component was formulated by Brandt (1987) in his work on the PRCA. Here he combines two linear functions to determine non-linear relations (Fig. 1(b)). Penalty factors have a steeper slope on the left side, where the poor performance is located, and a slighter slope on the right; in the case of reward factors, the situation is vice versa. This method is more suitable if a service performance should be levied.

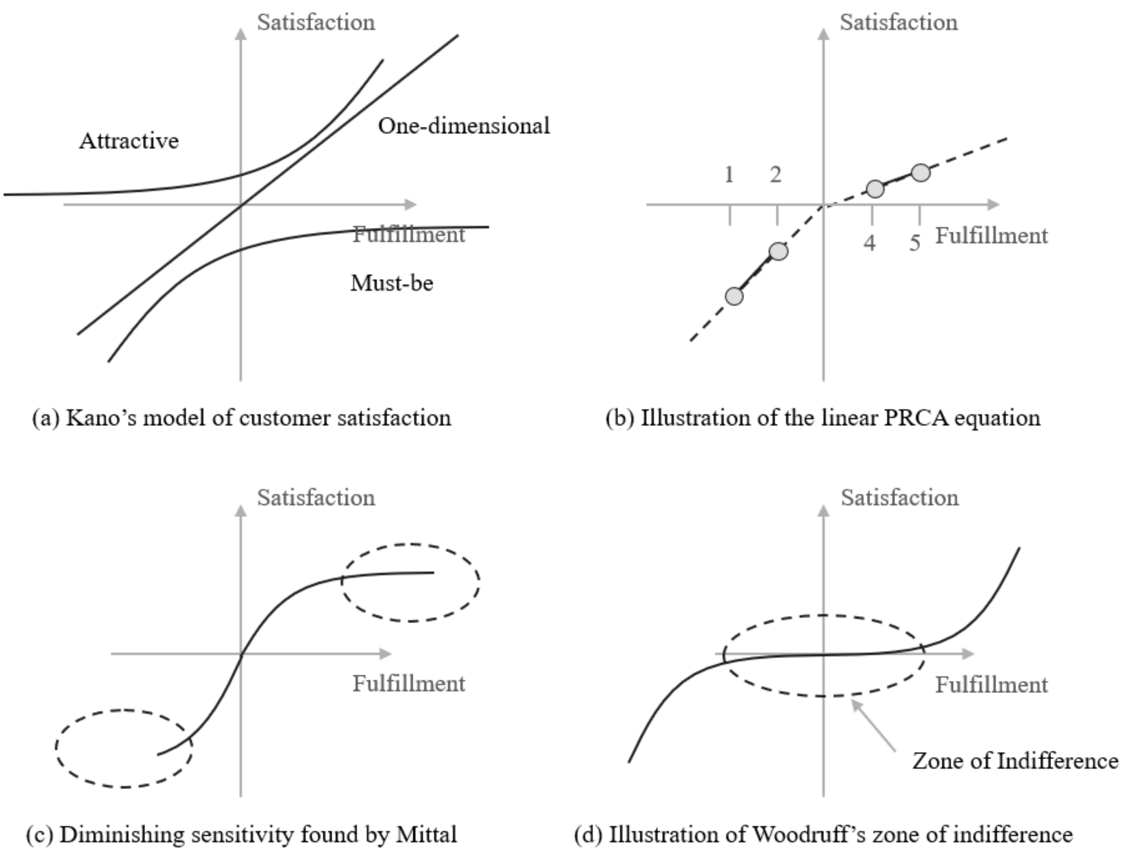


Fig. 1 Asymmetric relationships between fulfillment and satisfaction

2.2 Categorizing Service Components Using PRCA

As already mentioned, the PRCA is widely used in practical work to determine Kano's character (Tahir Albayrak and Meltem Caber 2013). The answers to the service fulfillment are usually queried based on a 5 point Likert scale, which has one middle option. Then for each (service) component, the PRCA fits a multiple linear regression using dummy variables to estimate the beta coefficients for penalty and reward. The dummy variable for penalty x_p is true for all answers "worse" or "much worse than expected," the dummy variable x_r representing reward is true for all answers "better" or "much better than expected." The middle option is not considered. The regression equation is

$$\hat{y} = b_p x_p + b_r x_r + b_0 \tag{1}$$

where $\hat{y}$ is the dependent variable for the overall satisfaction, b_0 y-intercept (constant term), and b_p and b_r are the beta coefficients for penalty and reward, synonymous with the slope.

To classify the components, Brandt (1987) proposed using the beta coefficients. A reward factor is given when the beta coefficient b_r is high, and b_p is low, for a penalty factor vice versa. Fuchs and Weiermair (2004) and Lin et al. (2010) suggest using the significance to classify the components. They call a significant b_p and an insignificant b_r Basic factor, an insignificant b_p, and a significant b_r Excitement factor, and add a third classification Performance factor for components where both beta coefficients are significant. Gierl and Bartikowski (2003) differentiate Brandt's classification into four classes. They use the strength of both beta coefficients combined to classify Satisfiers (high reward, low penalty), Criticals (high reward, high penalty), Neutrals (low reward, low penalty), and Dissatisfiers (low reward, high penalty).

Despite the different ways of defining the characters, the definition of which answers are considered to be recoded as high or low performance is also handled differently as shown in Table 1.

Table 1 Overview of the different recodings used in PRCA

Low	*High*	*Authors*	*Area of Research*
1, 2	4, 5	Lin et al. (2010)	Customer satisfaction with the online tax declaration services
1, 2	5	Matzler and Sauerwein (2002)	Customer satisfaction with the internal computer services of a hospital IT department
		Fuchs and Weiermair (2004)	Tourists' satisfaction with destination quality
		Alegre and Garau (2011)	Tourist satisfaction at sun and sand destinations
1	5	Mikulić and Prebežac (2008)	Passenger satisfaction with services at a major Croatian airport
		Mikulić and Prebežac (2011)	Passenger satisfaction with an international airport
		Back (2012)	Key drivers of customer satisfaction in Korean restaurants
		Coghlan (2012)	Tourists' satisfaction with destination attributes

Tahir Albayrak and Meltem Caber (2013), Table 1 modified

2.3 Categorizing Service Components: A New Approach

It is difficult for companies to get a complete picture of their customers' satisfaction. In order to be able to record the cause-effect relationships according to Kano and measure them, e.g., using PRCA, data from disappointed customers is also necessary in order to be able to make valid statements for penalty factors. In practice, the answers are not equally distributed; usually satisfied customers are overrepresented. On the one hand, this is because successful enterprises need content customers for their economic survival and, on the other hand, because dissatisfied customers hardly react and also for market research purposes are no longer accessible (Goodman et al. 1987, p.°169). In the TARP study of 1979 (Grainer et al. 1979) one proceeded from up to 50% Non-Complainers, so it was recognized in a more recent study 50-80% in the USA (Goodman et al. 2000) and Richins (1987) that especially in the case of not minor errors a supplier change is preferable to a complaint. For the service sector, Stauss (1989) also expects a higher proportion of Non Complainers due to the special characteristics.

3. Empirical Comparison

3.1 Data Collection

Two samples (study 1 with n = 480 and study 2 with n = 500) were analyzed. The qualitative data sets were collected via computer-assisted telephone interviews (CATI), each by using the same standardized cascaded (the respondents had to have used the service, self-assessed) questionnaire. To obtain the performance for a service component listed in Table 2: “You said you have the goods from the assortments… ordered by phone. How do you rate the telephone ordering process?”. Response option: (1) “Much worse than expected,” (2) “Worse than expected,” (3) “Neither good or bad,” (4) “Better than expected” or (5) “Much better than expected.” To receive the overall service satisfaction: “When you think of all the services we have discussed so far, how satisfied are you with them overall?”. Response option on a scale from (1) “very dissatisfied” to (5) “very satisfied.”

3.2 Categorizing Service Components Using PRCA

To show the differing results, we calculated the multiple linear regression for all three popular PRCA classification approaches for the question “[…] How do you rate the telephone ordering process?” (independent variable) and “[…] how satisfied are you with them overall?” (dependent variable) for all people, who used this service (n = 319, Table 3). All results can be found in Table 5 and Table 6. When both ends of the response scale are taken into account (12-45) (for penalty (1) “Much worse than expected” and (2) “Worse than expected”) and for reward ((4) “Better than expected” and (5) “Much better than expected”) we find that both beta coefficients are significant and have the same strength; therefore they are classified as One-dimensional. If we omit in addition the response option (4) by recoding (12-5), we see that two beta coefficients are significant and have the same strength, again One-dimensional. For the last approach, where just the ends of the rating scales enter into the dummy regression (1-5), we get a different picture: Only the reward beta coefficient is significant, then classified as Attractive—the poor adj. R^2 is also caused by the distinctive left-skewed distribution in both variables (dependent and independent).

Table 2 Service components queried in the two surveys

Phase	*Component*
Presales	Info Delivery Options Online Shop
	Info Delivery Options Catalogue
	Info Payment Methods Online Shop
	Info Payment Methods Catalogue
	Service Information at the Article in the Online Shop
	Service Information at the Article in the Catalogue
	Accuracy of Delivery Time Online Shop
	Accuracy of Delivery Time Catalogue
	Info Returns in the Online Shop
	Info Returns in the Catalogue
Ordering	Telephone Order Process
	Order Process Online Shop
Fulfillment	Delivery Time
	Reliability of Delivery Information
	Delivery to your Home
	Delivery to Another Address
	24-hour Delivery
	Delivery at the Desired Date
	Order Tracking and Tracing
	Delivery to the Parcel-Shop
	Delivery 2-man Team
	Simplicity of Bank Transfer
	Processing of the Instalment Purchase
	Satisfaction with the Telephone Complaint
	Satisfaction with the E-mail Complaint
Returning	Processing of the Return Shipment
	Return in the Parcel Shop
	Speed of the Credit Memo

Table 3 Results of the Different Three PRCA Approaches w.r.t. the Service Component "Telephone Order Process" in Study 1

Abbr.	*Low*	*High*	*R²*	*Adj. R²*	*Beta Penalty*	*Sig.*	*Beta Reward*	*Sig.*	*Class*
12-45	1, 2	4, 5	0.058	0.052	-0.685	0.044	0.333	0.000	O
12-5	1,2	5	0.056	0.050	-0.776	0.022	0.328	0.001	O
1-5	1	5	0.047	0.041	-1.260	0.124	0.344	0.000	A

Referring to Table 1: 12-45 = recoding answers 1-2 as dummy variable for penalty and 4-5 for reward, 12-5 = recoding just 1-2 and 5, 1-5 recoding only 1 and 5; O = One-dimensional, A = Attractive

3.3 Categorizing Service Components Using the New Approach

In order to find an equation other than in the PRCA that can reproduce the curves described in section 2.1, there are basically only two possibilities: using a piecewise function or a 3rd-degree polynomial. For piecewise functions, a positive or negative quadratic term is added to a linear term. In order to use this procedure correctly, the exact position of the joint of the functions must first be determined for each component. In our practical work, we have found out that the point at which the non-linear changes into the linear context (inflection point) is not always to be found in the middle of the scale (Fig. 2). We, therefore, propose the use of a 3rd degree polynomial, aka cubic term (CT), which can take on all three forms (Must-be, One-dimensional, and Attractive) for the domain and is not limited to a fixed inflection point at the same time. In our opinion, a characterization using a matrix analogous to Kano's evaluation chart is out of the question, as this would first require scaling in the data, especially in the case of skewed distributions; these would first have to be centered. The curve can be determined much more easily by the slope of the individual data points relatively in space. A high gradient indicates a high significance for the overall satisfaction; a gradient close to zero, on the other hand, indicates a low influence, also referred to as indifferent. In this context, it is not possible to determine the quality of fit using R^2, for example. Using the example of a Must-be component, a positive correlation can be determined for the left-hand side, but not for the right-hand side, since here, no correlation can be found. This is at the expense of the R^2, which represents the goodness of fit for all values.

Therefore, the highest R^2 would be found at a One-dimensional component, the lowest at an indifferent component. The interpretation of the cubic terms can be done graphically as well as via a table of values. In a direct comparison to PRCA, there is no need to recode the data into dummy variables or define high and low (Table 1).

A 3rd-degree polynomial regression fits a non-linear relation between the independent variables (service performance) denoted as x and the dependent variable (overall customer satisfaction) denoted as y using the method of least squares. The beta values (b_3, b_2, b_1, b_0) increase or decrease the conditional expectation of y.

(2) $$\hat{y} = b_3 x^3 + b_2 x^2 + b_1 x + b_0$$

As in the PRCA, we calculate the cubic regression for each service component, only including answers by people who have used the service. Using the example above, we arrive at this equation:

(3) $$\hat{y} = 0.03x^3 - 0.384x^2 + 1.807x + 1.399$$

with adj. R^2: .068.

If you compare the graph (Fig. 2(a)) with Kano‘s chart, you would categorize it as a Must-Be and not as One-dimensional or even Attractive, as the PRCA has shown in Table 3. We recommend going through these steps to apply CT properly:

1) Define the range for which values the cubic equation applies (domain). In some cases, you won’t find values for your independent variable on the lower or upper end of the rating scale. In this instance, the cubic regression can only speak for the given values

2) Calculate the derivate function of the cubic term to get the slope:

$$\hat{y}' = 0.09x^2 - 0.768x + 1.807$$

3) Calculate a table of values

x	*1*	*2*	*3*	*4*	*5*
$\hat{y}'$	1.13	0.63	0.31	0.18	0.22

A good example of Must-be can be seen in the component: “Accuracy of delivery time online shop” (n=96, study 1). Due to the non-linear change in the slope on the left-hand side (Table 4, Fig. 2(b))

Table 4 Table of values showing the slopes for the three ideal-typical examples

x	*1*	*2*	*3*	*4*	*5*
Must-be	3.4	1.3	0.2	0.0	0.6
One-dimensional	0.2	0.2	0.2	0.2	0.1
Attractive	N/A	-0.1	0.0	0.3	0.8

(4) $$\hat{y} = 0.148x^3 - 1.683x^2 + 6.279x - 3.282$$

For One-dimensional, the component “Delivery to your home” (n=511, study 1) is a good object of study. The slope remains constant through the whole domain shown in the graph and table of values (Table 4, Fig. 2(c)).

(5) $$\hat{y} = -0.009x^3 + 0.071x^2 + 0.052x + 3.569$$

And finally, the component “Order process online shop” (n=275, study 2) represents an Attractive factor. The slope changes non-linear on the right-hand side (Table 4, Fig. 2(d)). Another indicator for Attractive is that you find no values for “much worse than expected” (here: 1), which strengthens the findings of Rust and Oliver (2000).

$$\hat{y} = 0.025x^3 - 0.110x^2 + 0x + 4.399 \tag{6}$$

3.4 Comparison

As can be seen in Tables 5 and 6, the results within the different applications of the PRCA are very different (e.g., Table 5: “Speed of Credit Memo”) or congruent (Table 5: “Processing of the Return Shipment”). The categorization, according to CT, is similar in some cases to the PRCA (12-45), probably also because CT includes all response options in the regression. Depending on the recoding approaches, the variance is more or less lost so that only the ends of the scale are used in the PRCA approach (1-5). This leads to problems, especially when only a few answers fall into this range anyway (skewed distribution due to Non-Complainers, among other things). In addition, the PRCA tacitly assumes that inflection points are always to be found in the middle of the scale. Fig. 2 shows that this assumption is not always true. In the evaluation of all answers, we have often encountered inflection points outside the center of the scale, which is not sufficiently recognized by the PRCA. All in all, we are of the opinion that a determination of Kano’s model should preferably be done with CT because no dummy variables are formed, all responses are included in the regression, and any inflection points that may occur can lie outside the center of the scale. CT can be evaluated graphically or with a table of values.

In addition, we were also able to determine a lifecycle in the component categorization (Kano 2001; Fundin 2005). For example, the character of the “Telephone Order Process” changed from One-dimensional (study 1) to Must-be (study 2) or “Simplicity of Bank Transfer” from Attractive to Must-be. But also opposite effects were observed in “Order Process Online Shop” from Must-be (study 1) to Attractive (study 2), “Delivery Time” Must-be to One-dimensional or “Speed of the Credit Memo” Must-be to Attractive, which indicate that the company has actively worked on service quality or that customer expectations have changed.

Table 5 Results of the different PRCA approaches and cubic term, study 1

Component	*N*	*12-45*	*12-5*	*1-5*	*CT*
Info Delivery Options Online Shop	113	M	O	O	M
Info Delivery Options Catalogue	12	I	A	O	M
Info Payment Methods Online Shop	92	I	I	M	M
Info Payment Methods Catalogue	120	A	A	O	I
Service Information at the Article in the Online Shop	80	I	A	O	I
Service Information at the Article in the Catalogue	102	I	A	O	I
Accuracy of Delivery Time Online Shop	96	M	M	M	M
Accuracy of Delivery Time Catalogue	84	O	M	M	M
Info Returns in the Online Shop	62	I	I	M	I
Info Returns in the Catalogue	85	I	I	M	A
Telephone Order Process	323	O	O	A	O
Order Process Online Shop	219	A	A	O	M
Delivery Time	521	O	O	O	M
Reliability of Delivery information	503	O	O	O	M
Delivery to your Home	510	O	O	A	O
Delivery to Another Address	31	I	I	I	I
24-hour Delivery	40	M	M	M	M
Delivery at the Desired Date	35	I	I	M	I
Simplicity of Bank Transfer	388	A	A	O	A
Processing of the Instalment Purchase	103	A	A	A	O
Processing of the Return Shipment	283	A	A	A	A
Speed of the Credit Memo	81	A	I	M	M

M = Must-be, O = One-dimensional, A = Attractive, I = Indifferent

Table 6 Results of the different PRCA approaches and cubic term, study 2

Component	*N*	*12-45*	*12-5*	*1-5*	*CT*
Info Delivery Options Online Shop	128	I	A	O	A
Info Delivery Options Catalogue	118	A	I	I	I
Info Payment Methods Online Shop	109	I	I	M	M
Info Payment Methods Catalogue	107	A	I	M	M
Service Information at the Article in the Online Shop	103	M	O	O	O
Service Information at the Article in the Catalogue	76	A	I	M	O
Accuracy of Delivery Time Online Shop	131	A	O	A	M
Accuracy of Delivery Time Catalogue	69	I	I	I	M
Info Returns in the Online Shop	78	I	A	A	O
Info Returns in the Catalogue	57	I	I	M	M
Telephone Order Process	319	O	O	A	M
Order Process Online Shop	274	A	A	O	A
Delivery Time	570	O	O	O	O
Reliability of Delivery information	557	O	O	O	M
Delivery to your Home	545	O	O	A	M
Delivery to Another Address	19	I	I	I	O
24-hour Delivery	35				
Delivery at the Desired Date	14				
Simplicity of Bank Transfer	425	O	O	O	M
Processing of the Instalment Purchase	118	I	I	M	O
Processing of the Return Shipment	283	A	A	O	A
Speed of the Credit Memo	124	A	A	O	A

M = Must-be, O = One-dimensional, A = Attractive, I = Indifferent

4. Conclusions and Outlook

The study confirms a non-linear relationship between service performance and satisfaction. To classify the non-linear relations, according to Kano's Model, CT should be considered rather than any variation of the PRCA. CT does not share the same limitations of the PRCA, outperforming the PRCA approach, and sometimes delivering diametrically different interpretations. Choosing the

PRCA results instead of the CT ones can lead to wrong business decisions with far-reaching consequences. If a Must-be component is misinterpreted as Attractive, a completely different strategy is applied. Also, validation on other data is desirable.

The two surveys needed to perform a PRCA and CT are based on respondents who had used one or more specific services. Customer journeys usually have many different touchpoints and are unique, so it is impossible to fit a regression that includes all or even some important service components to overall customer satisfaction. In addition, every touchpoint can have a different influence on satisfaction, and critical incidents are recognized more often than, for example, a good performance in a Must-be setting. As a result, there are no data points for poor service performance because the company is managing these services very well. In this case, PRCA and CT can only speak for the given values in the surveys. Referring to point 2.3, respondents who are willing to take part in a survey are more likely to be loyal customers. To obtain more variance and get a more valid result, you need answers from Non-Complainers and lost customers as well.

References

Alegre, Joaquin; Garau, Jaume (2011): The factor structure of tourist satisfaction at sun and sand destinations. In: *Journal of Travel Research* 50 (1), S. 78–86.

Anderson, Eugene W.; Fornell, Claes; Lehmann, Donald R. (1994): Customer satisfaction, market share, and profitability: Findings from Sweden. In: *Journal of Marketing* 58 (3), S. 53–66.

Anderson, Eugene W.; Mittal, Vikas (2000): Strengthening the satisfaction-profit chain. In: *Journal of Service Research* 3 (2), S. 107–120.

Anderson, Eugene W.; Sullivan, Mary W. (1993): The antecedents and consequences of customer satisfaction for firms. In: *Marketing Science* 12 (2), S. 125–143.

Back, Ki-Joon (2012): Impact-range performance analysis and asymmetry analysis for improving quality of Korean food attributes. In: *International Journal of Hospitality Management* 31 (2), S. 535–543.

Baier, Daniel; Rese, Alexandra; Nonenmacher, Nikita; Treybig, Steve; Bressem, Benjamin (2019): Digital Technologies for Ordering and Delivering Fashion: How Baur Integrates the Customer's Point of View. In: Digitalization Cases: Springer, S. 59–77.

Baier, Daniel; Rese, Alexandra; Roeglinger, Maximilian (2018): Conversational User Interfaces for Online Shops? A Categorization of Use Cases.

Berger, Charles (1993): Kano's methods for understanding customer-defined quality. In: *Center for Quality Management Journal* 2 (4), S. 3–36.

Bernhardt, Kenneth L.; Donthu, Naveen; Kennett, Pamela A. (2000): A longitudinal analysis of satisfaction and profitability. In: *Journal of Business Research* 47 (2), S. 161–171.

Bolton, Ruth N. (1998): A dynamic model of the duration of the customer's relationship with a continuous service provider: The role of satisfaction. In: *Marketing Science* 17 (1), S. 45–65.

Brandt, Randall D. (1987): A procedure for identifying value-enhancing service components using customer satisfaction survey data. In: *Add Value to Your Service, Chicago: American Marketing Association*, S. 61–65.

Cheung, Christy M. K.; Lee, Matthew K. O. (2005): The asymmetric effect of website attribute performance on satisfaction: an empirical study. In: *Proceedings of the 38th annual Hawaii international conference on System Sciences*. IEEE, 175c-175c.

Cheung, Christy M. K.; Lee, Matthew K. O. (2009): User satisfaction with an internet-based portal: An asymmetric and nonlinear approach. In: *Journal of the American Society for Information Science and Technology* 60 (1), S. 111–122.

Coghlan, Alexandra (2012): Facilitating reef tourism management through an innovative importance-performance analysis method. In: *Tourism Management* 33 (4), S. 767–775.

Daniel Kahneman; Amos Tversky (1979): Prospect Theory: An Analysis of Decision under Risk. In: *Econometrica* 47 (2), S. 263–291.

Finn, Adam (2011): Investigating the non-linear effects of e-service quality dimensions on customer satisfaction. In: *Journal of Retailing and Consumer Services* 18 (1), S. 27–37.

Fuchs, Matthias; Weiermair, Klaus (2004): Destination benchmarking: An indicator-system's potential for exploring guest satisfaction. In: *Journal of Travel Research* 42 (3), S. 212–225.

Fundin, Anders (2005): Dynamics of quality attributes over life cycles of goods and services.

Gierl, Heribert; Bartikowski, Boris (2003): Ermittlung von Satisfiers, Dissatisfiers und Criticals in der Zufriedenheitsforschung. In: *der markt* 42 (1), S. 14–34.

Gomez, Miguel I.; McLaughlin, Edward W.; Wittink, Dick R. (2004): Customer satisfaction and retail sales performance: an empirical investigation. In: *Journal of Retailing* 80 (4), S. 265–278.

Goodman, John; O'Brien, P.; Segal, Eden (2000): Selling quality to the CFO. In: *Quality Progress, March*.

Goodman, John A.; Malech, Arlene R.; Marra, Theodore R. (1987): Beschwerdepolitik unter Kosten/Nutzen-Gesichtspunkten-Lernmöglichkeiten aus den USA. In: *Verbraucherzufriedenheit und Beschwerdeverhalten, Frankfurt/New York*, S. 165–202.

Grainer, Marc A.; McEvoy, Kathleen A.; King, Donald W. (1979): Consumer problems and complaints: A national view. In: *ACR North American Advances*.

Herzberg, F.; Mausner, B.; Snyderman, B. (1959): The motivation to work. In: *New York: Wiley*.

Högström, Claes (2011): The theory of attractive quality and experience offerings. In: *The TQM Journal* 23 (2), S. 111–127.

Kano, Noriaki (1968): Concept of TQC and its Introduction. In: *Kuei* 35 (4), S. 20–29.

Kano, Noriaki (1984): Attractive quality and must-be quality. In: *Hinshitsu (Quality, The Journal of Japanese Society for Quality Control)* 14, S. 39–48.

Kano, Noriaki (1987): Total quality creation. In: *ICQCC Tokyo Proceeding*.

Kano, Noriaki (1995): Upsizing the organization by attractive quality creation. In: Total Quality Management: Springer, S. 60–72.

Kano, Noriaki (2001): Life cycle and creation of attractive quality. In: *Proceedings of the 4th QMOD Conference*, Linkoping, S. 18–36.

Lin, Shu-Ping; Yang, Chen-Lung; Chan, Ya-hui; Sheu, Chwen (2010): Refining Kano's 'quality attributes-satisfaction'model: A moderated regression approach. In: *International Journal of Production Economics* 126 (2), S. 255–263.

Löfgren, Martin; Witell, Lars (2005): Kano's theory of attractive quality and packaging. In: *Quality Management Journal* 12 (3), S. 7–20.

Mark C. Lee; John F. Newcomb (1997): Applying the Kano Methodology to Meet Customer Requirements: NASA's Microgravity Science Program. In: *Quality Management Journal* 4 (3), S. 95–106.

Matzler, Kurt; Bailom, Franz; Hinterhuber, Hans H.; Renzl, Birgit; Pichler, Johann (2004): The asymmetric relationship between attribute-level performance and overall customer satisfaction: a reconsideration of the importance-performance analysis. In: *Industrial Marketing Management* 33 (4), S. 271–277.

Matzler, Kurt; Sauerwein, Elmar (2002): The factor structure of customer satisfaction: An empirical test of the importance grid and the penalty-reward-contrast analysis. In: *International Journal of Service Industry Management* 13 (4), S. 314–332.

Mikulić, Josip; Prebežac, Darko (2008): Prioritizing improvement of service attributes using impact range-performance analysis and impact-asymmetry analysis. In: *Managing Service Quality: An International Journal* 18 (6), S. 559–576.

Mikulić, Josip; Prebežac, Darko (2011): Rethinking the importance grid as a research tool for quality managers. In: *Total Quality Management & Business Excellence* 22 (9), S. 993–1006.

Mittal, Vikas; Ross Jr, William T.; Baldasare, Patrick M. (1998): The asymmetric impact of negative and positive attribute-level performance on overall satisfaction and repurchase intentions. In: *Journal of Marketing* 62 (1), S. 33–47.

Oliver, Richard L.; Rust, Roland T.; Varki, Sajeev (1997): Customer delight: foundations, findings, and managerial insight. In: *Journal of Retailing* 73 (3), S. 311–336.

Richins, Marsha L. (1987): A multivariate analysis of responses to dissatisfaction. In: *Journal of the Academy of Marketing Science* 15 (3), S. 24–31.

Rust, Roland T.; Oliver, Richard L. (2000): Should we delight the customer? In: *Journal of the Academy of Marketing Science* 28 (1), S. 86.

Sauerwein, Elmar (2000): Das Kano-Modell der Kundenzufriedenheit. In: Das Kano-Modell der Kundenzufriedenheit: Springer, S. 27–55.

Stauss, Bernd (1989): Beschwerdepolitik als Instrument des Dienstleistungsmarketing. In: *Jahrbuch der Absatz-und Verbrauchsforschung* 35 (1), S. 41–62.

Szymanski, David M.; Henard, David H. (2001): Customer satisfaction: A meta-analysis of the empirical evidence. In: *Journal of the Academy of Marketing Science* 29 (1), S. 16–35.

Tahir Albayrak; Meltem Caber (2013): Penalty–Reward-Contrast Analysis: a review of its application in customer satisfaction research. In: *Total Quality Management & Business Excellence* 24 (11-12), S. 1288–1300.

Ting, Shueh-Chin; Chen, Cheng-Nan (2002): The asymmetrical and non-linear effects of store quality attributes on customer satisfaction. In: *Total Quality Management & Business Excellence* 13, S. 547–569.

Woodruff, Robert B.; Cadotte, Ernest R.; Jenkins, Roger L. (1983): Modeling consumer satisfaction processes using experience-based norms. In: *Journal of Marketing Research* 20 (3), S. 296–304.

Chapter 3

New Insights in Online Fashion Retail Returns from a Customers' Perspective and Their Dynamics

Björn Stöcker, Daniel Baier and Benedikt Brandt

Abstract

Returns are an inconvenient problem in the mail-order business, not only for the merchant but also for the customer. However, we do not consider returns to be generally bad, but rather an integral part of the business model. Therefore, we investigate potentially suitable measures to avert or avoid returns in the pre-purchase, purchase and post-purchase phases. We look at current and technological developments in return management and the most critical drivers for fashion assortment returns on a holistic view of the issue and target all three purchase phases.

The resulting measures were assessed via an online questionnaire with 8,393 participants (customers of a German fashion online retailer) to impact customer satisfaction using Kano's method. There are clear measures that promise high customer satisfaction (such as 360° view) and a clear hierarchy regarding monetary and non-monetary measures. By applying the segmented Kano perspective, we found customer segments, which are different in their expectations towards returns. That allowed us to conclude dynamics regarding return management. This assessment is followed by discussing the results, conclusions, and indications for further research fields.

This chapter is under review in:

Journal of Business Economics

1. Introduction

While serving consumers online provides multiple benefits for online retailers (e.g., reaching consumers worldwide), it is also tied to some disadvantages inherent to distance trading. Especially product (fit) uncertainty (Hong and Pavlou 2014) and the missing touch and feel of products (Shulman et al. 2011) result in large amounts of product returns. These product returns are not only causing enormous costs for online shop operators (Samorani et al. 2019; Yan and Pei 2019) but additionally negatively affect the environment (Dutta et al. 2020; Pålsson et al. 2017). The number of returns shows to be very high in the online fashion business in particular, due to its less standardized products (Difrancesco et al. 2018; Saarijärvi et al. 2017), the need for clothing to fit correctly (Gallino and Moreno 2018; Gelbrich et al. 2017) and the importance of apparel's texture (Ofek et al. 2011). Since handling the return policy more or less lenient in this business will trigger higher purchase frequencies or prevent consumers from buying products (Hjort and Lantz 2016; Janakiraman et al. 2016), it is crucial to ascertain the golden mean for managerial implications.

Previous studies focused on finding optimal countermeasures for keeping return rates low without scaring off potential customers, whereas we contribute to the literature by examining the problem of returns holistically. Therefore, we extend the two-step decision perspective from Wood (2001), according to which online purchase decisions are divided into the (first) decision for or against a purchase, and the (second) decision for against keeping the product, by analyzing measures to prevent product returns in three stages. These measures comprise supporting consumers searching for fashion products (pre-purchase stage), assistance in the ordering process (purchase stage), as well as strategies inducing consumers to keep the product (post-purchase stage). While the vast majority of literature focuses on preventing returns either before or after the purchase, we enable a direct comparison of measures for reducing returns by investigating all three stages with the same methodological approach. We use Kano's „Theory of Attractive Quality“ (Kano et al. 1984) as a basis, from which we have the respondents categorize the measures. Besides, to the best of the authors' knowledge, we are the first to apply (segmented) Kano's method in product returns, thus revealing those return measures that increase customers' satisfaction the most.

Furthermore, we address potential solutions for product returns by implementing the most recent technological advances, such as virtual fitting of articles or 360° views of the products. Hence, we want to shed light on how consumers evaluate measures for preventing product returns in the context of online fashion shops at each of the three stages and to what extent they affect consumers' satisfaction. By answering this question, we cover recently postulated research gaps (Janakiraman et al. 2016; Samorani et al. 2019) and indicate how managers could efficiently allocate financial budgets regarding their return policy.

Therefore, this study is structured as follows: first, we illustrate return management, its most recent developments, and technological improvements, as well as drivers of returns. We then describe our methodical approach leading to the results yielded. After discussing these, we end with a conclusion and directions for future research.

2. Theoretical Background

With ever-increasing numbers of online shopping orders, the issue of product returns also becomes more critical. Even if the current return ratio remains constant, the consequence will negatively affect the environment heavily (Dutta et al. 2020; Pålsson et al. 2017). Furthermore, product returns constitute a cumbersome, unpleasant task for companies and consumers likewise. As the e-commerce industry still struggles to provide sufficient and appropriate product information for customers to prevent (or at least reduce) returns (Gelbrich et al. 2017), and thus might not be able to offer suitable solutions soon, it is essential to explore product returns in comprehensive depth and based on recent technological advancements. Following the theoretical framework of the Confirmation-Disconfirmation paradigm in the context of products bought online (Hong and Pavlou 2014), the satisfaction with the delivered product (post-purchase) might be (1) lower than expected, resulting in a negative confirmation, (2) as expected resulting in zero (dis)confirmation, or (3) higher than expected resulting in positive confirmation.

2.1 Return Management and Recent Developments

The emergence of a return is to be understood due to a comparison of expectations (while shopping online) and reality (when receiving the product), as illustrated in Figure 1. In the context of fashion, the expectations regarding the nature of the article (correct article) and the fit (correct fit) should be understood as a logical consequence, whereby the comparison of the expectations to the actual product can be moderated by curating the offer, e.g., through personal or personalized outfit recommendations. Resolving the information gap then leads to satisfaction or dissatisfaction with the ordered article. Nevertheless, it can be assumed that satisfaction alone does not directly affect the return behavior. A customer can be satisfied with a delivered article but still return it (selection order of several sizes or budget reached). It is also conceivable that an unsatisfied customer does not make a return but avoids buying a product from the supplier/manufacturer as a result. The influence of perceived service quality and its influence on the return behavior (e.g., delivery time) was not considered in this study.

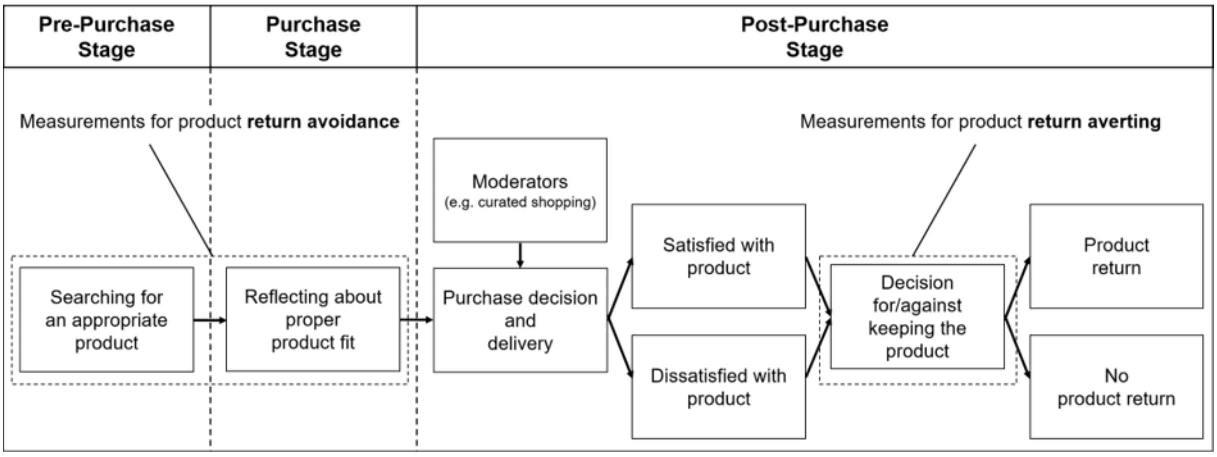


Fig.1 Pre- and post-purchase stages with corresponding return prevention starting points

To categorize product returns properly, we refer to product returns *before* the purchase decision as „return avoidance," whereas those return measures *after* the purchase decision will be named „return averting." In the second case, the aim is no longer to influence an article's expectations but to negotiate with the customer about the intended return. This negotiation can be done with, for example, money or an appeal. We assume that it is easier to negotiate with a customer satisfied with the article than with dissatisfied customers. In the latter case, the company must also consider whether suppressing the return is beneficial for the customer relationship or conceptualizing the offer should avoid lasting customer annoyance. In general, these measures should be applied with caution because once customers have understood this mechanism, they could actively use it to their advantage and change their ordering and purchasing behavior in this direction (Gelbrich et al. 2017).

The return literature dealing with these issues could be segmented into different groups based on their approach (Table 1). While some studies model different scenarios based on researchers' assumptions (Difrancesco et al. 2018; Letizia et al. 2018; Li et al. 2019; Dutta et al. 2020; Ülkü and Gürler 2018) or founded on observable online shopping data (Samorani et al. 2019; Gallino and Moreno 2018; Sahoo et al. 2018; Rao et al. 2018; Lohse et al. 2017; Walsh et al. 2016; Hjort and Lantz 2016; Petersen and Kumar 2015; Minnema et al. 2016), we analyze measures of return avoidance and averting. By focusing on the customer's voice, as finally, customers' evaluation contributes to a more or less successful implementation of these measures. Thus, we conducted a literature review about recent articles (published between 2015-2020) that either include „product return", „return prevention", „reverse logistics", or „return policy" in common scientific databases. After screening them by abstracts, we highlight those incorporating customers' viewpoints derived from survey-based investigations.

It becomes evident that most studies investigating return measures from a customer's viewpoint explore either the purchase or the post-purchase (returning) stage, and thereby not allowing a direct comparison of the effectiveness of the measures analyzed. In the same vein, the meta-analytic review by Janakiraman et al. (2016, p. 234) concludes that „[p]rior research has largely examined these effects separately." In contrast, studies interviewing the same respondents on product return prevention measures for both purchase and post-purchase are very scant. To the best of our knowledge, we are the first to analyze return prevention measures for all three stages by applying the Kano method.

Table 1 Recent studies investigating returns from a consumers' viewpoint

Author(s) (Year)	*Method(s) of Investigation*	*Stages Analyzed*
Shulman et al. (2015)	ANCOVA (n=420)	Purchase and Post-Purchase
Singh and Pandey (2015)	EFA (n=347)	Purchase
Seo et al. (2016)	ANOVA (n=100; n=113; n=250)	Purchase
Gelbrich et al. (2017)	ANOVA and ANCOVA (n=217; n=138)	Post-Purchase
Lee and Yi (2017)	ANOVA (n=78; n=82; n=107)	Post-Purchase
Saarijärvi et al. (2017)	Semi-structured interviews (n=21)	Post-Purchase
Oghazi et al. (2018)	SEM (n=730)	Purchase
Pei and Paswan (2018)	SEM (n=400)	Post-Purchase
Zhou et al. (2018)	ANOVA and SEM (n=320; n=108)	Post-Purchase

Note: ANCOVA= Analysis of Covariance; EFA=Exploratory Factor Analysis; ANOVA= Analysis of Variance; SEM=Structural Equation Modelling

2.2 Drivers of Returns and Potential Solutions

Whether to buy online instead of in a store also depends on the disadvantages of the mail-order business (Hong and Pavlou 2014; Shulman et al. 2011), which are common knowledge. If someone orders online, they have already familiarized/acquainted with it in advance (Ülkü and Gürler 2018) and might even take advantage of vendors' lenient return policy (Pei and Paswan 2018).

Reasons for product returns are multi-faceted and very individualistic in the field of fashion in particular, but not all cases of product returns can be prevented. Based on a recent investigation with n=1,024 respondents (ibi research 2017), the drivers of product returns reveal to be product did not fit (62%), consumers did not like the product (39%), the product was defective or delivered in damaged conditions (30%), the product was not as described (30%). Followed by multiple variants were ordered (20%), wrong delivery (7%), delivery took too long (5%), the product was found cheaper in another shop (2%) or other reasons (2%), which is comparable to prior investigations (Lee 2015; Gelbrich et al. 2017). These drivers identified (Table 2) could be condensed into an information gap related return reasons and those caused by online shopping operators' service. However, in some cases, customers return articles due to consumer behavior related causes, such as impulsive purchases (Ülkü and Gürler 2018), so-called „showrooming“ behavior (Bell et al. 2018), or not fulfilled returns, which might result in dissatisfaction. Besides this categorization, ordered products were intended to be worn as a set and cannot be delivered or combined fall in-between consumer behavior and fulfillment/service reasons.

Table 2 Three main categories for product return reasons

Information Gap	*Fulfillment/Service*	*Consumer Behavior*
Insufficient visualization (Gallino and Moreno 2018)	*Delayed delivery*	Impulsive purchases
Misleading product description	*Wrong delivery*	Planned product return (showrooming)
Price-performance ratio/quality	*Defective/damaged product*	Not fulfilled returns result in dissatisfaction
Multiple variants in different sizes	*Products ordered to wear it as a set cannot be delivered/combined*	
Uncertainty about outfit combinations (Shulman et al. 2015)		

Note: Reasons that can be influenced by companies

Based on these reasons, we collected potential measures for the three stages (Fig. 1) in Table 3. These are substantiated based on literature and illustrated by practical examples, representing the measures used in our investigation. (As we intend to explore customers' viewpoint for technological-advanced and state-of-the-art measures, some of the items applied have not yet been investigated in established journals. Apart from that, we focused on rewarding rather than sanctioning measures. Most online retailers try to avoid the adverse effects of a less lenient return policy, such as ordering elsewhere (Gelbrich et al. 2017). This avoidance is in line with the operant conditioning theory (Skinner 1965), where the intended customer behavior (from a retailer's perspective) is assumed to occur more frequently when this behavior is linked to a pleasant consequence („positive reinforcement"). This theory has been applied in many areas of consumer behavior research (Wells 2014), such as online product selections (Perotti et al. 2003), corporate behavior (Vella and Foxall 2013), the effectiveness of TV commercials (Nathan and Wallace 1971), and even in the context of product returns (Gelbrich et al. 2017).

Hence, we also incorporate recent measures yet only discussed in blogs and contained in market research reports.) Additionally, we assume an influence on the categorization by the market standard (MS) and the degree of user integration (DoIU).

Unfortunately, there are no relevant publications on the market standard or the diffusion of the measures. We have decided to rate the market standard in three dimensions: 1=very common, 2=partly common, 3=very rare/not (yet) existing. For this purpose, we went among others through the top 20 German fashion online stores in 2018 to be consistent with the customers surveyed, who also live in Germany. In our assessment, only three measures can be considered very common (MS=1): „360° view", whereby we have also included an all-around photo series. Personalized newsletters were also offered by all providers, although not every newsletter contained a personalized element. We categorized measures as partially common (MS=2) if they were not shown consistently or only for selected articles in the top 10 providers, which was the case with „catwalk videos" or „information model size." For measures that were hardly shown (MS=3), we had to search outside the top 20. In general, it can be said that measures from the post-purchase phase are hardly widespread (and challenging to investigate from an outside position), probably also because the return behavior after the purchase should no longer be a topic of discussion. Bonus points for retained goods are an exception.

For the Degree of User Interaction (DoUI), we have also decided on three categories: (○=no user interaction needed, ◑=user interaction needed, but can still be used without, ●=can only be accomplished by integrating the user. Many of the measures do not rely on active user participation. We have assigned „No user interaction needed“ if, on the one hand, no direct interaction is required, and the result does not change with even partial user interaction (e.g., „Size advice - figure types“). This category is followed by measures that deliver results even without user input, but user interaction leads to improved results (e.g., „Favorite article for comparison“). The highest requirements are measures that can only be achieved together with the user. These include virtual try-on or self-measure.

With our study, we would like to give for the first time a holistic view of the returns management in fashion online retail. In this context, we examine measures already established on the market and innovative approaches that have only become possible with the last few years' technological developments. We aim to identify particularly desirable measures and obtain general conclusions about returns management through generalization. In the literature, only partial aspects of returns management are examined. Our contribution consists of comparing measures from all three purchase phases and incorporating currently available technological measures.

Table 3 Investigated measures derived from practical applications and current literature

	Measures applied	*Practical example(s)*	*Literature*	*MS*	*DoUI*
Pre-Purchase	Personalized newsletter *Personalized newsletter only showing items which are chosen from a vendor's algorithm to fit best*	Stitch Fix, Breuninger	Deges 2017; Gehrckens and Boersma 2013; Kreutzer 2018	1	○
	Online shop as social platform *Online shop as a social platform: shop with a friend, who can help to choose the right items*	Geox, Nike	Haug and Küper 2010	3	◑
	360° view *Display items with a virtual 360º view of the article*	New Balance	Deges 2017; Heinemann 2019; Melchior 2018a	1	○
	Outfit recommendations from influencers *Revealing a detailed outfit recommendation from influencers*	About You, Zalando, Otto, Orsay	Heinemann 2019; Melchior 2018b	3	○
	Photos from social networks *Integration of photos taken by customers or followers from social networks*	Target, Adidas, Nike, George (Asda)	Haug 2013; Heinemann 2019; Melchior 2018b	3	○
	Presentation via (catwalk) video *Presentation of the article using a video/catwalk video*	Asos, Zalando	Heinemann 2019; Melchior 2018a	2	○
Pre- or Purchase	Commenting on reviews *Commenting on existing customer reviews to get better impressions*	Walmart, Otto, Zalando, Marks and Spencer	Deges 2017; Heinemann 2019	2	◑
	Curated Shopping *A styling expert picks the items based on preferences or shopping history*	Stitch Fix, Outfittery, Kisura, Zalando	Holland and Bolz 2017	3	◑
	Reward of article ratings *Get rewarded for writing reviews on purchased items*	Amazon, Shopify	Burton and Khammash 2010	3	●
Purchase	Virtual fitting of articles *Using virtual reality to see how the item could look on oneself*	Mister Spex, Otto	Deges 2017; Walsh and Möhring 2015	3	●
	One model wears all sizes *The same model wears all sizes for comparison*	C'est normal	Deges 2017; Heinemann 2019	3	○
	Find out individual size *Find out one's size via self-measurement using an interactive online tool*	Kohl's, Mytheresa, Quiz	Deges 2017; Heinemann 2019	2	●
	Favorite article for comparison *Compare the size of a new item with the size of a favorite item*	Next (Bra Size), Thirdlove, Quiz	Deges 2017; Heinemann 2019	3	◑
	Size recommendation - previous purchases *Size recommendation from the vendor based on previous purchases and returns*	Zalando	Deges 2017; Heinemann 2019	3	○
	Size advice - figure types *Size advice based on which figure type is most similar to oneself*	About You, Sizeable, The Yes	Deges 2017; Heinemann 2019	2	○
	Information model size *Information on the size of the model who wears the item*	Asos, Nelly, Target, Pretty Little Thing, River Island	Deges 2017; Heinemann 2019	2	○
	Self-measurement via webcam *Via webcam: using a special suit or object for comparison while standing in front of a camera*	Upcload	Walsh and Möhring 2015	3	●
	Assisted shopping *Real-time guidance from the vendor to help to choose sizes or colors*	John Lewis, BAUR	Heinemann 2019	3	◑
Post-Purchase	Discount on current order *Keep the item and receive a discount on the current order*	Amazon Prime Wardrobe	IFH Köln and AZ Direct 2016	3	○
	Discount on next order *Keep the item and receive a discount on the next order*	Bonprix	Deges 2017; IFH Köln and AZ Direct 2016	3	○
	Return impact information *For the considered or taken return*	Mirapodo, Zalando	Deges 2017	3	○
	Bonus points for purchases *Receive bonus points for kept purchases*	Adidas	IFH Köln and AZ Direct 2016	2	○
	Bonus points for non-return orders *Receive bonus points for intended, but not returned, orders*	No common example known	IFH Köln and AZ Direct 2016	3	○
	Support for social projects *The vendors support social projects in exchange for a not carried out return*	BAUR	Buxel and Weidlich 2010	3	○
	Display of the return behavior *Displaying the own return behavior to customers*	Zalando Lounge	Decker 2018; Deges 2017	3	○
	Waiver of shipping costs *For no return*	Amazon	IFH Köln and AZ Direct 2016	3	○

Note: MS=Market Standard (1=very common, 2=partly common, 3=very rare/not (yet) existing); DoUI=Degree of User Interaction (○=no user interaction needed, ◑=user interaction needed, but can still be used without, ●=can only be accomplished by integrating the user)

2.3 The relationship between expectation fulfillment and satisfaction

The effect of the individual (service) attributes on customer satisfaction is not always linear (Kano et al. 1984; Shahrestani et al. 2020; Shahin et al. 2017) and changes over time (Kano 2001). We would like to provide an informative insight into the different measures' expected effects with our work from a customers' perspective. For this purpose, various approaches are available (Mikulić and Prebežac 2011). Kano's model (Kano et al. 1984) is a proper way to capture effects in the design stage of a product or service and later to derive managerial strategies. Therefore, we will use Kano's method for our investigations.

In the literature, Kano's model is not precisely distinguished. Therefore, the following shall apply to this work: Kano's model (Matzler 2003, p. 341) is the term used to describe the work of Kano (1968, 1987, 1995, 2001); Kano et al. (1984), which is often referred to as „Theory of Attractive Quality." Kano describes that the relationship between expectation fulfillment and customer satisfaction is not always linear. It should serve us as a theoretical concept for the multi-factor structure in customer satisfaction. Kano's model is in contrast to the Kano method. It describes a procedure that can be used for categorization. We will come to this in the next chapter.

According to Kano et al. (1984) and Kano (2001), there are four primary patterns for cause-effect relationships: Must-be, One-dimensional, Attractive, and Indifferent (Fig. 2) supplemented by two relatively rare and theoretical cases (Matzler et al. 1996; Mikulić and Prebežac 2011; Nilsson-Witell and Fundin 2005) from which strategies for companies are derived.

- Must-be (M) items are items for which poor performance has the strongest effect on customer satisfaction in its entirety; meeting or even exceeding expectations cannot increase overall customer satisfaction. Strategy: Securing primary performance via, e.g., service level agreements, following no further investment.

- One-dimensional (O) items are items with a direct influence on overall satisfaction for good and bad fulfillment. Strategy: Ensure primary performance and increase it further.

- Attractive (A) items are usually not expected by the customer and, if present, lead to an improvement in satisfaction. Absence or poor performance does not affect overall satisfaction. Strategy: If the necessary services (M and O) are acceptable, they can differentiate in the market.

- Indifferent (I) items have no neither positive nor negative influence on customer satisfaction. Strategy: Avoid Investments.

- Reverse (R) items lead to a decline in satisfaction when present, but their absence leads to an improvement. Strategy: Not only should any investment be avoided, but consideration should also consider whether a consciously externally communicated demarcation can be perceived as Attractive.

- Questionable (Q) items are forfeited if none of the five correlations listed could be determined; subsequently, no general strategy applies.

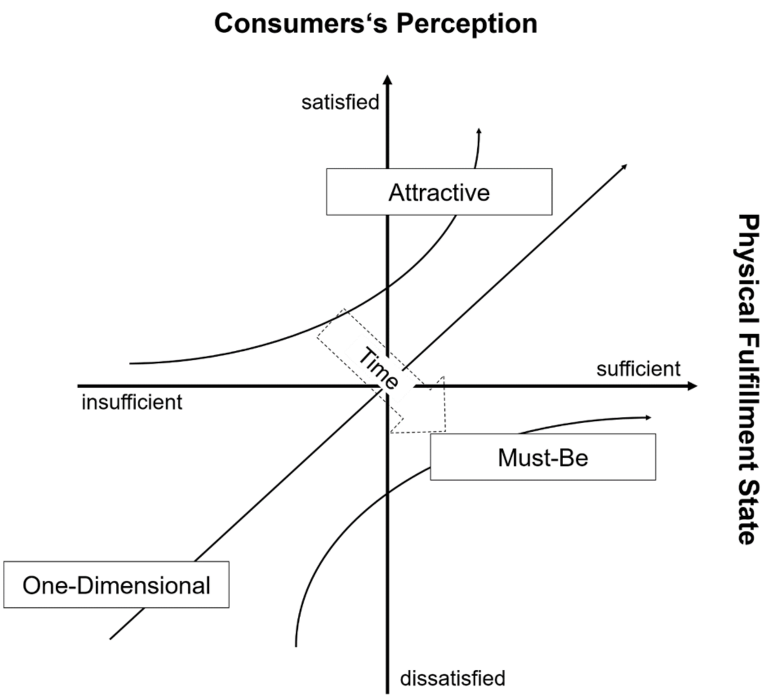


Fig. 2 Kano’s model (Kano et al. 1984) with the illustration of its life cycle (Kano 2001)

Kano (2001) also addresses a dynamic change over time. In his view, a successful quality element of a product or service passes through this sequence or lifecycle: Indifferent -> Attractive -> One-dimensional -> Must-be. Nevertheless, also, other sequences can be found. Nilsson-Witell and Fundin (2005) have shown that when an adoption level is taken into account, the answers can be categorized differently. For example, one service studied was during introduction I and later A. Respondents referred to as early adopters already categorized this service as O or even M instead of A. Further studies have also shown in time series comparisons that the attributes change dynamically over time. Hölzing (2008) examined services for people with diabetes at an interval of 6 months (2005, 2006), Raharjo et al. (2010) for characteristics of notebooks with ten data points at a 2-month rhythm, Löfgren et al. (2011) quality attributes of commodity packaging (2003, 2009) and Stöcker and Nasseri (2020) touchpoint satisfaction of customers of an eCommerce retailer (2011, 2013).

2.4 Hypothetical Framework

We will now derive our hypotheses about the measures presented in Table 3. These can be divided into two main groups: Characteristics that concern the measure itself (time effects, type of incentive, and user interaction) and variations in customer attributes (age and order frequency).

Measure-related Hypotheses

As illustrated in Figure 2, new service features will first be evaluated as Attractive and perceived as One-dimensional with a linear increase regarding satisfaction and, finally, Must-be dimension (Kano 2001). However, online shopping operators need to consider the features' adoption rates in terms of time and incorporate the potential competitive advantage by being the first to offer specific measures. According to the law of differentiation dynamics, the prospective competitive advantage will diminish if competitors are already providing such features (Rudolph and Becker 2003). While some measures for (those with high levels of the market standard, see Table 3) are already widely implemented in online shops, others are still in an evolving stage with only a few practical examples existing. Therefore, we assume:

H_1: Measures with a low level of market standard are more frequently categorized as I and A instead of O and M than those representing a high level of market standard.

Within the post-purchase stage measures, those related to compensation or rewards might be perceived as positive, as they will trigger reinforcement according to the operant conditioning theory (Skinner 1965). Hence, they result in higher consumer satisfaction than other sanctioning consumers (such as displaying return behavior or return impact information). So, measures that reward consumers seem to pay off more than sanctioning them (IFH Köln and AZ Direct 2016; Gelbrich et al. 2017). Although we have excluded re-purchase behavior from this study, it should be evident that, especially in a buyer's market with many suppliers, respectively, a negative sanction leads to customers' churn. Therefore the implementation of these measures must follow with great sensitivity. Accordingly, we hypothesize that monetary measures („Discount on next order“, „Discount on current order“, „Bonus points“, and „Waiver of shipping costs“) will result in a higher increase in customer satisfaction, especially in contrast to measures sanctioning customers (display of the return behavior; return impact information).

H_{2a}: Monetary measures have the strongest positive influence on customer satisfaction (CS+)

H_{2b}: Monetary measures have the strongest positive influence on customer satisfaction (CS+) compared to all measures sanctioning customers.

In the measures described for the avoidance of returns, some can only succeed with the user’s active collaboration (see Table 3, column DoUI). Here, such measures’ success depends on customers’ willingness to engage in these measures (Lai et al. 2014). Since the fashion market is a buyer’s market, we assume that these are less appealing.

H_3: Measures that require the direct engagement of users are less frequently categorized as A, O, and M compared to other measures.

Customer-related Hypotheses

Technical innovations undergo the life cycle, according to Kano (2001). We assume that newer measures, which cannot be considered the market standard, are preferred more by younger than older customers. In this study, the measures “virtual fitting of articles,” “self-measurement via webcam,“ „curated shopping,“ „assisted shopping,“ and „online shop as a social platform.“ This effect is exceptionally actual for the millennial generation, who possess excellent technological skills (Ladhari et al. 2019).

H_4: Innovative measures positively influence customer satisfaction (CS+) by younger customers.

As purchase frequency can be a moderator concerning the categorization of the measures (Gelbrich et al. 2017), we assume purchase frequency also moderates return averting measures. Meanwhile, customers with high shopping frequency are used to handle product returns as part of the shopping online (Ülkü and Gürler 2018). Hence, they easily hazard the related consequences and sometimes even take advantage of a merchant’s lenient return policy (Pei and Paswan 2018). Therefore, we expect:

H_5: Customers with a high purchase frequency tend to categorize the queried measures in the three purchase stages as A and O.

This study aims to describe the return process as a holistic problem, gain insights into current and innovative measures in return management, and map their effect on customer satisfaction. With the hypotheses that have been formulated, we try to determine structures within the individual measures, which can later be generalized. Using the segmented Kano perspective, we also investigate whether the answers already show signs of a life cycle for the measures. For this purpose, we use a structured questionnaire, which also includes questions on buying and return behavior. Thus we hope to isolate additional descriptive characteristics that can profile our findings even more precisely.

3. Research Design

To collect the customers' voice, we decided to use an online questionnaire sent to all customers. In this questionnaire, we asked one functional and one dysfunctional question for each measure; these questions were combined in the evaluation.

3.1 Survey and Descriptive Statistics

While many studies in return management literature applying self-report surveys suffer from acquiring an adequate sample and use student samples instead (Pei and Paswan 2018; Oghazi et al. 2018; Gelbrich et al. 2017), we want to overcome this issue by enquiring about actual customers from a leading online shop in Germany. This approach provides multiple advantages. First, in contrast to students, actual customers exhibit higher income levels and, therefore, higher purchase power (Iyer and Eastman 2006), leading to more realistic responses regarding price issues. Second, even though elderly consumers represent a fast-growing segment in e-commerce, literature on consumers' online shopping behavior older than 50 years is still very scant (Lian and Yen 2014) and should be examined. Third, while students' answers for hypothetical scenarios might not reveal their actual shopping and return behavior, we expose our questions within the determined online shop's framework addressing this specific online shop's customers, which results in more realistic findings. For our research, we had the opportunity to contact customers of BAUR Versand (baur.de), a top 10 online retailer for fashion in Germany (EHI Retail Institute 2019). BAUR's product range focuses on fashion, shoes, and home, including furniture, and concentrates primarily on female customers between 40 and 55. BAUR relies primarily on well-known brands, and around 90% of the business volume is handled via the online shop.

The invitation to participate in the survey was sent by e-mail on December 14, 2018, to all BAUR customers providing the opportunity to answer the questionnaire until January 18, 2019. A raffle of 15 shopping vouchers worth EUR 20 for the BAUR online shop was announced among all participants in the invitation. To not overstrain respondents with the very time-consuming questionnaire, three surveys with different clusters of measures were used, randomly assigned to the e-mail addresses. All questionnaires had the same structure and differed only in the return measures exposed using the Kano methodology (survey 1: 10 measures, survey 2: 11, and survey 3: 9, see Table 4). In the beginning, the aim and purpose of the study were explained. It was pointing out that this was a joint research project of BAUR and students of a near-by University. The initial questions on the current ordering and returns behavior were subsequently asked (no further validation via the customer database). The self-assessment of the respondents serves, on the one hand, as an icebreaker question; on the other hand, the respondent should reflect his or her return behavior at this point and thus form the basis for further answers. They were following these questions by the evaluation of one of the three clusters of measures. The Kano questioning technique, unusual for many respondents, was first introduced using an example. Finally, presenting the questions on socio-demographics and space for comments and the opportunity to participate in the raffle. Pretests helped to test the comprehensibility of the questions and the structure during the questionnaire development.

For describing the respondents in more detail in the following analysis, other characteristics were queried: a) On the one hand, the current ordering behavior, whereby the ordering frequency, the average expenditure on fashion, for who is mainly purchased, where individual product ranges are purchased preferentially (online or offline), whether these purchases are mainly spontaneous or planned and how fashion buying online is generally perceived. Afterward, b) the current return behavior: how often a return took place, the reasons for it, how complex a return is perceived, and whether the return behavior differs between orders from different shops. Finally, in addition to age and gender, c) the residence place's size was also surveyed to detect any differences in an assumed imbalance of supply.

A total of 8,393 complete questionnaires were evaluated (survey 1: n=2,789 completion rate 68%, survey 2 n=2,855 completion rate 70%; survey 3 n=2,749 completion rate 64%). The three samples are structured as follows about their purchasing behavior and socio-demographic characteristics (for full detail, see appendix 1).

The majority of customers order fashion online between once a month (30.6%) and once a quarter (32.6%). At the same time, 85.8% of those surveyed stated that they spend up to 150 EUR. Regarding their shopping behavior, 22.8% describe themselves as planning, 36.0% as partly/partially planning, and 41.1% as browsing and discovering. Only very few of the respondents (8.0%) answer that they avoid online shopping when possible. Besides, the vast majority (62.0%) answered that they love buying fashion online. Concerning the number of returns, customers state that they have also returned in 32.6% of (all) orders transacted.

Regarding the reasons for a fashion return, 87.1% of the respondents answered with „Item does not fit," 45.9% with „I do not like this item," 41.6% ordered several sizes to choose from, 21.0% „not as described" and 4.2% bought more to choose from at home due to a promotional measure. In the upper third of the scale, 55.4% rate a fashion return's effort as "not elaborate." Here too, bias is to be assumed from the survey of active online shoppers. When asked whether the return behavior differs among different providers, 52.4% explicitly answered "no," while 77.7% of the answers tended to be "no" in the first half of the 6-point Likert scale.

Among the respondents, 79.8% are female, 29.1% are between 29 and 44 years old, 32.9% are between 45 and 54 years old, and 38.1% are older than 55, slightly above average in small and medium-sized cities (5 to 100 thousand inhabitants) and firmly below average in cities with millions of inhabitants.

3.2 Categorization of the Measures

In order to determine the cause-effect relationships for each item in Table 3, two questions were asked: the functional ("imagine that ... has [item] ...") (Kano et al. 1984; Matzler et al. 1996; Mikulić and Prebežac 2011) and dysfunctional (“imagine that ... has not [item] ...”) questions (Berger et al. 1993; Matzler et al. 1996; Nilsson-Witell and Fundin 2005). The answer is given on an ordinal scale with a middle option (“(1) I like it that way”, “(2) It must be that way”, “(3) I am neutral”, “(4) I can live with it that way”, „(5) I dislike it that way”). Nature can be determined via the Kano table from combining the two answers to the two questions (Table 4).

Table 4 Kano table: Categories derived from answers to the (dys-)functional questions (Kano et al. 1984)

		Dysfunctional Question				
		(1) Like	*(2) Must be*	*(3) Neutral*	*(4) Live with*	*(5) Dislike*
Functional Question	(1) Like	Q	A	A	A	O
	(2) Must be	R	I	I	I	M
	(3) Neutral	R	I	I	I	M
	(4) Live with	R	I	I	I	M
	(5) Dislike	R	R	R	R	Q

Note: A=Attractive; I=Indifferent; M=Must-Be; O=One-Dimensional; Q=Questionable; R=Reverse.

The characteristic is now derived from the Kano table. If all survey results of one question are plotted as value pairs in a coordinate system, the characteristic Kano curves are obtained (see Fig. 2).

In the literature, however, another approach is also common. In this case, no curves are shown; the character is reflected here in the position of the individual measures in the respective quadrants. This approach presents the positive and negative impact on customer satisfaction as two coefficients (Berger et al. 1993; Shahin et al. 2013; Shahin and Zairi 2009). Assuming a positive factor on the customer satisfaction (CS^+) for answers falling into classes A and O, a negative factor (CS^-) for O and M. Answer combinations from classes Q and R are not considered. The results are then displayed graphically in a coordinate system representing the two axes CS^+ and CS^- orthogonally. The two coefficients tell us how often an item has been categorized into the mentioned groups. For CS+, the mentions are counted positively influencing satisfaction when the expectation is fulfilled positively (A and O) and for CS- those where a negative fulfillment negatively influences satisfaction (O and M). A high value consequently shows a high correlation with customer satisfaction. Since Kano's categorization can only be interpreted by translating the terms into one of the corresponding curves, the coefficients dispense with this step. The positive as well as the negative effect can be read off directly.

$$CS^{+} = \frac{\#A + \#0}{\#A + \#O \#M + \#I}$$

$$CS^{-} = \frac{\#O + \#M}{\#A + \#O \#M + \#I}$$

#A, #I, #M and #O represent the response frequencies of the categories or the number of responses categorized as A, I, M, or O. The indices are between 0 and 1 and -1, respectively, and reflect the impact on satisfaction. From the location of the points, their categorization is again apparent. The coordinate system is, therefore, divided into four quadrants:

$$Attractive, if \begin{cases} 0.5 \leq CS^{+} \leq \ 1 \quad and \\ 0 \geq CS^{-} > -0.5 \end{cases}$$

$$Indifferent, if \begin{cases} 0 \leq CS^{+} < \ 0.5 \ and \\ 0 \geq CS^{-} > -0.5 \end{cases}$$

$$Must\text{-}be, if \begin{cases} 0 \leq CS^{+} < 0.5 \, and \\ -0.5 \geq CS^{-} \geq -1 \end{cases}$$

$$One\text{-}dimensional, if \begin{cases} 0.5 \leq CS^{+} \leq 1 \;\; and \\ -0.5 \geq CS^{-} \geq -1 \end{cases}$$

If points are close to the origin, no influence can be proven at all. If a point lies precisely in the middle, a positive and, at the same time, the negative influence is detectable in 50% of the respondents. The coordinate system position can now be determined (see values in Table 5) using the formulas described in Chapter 3 or read directly from Table 5. For example, the item „Discount on next order“ has a CS^{+} of 0.8 and a CS^{-} of -0.23. It can therefore be found in quadrant „A“ in the upper left corner.

Also, the Total Strength (TS) represents the number of mentions categorized as A, M, or O compared to all mentions. Items with a high TS also have a strong influence (positive or negative) on total customer satisfaction. The TS serves to prioritize the individual items concerning their effect on customer satisfaction. Improvements to items with a high TS should have a high impact on the change in customer satisfaction.

$$Total\ Strength = \frac{\#A + \#M + \#O}{\#A + \#O + \#M + \#I + \#Q + \#R}$$

Recently, other papers apply an additional variant to the method described above. The Segmented Kano perspective descends one level deeper by searching for clusters within the answers. The new approach makes it possible to identify different customer segments with different expectations, otherwise not visible in the aggregated form. For this purpose, the answers enter the functional and dysfunctional question as a metric feature into the cluster analysis (Baier and Rese 2018) or using one-mode non-metric cluster analysis concerning the derived categories (Rese et al. 2019). The number of clusters is then determined iteratively under the observation of the Bayesian Information Criteria (BIC) concerning the likeness functions.

4. Findings

Table 5 displays the overall assessment of the measures based on Kano's model, indicating category frequencies, the total strength (TS), the customer satisfaction index CS^+, and the customer dissatisfaction index CS^-.

The surveyed measures' results are evaluated solely as A (not expected, but if there is a positive influence on overall satisfaction) or I (no evident influence on overall satisfaction). Regarding the category Attractive, the measures "360° view", „Discount on current order," „Discount on next order," „Bonus points for purchases," and „Waiver of shipping costs" stand out. More than 50% of the respondents rated these measures as Attractive, suggesting that these measures could substantially contribute to customer satisfaction. In contrast, the measures' Curated shopping", „Assisted shopping," „Commenting on reviews," „Online shop as a social platform," „Photos from social networks," „Outfit recommendations from influencers" and „Return impact information" are also Indifferent to more than 50% of the mentions. Here, no influence on customer satisfaction is expected when implementing the measures. None of the measures can be described as M. The only measure that could be considered One-dimensional is „Waiver of shipping costs." Here, the closest mentions are for A 1,373 and O 900. Measures are categorized as Reverse if their interrelationship towards satisfaction is precisely the opposite. An exemplary implementation has a negative effect and a bad one, a positive effect on satisfaction. Here the measure webcam size is particularly striking. Categorized as Indifferent, with 1,125 mentions, but with 996 mentions, it is also *very close to Reverse*.

Table 5 Overall assessment of possible measures

	Measure	*Overall Category Frequencies (n=2,792)*						*TS*	*CS^+*	*CS^-*
		#A	#I	#M	#O	#Q	#R			
Survey 1	360° view	**1,481**	605	91	532	67	16	56%	0.74	-0.23
	Presentation via catwalk video	1,134	**1,221**	14	222	44	157	42%	0.52	-0.09
	Virtual fitting of articles	1,030	**1,183**	15	212	48	304	38%	0.51	-0.09
	Curated shopping	645	**1,516**	29	151	42	409	25%	0.34	-0.08
	Assisted shopping	579	**1,486**	41	131	47	508	23%	0.32	-0.08
	Commenting on reviews	605	**1,633**	61	196	32	265	25%	0.32	-0.10
	Online shop as social plat-form	296	**1,605**	23	69	53	746	12%	0.18	-0.05
	Personalized newsletter	659	**1,388**	54	170	53	468	26%	0.37	-0.10
	Photos from social net-works	699	**1,334**	33	122	58	546	26%	0.38	-0.07
	Outfit recommendations from influencers	424	**1,452**	41	113	61	701	18%	0.26	-0.08
		Overall Category Frequencies (n=2,855)								
Survey 2	Presentation via video	**1,351**	992	34	353	58	67	49%	0.62	-0.14
	360° view	**1,473**	462	165	676	60	19	60%	0.77	-0.30
	Find out individual size	**1,330**	701	122	576	61	65	52%	0.70	-0.26
	Information model size	**1,147**	852	181	541	66	68	48%	0.62	-0.27
	One model wears all sizes	**1,026**	930	50	403	81	365	38%	0.59	-0.19
	Favorite article for compar-ison	1,134	**1,143**	58	340	63	117	42%	0.55	-0.15
	Size advice - figure types	**1,258**	722	128	614	61	72	50%	0.69	-0.27
	Photos from social net-works	562	**1,392**	31	111	71	688	21%	0.32	-0.07
	Virtual fitting of articles	881	**1,203**	32	263	57	419	32%	0.48	-0.12
	Size recommendation - previous purchases	**1,290**	869	70	336	77	213	48%	0.63	-0.16
	Self-measurement via webcam	494	**1,125**	14	114	112	996	18%	0.35	-0.07
		Overall Category Frequencies (n=2,749)								
Survey 3	Discount on current order	**1,459**	522	65	550	60	93	58%	0.77	-0.24
	Discount on next order	**1,540**	477	38	568	50	76	59%	0.80	-0.23
	Return impact information	364	**1,517**	147	235	60	426	22%	0.26	-0.17
	Bonus points for purchases	**1,484**	542	43	584	47	49	56%	0.78	-0.24
	Bonus points for non-re-turn orders	**1,299**	771	50	474	52	103	51%	0.68	-0.20
	Reward of article ratings	**1,222**	805	48	537	44	93	47%	0.67	-0.22
	Support for social projects	**1,096**	802	59	643	52	97	43%	0.67	-0.27
	Display of the return be-havior	809	**1,336**	27	260	39	278	31%	0.44	-0.12
	Waiver of shipping costs	**1,373**	265	110	900	62	39	56%	0.86	-0.38

Note: The Most Frequent Category is Marked in Bold. A=Attractive; I=Indifferent; M=Must-Be; O=One-Dimensional; Q=Questionable; R=Reverse; TS=Total Strength; CS^+=Customer Satisfaction Index; CS^-= Customer Dissatisfaction Index.

Considering the categorization of the measures and the market standard's degree (MS) that we have assumed (H_1), no consistent picture emerges (Table 6). The measures investigated are distributed equally between I and A, depending on the market standard's level. Interestingly, even measures that have been in the market for a long time and established are only categorized as A.

Table 6 Categorization of measures concerning the assumed market standard

	Very common	*Partly common*	*Very rare/not (yet) existing*
I	Personalized newsletter	Presentation via cat-walk/item video, commenting on reviews, photos from social networks	Virtual fitting of articles, photos from social networks, curated shopping, assisted shopping, online shop as a social platform, outfit recommendations from influencers, favorite article for comparison, self-measurement via webcam, *return impact information, display of the return behavior*
	1 item	4 items	10 items
A	360° view, size advice – figure types,	find out individual size, Information model size, *bonus points for purchase*	One model wears all sizes, size recommendation – previous purchases, reward for article ratings, d*iscount on current order*, *discount on next order, bonus points for purchases, support for social projects, Waiver of shipping costs*
	2 items	3 items	8 items
O			
M			
R			
Sum	3 items	7 items	18 items

Note: measures in italics refer to return averting.

Nor can a uniform picture be formed for the degree of required user interaction (H_3, Table 7). Suppose we additionally exclude return averting measures, which do not prevent returns in the narrower sense but negotiate the conditions under which the customer would refrain from returning. In that case, a similar distribution between I and A can be observed here as well. Consequently, this hypothesis must also be rejected.

Table 7 Categorization of measures concerning the assumed degree of user interaction

	No user interaction needed	*User interaction needed, but can still be used without*	*Can only be accomplished by integrating the user*
I	Presentation via catwalk/item video, personalized newsletter, photos from social networks, outfit recommendations from influencers, *return impact information*, *display of the return behavior*	Curated shopping, assisted shopping, commenting on reviews, online shop as a social platform, favorite article for comparison	Virtual fitting of articles, self-measurement via webcam
	7 items	5 items	2 items
A	360° view, information model size, one model wears all sizes, size advice - figure type, photos from social networks, size recommendation – previous purchases, *discount on current order*, *discount on next order*, *bonus points for purchase*, *bonus points for non-return orders*, *support for social projects*, *waiver of shipping costs*		Find out individual size, reward of article ratings
	12 items		2 items
O			
M			
R			
Sum	19 items	5 items	4 items

Note: measures in italics refer to return averting.

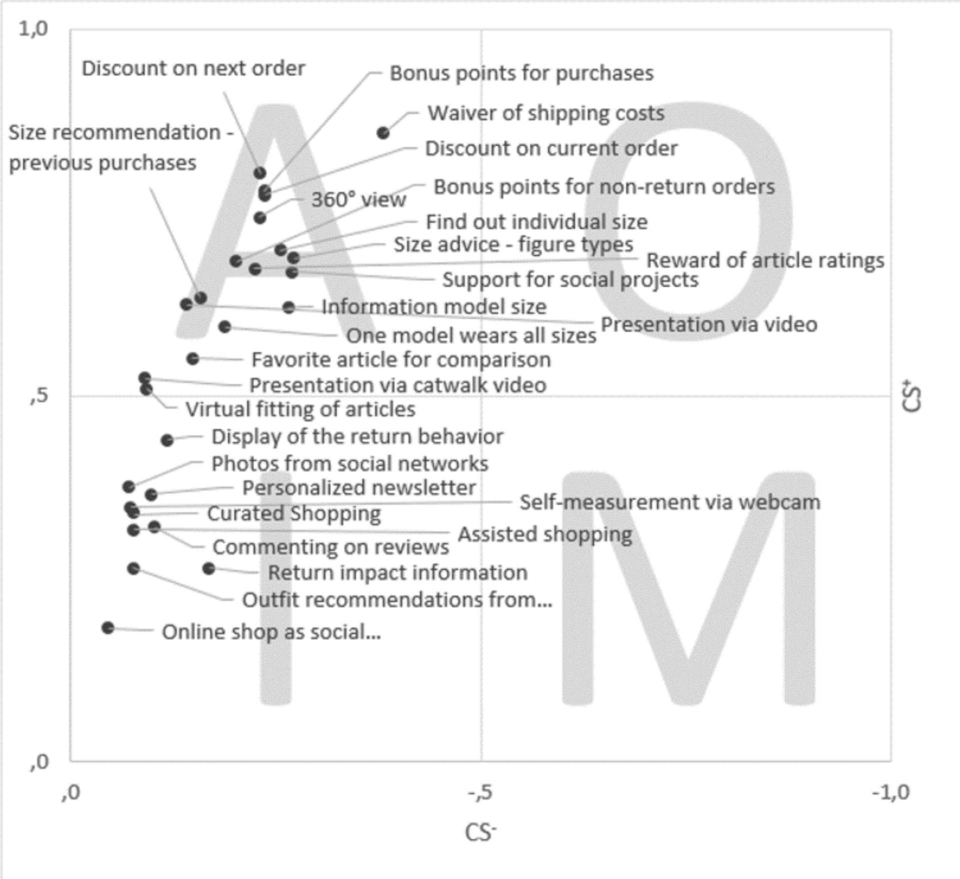


Fig. 3 Depiction of the overall assessment of possible measures (n= 8,396)

Figure 3 shows all measures based on their impact coefficients CS^+ and CS^-. Here, too, the same picture emerges. All measures presented are located in the two quadrants I and A. Furthermore, an exciting pattern becomes visible: most measures with a monetary reward show the most considerable positive impact („Waiver of shipping costs," „Discount on next order," „Discount on current order," „Bonus points for purchases"). As we stated in $H_{2a,}$ monetary measures have an above-average positive influence on customer satisfaction than the other examined measures and measures sanctioning the customer (H_{2b}). These are followed by measures that primarily result in an improvement of the presentation *by the vendor* („360° view ", „Find out individual size," „Size advice - figure types," „Size recommendation - previous purchases," „Presentation via video," „One model wears all sizes," „Size recommendation - previous purchases,"). A third block can be seen in the Indifferent quadrant. This cluster contains measures that either include external content in the shop („Online shop as a social platform," „Outfit recommendations from influencers," „Commenting on reviews," „Photos from social networks"), require the customer to be involved („Assisted shopping," „Self-measurement via webcam," „Photos from social networks") or reflect their return behavior. Since our study examines all stages of the purchasing process, we can clearly show this hierarchy of measures at this point.

We then have examined all the proposed measures regarding dependencies (linear or segmental) in the answering behavior to their age (H_4) and shopping frequencies (H_5) with no significant differences found.

We apply the before-mentioned segmented Kano perspective to reveal more meaningful insights based on the overall results and derive more clear implications. We have used the well-known two-step clustering approach, according to Chiu et al. (2001). In each record, each measure is categorized according to Kano's evaluation table. For the resulting nominal data matrix, independent multinomial distribution of the categories over the clusters' attributes is assumed. The optimal number of clusters is now determined iteratively, taking into account the BIC. In this case, three clusters have proven to be ideal.

From the results in Table 8 and Figs. 4 to 6, initial findings can already be deduced. A closer look reveals that the three surveys' segments follow similar patterns: each segment can be assigned to a quadrant. We named segments primarily in I „Indifferents", those in A „Enthusiastics" and in O „Demanders." A segment in M we would call „Taken-for-granteds". Nilsson-Witell and Fundin (2005) have found a similar starting position in their study for „e-service. " When introduced perceived as Indifferent, it became later Attractive. They investigated the Attractive with a technology adoption level. They found segments in O and M, which they refer to as „early adopters, "a term also used to the diffusion of innovations theory Rogers (1962).

Table 8 Segment-specific category frequencies for each of the three surveys

	Measure	*Segment-Specific Category Frequencies („Indifferents“: n=1,498 / „Enthusiastics“: n=1,005 / „Demanders“: n=289)*					
		#A	#I	#M	#O	#Q	#R
Survey 1	360° view***	29 / 241 / 25	**871** / **634** / **108**	5 / 5 / 19	3 / 3 / 71	9 / 5 / 35	581 / 117 / 31
	Presentation via catwalk video***	53 / 346 / 26	**860** / **484** / **100**	12 / 10 / 13	8 / 34 / 63	22 / 4 / 39	543 / 127 / 48
	Virtual fitting of articles***	106 / 410 / 25	**1,012** / **475** / 77	5 / 29 / 21	11 / 58 / **80**	10 / 1 / 38	354 / 32 / 48
	Curated shopping***	101 / 450 / 28	**953** / **451** / 82	16 / 7 / 18	17 / 28 / **86**	11 / 3 / 33	400 / 66 / 42
	Assisted shopping***	150 / **472** / 25	**896** / 389 / 90	9 / 23 / 20	7 / 27 / **105**	10 / 8 / 28	426 / 86 / 21
	Commenting on reviews***	160 / **514** / 35	**1,064** / 414 / 94	20 / 4 / 12	39 / 35 / **99**	5 / 4 / 26	210 / 34 / 23
	Online shop as social platform***	169 / **524** / 18	**921** / 361 / 78	18 / 3 / 13	28 / 35 / **115**	5 / 10 / 40	357 / 72 / 25
	Personalized newsletter***	328 / **651** / 51	**859** / 251 / 73	4 / 2 / 9	40 / 57 / **115**	16 / 5 / 27	251 / 39 / 14
	Photos from social networks***	398 / **690** / 46	**903** / 247 / 71	4 / 3 / 7	54 / 49 / **119**	10 / 0 / 34	129 / 16 / 12
	Outfit recommendations from influencers***	**692** / **767** / 22	529 / 55 / 21	44 / 29 / 18	214 / 139 / **179**	8 / 11 / 48	11 / 4 / 1
	Measure	Segment-Specific Category Frequencies („Indifferents“: n=802 / „Enthusiastics“: n=1,427 / „Demanders“: n=626)					
Survey 2	Presentation via video***	162 / **1,008** / 181	**504** / 362 / 126	5 / 2 / 27	37 / 35 / **281**	52 / 2 / 4	42/ 18 / 7
	360° view***	291 / **1,053** / 129	**291** / 121 / 50	39 / 64 / 62	110 / 185 / **381**	56 / 1 / 3	15 / 3 / 1
	Find out individual size***	130 / **1,093** / 107	**488** / 171 / 42	32 / 31 / 59	53 / 118 / **405**	52 / 5 / 4	47 / 9 / 9
	Information model size***	136 / **913** / 98	**473** / 302 / 77	40 / 46 / 95	52 / 147 / **342**	51 / 9 / 6	50 / 10 / 8
	One model wears all sizes***	144 / **777** / 105	**432** / 378 / 120	9 / 9 / 32	35 / 71 / **297**	65 / 6 / 10	117 / 186 / 62
	Favorite article for comparison***	82 / **891** / 161	**553** / 464 / 126	14 / 8 / 36	20 / 32 / **288**	57 / 2 / 4	76 / 30 / 11
	Size advice - figure types***	154 / **982** / 122	**449** / 222 / 51	25 / 41 / 62	71 / 166 / **377**	53 / 4 / 4	50 / 12 / 10
	Photos from social networks***	38 / 395 / 129	438 / **708** / 246	9 / 5 / 17	6 / 10 / 95	58 / 3 / 10	253 / 306 / 129
	Virtual fitting of articles***	59 / **679** / 143	**495** / 549 / 159	8 / 2 / 22	12/ 21 / **230**	51 / 0 / 6	177 / 176 / 66
	Size recommendation – previous purchases***	131 / **985** / 174	**449** / 318 / 102	11 / 12 / 47	22 / 32 / **282**	62 / 11 / 4	127 / 69 / 17
	Self-measurement via webcam***	33 / 343 / 118	287 / **629** / **209**	7 / 1 / 6	5 / 9 / 100	86 / 14 / 12	**384** / 431 / 181

	Measure	Segment-Specific Category Frequencies („Indifferents“: n=757 / „Enthusiastics“: n=1,119 / „Demanders“: n=873)					
Survey 3	Discount on current order***	286 / **1,026 /** 147	**415** / 49 / 58	5 / 3 / 57	13 / 27 / **510**	3 / 11 / 46	35 / 3 / 55
	Discount on next order***	340 / **1,045** / 155	**377** / 29 / 71	3 / 4 / 31	13 / 34 / **521**	2 / 0 / 48	22 / 7 / 47
	Return impact information***	49 / 237 / 78	**527** / **643** / **347**	20 / 35 / 92	14 / 62 / 159	1 / 6 / 53	146 / 136 / 144
	Bonus points for purchases***	311 / **979** / 194	**382** / 56 / 104	16 / 5 / 22	29 / 75 / **480**	5 / 1 / 41	14 / 3 / 32
	Bonus points for non-return orders***	156 / **1,010** / 133	**553** / 72 / 146	5 / 3 / 42	8 / 30 / **436**	1 / 2 / 49	34 / 2 / 67
	Reward of article ratings***	198 / **835** / 189	**466** / 192 / 147	13 / 3 / 32	31 / 73 / **433**	6 / 1 / 37	43 / 15 / 35
	Support for social projects***	260 / **692** / 144	**391** / 201 / 210	11 / 10 / 38	69 / 197 / **377**	6 / 4 / 42	20 / 15 / 62
	Display of the return behavior***	90 / **506** / 213	**542** / 501 / **293**	5 / 0 / 22	10 / 40 / 210	2 / 3 / 34	108 / 69 / 101
	Waiver of shipping costs***	**387** / **833** / 153	202 / 33 / 30	33 / 26 / 51	107 / 219 / **574**	14 / 3 / 45	14 / 5 / 20

Note: The Most Frequent Category per Segment Marked in Bold. A=Attractive; I=Indifferent; M=Must-Be; O=One-Dimensional; Q=Questionable; R=Reverse; Differences Across Segments are Analyzed Using the Chi-squared Test of Independence with ***: p<0.01; **: p<0.05, *: p<0.1.

These segments can again be depicted graphically, where each graph represents one of the three surveys. In the graphs, there are three data points (segments) for each measure. Different symbols indicate the affiliation to the respective segment. For the sake of clarity, we have refrained from displaying all 84 data points in one graph. Therefore, we have staggered the graphs according to the surveys. We see the more excellent information value directly comparing the clusters' positions to each other for each item, and also, the clusters were calculated independently for each survey.

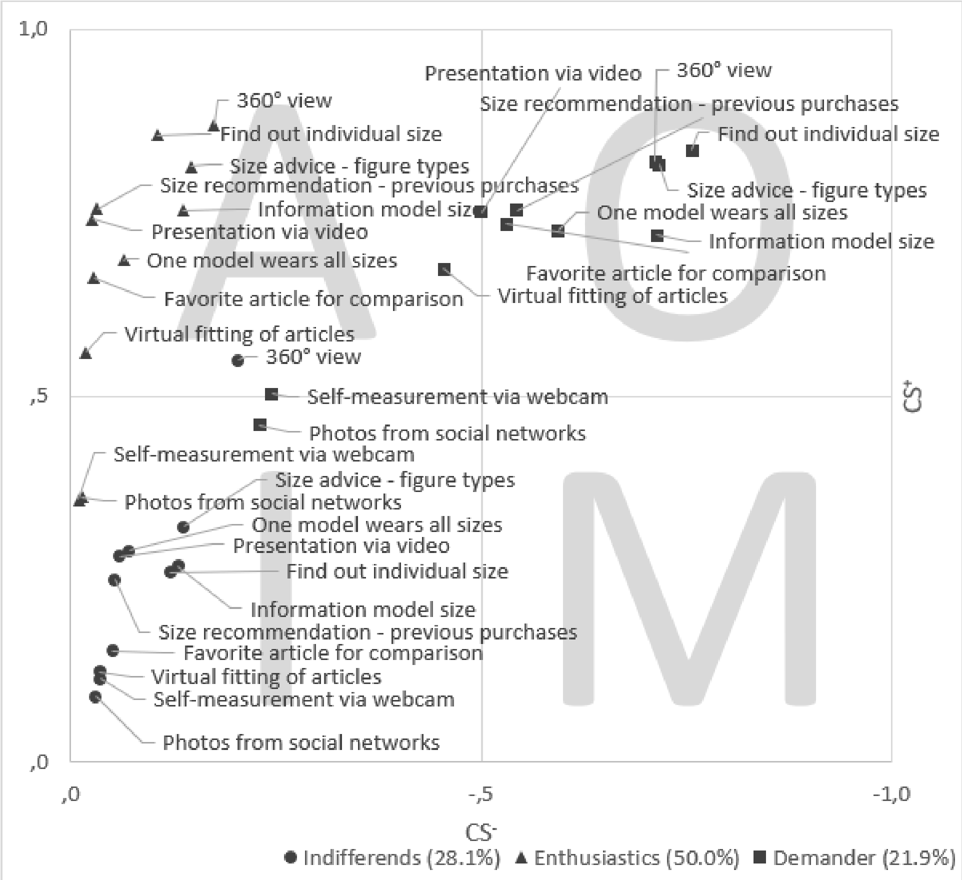


Fig. 4 Depiction of the assessment of possible measures survey 1 (n=2,792)

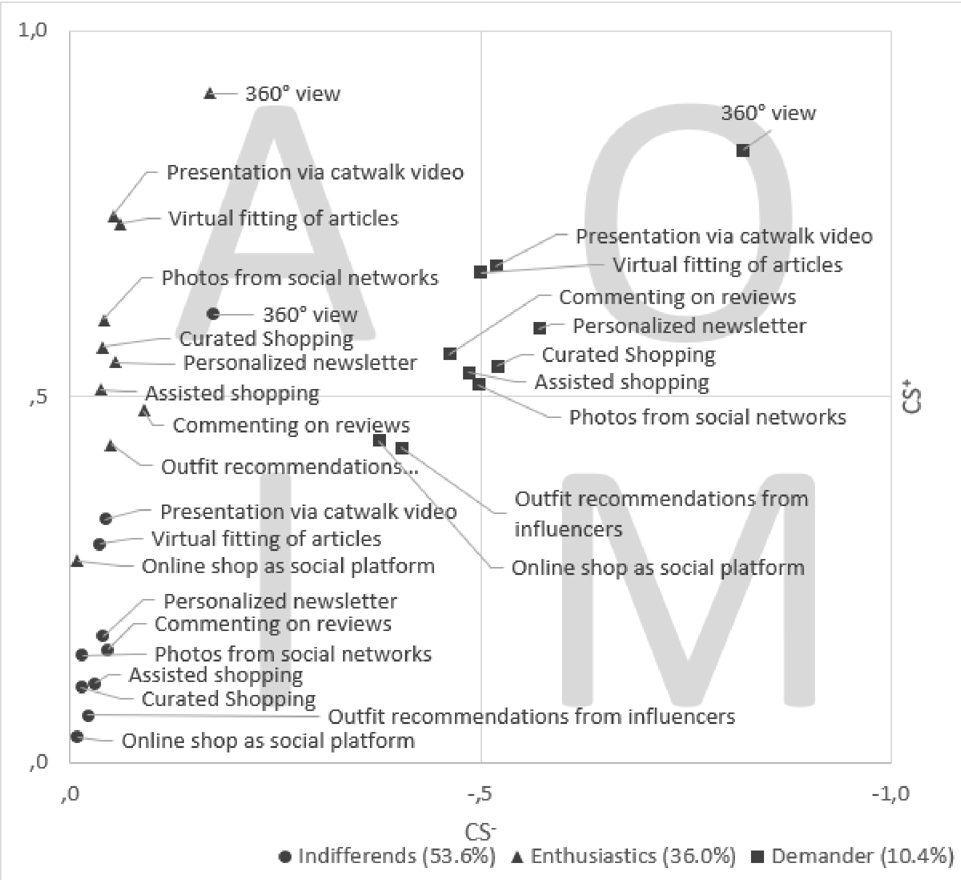


Fig. 5 Depiction of the assessment of possible measures survey 2 (n=2,855)

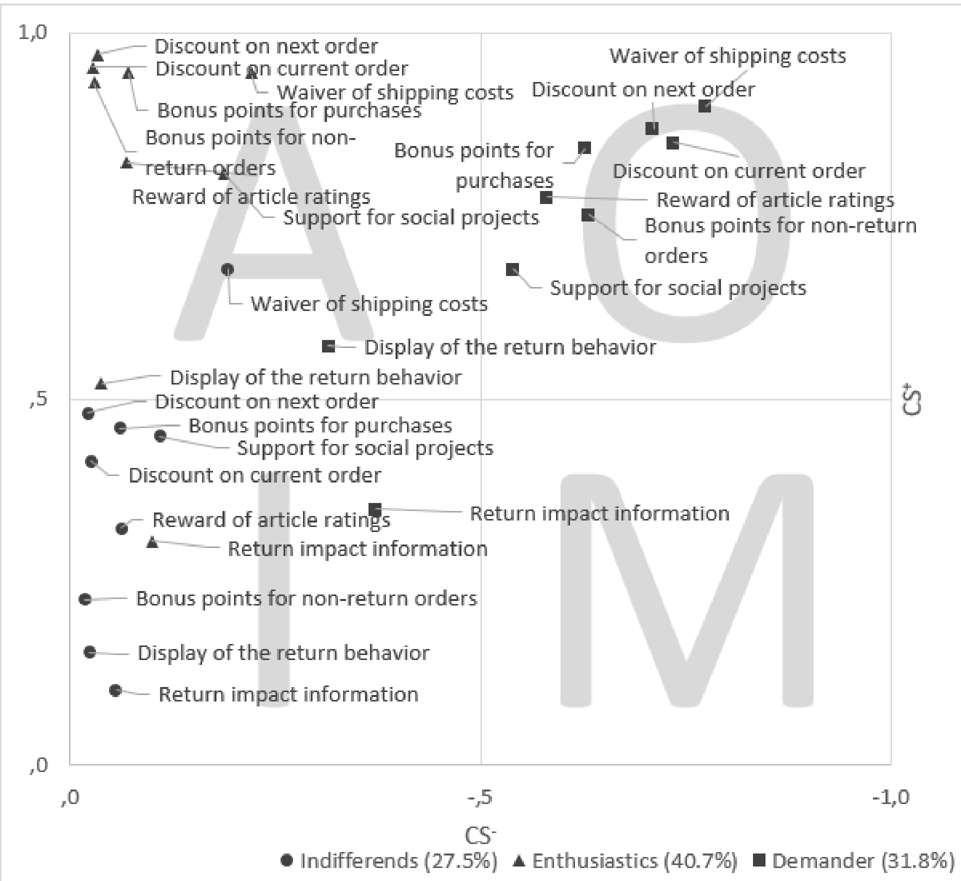


Fig. 6 Depiction of the assessment of possible measures survey 3 (n=2,749)

Analogous to Kano's Life Cycle Theory, there are already segments with statistically significant differences in the lifecycle. In the segmented Kano perspective, not the measures differ among themselves, but the persons confronted with the measures. It is possible to see in Figures 4 to 6 that a general market standard apparently has no influence or otherwise cannot be determined in the first place. Instead, it seems that customers have individual expectations regarding the measures, referring again to H_1,

Naturally, measures that show the highest and lowest influences aggregated (Table 5) also show the strongest or lowest influences in relative terms for the individual segments. To have a strong influence overall, many respondents have to answer in the same way, which is also the case after dividing into the three segments. Therefore, measures with a very high or meager impact on customer satisfaction are found in similar (relative) positions after splitting into segments.

To provide a more detailed characterization of the segments, we also investigated them regarding their buying and return behavior and socio-demographic. Unfortunately, no significant differences were found here either.

5. Discussion

5.1 Theoretical Contribution

While the majority of previous studies analyzed return behavior either before/during, or after the purchase decision (Janakiraman et al. 2016), we contribute to the literature by expanding the view on return management to a 3-stage approach, which is investigated in the pre-purchase, post-purchase, and purchase stage based on a large data pool of actual customers (n=8,396). This holistic approach reveals that return measures in the post-purchase and actual purchase stage are more applicable to increase consumers' satisfaction than those related to the pre-purchase stage. For this purpose, we have extended the already existing approaches in Fig. 1. Besides, to the best of our knowledge, this study represents the first to analyze consumer return behavior by applying the Kano method. Hence, we enable an overview of product return avoidance and averting measures to satisfy consumers the most. This juxtaposition shows for the first time how strongly monetary approaches differ from the remaining measures. Without a combination of the three stages, this finding would not have been possible.

5.2 Managerial Implications

Returns in the mail-order business, especially fashion, are a great nuisance for the customer, the company, and the environment. However, not all of the proposed measures can be effectively implemented by a company. It is, therefore, essential to focus on a few but effective measures. Our paper offers new insights in this respect. As one might expect, measures that positively sanction customers are prevalent. Since these, in turn, actively influence pricing policy, such measures must be weighed up carefully. The next exciting group includes measures that aim at improving the presentation of merchandise without requiring further effort from the customer. The „360° view“ stands out in particular. This measure has even prevailed over more elaborate presentations such as „model type photos,“ „presentation via catwalk videos,“ „virtual fitting of articles,” presentation via video,“ or „information model size.“

The second important finding is that the measures were categorized as exclusively Indifferent or Attractive on the overall level. A lack of or poor performance in these measures still has little effect on satisfaction. Two conclusions can be drawn from this: either returns are hardly an issue for the respondents. More than half of the respondents answered that they do not consider returns to be costly (the bias is that only active mail-order customers participated in the survey). The other conclusion may be the real market standard in terms of avoiding and averting returns is still deficient, and so are the expectations.

However, a more precise segmentation into clusters already reveals the first One-directional measures. Thus, there are already customers whose expectations are significantly higher and whose absence or poor performance leads to dissatisfaction. In the sense of early strategic detection, this customer group should be observed more closely. If this group grows significantly over time, investment in return management is no longer just nice to have but essential for customer satisfaction. Ultimately, it can be assumed that measures will migrate to the Must-be quadrant in some time, which means that investments in this area will not even increase satisfaction but will only prevent dissatisfaction. Unfortunately, we were not able to describe these clusters more precisely with the customer characteristics queried. Further research in this area would, therefore, be highly desirable.

In brief, this means that customer expectations in return management are generally still in a very early life cycle stage. However, this does by no means mean that this is unimportant. On the contrary, vendors can set themselves ahead of the competition and gain a competitive advantage by, for example, improving the presentation of their products. Our categorization of the market standard for Germany clearly shows how few measures can already be considered established. A similar situation would be observed in the European region. The online market, which is still growing dynamically, will also be joined by different groups of consumers who can no longer be described as early movers. Here the demands will change even more significantly, also concerning returns management. The current social discussion in Western countries is also bringing the environmental impact of human activity more into focus. Here, too, vendors can already differentiate themselves from the competition today and use the first-mover advantage for themselves.

Although monetary incentives such as vouchers or discounts promise a high impact, these mechanisms are usually easy to comprehend. On the other hand, we are firmly convinced that a focus on monetary incentives alone does not represent a differentiating feature and can also be easily copied by the competition.

Finally, it should be noted that some measures, albeit unintentionally, can have a negative impact on repeat purchase behavior. A test and learning approach is, therefore, advisable here.

5.3 Limitations and Future Research

This paper is limited in some respects. First of all, the respondents are active, mail-order customers acquired via newsletters. Potential customers who, for example, do not buy by mail order at all due to the problem of returns are not present. Neither the age nor gender structure is representative of Germany. It is also conceivable that BAUR customers differ from other mail order customers in their attitude to, among other things, new technologies, precisely because in the survey, mainly women with an aging focus over 45 years answered (Appendix 1). They differ significantly from the millennial generation regarding their technological skills (Ladhari et al. 2019).

Secondly, the survey was conducted in the German market. Thus, no assertions about possible cultural influences are possible, nor can the industry structure be transferred to other markets without adjustments. Competition may be more intense or extensive, which also affects expectations. For instance, based on the cultural dimension (Hofstede 1980), Germans are assumed to be more likely to avoid uncertain outcomes (rating: 65), compared to, e.g., Americans (rating: 46) or Chinese (rating: 30; Hofstede Insights 2020). Hence, return avoidance and return averting measures can be expected to be of higher interest among Germans to avoid such potentially wrong decisions.

The formed segments strongly indicate a dynamic over time, but unfortunately, could not be described in more detail using the other characteristics that were queried. Therefore, no further contribution could be made here about the presumed adoption behavior, leaving space for further investigations.

There is also strong evidence in the literature that a very restrictive or inconvenient can also affect purchase and re-purchase behavior, especially in a competitive environment like fashion retail. In our view, this field also still receives little attention in research.

Finally, Kano’s method has its limits. Especially in innovation research, many measures are categorized as Indifferent or Attractive. On the other hand, the method can only indicate the current status without providing direct trends for individual attributes’ future progression. A lifecycle is only determined retrospectively. Especially in very dynamic markets such as (fashion) eCommerce, new features are often simply trialed without the need to go through a classic life cycle. Therefore, the context of the featured solutions must always be considered. Also, the special questioning is quite time-consuming and requires a high level of concentration in answering it, diminishing long surveys.

6. Conclusion

We wanted to investigate the most effective strategies to counteract returns from a customers' standpoint in our work. Based on a newly developed three stages process purchase model, a view of several measures have been investigated towards their potential impact on customer satisfaction. Using the Kano method and its subsequent segmented Kano perspective, exciting results were obtained. Among other things, we were able to show that an improvement in the presentation of the products on offer is generally an excellent choice for counteracting returns and that different expectations regarding return management can already be observed today. We thus confirm prior findings, revealing that enhanced product presentation features, such as zooming (De et al. 2013), or in our case a 360° perspective, paves the way for fewer returns or higher customer satisfaction, respectively.

Similarly, photos from social networks or, more generally, alternative product photos are perceived indifferently or might even result in more returns (De et al. 2013). Moreover, we validated that offering virtual fit information enables declined returns (Gallino and Moreno 2018), as virtual reality tools lead to increased customer satisfaction. Generally, our insights emphasize monetary gratifications to represent the measures increasing customer satisfaction the most, which contradicts elder findings derived from online shop return rates below the usual average in the fashion industry (Walsh and Möhring 2015). Besides the nature of gratification and contrast to previous literature, our holistic perspective demonstrated that measures from the post-purchase stage are most likely to increase customer satisfaction, as five measures are among the eight most practical measures (highest CS^+). With this work, we hope to have provided valuable insights into the avoidance and prevention of returns, leading to a reduction of returns in practice.

References

Baier, Daniel; Rese, Alexandra (2018): Conversational user interfaces for online shops: A segmented Kano perspective. In *5th German-Japanese Symposium on Classification and Related Techniques (GJSCRT2018)*, Dortmund, Germany, July 1-3, 2018.

Bell, David R.; Gallino, Santiago; Moreno, Antonio (2018): Offline Showrooms in Omnichannel Retail: Demand and Operational Benefits. In *Management Science* 64 (4), pp. 1629–1651. DOI: 10.1287/mnsc.2016.2684.

Berger, Charles; Blauth, Robert; Boger, David; Bolster, Christopher; Burchill, Gary; DuMouchel, William et al. (1993): Kano's Methods for Understanding Customer-defined Quality. In *Center for Quality of Management Journal* 2 (4), pp. 3–36.

Burton, Jamie; Khammash, Marwan (2010): Why do people read reviews posted on consumer-opinion portals? In *Journal of Marketing Management* 26 (3-4), pp. 230–255. DOI: 10.1080/02672570903566268.

Buxel, H.; Weidlich, T. (2010): Werben mit dem guten Zweck-Akzeptanz karitativer Marketingkonzepte. In *Ergebnisse einer empirischen Untersuchung am Beispiel Krombacher „Regenwaldprojekt "und der Aktion „Gesundheit für Kinder in Afrika "von Actimel/Danone, Münster*.

Chiu, Tom; Fang, DongPing; Chen, John; Wang, Yao; Jeris, Christopher (2001): A robust and scalable clustering algorithm for mixed type attributes in large database environment. In *Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 263–268.

De, Prabuddha; Hu, Yu; Rahman, Mohammad S. (2013): Product-Oriented Web Technologies and Product Returns: An Exploratory Study. In *Information Systems Research* 24 (4), pp. 998–1010. DOI: 10.1287/isre.2013.0487.

Decker, Hanna (2018): Was tun mit Retouren? Edited by Frankfurter Allgemeine Zeitung. Available online at https://www.faz.net/aktuell/wirtschaft/digitec/amazon-und-die-retouren-wie-machen-es-otto-und-zalando-15638099.html, checked on 5/12/2020.

Deges, Frank (2017): Retourenmanagement im Online-Handel. Wiesbaden: Springer Fachmedien Wiesbaden.

Difrancesco, Rita Maria; Huchzermeier, Arnd; Schröder, David (2018): Optimizing the return window for online fashion retailers with closed-loop refurbishment. In *Omega* 78, pp. 205–221. DOI: 10.1016/j.omega.2017.07.001.

Dutta, Pankaj; Mishra, Anurag; Khandelwal, Sachin; Katthawala, Ibrahim (2020): A multiobjective optimization model for sustainable reverse logistics in Indian E-commerce market. In *Journal of Cleaner Production* 249, p. 119348. DOI: 10.1016/j.jclepro.2019.119348.

EHI Retail Institute (2019): Top 100 umsatzstärkste Onlineshops in Deutschland. Available online at https://www.ehi.org/de/top-100-umsatzstaerkste-onlineshops-in-deutschland/, updated on 9/9/2019, checked on 5/17/2020.

Gallino, Santiago; Moreno, Antonio (2018): The Value of Fit Information in Online Retail: Evidence from a Randomized Field Experiment. In *M&SOM* 20 (4), pp. 767–787. DOI: 10.1287/msom.2017.0686.

Gehrckens, Mathias; Boersma, Thorsten (2013): Zukunftsvision Retail – Hat der Handel eine Daseinsberechtigung? In Gerrit Heinemann, Kathrin Haug, Mathias Gehrckens (Eds.): Digitalisierung des Handels mit ePace. Wiesbaden: Springer Fachmedien Wiesbaden, pp. 51–74.

Gelbrich, Katja; Gäthke, Jana; Hübner, Alexander (2017): Rewarding customers who keep a product: How reinforcement affects customers‘ product return decision in online retailing. In *Psychology & Marketing* 34 (9), pp. 853–867.

Haug, Kathrin (2013): Digitale Potenziale für den stationären Handel durch Empfehlungsprozesse, lokale Relevanz und mobile Geräte (SoLoMo). In Gerrit Heinemann, Kathrin Haug, Mathias Gehrckens (Eds.): Digitalisierung des Handels mit ePace. Wiesbaden: Springer Fachmedien Wiesbaden, pp. 27–50.

Haug, Kathrin; Küper, Jérémy (2010): Das Potenzial von Kundenbeteiligung im Web-2.0-Online-Shop. In Gerrit Heinemann, Andreas Haug (Eds.): Web-Exzellenz im E-Commerce. Wiesbaden: Gabler, pp. 115–133.

Heinemann, Gerrit (2019): Der neue Online-Handel. Wiesbaden: Springer Fachmedien Wiesbaden.

Hjort, Klas; Lantz, Björn (2016): The impact of returns policies on profitability: A fashion e-commerce case. In *Journal of Business Research* 69 (11), pp. 4980–4985. DOI: 10.1016/j.jbusres.2016.04.064.

Hofstede, Geert (1980): Culture‘s consequences. International differences in work-related values. 1. print. Beverly Hills u. a.: Sage Publications (Cross-cultural research and methodology series, 5).

Hofstede Insights (2020): Uncertainty Avoidance among China, Germany and the United States. Available online at https://www.hofstede-insights.com/country-comparison/china,germany,the-usa/.

Holland, Heinrich; Bolz, Lea Alice (2017): Betreutes Einkaufen im Internet: Curated Shopping. So funktioniert das Trend-Geschäftsmodell mit hohem Erfolgspotential. Available online at https://www.marketing-boerse.de/fachartikel/details/1708-betreutes-einkaufen-im-internet--curated-shopping/137243, checked on 5/12/2020.

Hölzing, Jörg A. (2008): Die Kano-Theorie der Kundenzufriedenheitsmessung. Eine theoretische und empirische Überprüfung. Zugl.: Mannheim, Univ., Diss., 2007. 1. Aufl. Wiesbaden: Gabler Verlag / GWV Fachverlage GmbH Wiesbaden (Gabler Edition Wissenschaft). Available online at http://gbv.eblib.com/patron/FullRecord.aspx?p=749291.

Hong, Yili; Pavlou, Paul A. (2014): Product Fit Uncertainty in Online Markets: Nature, Effects, and Antecedents. In *Information Systems Research* 25 (2), pp. 328–344. DOI: 10.1287/isre.2014.0520.

ibi research (2017): Trends und Innovationen beim Versand. Was erwartet der Kunde? Regensburg: ibi research an der Universität Regensburg GmbH.

IFH Köln; AZ Direct (2016): Belohnen statt Bestrafen. – So gibt es weniger Retouren beim Fashion-Kauf. Available online at https://www.ifhkoeln.de/pressemitteilungen/details/belohnen-statt-bestrafen-so-gibt-es-weniger-retouren-beim-fashion-kauf/, checked on 5/12/2020.

Iyer, Rajesh; Eastman, Jacqueline K. (2006): The Elderly and Their Attitudes Toward the Internet: The Impact on Internet Use, Purchase, and Comparison Shopping. In *Journal of Marketing Theory and Practice* 14 (1), pp. 57–67. DOI: 10.2753/MTP1069-6679140104.

Janakiraman, Narayan; Syrdal, Holly A.; Freling, Ryan (2016): The Effect of Return Policy Leniency on Consumer Purchase and Return Decisions: A Meta-analytic Review. In *Journal of Retailing* 92 (2), pp. 226–235. DOI: 10.1016/j.jretai.2015.11.002.

Kano, Noriaki (1968): Concept of TQC and its Introduction. In *Kuei* 35 (4), pp. 20–29.

Kano, Noriaki (1987): Total quality creation. In *ICQCC Tokyo Proceeding*.

Kano, Noriaki (1995): Upsizing the organization by attractive quality creation. In: Total Quality Management: Springer, pp. 60–72.

Kano, Noriaki (2001): Life cycle and creation of attractive quality. In *Proceedings of the 4th QMOD Conference*. Linkoping, Sweden, 12.-14.09.2001, pp. 12–14.

Kano, Noriaki; Seraku, Nobuhiko; Takahashi, Fumio; Tsuji, Shinichi (1984): Attractive quality and must-be quality. In *Hinshitsu (Quality, The Journal of Japanese Society for Quality Control)* 14, pp. 39–48.

Kreutzer, Ralf T. (2018): E-Mail-Marketing kompakt. Wiesbaden: Springer Fachmedien Wiesbaden.

Ladhari, Riadh; Gonthier, Jessica; Lajante, Mathieu (2019): Generation Y and online fashion shopping: Orientations and profiles. In *Journal of Retailing and Consumer Services* 48, pp. 113–121. DOI: 10.1016/j.jretconser.2019.02.003.

Lai, Kee-hung; Wong, Christina W.Y.; Venus Lun, Y. H. (2014): The role of customer integration in extended producer responsibility: A study of Chinese export manufacturers. In *International Journal of Production Economics* 147, pp. 284–293. DOI: 10.1016/j.ijpe.2013.06.028.

Lee, Dong Hwan (2015): An alternative explanation of consumer product returns from the postpurchase dissonance and ecological marketing perspectives. In *Psychology & Marketing* 32 (1), pp. 49–64.

Lee, Shinhyoung; Yi, Youjae (2017): “Seize the Deal, or Return It Losing Your Free Gift”: The Effect of a Gift-With-Purchase Promotion on Product Return Intention. In *Psychology & Marketing* 34 (3), pp. 249–263.

Letizia, Paolo; Pourakbar, Morteza; Harrison, Terry (2018): The Impact of Consumer Returns on the Multichannel Sales Strategies of Manufacturers. In *Production and Operations Management* 27 (2), pp. 323–349. DOI: 10.1111/poms.12799.

Li, Guo; Li, Lin; Sethi, Suresh P.; Guan, Xu (2019): Return strategy and pricing in a dual-channel supply chain. In *International Journal of Production Economics* 215, pp. 153–164.

Lian, Jiunn-Woei; Yen, David C. (2014): Online shopping drivers and barriers for older adults: Age and gender differences. In *Computers in Human Behavior* 37, pp. 133–143. DOI: 10.1016/j.chb.2014.04.028.

Löfgren, Martin; Witell, Lars; Gustafsson, Anders (2011): Theory of attractive quality and life cycles of quality attributes. In *TQM* 23 (2), pp. 235–246. DOI: 10.1108/17542731111110267.

Lohse, Tobias; Kemper, Jan; Brettel, Malte (Eds.) (2017): How online customer reviews affect sales and return behavior–an empirical analysis in fashion ecommerce. European Conference On Information Systems (ECIS): Association For Information System (AIS).

Matzler, Kurt (2003): Kundenzufriedenheit: Prospect Theory oder Kano-Modell. In *Zeitschrift für Betriebswirtschaft* 73 (4), pp. 341–344.

Matzler, Kurt; Hinterhuber, Hans H.; Bailom, Franz; Sauerwein, Elmar (1996): How to delight your customers. In *Journal of Product & Brand Management*.

Melchior, Laura (2018a): Die 3 wichtigsten E-Commerce-Trends für 2019. Available online at https://www.internetworld.de/e-commerce/online-handel/3-wichtigsten-e-commerce-trends-2019-1592762.html, checked on 5/12/2020.

Melchior, Laura (2018b): So erreichen Werbetreibende Millenials am besten. Available online at https://www.internetworld.de/social-media/zahlen-studien/so-erreichen-werbungtreibende-millennials-am-besten-1641239.html, checked on 5/12/2020.

Mikulić, Josip; Prebežac, Darko (2011): A critical review of techniques for classifying quality attributes in the Kano model. In *Managing Service Quality: An International Journal*.

Minnema, Alec; Bijmolt, Tammo H. A.; Gensler, Sonja; Wiesel, Thorsten (2016): To keep or not to keep: Effects of online customer reviews on product returns. In *Journal of Retailing* 92 (3), pp. 253–267.

Nathan, Peter E.; Wallace, Wallace H. (1971): An operant behavioral measure of TV commercial effectiveness. In: Consumer Behavior. Boston: Houghton Mifflin, pp. 78–89.

Nilsson-Witell, Lars; Fundin, Anders (2005): Dynamics of service attributes: a test of Kano‘s theory of attractive quality. In *International Journal of Service Industry Management* 16 (2), pp. 152–168. DOI: 10.1108/09564230510592289.

Ofek, Elie; Katona, Zsolt; Sarvary, Miklos (2011): “Bricks and clicks”: The impact of product returns on the strategies of multichannel retailers. In *Marketing Science* 30 (1), pp. 42–60.

Oghazi, Pejvak; Karlsson, Stefan; Hellström, Daniel; Hjort, Klas (2018): Online purchase return policy leniency and purchase decision: Mediating role of consumer trust. In *Journal of Retailing and Consumer Services* 41, pp. 190–200.

Pålsson, Henrik; Pettersson, Fredrik; Winslott Hiselius, Lena (2017): Energy consumption in e-commerce versus conventional trade channels - Insights into packaging, the last mile, unsold products and product returns. In *Journal of Cleaner Production* 164, pp. 765–778. DOI: 10.1016/j.jclepro.2017.06.242.

Pei, Zhi; Paswan, Audhesh (2018): Consumers‘ legitimate and opportunistic product return behaviors in online shopping. In *Journal of Electronic Commerce Research* 19 (4), pp. 301–319.

Perotti, Victor; Sorce, Patricia; Widrick, Stanley (2003): An Exploratory Study of Operant Conditioning Theory as a Predictor of Online Product Selection. In *Journal of Electronic Commerce in Organizations* 1 (1), pp. 42–54. DOI: 10.4018/jeco.2003010103.

Petersen, J. Andrew; Kumar, V. (2015): Perceived risk, product returns, and optimal resource allocation: Evidence from a field experiment. In *Journal of Marketing Research* 52 (2), pp. 268–285.

Raharjo, Hendry; Brombacher, Aarnout C.; Goh, T. N.; Bergman, Bo (2010): On integrating Kano‘s model dynamics into QFD for multiple product design. In *Quality and Reliability Engineering. International* 26 (4), pp. 351–363. DOI: 10.1002/qre.1065.

Rao, Shashank; Lee, Kang Bok; Connelly, Brian; Iyengar, Deepak (2018): Return time leniency in online retail: A signaling theory perspective on buying outcomes. In *Decision Sciences* 49 (2), pp. 275–305.

Rese, Alexandra; Schlee, Tobias; Baier, Daniel (2019): The need for services and technologies in physical fast fashion stores: Generation Y's opinion. In *Journal of Marketing Management* 35 (15-16), pp. 1437–1459.

Rogers, Everett M. (1962): Diffusion of innovations. New York: Free Press of Glencoe.

Rudolph, Thomas; Becker, Kalle (2003): Efficient Differentiation: A systematic approach for retailers to appear unique. In *The European Retail Digest* 38, pp. 80–85.

Saarijärvi, Hannu; Sutinen, Ulla-Maija; Harris, Lloyd C. (2017): Uncovering consumers' returning behaviour: a study of fashion e-commerce. In *The International Review of Retail, Distribution and Consumer Research* 27 (3), pp. 284–299. DOI: 10.1080/09593969.2017.1314863.

Sahoo, Nachiketa; Dellarocas, Chrysanthos; Srinivasan, Shuba (2018): The impact of online product reviews on product returns. In *Information Systems Research* 29 (3), pp. 723–738.

Samorani, Michele; Alptekinoğlu, Aydın; Messinger, Paul R. (2019): Product Return Episodes in Retailing. In *Service Science* 11 (4), pp. 263–278. DOI: 10.1287/serv.2019.0250.

Seo, Joon Yong; Yoon, Sukki; Vangelova, Milena (2016): Shopping plans, buying motivations, and return policies: impacts on product returns and purchase likelihoods. In *Marketing Letters* 27 (4), pp. 645–659.

Shahin, Arash; Mohammadi, Somaye; Harsij, Hossein; Rahbar Qazi, Mahmoud Reza (2017): Revising satisfaction and dissatisfaction indexes of the Kano model by reclassifying indifference requirements. In *TQM* 29 (1), pp. 37–54. DOI: 10.1108/TQM-05-2015-0059.

Shahin, Arash; Pourhamidi, Masoud; Antony, Jiju; Hyun Park, Sung (2013): Typology of Kano models: a critical review of literature and proposition of a revised model. In *International Journal of Quality & Reliability Management* 30 (3), pp. 341–358. DOI: 10.1108/02656711311299863.

Shahin, Arash; Zairi, Mohamed (2009): Kano model: A dynamic approach for classifying and prioritising requirements of airline travellers with three case studies on international airlines. In *Total Quality Management & Business Excellence* 20 (9), pp. 1003–1028. DOI: 10.1080/14783360903181867.

Shahrestani, Vaez Hossein; Shahin, Arash; Teimouri, Hadi; Shaemi Barzoki, Ali (2020): Revising the Kano model for designing an employee compensation system. In *TQM* 32 (1), pp. 78–91. DOI: 10.1108/TQM-05-2019-0153.

Shulman, Jeffrey D.; Coughlan, Anne T.; Savaskan, R. Canan (2011): Managing Consumer Returns in a Competitive Environment. In *Management Science* 57 (2), pp. 347–362. DOI: 10.1287/mnsc.1100.1274.

Shulman, Jeffrey D.; Cunha, Marcus; Saint Clair, Julian K. (2015): Consumer Uncertainty and Purchase Decision Reversals: Theory and Evidence. In *Marketing Science* 34 (4), pp. 590–605. DOI: 10.1287/mksc.2015.0906.

Singh, Gaganpreet; Pandey, Neeraj (2015): Leveraging return policy for price premium. In *Journal of Revenue and Pricing Management* 14 (4), pp. 276–292.

Skinner, Burrhus Frederic (1965): Science and human behavior. 1. ed., 14. print. New York NY: Free Press (A Free Press paperback: Psychology).

Stöcker, Björn; Nasseri, Aydin (2020): Penalty Reward Contrast Analysis (PRCA) for Categorizing Service Components: A New Approach. In *Archives of Data Science, Series A (Online First)* 6 (2).

Ülkü, M. Ali; Gürler, Ülkü (2018): The impact of abusing return policies: A newsvendor model with opportunistic consumers. In *International Journal of Production Economics* 203, pp. 124–133.

Vella, Kevin J.; Foxall, Gordon R. (2013): The Marketing Firm: Operant Interpretation of Corporate Behavior. In *TPR* 63 (2), pp. 375–402. DOI: 10.11133/j.tpr.2013.63.2.011.

Walsh, Gianfranco; Albrecht, Arne K.; Kunz, Werner; Hofacker, Charles F. (2016): Relationship between online retailers' reputation and product returns. In *British Journal of Management* 27 (1), pp. 3–20.

Walsh, Gianfranco; Möhring, Michael (2015): Wider den Retourenwahnsinn. In *Harvard Business Manager* 3, pp. 6–10.

Wells, Victoria K. (2014): Behavioural psychology, marketing and consumer behaviour: a literature review and future research agenda. In *Journal of Marketing Management* 30 (11-12), pp. 1119–1158. DOI: 10.1080/0267257X.2014.929161.

Wood, Stacy L. (2001): Remote Purchase Environments: The Influence of Return Policy Leniency on Two-Stage Decision Processes. In *Journal of Marketing Research* 38 (2), pp. 157–169. DOI: 10.1509/jmkr.38.2.157.18847.

Yan, Ruiliang; Pei, Zhi (2019): Return policies and O2O coordination in the e-tailing age. In *Journal of Retailing and Consumer Services* 50, pp. 314–321. DOI: 10.1016/j.jretconser.2018.07.006.

Zhou, Wenyan; Hinz, Oliver; Benlian, Alexander (2018): The impact of the package opening process on product returns. In *Business Research* 11 (2), pp. 279–308. DOI: 10.1007/s40685-017-0055-x.

Appendix 1

	Aspect	*Specification*	*Survey 1*	*Survey 2*	*Survey 3*
Shopping behavior	Order frequency	Several times a month	508	511	498
		monthly	877	862	829
		About once every three months	901	946	887
		About once every six months	306	310	316
		Less frequently	200	226	219
	average spending on fashion	< 49 EUR	585	659	638
		50 - 99 EUR	998	1,052	1,038
		100 - 149 EUR	641	627	562
		150 - 199 EUR	244	221	190
		200 - 249 EUR	100	71	92
		>= 250 EUR	64	74	74
		I will not tell	160	151	155
	who is shopped for (multiple responses)	For me	2,552	2,605	2,314
		My partner	1,326	1,334	965
		My children	942	963	692
		My grandchildren	289	283	222
		My Parents	175	189	147
		Friends	134	103	78
		other relatives (e.g., sister, aunt)	192	163	148
	where to buy what (online/offline/indifferent)	Tops (sweaters, T-shirts, blouses)	1,615/233/944	1,616/252/987	1,597/215/937
		Trousers	1,319/719/754	1,345/735/775	1,288/728/733
		Lingerie	1,377/532/883	1,383/581/891	1,310/561/878
		Swimwear	1,374/522/896	1,404/539/912	1,306/557/886
		Jackets and coats	1,184/677/931	1,119/729/1007	1,117/709/923
		Dresses	1,110/520/1162	1,085/546/1224	1,031/492/1,226
		Shoes	1,068/719/1005	1,044/767/1,044	1,009/722/1,018
		Skirts	921/551/1320	892/578/1385	842/555/1,352
		Jewelry and accessories (e.g., scarves, caps, bags)	822/668/1,302		771/694/1,284
		Blazers and suits	803/832/1,157	737/918/1,200	701/857/1,191
	Self-assessment of buying behavior (impulsive/disciplined)	< 33.3%	711	561	632
		33.3% - 66.6%	955	1,109	942
		>66.6%	1,126	1,128	1,175
	Perception shopping fashion online (avoidance - love)	< 33.3%	244	199	227
		33.3% - 66.6%	830	889	784
		>66.6%	1,718	1,767	1,697

Return behavior	Return rate of fashion orders		32.9% (2,792)	32.0% (2,855)	33.0% (2,666)
	Reasons for returns (multiple)	Item does not fit	2,396	2,556	2,365
		I don't like this item	1,301	1,293	1,262
		Several sizes ordered for choice	1,209	1,273	1,012
		Items do not correspond to the description	529	593	641
		I have bought more due to a voucher and selected at home	119	148	89
	Elaboration of a return	< 33.3%	420	626	674
		33.3% - 66.6%	656	777	593
		> 66.6%	1,716	1,452	1,482
	Vendor-specific differences in returns	1 doesn't differ at all	1,200	1,784	1,414
		2	314	285	257
		3	533	320	412
		4	343	205	312
		5	159	112	121
		6 differs very much	243	139	233
Socio-demography	Sex	Female	2,253	2,283	2,160
		Male	522	545	547
		I will not tell	17	27	42
	Age	< 20 years	9	3	10
		20 - 24 years	62	51	41
		25 - 29 years	101	96	94
		30 - 34 years	167	164	151
		35 - 39 years	234	223	212
		40 - 44 years	267	280	275
		45 - 49 years	383	367	369
		50 - 54 years	559	575	506
		55 - 59 years	412	484	464
		>= 60 years	598	612	627
	Size Residence in thousands	< 2 inhabitants	491	505	468
		2 - 5 inhabitants	383	399	419
		5 - 20 inhabitants	600	700	577
		20 - 100 inhabitants	644	600	595
		100 - 1,000 inhabitants	506	471	513
		> 1,000 inhabitants	168	180	177

Chapter 4

Maximizing Profit from Direct Marketing Campaigns: Profit Uplift Modeling Approaches for Online Shops

Daniel Baier and Björn Stöcker

Abstract

Nowadays, uplift modeling approaches are widespread in direct marketing. They predict the incremental response of a customer to a campaign. So, they allow for scoring the customers in advance and, e.g., to better focus on customers that will only purchase due to a contact. However, up to now, only approaches with binary response model outcomes (e.g., visit of a website or conversion) and with continuous revenue outcomes (e.g., money spend) have been proposed in the literature. In this paper, we discuss their shortcomings and how they can be adapted to model profit uplift. We apply the Heckman sample selection model, the zero-inflated negative binomial regression model, and random forest-based regression for this purpose. Two datasets demonstrate the usefulness of these approaches. One dataset describes recent discount offers of a large German online fashion retailer. Another is the well-known Hillstrom dataset that reflects two Email campaigns for men's and women's merchandising. The results are promising.

This chapter is under review in:

Journal of Business Economics

1. Introduction

Before launching a direct marketing campaign, often, a (small) sample of customers is contacted, and their response (e.g., purchased or not) and available customer data (e.g., past information and buying behavior) is used to build a predictive response model. Then, this model is used to select likely responders to a campaign.

However, this classical response modeling approach has two shortcomings: First, the response model selects customers who respond, maybe regardless of the campaign. This would be waste of money in case, e.g., a discount is offered. Second, most response models only predict binary outcomes (purchase or not), not the more informative continuous outcome (revenue or profit generated). Both shortcomings restrict the usefulness of this approach when maximizing profit.

In this paper, we propose new approaches that overcome these problems. They connect findings from the field of uplift modeling (e.g., Radcliffe and Surry 1999, Radcliffe and Surry 2011, Kane et al. 2014, Rudaś and Jaroszewicz 2018, Gubela et al. 2020) with findings from the field of sample selection (see, e.g., Heckman 1979) and zero-inflated regression (see, e.g., Lambert 1992, Ridout et al. 2001). Uplift models focus – in contrast to response models – on the incremental response to a treatment. For each customer, they predict the effect of being treated. The Heckman sample selection and the zero-inflated regression models are designed to overcome the problem that a continuous outcome (revenue or profit) only is available in the few cases when a customer responds to a campaign. We discuss the new approaches in the paper. The well-known Hillstrom direct marketing campaign dataset (with data from n=64,000 customers) and a new direct marketing campaign dataset from a major German online retailer (with data from n=155,388 customers) demonstrate their usefulness. We show that the new approaches are well suited to select "best" customers as targets and to improve profit from direct marketing campaigns.

The paper is organized as follows: In section 2, we discuss related approaches and their shortcomings. In section 3, we introduce the new profit uplift modeling approaches. Then, in section 4, we discuss the application of the new approach to the new dataset and in section 5 to the well-known Hillstrom dataset. The paper closes with conclusions and outlook.

2. Background and Related Work

Testing and predictive modeling are assumed to be the analytical cornerstones of today's direct marketing (Blattberg et al. 2008). The modeling process usually consists of the following five steps: (1) Define the managerial problem in terms of a campaign and its intended effects, (2) translate this description to a predictive model with treatment, responses, and potential predictors, (3) sample customers for collecting responses, (4) calibrate and validate the predictive model, (5) apply the model to all customers and select "best" customers according to the predictive model. Typical managerial problems are the selection of targets for an acquisition campaign at hand, the deciding on customers to receive a catalog or inlay or the identification of promising customers to be invited to a customer tier program. Outcomes are the response to the treatment, predictors customer characteristics, and variables that describe past information and buying behavior in the customer database (see, e.g., Blattberg et al. 2008 for an overview).

Let be Y_i the binary ($Y_i \in \{0,1\}$) or continuous ($Y_i \in \mathbb{R}$) outcome for customer i $(i = 1, \ldots, n)$ in the customer sample, $\boldsymbol{x}_i$ $(\boldsymbol{x}_i, \in \mathbb{R}^m)$ customer i's values for the m predictors, and τ_i the indicator whether the customer I received the treatment (=1) or not (=0). Then, the main goal for a traditional response model would be to predict the following score for selecting targets

$$(1) \qquad Response_i = \mathrm{E}\left(Y_i \mid \boldsymbol{x}_i, \tau_i = 1\right)$$

Depending on the scale of the outcome (e.g., binary or continuous), the above response model can be easily estimated (e.g., using logistic regression or linear regression models), using the data of all treated customers for model calibration and the whole customer database for prediction. The customers with the highest (response) scores are targets for the campaign. However, this response modeling approach has one major shortcoming: It favors customers who respond most likely, but it doesn't take into account that some of them would also respond if not treated. When the treatment is a discount, a voucher, a catalog, an inlay, or one has to deal with postage, this could result in a waste of money for the company.

Therefore, recently, uplift models have been proposed: Responses are again collected from a sample of treated customers (the treatment group), but also from a sample of not treated customers (the control group). An uplift model now predicts the difference in the response of a customer if treated ($\tau_1 = 1$) and if not treated ($\tau_1 = 0$), the so-called uplift score

$$Uplift_i = \mathrm{E}\left(Y_i \mid \boldsymbol{x}_i, \tau_i = 1\right) - \mathrm{E}\left(Y_i \mid \boldsymbol{x}_i, \tau_i = 0\right) \tag{2}$$

Terms like differential response (e.g., Radcliffe and Surry 1999), true lift (Lo 2002), or uplift (Radcliffe and Surry 2011) are used for the same idea. Formula (2) allows to estimate the effect of the treatment and enables the company to select customers where the treatment has an impact.

However, when trying to estimate the parameters of this model, a problem arises from the fact that per customer, only one of these two responses are observable: A customer is part of the treatment group ($\tau_i = 1$) or part of the control group ($\tau_i = 0$), not in both groups. Consequently, an uplift model cannot be estimated directly when using formula (2). Instead, one straightforward idea is to develop two separate models (the so-called two model approach): A first model is derived similar to formula (1), basing the treatment group. This model predicts the outcome if treated in terms of $\boldsymbol{x}_i$ for all customers. A second model is derived using the control group. This model predicts the outcome if not treated in terms of $\boldsymbol{x}_i$ for all customers (see Radcliffe and Surry 1999), the difference between the predictions of the two models is the uplift. An alternative solution is the so-called interaction model proposed by Lo (2002): An interaction (response) model uses the treatment (τ_i) and interactions between the predictors and the treatment as additional predictors. The interaction model can be calibrated on the treatment and the control group simultaneously. Then, for all customers, predictions for all customers are derived via formula (2) by setting the treatment for all customers to 1 in the first summand and 0 in the second.

Table 1 Uplift modeling approaches (only new approaches are reflected, CART=classification & regression tree, DT=decision tree, MLP=multilayer perceptron, RF=random forest, SVM=support vector machine)

Approach	*Outcome*	*Algorithm*	*Reference*
Differential response analysis: Modeling true response	Binary	DT	Radcliffe and Surry 1999
Incremental value modeling	Binary	DT	Hansotia and Rukstales 2002
The true lift model	Binary	Logistic regression	Lo 2002
Influential marketing: A new direct marketing strategy	Binary (transformed)	Association rules, DT, Logistic regression	Lai 2006
Using control groups to target on predicted lift	Continuous	CART	Radcliffe 2007
Uplift modelling with significance-based uplift trees	Continuous	CART	Radcliffe and Surry 2011
DTs for uplift modeling with multiple treatments	Multiple binary	DT	Rzepakowski and Jaroszewicz 2012
Support vector machines for uplift modeling	Binary (transformed)	SVM	Zaniewicz and Jaroszewicz 2013
Uplift random forests	Binary (transformed)	Causal conditional inference tree / RF	Guelman et al. 2015
Mining for truly responsive customers	Binary (transformed)	Logistic regression	Kane et al. 2014
Ensemble methods for uplift modeling	Binary (transformed)	Ensemble methods	Sołtys et al. 2015
L_p-support vector machines for uplift modeling	Binary (transformed)	SVM	Zaniewicz and Jaroszewicz 2017
Revenue uplift modeling	Continuous (transformed)	Linear regression	Rudaś and Jaroszewicz 2018
Revenue uplift modeling	Continuous (transformed)	Lasso, Ridge, and Theilsen regression, MLP, RF	Gubela et al. 2020
Profit uplift modeling	Continuous (transformed)	Heckman sample selection, Zero-inflated NB, RF	This paper

Over the years, a large number of uplift modeling approaches and algorithms to estimate their parameters have been proposed. Table 1 gives an overview. As one can easily see, most of them aim at predicting uplifts for binary outcomes, e.g., indicators for visit, conversion, or purchase. Here, logistic regression or decision trees can be easily applied to estimate model parameters. However, more recently, also revenue uplift modeling approaches have become popular (Gubela et al. 2020; Rudaś and Jaroszewicz 2018). The main idea behind this popularity is that the revenue uplift more closely relates to economic goals than the visit or purchase uplift. However, in the next section, we will see that even with a revenue uplift modeling approach, suboptimal sortings of the customers could be generated.

Another interesting aspect in Table 1 is that the most recent uplift modeling approaches rely on transformed outcomes for parameter estimation. This transformation was introduced for binary outcomes in a seminal paper (Lai 2006) and later extended to continuous outcomes (Gubela et al. 2020; Rudaś and Jaroszewicz 2018). The main idea behind is to transfer as much information as possible from the observed responses in the two groups into the dependent variable and so being able to directly estimate the uplift model parameters. So, e.g., Rudaś and Jaroszewicz (2018) – following the proposal of Lai (2006) for binary outcomes – proposed to estimate their revenue uplift model

$$Uplift_i = \mathrm{E}\left(Z_i \mid \boldsymbol{x}_i\right) \tag{3}$$

directly using transformed continuous revenue outcomes

$$Z_i \begin{cases} +\dfrac{1}{q^T} Y_i & if \quad \tau_i = 1 \wedge Y_i > 0 \\ \quad 0 & if \quad Y_i = 0 \\ -\dfrac{1}{q^C} Y_i & if \quad \tau_i = 0 \wedge Y_i > 0 \end{cases} \tag{4}$$

q^T and q^C are the fractions of the treatment group and the control group in the customer sample. Rudaś and Jaroszewicz (2018) discuss in their paper that this weighting facilitates unbiased estimation of the model parameters when relying on linear models. The main idea behind the positive weighting of the observed revenues in the treatment sample and the negative weighting of the observed revenues in the control sample is that so the best possible information is forwarded to parameter estimation. It is assumed that the purchasers in the treatment group generate probably a (low to high) positive revenue uplift. Likewise, it is assumed that the purchasers in the control group generate probably a (low to high) negative revenue uplift.

Another major problem with uplift modeling approaches is to validate their predictions at the customer level since for these predictions – as mentioned above – no observations exist. The widespread solution for this problem is to develop so-called Qini curves and calculate the so-called Qini coefficient Q (Radcliffe 2007; Radcliffe and Surry 2011): The customers are sorted according to a descending uplift score and partitioned into deciles (or other partitions) with similar scores. Then, within the deciles, averaged responses of the customers from the treatment group and the averaged responses from the control group are calculated, and so "observed" average uplifts are available via their difference. Figure 1 shows the typical results for such a validation of an uplift model applied to a sample dataset. In both diagrams, the customers are sorted according to descending uplift predictions from left to right, and deciles are formed. In the right diagram, one can see the calculated average uplift per decile, as discussed above. In the left diagram, from decile to decile, the average uplift is plotted cumulatively, which means that for the first decile, the values in the left and right diagram are identical, but from then, aggregated values for deciles are plotted in the left diagram. The last value in the left diagram (with value 0.045) of this so-called Qini curve reflects the uplift across all deciles (the target population). For comparisons, also the Qini curve for a random uplift model is plotted in the left diagram. Its incremental uplift curve connects the zero point with the average uplift across the target population (0.045). The quality of an uplift model is judged by its ability to sort customer deciles according to decreasing "observed" uplifts in the right diagram but also by calculating the related area between the Qini curve for this model and the line for the random model that reflects this perfect sorting. In Figure 1, this value – the so-called Qini coefficient Q – is 0.0296 and could serve for comparisons with other uplift models (the random model has Q=0, the maximum is data-dependent). It should be noted that these two diagrams can be generated for uplift models with binary response outcomes but also for uplift models with continuous outcomes (as in our new profit uplift modeling approach discussed in the following section).

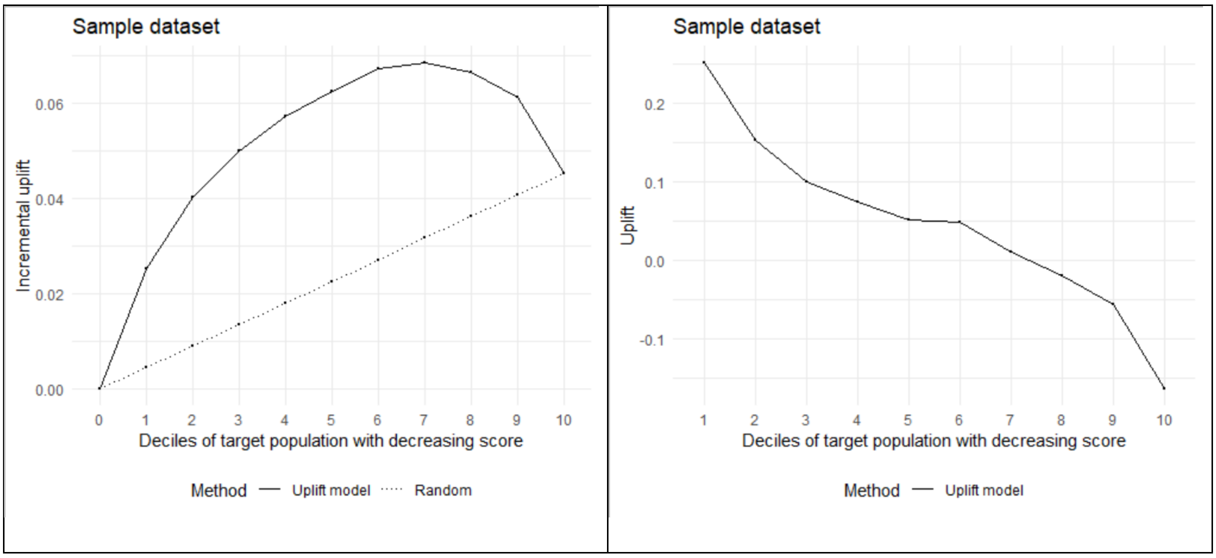


Fig. 1 Qini curve (left) and mean uplifts (right) for a sample dataset. The area between the Qini curves of an uplift model and a random model is the Qini coefficient Q (here: Q=0.0296), which can be used for model selection

3. A new Profit Modeling Approach for Online Shops

As already discussed, most uplift modeling approaches model binary outcomes. Only recently, continuous outcomes have received more interest, e.g., in the papers by Rudaś and Jaroszewicz (2018) and Gubela et al. (2020). This is surprising since from the beginning of the development of the uplift modeling approaches; also datasets with continuous outcomes have been made available. So, e.g., the famous Hillstrom dataset (Radcliffe 2008) – that is often seen as the standard dataset in uplift modeling and has been used in many papers when models were compared or introduced – contains as binary outcomes the visit of the website (=1: yes, =0: no) and the purchase information (=1: yes, =0: no) but also the revenue generated by this purchase (spend in $). However, maybe since the share of purchasers in this dataset was very low (0.9%), the revenue uplift being very low, and revenues concentrate on few purchasers, this continuous outcome didn't stimulate the community to develop continuous outcome uplift models. Even in the newer and methodologically advanced paper by Rudaś and Jaroszewicz (2018) this dataset is only used as a basis for a simulation at the end of the paper. The main methodological progress in revenue uplift modeling in their paper was demonstrated by using synthetic data. However, recently, Gubela et al. (2020) have demonstrated in their paper with large real-world datasets (nearly 3 million sessions from visits at 25 European online shops) that revenue uplift modeling provides further insights.

This superiority of a continuous outcome uplift modeling can also been seen when reflecting the assumed behavior of a small sample of customers, as shown in Table 2. Here, for 12 customers, their (assumed) purchases as well as (assumed) revenues and profits generated by them are given in case of a direct marketing campaign with a discount offer of d=20% and a profit margin of m=30%. Profits were calculated for the customers in the treatment group as 10% (=(m-d)) of the revenue and for the control group as 30% (=m) of the revenue.

One can easily see that the 12 customers reflect a typical behavior: They show – on average – a purchase outcome uplift (8%) when offered a discount, they generate a higher revenue when a discount is offered (+69 €), but it is not useful to offer the discount to all customers since the profit uplift – on average – is negative (-9 €). Only five customers (1, 2, 3, 4, and 5) show a profit uplift, which means that only these four customers should be offered the discount. The customer sorting according to the purchase outcome and the revenue outcome is different: The three customers with the highest revenue uplift show a purchase uplift of 0. However, as also can be seen in Table 2, both sortings considerably differ from the sorting according to the profit uplift: If the customers were targeted according to their revenue uplift, customers with a positive profit uplift but also with negative profit uplift would receive a discount offer. It should be noted that this difference in sorting heavily relies on the ability of discounts to generate additional revenues but also on the fact that in online shops, high discounts are widespread but would lead to losses if granted to all customers. Moreover, it should be mentioned that Table 2 reflects an ideal situation in-so-far that from each customer, two observations are available – the purchase and revenue outcome with and without treatment – which in reality would not be possible.

Table 2 Sample of customers with purchase, revenue, uplift if treated or not (with margin m=0.3, discount d=0.2)

Customer	*Purchase*			*Revenue*			*Profit*		
	if treated	if not tr.	uplift	if treated	if not tr.	uplift	if treated	if not tr.	uplift
1	1	0	1	160 €	0 €	160 €	16 €	0 €	16 €
2	1	1	0	300 €	60 €	240 €	30 €	18 €	12 €
3	1	0	1	40 €	0 €	40 €	4 €	0 €	4 €
4	1	0	1	30 €	0 €	30 €	3 €	0 €	3 €
5	1	0	1	20 €	0 €	20 €	2 €	0 €	2 €
6	1	1	0	70 €	40 €	30 €	7 €	12 €	-5 €
7	0	1	-1	0 €	20 €	-20 €	0 €	6 €	-6 €
8	0	1	-1	0 €	40 €	-40 €	0 €	12 €	-12 €
9	0	1	-1	0 €	60 €	-60 €	0 €	18 €	-18 €
10	1	1	0	400 €	200 €	200 €	40 €	60 €	-20 €
11	1	1	0	500 €	250 €	250 €	50 €	75 €	-25 €
12	1	1	0	270 €	290 €	-20 €	27 €	87 €	-60 €
Mean	75%	67%	8%	149 €	80 €	69 €	15 €	24 €	-9 €

After demonstrating the potential usefulness of profit uplift modeling approaches, now, they are discussed in detail. The main idea is to use formulae (2) or (3) and (4) for modeling continuous outcomes but to replace the observed revenue by derived profits and the revenue uplift predictions by profit uplift predictions. We follow Blattberg et al. (2008) as in section 2 and discuss the five steps of the predictive modeling process now in detail:

(1) **Define the managerial problem**: Online shops have a huge variety of potential offerings that could motivate their customers to purchase and/or to spend more. So, e.g., discount offerings are widespread. Gubela et al. 2020 mention in their eCommerce datasets discounts of 10% to stimulate a purchase during a website visit. Depending on the branch or the product group, the discounts offered to customers per mail, inlays, or newsletter could even be higher. So, e.g., in online fashion shops, discount offerings of 20% are quite common. Moreover, in online furniture shops, even discounts up to 50% and more are frequent. Alternative purchase stimuli are, e.g., vouchers, attached gifts, bonus programs, tombolas, and raffles. However, since profit margins for online shops are typically low (e.g., between 5% and 15%, sometimes up to 40%), these discounts, vouchers, and gifts are double-edged swords: They could generate more revenue but at the same time reduce profit at the customer level dramatically. Consequently, a scoring system is needed that relates offerings, (past) information, and shopping behavior to profit uplift.

(2) **Translate the managerial problem to a predictive model**: As Blattberg et al. (2008, p.°250) discuss in their overview, widespread and useful predictors for binary and continuous response outcomes (e.g., visit, purchase, revenue, profit) in database marketing response are

- customer characteristics (demographics, lifestyle, psychographics),
- previous behavior (purchases and responses to previous marketing efforts, typically described using recency, frequency, and monetary value (RFM) variables), and
- previous marketing (efforts targeted at the customer, including catalogs, Emails, discounts).

Similar variables have been used in the uplift modeling literature. So, e.g., the Hillstrom dataset embodies as customer characteristics the living environment (rural, suburban, urban), as previous behavior recency (time since last purchase), history (money spend in the last year), men's and women's (indicators for product categories bought in the last year), and newbie (indicates a first purchase in the last year), and as previous marketing the used shopping channels. Additionally, nowadays, for online shops, variables that describe the online information behavior are tracked and used, e.g., the duration and recency of shop visits or the number of page views (see, e.g., Gubela et al. 2020). These predictors should also be used in our profit uplift modeling approaches, if available. As outcome of our predictive model – in contrast to the already published revenue uplift modeling approaches – we define the profit outcome (in case of an indirect estimation and prediction similar as in formula 2) or the profit uplift outcome (in case of a direct estimation and prediction similar as in formulae 3 and 4 with transformed response outcomes) at the customer level. The calculation of the profit outcomes from the revenue outcomes depends in the treatment group from the offering to the customer (discount, bonus, vouchers, attached gifts, bonus programs, tombolas or raffles) and the margin; in the control group, it only depends from the margin. Especially the calculation of the latter is a critical point since, in online shopping, the clear allocation of item-related costs to purchase is difficult since besides the supply costs also return, damage, loss, and other aspects would have to be taken into account. Nevertheless, we follow the argumentation by Blattberg et al. 2008 and Gubela et al. 2020 in their determination of cut-off points for customers where average values across product categories or shops were used (and are available in online shops).

(3) **Sample customers for collecting responses**: As usual in uplift modeling, the dependency of outside-effects can be reduced if the treatment and the control groups are random samples of the customer base, ideally balanced with respect to selected predictors (e.g., recency, frequency, monetary value). Moreover, since revenue- or profit-generating responses to direct marketing campaigns typically are rather low (e.g., 0.9% purchasers in the Hillstrom dataset), the drawing of large samples is necessary to develop stable models.

(4) **Calibrate and validate the predictive model**: The small percentage of purchasers in the treatment group and the control group reduces the number of applicable models and parameter estimation algorithms considerably. In fact, the revenue- or profit-generating response can be seen – simplified – as a two-stage process that has to be modeled: In a first stage (few) customers

decide to purchase items (independent of being treated or not): We have a traditional response model with binary outcome. In the second stage, only the revenue or profit-generating behavior of the purchasers is modeled: Here, we have a revenue or profit uplift model. The predictors in both model stages could be the same or different ones. If we use formulae (3) and (4) for this purpose (the direct uplift modeling approach), we observed negative and positive revenue or profit outcomes that can be transformed in "normal shape" by a Box-Cox-transformation. If we use formula (2) for this purpose (the indirect modeling approach with an interaction uplift model) and transform the revenue or profit data to counts (e.g., rounded cents to preserve variability), we observe non-negative count data with "negative binomial shape." For both two-stage modeling cases, well-known parameter estimation procedure exist:

- In the first case with revenue or profit outcomes of the purchasers in "normal shape," Heckman's sample selection model (Heckman 1979) – also called Tobit-2 model – can be applied (Toomet and Henningsen 2008). This model can be described by two equations:

$$(5) \qquad \begin{aligned} Y_i^{S^*} &= \beta^{S^{*\prime}} \mathbf{x}_i + \varepsilon_i^{S^*} \\ Y_i^{O^*} &= \beta^{O^{*\prime}} \mathbf{x}_i + \varepsilon_i^{O^*} \end{aligned} \quad with \quad \begin{pmatrix} \varepsilon_i^{S^*} \\ \varepsilon_i^{O^*} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & \sigma^2 \end{pmatrix} \right)$$

where $Y_i^{S^*}$ represents the selection tendency (here: purchasing tendency) for individual i and $Y_i^{O^*}$ The latent (revenue or profit) outcome. We observe the binary outcome Y_i^S and – for the selected cases (the purchasers) – the continuous outcome Y_i^O as follows

$$
(6) \qquad Y_i^S = \begin{cases} 0 & if\ Y_i^{S^*} < 0 \\ 1 & otherwise, \end{cases} \qquad Y_i^O = \begin{cases} 0 & if\ Y_i^S = 0 \\ Y_i^{O^*} & otherwise. \end{cases}
$$

The conditional regression estimation proposed by Heckman (1979) applies the so-called Heckman correction (inverse of Mill's ratio) to eliminate the sample selection effect.

- For the second case – revenue or profit outcomes converted to count data, count models can be applied. Here, again, the (few) purchasers induce many zeros in the count outcomes, which can be reflected again by a two-equation model, the so-called Zero-inflated Poisson regression model by Lambert 1992 or – as used in our paper – the more flexible Zero-inflated negative binomial regression model by Ridout et al. 2001. In both models, the standard Poisson or Negative binomial model regression model is used as a second stage model and again combined with a selection model for non-negative counts.

As a third alternative, a random forest regression model can be applied (e.g., CART by Breiman 2001) according to formula (3) and (4) since random forests are known in machine learning for being very robust against skewed distributed data with few purchasers and consequently few positive (revenue or profit) outcomes. As usual in predictive modeling, a partitioning of the data in train and (hold out) test data is needed to control the predictive validity in the test data (here: with respect to Qini coefficients). Also, a partitioning of the train data in calibration and validation data to tune hyperparameters of the algorithms within the train is widespread. According to many authors in the uplift modeling literature (e.g., Devriendt et al. 2018, Gubela et al. 2020), here, especially the preprocessing of the predictors and a selection of not too much (say 5 to 15 according to Devriendt et al. 2018) predictors are important for calibration and validation.

(5) **Apply the model to all customers and select "best" customers**: The calibrated, validated, and tested profit uplift model is then used to score the customers and to select profitable ones for the direct marketing campaign. Since the predicted score, the profit uplift per customer is informative; usually, a concentration on customers with scores larger than 0 could be a standard strategy.

In the following two sections, the discussed three new profit uplift modeling approaches (based on Heckman's sample selection model, zero-inflated negative binomial regression, and random forest) are applied to demonstrate their usefulness.

4. Application to Direct Marketing Campaigns of a German Online Shop

4.1 Company, Campaigns, Descriptive Uplift Statistics, and Preprocessing of the Data

The data for the first application was provided by one of the pioneers in the mail order business in Germany, the BAUR group. Today, website www.baur.de is one of the ten largest online shops in Germany. Clear customer and service orientation, high-quality standards, and a constantly up-to-date range of items in the fashion, shoes, and furniture product range are assumed to be key success factors (see Baier et al. 2019). The company mainly focuses on customers aged 40 to 55 and offers well-known brands as well as exclusive fashion branded by BAUR. The online market presence – which represents around 90% of the business volume – is supported by catalogs that focus on seasonal or special fashion topics. Like many other online shops, scoring systems are used to select customers for direct marketing campaigns. The development of an effective scoring system is an ongoing central challenge for the company. Therefore, on a regular basis, tests are performed: Random samples of customers are divided into treatment and control groups according to balanced designs. Then, the customers of the treatment groups are offered discounts (e.g., by mails), and the purchases of the customers of both groups are tracked in the follow-up weeks and used to refine the scoring system.

The provided data reflects two recent tests. Altogether 155,388 selected different customers were divided up into treatment and control groups. The customers in the treatment groups received a 20% discount offer for the next order; the purchases of both groups were tracked in the follow-up weeks. Table 3 reflects the descriptive uplift statistics of these two tests. As one can easily see, the sampling resulted in equally large treatment and control groups. It should be mentioned that for both tests, the samples were selected randomly out of the company's customer base (without overlap) and that the dividing up of the two samples into treatment and control groups was performed in a balanced manner with respect to pre-defined variables that describe the customers' past information and buying behavior, e.g., their purchase volume in the last two years, their usage of the website, as well as the recency of their visits and purchases.

Table 3 Descriptive uplift statistics of the BAUR dataset (with margin m=0.3 and discount d=0.2)

Group	*Share (%)*	*Custo-mers*	*Pur-chasers*	*Purch. rate (%)*	*Purch. upl. (%)*	*Rev./ purch. (€)*	*Rev./ cust. (€)*	*Rev./ purch. upl.(€)*	*Profit/ purch. (€)*	*Profit/ cust.(€)*	*Profit/ purch. upl.(€)*
Treat.	49.97	77,648	9,133	11.76	5.02	183.04	21.53		18.30	2.15	
Control	50.03	77,740	5,244	6.75		156.32	10.55	10.98	46.90	3.17	-1.01
Total		155,388	14,377	9.25		173.92	16.04		28.73	2.66	

A closer look into Table 3 shows that the two tested campaigns were very successful with respect to purchase rates as well as revenue per purchase and revenue per customer: Whereas only 6.75% of the customers in the control groups purchased in the two weeks after the campaign, 11.76% in the treatment groups did so. The purchasers in the treatment groups bought on average items worth 183.04 €, whereas in the control groups, the bought items per purchaser were only worth 156.32 €

on average. This difference is even more striking when taking all customers in the two samples into account (21.53 € per customer in the treatment groups vs. 10.55 € in the control groups).

However, Table 3 also shows a major problem with discount offers. Assuming a (disguised) margin of m=0.3 (30%) and a discount of d=0.2 (20%), the profit per purchaser and the profit per customer in the treatment group (10% of the averaged revenue) is clearly lower than the ones in the control group (30% of the averaged revenue). This results in an overall profit per purchase uplift of the tests of -1.01 €: Offering the discount to all customers in the company's customer base seems to increase the overall revenue, but it would decrease the overall profit. So, a concentration on customers with positive uplift predictions and the development of a predictive scoring system is necessary.

The provided data from the two tests were randomly partitioned into a train set (~70% or 77,617 customers) and a hold-out test set (~30% or 77,771 customers). Additionally, for parameter tuning, the train set was randomly partitioned into a calibration set (~40% or 44,353 customers) and a validation set (~30%, 33,264). For all customers, besides the above-discussed variables that describe the belonging to the treatment and to the control groups, the purchase information, and the generated revenue, altogether 472 metric variables with a non-zero variance that describe their past information and buying behavior were available. Table 4 gives a short description of the 472 variables.

Based on the train set, the 472 variables were preprocessed by setting means to zero, setting standard deviations to 1, and applying a Box-Cox-transformation to transform skewed distributed variables into "normal shape." Moreover, since the variables were highly correlated and – according to Devriendt et al. 2018 – the "best" number of predictors for uplift models has proven to be low (say 5 to 15), the variables were transferred to principal components. Here, the first 88 principal components accounted for 95%, the first 55 for 90%, and the first 20 for 75% of the variance in the transformed training data. The same preprocessing (including the transformation into principal components) was applied to the test set, using the transformation parameters and coefficients derived from the train set.

Table 4 472 Variables of the BAUR dataset that describe past information and buying behavior

Variable category	*Number of variables*	*Description*
Recency	23	Variables that count days since last order (w.r.t. discount types, item categories, and time slots)
Frequency	193	Variables that count past orders (w.r.t. discount types, item categories, and time slots)
Monetary value	191	Variables that reflect past revenues (w.r.t. discount types, item categories, and time slots)
Shop visit	14	Variables that describe the online information behavior (w.r.t. number of visits, visit duration, basket size, and value across time slots and item categories)
Sensitivity to recommendations	3	Variables that describe the number of orders and their value due to recommendations (w.r.t. time slots)
Sensitivity to discounts	26	Variables that describe the share of orders with discounts to all orders in the past (w.r.t. discount types, item categories, and time slots)
Return behavior	22	Variables that describe the number of returns and their value (w.r.t. time slots)

4.2 Applying the Profit Uplift Modeling Approaches

As described in section 3, three profit uplift modeling approaches were used for training and testing a scoring system:

- **Heckman**: The Heckman selection model (Heckman 1979) is estimated based on the binary outcome (purchase) and – in case of a predicted purchase – on the profit uplift. For parameter estimation, first, the profit for all purchasers is derived from the observed revenues multiplying by the margin (m=0.3) for the purchasers in the control group and by the margin-discount (m-d=0.1) for the purchasers in the treatment group. Then, the profit outcome is transformed to "observed" profit uplift outcomes according to formula (4), and the

Heckman selection model is estimated. Finally, profit uplift predictions can be directly derived for all customers using formula (3) via formulae (5) and (6). The R package sampleSelection is applied.

- **RF**: The random forest model is widespread in uplift modeling. We use the ranger implementtation in R (Wright and Ziegler 2017) of the classical approach for modeling continuous outcomes by regression (Breiman 2001). The same continuous outcome, as with the Heckman selection model, is used for parameter estimation. Again, predictions for the profit uplift outcome can be derived for all customers directly from the predictors according to formula (3).

- **Zeroinfl**: The zero-inflated Poisson regression model (Lambert 1992) and its zero-inflated negative binomial regression model alternative (Ridout et al. 2001) assume non-negative count data as input. Therefore, first, the observed profit has to be converted to cents (to preserve variability) and to be rounded. Also, as discussed in section 3, an interaction model is needed that includes the treatment indicator (1 for customers in the treatment group, 0 for the others) and its interactions with the other predictors. The estimated interaction model is then used for predicting the profit uplift as the difference between the profit uplift when the treatment indicator is set to 1 and the profit uplift when it is set to 0 according to formula (2). In our applications, we use the zero-inflated negative binomial regression model due to overdispersion in the training dataset. The R package pscl is applied.

Before estimating the models based on the train data and comparing the results on the test data – as usual in machine learning – reflections on performance evaluation and parameter tuning are necessary. As already discussed in section 3, the incremental profit uplift curve and the derived profit Qini coefficient are suitable measures for this purpose. Since only one observation per customer is available in the data (profit if treated or profit if not treated due to the belonging to the treatment or the control group), for calculating uplifts a grouping of customers and comparing average profits of treated and not treated customers in each group is needed. This grouping is based on sorting the customers according to the developed scoring system (starting with the customers where we assume the highest profit uplift) and forming quantiles (usually deciles) of the sorted customers. Basing on these groupings, now, the incremental profit uplift across the quantiles can be plotted (the incremental profit uplift curve), and the area between this curve and a curve derived by random sorting (the profit Qini coefficient Q) can be calculated and used for selecting "best" scoring systems.

Figure 2 shows the profit Qini coefficients for the three discussed models when estimated with varying numbers of predictors on the basis of the calibration subsample of the train data and used for predictions on the basis of the validation sample. It can be easily seen that the profit Qini coefficients are low with small numbers of predictors as well as with high numbers of predictors. These findings are consistent with the findings of Devriendt et al. 2018, who found in their comparison of binary uplift models that 5 to 15 predictors typically provide the best results. Against this background, we decided to use 20 predictors in the following for training and testing the three final profit uplift models.

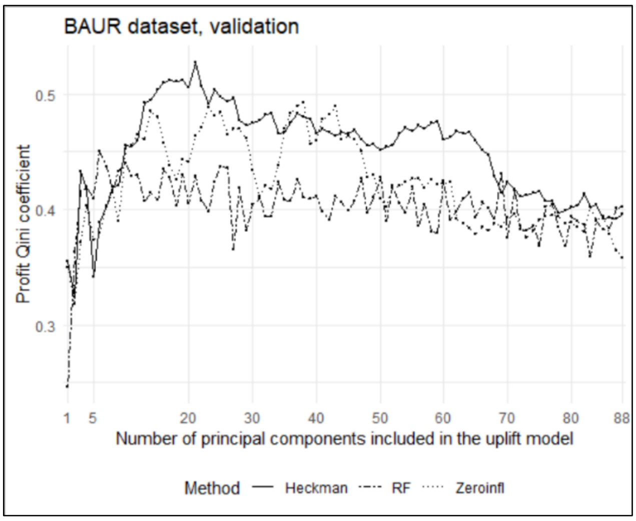


Fig. 2 Profit Qini coefficients for the validation set (30% of the BAUR dataset) based on training the profit uplift modeling approaches on the calibration set (40% of the BAUR dataset)

Table 5 Results of the application of profit uplift modeling approaches to the BAUR dataset (20 principal components) for the train set (70% of the data, used for model training) and for the test set (30% of the data)

Profit uplift modeling approach	*Profit Qini coefficient for the train set*	*Profit Qini coefficient for the test set*
Heckman	0.4922	**0.4298**
RF	**0.5538**	0.3436
Zeroinfl	0.4614	0.4062

In Table 5 and Figure 3, the results of this modeling are illustrated. The results reflect, to some extent, the results of parameter tuning: The Heckman model performs best with respect to the hold-out test set, followed by the Zeroinfl model and RF. However, as Figure 3 demonstrates, all three models provide quite similar results, which is – to some extent – surprising since the modeling assumptions (count data vs. "normal shape," direct model vs. difference of two predictions based on the interaction model) and the estimation algorithms (one-step vs. two-step models) are very different. The application shows that it seems to be possible that – besides already existing binary uplift and revenue uplift models, it is also possible to estimate profit uplift models which show clear, practical advantages because they model for sorting a score which reflects the – for companies – most important criteria, the profit uplift.

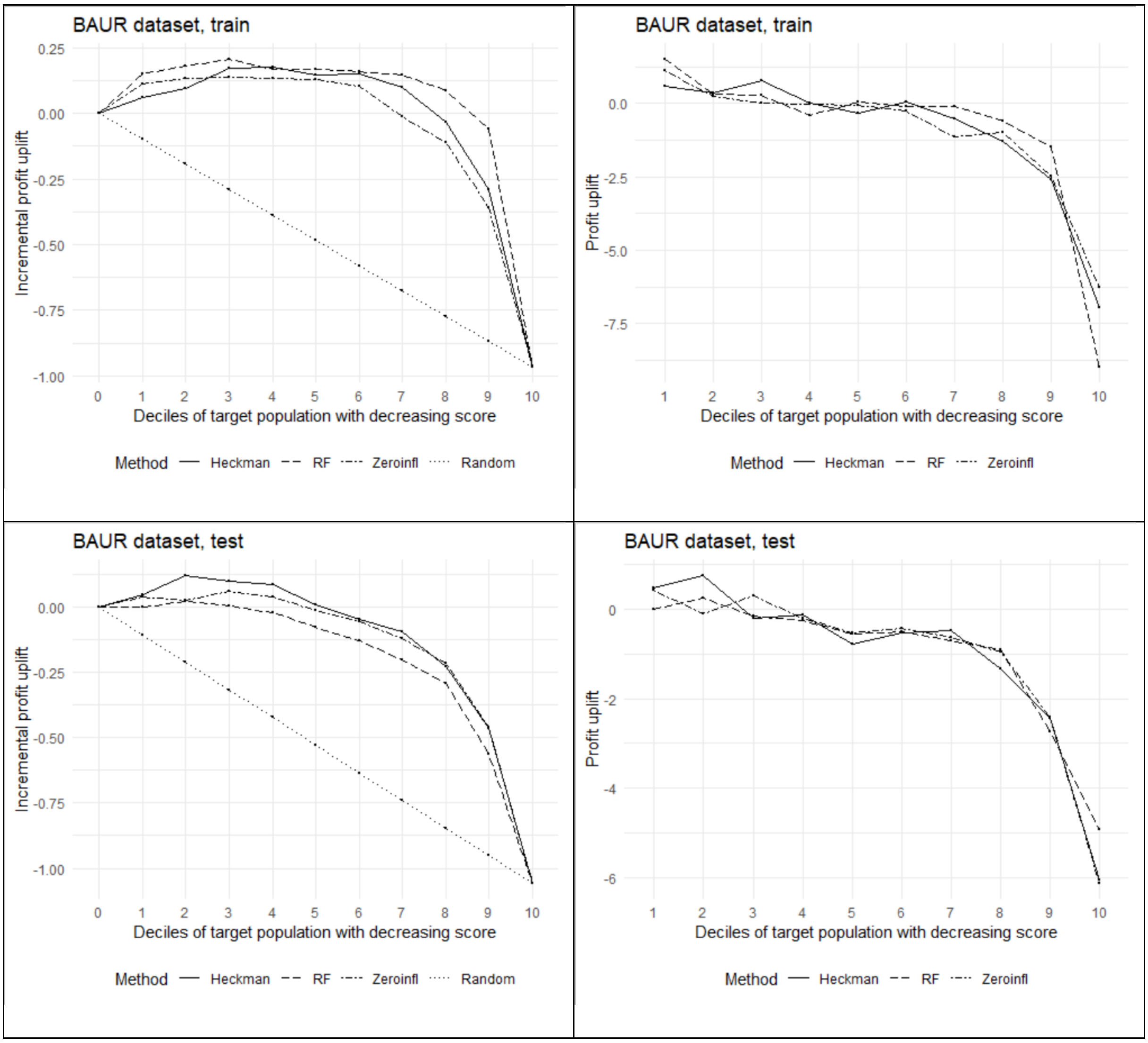


Fig. 3 Results of the application of profit uplift modeling approaches to the BAUR dataset (20 principal components) for the train set (70% of the data, used for model training) and for the test set (30% of the data)

5. Application to the Hillstrom Dataset

In order to demonstrate that the proposed profit uplift modeling approaches are applicable, also a standard dataset from the uplift modeling literature is analyzed, the Hillstrom dataset (Radcliffe 2008). The dataset was made available by Kevin Hillstrom through his MineThatData blog and described a sample of 64,000 customers which had been divided up into three nearly equally sized subsamples, two of them contacted via two direct marketing campaigns and one not contacted, serving as a control group (see the similar usage of this dataset in Rudaś and Jaroszewicz 2018). Table 6 summarizes the descriptive uplift statistics of this dataset, where the two treated subsamples are merged. As can easily be seen, the conversion rate is much lower as in the BAUR dataset (on average 1.07% in the treatment group) but nevertheless shows a conversion rate uplift compared to the control group (on average, an uplift of 0.50%). The revenue uplift per customer is 0.60$, but this uplift seems to be solely from the conversion rate uplift since the average revenue spends by a purchaser in the treatment group (117.00$) is only slightly higher than in the control group (114.00$). Again, as in the BAUR dataset, we assume that the campaign offers a 20% discount and that the margin for the retailer is 30%. With these assumptions (not part of the original communication of the dataset, just an assumption to be able to analyze the dataset with our profit uplift modeling approaches), the overall profit uplift per customer is negative (-0.07$). So, again we have to develop a scoring system that helps to restrict the direct marketing campaign to customers with a positive profit uplift prediction.

Table 6 Descriptive uplift statistics of the Hillstrom dataset (with margin m=0.3 and discount d=0.2)

Group	*Share (%)*	*Customers*	*Purchasers*	*Conv. rate (%)*	*Conv. upl. (%)*	*Rev./ conv. ($)*	*Rev./ cust. ($)*	*Rev./ cust. upl.($)*	*Profit/ conv. ($)*	*Profit/ cust. ($)*	*Profit/ cust. upl.($)*
Treat.	66.71	42,694	456	1.07	0.50	117.00	1.25		11.70	0.12	
Control	33.30	21,306	122	0.57		114.00	0.65	0.60	34.20	0.20	-0.07
Total		64,000	578	0.90		116.36	1.05		16.45	0.15	

The original dataset also contains potential predictors for this scoring system, as given in Table 7. The original eight potential predictors (in Table 7 described as variable categories) were scaled nominally (e.g., history_segment with 7 values or channel with three values) or metrically (e.g., recency or history). For our further analysis with the three models, we dummy-coded the nominally scaled potential predictors and so received in total 25 metrically scaled variables (see Table 7).

Table 7 Variables of the Hillstrom dataset that describe past buying behavior (-1 indicates that one indicator is dependent on the others and therefore, is omitted for estimation)

Variable category	*Number of variables*	*Description*
Recency	12 (-1)	Indicators for months since last purchase (1,…,12)
History_ segment	7 (-1)	Indicators for revenue categories last year ([0,100$), [100$,200$), [200$,350), [350$,500$), [500,750$), [750$,1000$), [1000$,)
History	1	Revenue generated last year (in $)
Men's	1	Indicator whether customer bought men's merchandise last year
Women's	1	Indicator whether customer bought women's merchandise last year
Zipcode	3 (-1)	Indicator whether the customer's zip code is rural, suburban, urban
Newbie	1	Indicator whether customer bought last year the first time
Channel	3 (-1)	Indicator whether customer bought last year via phone, web, both

As in section 4, the data were randomly partitioned into a train set (~70% or 44,800 customers) and a holdout test set (~30% or 19,200 customers), and the train set was preprocessed by setting means to zero, setting standard deviations to 1, and applying a Box-Cox-transformation to transform skewed distributed variables into "normal shape." The same preprocessing was applied to the test set, using the transformation parameters derived from the train set. Then the three models, as in section 4, were applied, which resulted in the profit Qini coefficients of Table 8 and the incremental profit uplift curves and profit uplifts across deciles of customers in Figure 4.

Table 8 Results of the application of profit uplift modeling approaches to the Hillstrom dataset (25 variables) for the train set (70% of the data, used for model training) and for the test set (30% of the data)

Profit uplift modeling approach	*Profit Qini coefficient for the train set*	*Profit Qini coefficient for the test set*
Heckman	0.0392	0.0120
RF	**0.0848**	**0.0139**
Zeroinfl	0.0430	0.0123

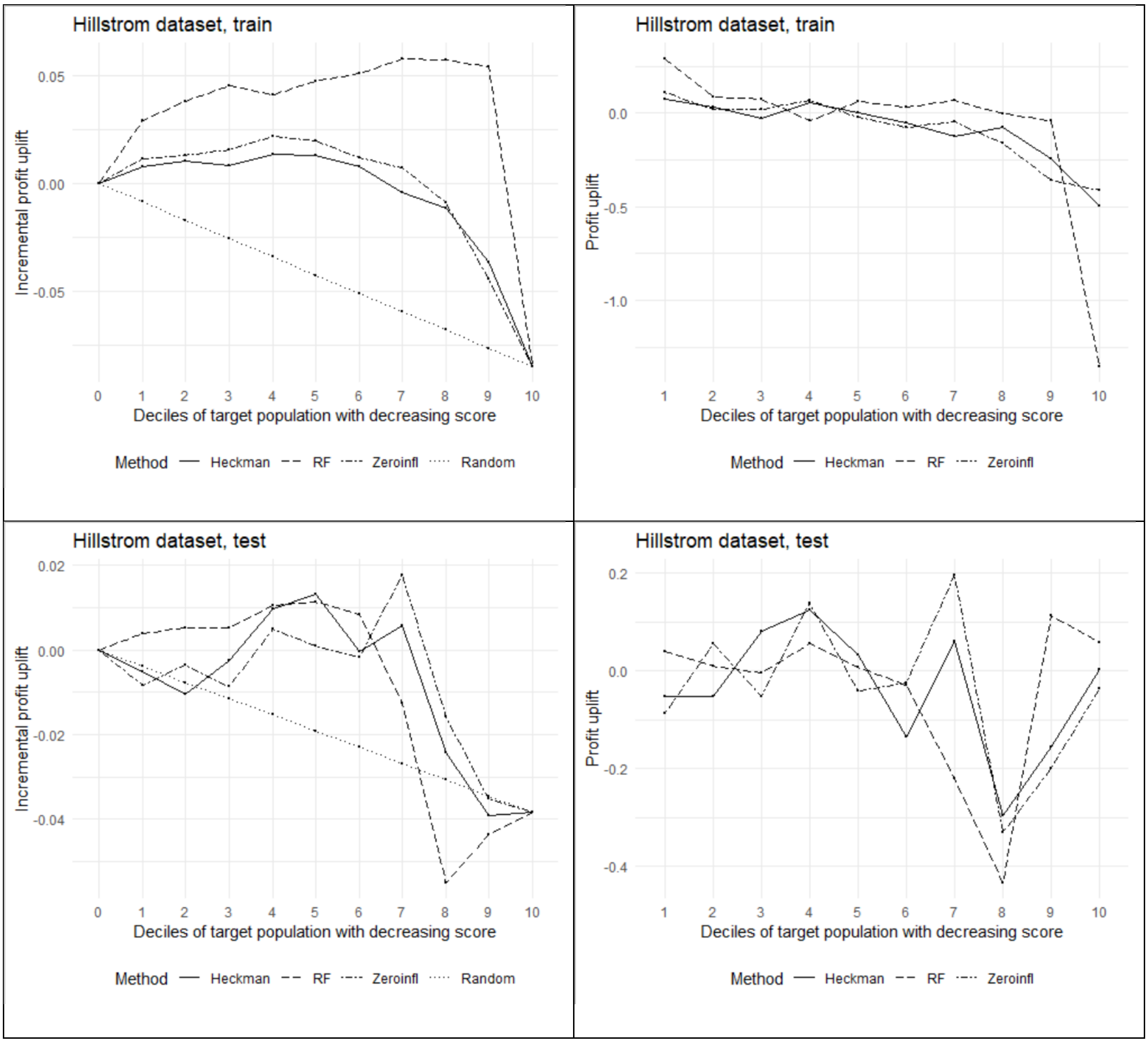


Fig. 4 Results of the application of profit uplift modeling approaches to the Hillstrom dataset

One can easily see that three models, again, show similar results with random forest providing the best performance but also that due to the few purchasers in the dataset with a high concentration of revenues on a few purchasers leads to a worse performance compared to the application of the BAUR dataset. The problem of the Hillstrom dataset when it comes to modeling metric outcomes has also been discussed by other authors; here, we refer to the analysis in the paper by Rudaś and Jaroszewicz (2018).

6. Conclusions and Outlook

In this paper, we introduced a new approach to uplift modeling, the so-called profit uplift modeling approach. In contrast to former revenue uplift modeling, these approaches directly model the individual profit uplift, sort the customers according to the important profit criteria, and don't need an unrelated second step to transform modeled revenues to aggregate profits. Three different approaches were applied to two available datasets: One based on the Heckman sample selection model, where the observed binary outcome (purchase or not) and the observed continuous outcome (positive profits for the treatment group, negative profits for the control group) is modeled, one based on the zero-inflated negative binomial model, where an interaction model is used to model the observed count data, and one using traditional random forest regression to predict individual profit uplifts. The three approaches are based on very different assumptions but nevertheless provide quite similar prediction results with a clear ordering of the customers according to their predicted profit uplift. These results support the meaningfulness of the approaches via cross-validation. Of course, further research is needed. So, e.g., the three approaches have to demonstrate their usefulness also with other datasets.

References

Baier D, Rese A, Nonenmacher N, Treybig S, Bressem B (2019) Digital technologies for ordering and delivering fashion: How Baur integrates the customer's point of view. In: Urbach N, Röglinger M (eds) Digitalization cases. Springer, Cham, pp 59–77

Blattberg RC, Kim B-D, Neslin SA (2008) Database marketing - Analyzing and managing customers. Springer, New York, NY

Breiman L (2001) Random forests. Machine Learning 45:5–32

Devriendt F, Moldovan D, Verbeke W (2018) A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping stone toward the development of prescriptive analytics. Big Data 6:13–41. https://doi.org/10.1089/big.2017.0104

Gubela RM, Lessmann S, Jaroszewicz S (2020) Response transformation and profit decomposition for revenue uplift modeling. European Journal of Operational Research 283:647–661. https://doi.org/10.1016/j.ejor.2019.11.030

Guelman L, Guillén M, Pérez-Marín AM (2015) Uplift random forests. Cybernetics and Systems 46:230–248. https://doi.org/10.1080/01969722.2015.1012892

Hansotia B, Rukstales B (2002) Incremental value modeling. Journal of Interactive Marketing 16:35–46. https://doi.org/10.1002/dir.10035

Heckman J (1979) Sample selection bias as a specification error. Econometrica 47:153–161

Kane K, Lo VSY, Zheng J (2014) Mining for the truly responsive customers and prospects using true-lift modeling: Comparison of new and existing methods. J Market Anal 2:218–238. https://doi.org/10.1057/jma.2014.18

Lai LY-T (2006) Influential marketing: a new direct marketing strategy addressing the existence of voluntary buyers. Master Thesis, School of Computing Science - Simon Fraser University

Lambert D (1992) Zero-inflated poisson regression, with an application to defects in manufacturing. Technometrics 34:1–14

Lo VSY (2002) The true lift model. ACM SIGKDD Explorations Newsletter 4:78–86

Radcliffe NJ (2007) Using control groups to target on predicted lift: Building and assessing uplift models. Direct Marketing Analytics Journal 1:14–21

Radcliffe NJ (2008) Hillstrom's MineThatData email analytics challenge: An approach using uplift modelling, Edinburgh, UK

Radcliffe NJ, Surry PD (1999) Differential response analysis: Modeling true response by isolating the effect of a single action. In: Proceedings of Credit Scoring and Credit Control VI. Credit Research Center, University of Edinburgh Management School

Radcliffe NJ, Surry PD (2011) Real-world uplift modelling with significance-based uplift trees, Technical Report TR-2011-1. Stochastic Solutions

Ridout M, Hinde J, Demétrio CG (2001) A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. Biometrics 57:219–223. https://doi.org/10.1111/j.0006-341X.2001.00219.x

Rudaś K, Jaroszewicz S (2018) Linear regression for uplift modeling. Data Min Knowl Disc 32:1275–1305. https://doi.org/10.1007/s10618-018-0576-8

Rzepakowski P, Jaroszewicz S (2012) Decision trees for uplift modeling with single and multiple treatments. Knowl Inf Syst 32:303–327. https://doi.org/10.1007/s10115-011-0434-0

Sołtys M, Jaroszewicz S, Rzepakowski P (2015) Ensemble methods for uplift modeling. Data Min Knowl Disc 29:1531–1559. https://doi.org/10.1007/s10618-014-0383-9

Toomet O, Henningsen A (2008) Sample selection models in R: Package sampleSelection. J. Stat. Soft. 27. https://doi.org/10.18637/jss.v027.i07

Wright MN, Ziegler A (2017) ranger: A fast implementation of random forests for high dimensional data in C++ and R. J. Stat. Soft. 77:1–17. https://doi.org/10.18637/jss.v077.i01

Zaniewicz L, Jaroszewicz S (2013) Support vector machines for uplift modeling. In: 2013 IEEE 13th International Conference on Data Mining Workshops. IEEE

Zaniewicz Ł, Jaroszewicz S (2017) L_p-support vector machines for uplift modeling. Knowl Inf Syst 53:269–296. https://doi.org/10.1007/s10115-017-1040-6

Chapter 5

A Better Understanding of Cost-related Dependencies in the Estimation of the Causal Effects in Direct Marketing Campaigns

Björn Stöcker and Daniel Baier

Abstract:

The economically optimal customer selection for a direct marketing campaign, such as a discount offer via a newsletter, is challenging. On the one hand, one experiences classically low responses to these campaigns (~2%). On the other hand, in A/B test scenarios, one repeatedly finds out that the control group also makes not inconsiderable sales, negatively influencing the campaign's profitability. The causal effect modeling tries to counteract this by only contacting customers who buy because of the treatment. The recent literature almost reduces this to a classification problem. In this article, we give a holistic view for the first time by looking at different cost structures, deriving new selection strategies, and validating them with new metrics on a real data set of an e-commerce retailer from Germany. We can show that the modeling's economic success depends significantly on the cost structure and the selection strategy.

This chapter is under review in:

International Journal of Research in Marketing

1. Introduction

The management of direct marketing campaigns involves an interesting question: Which customers should be addressed? Marketing campaigns have the goal of changing customers' behavior. This goal is usually relatively simple: the company wants the campaign to generate additional revenue with the best possible return on investment (ROI). Which customer is particularly suitable for the measure and whether there may also be customers that it is better not to address is the subject of this optimization. The scientific question is: can the behavior change be attributed to a specific measure, or would the customers' development not have occurred even without the campaign? In other words: is the marketing campaign causal for the behavior change? Consequently, only customers whose behavior changes positively as a result of the campaign should be addressed. Some important works have already been published on this topic - see chapter 2.3.

Our work focuses on a specific application at a large German e-commerce retailer for fashion. Customers regularly receive purchase incentives in the form of catalogs, coupons, and discounts. Again, it was observed that in randomized experiments with control groups, a considerable proportion of customers had ordered even without the campaign. However, the application of the previously known and published methodology had not led to a significant improvement in ROI compared to the previous application of a response score. However, there is extensive evidence in the literature that a response score performs worse than calculating the causal effect in the context of ROI.

Our analysis found a noteworthy part not been considered in the literature yet: how costs are incurred in a marketing campaign. Chapter 2.2 shows that the previous way of calculating the causal effect estimation for some cost types needs to be adjusted to ensure an ROI-optimal sorting of customers. Furthermore, we also deal with the target figure to be optimized. In our opinion, this is also insufficiently covered in the literature. What is the expression of the changed customer behavior, and what effects does this have on the ROI? Concretely: A widespread application of the method determines the difference based on the purchase probability. However, if a customer buys more due to the marketing campaign, he will fall through this grid.

To prove our theoretical considerations in practice, we introduce a new data set in chapter 3.1. It consists of real data from two marketing campaigns collected in a randomized trial, including a control group (~300k). The data includes past ordering behavior and target figures for the sales achieved during the campaign period and its costs. To consider the relationship between ROI and the different cost types, we add simulated costs to the data set. In the data set presented, the marketing campaign increases total purchases. To get an impression of how the results change in case of increased turnover per customer, we transformed the data set.

Based on these data sets, we develop two models that predict the behavior in case of a marketing campaign and without. Measuring the model quality also presented us with challenges. In our opinion, the methods used so far are not suitable to provide reliable results for different cost types. Therefore, in chapter 2.4, we generalize the widely used metrics, introduce new measures, and compare them with previous ones.

In chapter 3.5, we now perform the calculations on 100 randomly selected data splits to test our assumptions. A discussion of the results follows this.

2. Theoretical Framework and Literature Review

To approach the dependencies of the cost types' causal effect, we combine two frameworks: First, we focus on the decision rule when a customer should be contacted. For this purpose, we use the approach from cost and activity accounting, which classifies costs according to their characteristics. The second framework includes the statistical calculation of the causal effect as such. After combining both concepts, we look at the work published so far and point out the research gap.

2.1 Definition of the Decision Rule, the Emergence of Cost in Marketing Campaigns, and Their Influence on ROI

To select the right customers for a marketing campaign, we must first ask ourselves the question of the decision-making rule: on what basis do we decide whether to include a customer in a marketing campaign or not. The objectives of marketing campaigns, in general, can be extremely diverse.

Since we want to examine the effect of different cost types, we will consider the contribution margin for this work. The underlying decision rule is then: A customer i is to be selected for a campaign, e.g., treatment ($\tau_i \in \{0,1\}$) if the expected contribution margin (cm_i) is higher than the expected cm_i if he is not contacted, given the vector of predictor variables X_i.

(1) $$\mathrm{E}\left(cm_i \mid X_i, \tau_i = 1\right) \geq \mathrm{E}\left(cm_i \mid X_i, \tau_i = 0\right)$$

It is now possible for modeling purposes to either set the absolute contribution margin as a target figure or predict the turnover and reconcile it to the contribution margin through a fixed factor. There are different methods to include the marketing costs in the prediction. One could train the model on the contribution margin previously calculated in the training data, where the campaign costs have already been subtracted. Alternatively, one could represent this in the decision rule equation.

Every consideration has its advantages and disadvantages. The calculation of contribution margins is, by far, not standardized and highly individual depending on the company, market, and customer structure. For example, to determine an order's contribution margin in an online shop, one could take the selling price minus the goods' costs per item and add it up. Charges for storage, picking, packaging, and shipping could be subtracted from this margin. If several items from the same order are stored and therefore shipped from different locations, these costs could appear several times. Also, the costs for a possible return, the depreciation of the stock could have an influence, and so on. Moreover, this list can be continued at will. Strictly speaking, these internal company variables would have to be available for modeling to correctly predict the contribution margin.

For didactic reasons, we have decided in this paper to present the absolute contribution margin as the product of the predicted turnover and a constant relative margin (m), diminished by the marketing costs incurred. This definition will help us to understand better how campaign costs and relative margins are related.

To do so, we now add different cost types to Formula 1. Campaign costs are incurred in different ways and therefore depend on different calculation bases. Cost and activity accounting divides costs into two general groups: Costs that change with variable, so-called variable costs, and costs that do not depend on this, so-called fixed costs. We transfer this logic to campaign costs (Table 1). The optimization of campaigns needs to comprehend these different types of correlations because, from this understanding, different optimization strategies emerge.

Table 1 Different cost elements in direct marketing campaigns

Cost Type	*Example*
Fixed Costs / Cost per Contact	Catalogs, call charges cold calling, postage, and production for print-mailings
Response-fixed Costs	Vouchers (15€ for the purchase), free shipping, Add-ons or gifts to the purchase, follow up costs for the offer preparation per lead
Response-variable Costs	discounts (15% on the purchase), buy-one-get-one-free, quantity discount

Based on the different costs (Table 1), different approaches to optimizing the profitability of the campaign emerge:

- Fixed costs or costs per contact implies that contacting a customer costs a fixed rate. The optimization is to avoid probably unsuccessful contacts. Likewise, contacts with a high suspected probability of purchase would buy even without contact. Especially for campaigns with low response rates, the added value from calculating the causal effect appears very small. Let us assume a response rate of 3%: The most prominent effect on ROI is the exact prediction of the 97% futile costs by customers who do not buy. Even with 50% deadweight in the campaign, only 1.5% of the marketing costs are fine-tuned or sorted out in this case. Therefore, it can be expected that a response model will also deliver good results at low response rates.

- Cost per response (order or turnover), on the contrary, only becomes relevant if a purchase has taken place. Unlike costs per contact, it is not essential here, but the order, respectively, the turnover. As it is free to contact all persons, we singularly should contact those customers first, who would indicate the highest expected difference in spending behavior regarding the possible contact or differently expressed, the most positive causal effect in turnover.

2.1.1 Cost per contact and cost per order

If the marketing campaign runs on a fixed cost structure for each customer, the expected return in turnover (r_i) in the case of τ_i=1 multiplied with the relative margin (m) minus the variable costs (c_i) should be equal or higher than the turnover multiplied with the margin in τ_i=0.

(2) $$\mathrm{E}\left(r_i \mid X_i, \tau_i = 1\right) * m - c_i \geq \mathrm{E}\left(r_i \mid X_i, \tau_i = 0\right) * m$$

The break-even is reached when the difference between the expected turnover equals the quotient of variable costs and relative margin.

(3) $$\mathrm{E}\left(r_i \mid X_i, \tau_i = 1\right) - \mathrm{E}\left(r_i \mid X_i, \tau_i = 0\right) = \frac{c_i}{m}$$

Suppose the return is only a binary response. In that case, the equation can be simplified to the differences in the purchase probabilities, ascribed by Radcliffe and Surry (1999), and used in most papers.

(4) $$\Pr\left(r_i \mid X_i, \tau_i = 1\right) \geq \Pr\left(r_i \mid X_i, \tau_i = 0\right)$$

In this case, only, e.g., the submission of newsletter permission is known and nothing else. Therefore, no ROI can be determined here. Still, we would like to encourage every reader to think about how much, e. g. newsletter permission in general or even how much it could be worth for different customer qualities. So again, an ROI can be assumed.

2.1.2 Cost per turnover

Discounts differ from costs per contact or order in that their value is directly dependent on the turnover made. Here too, the discount (d) granted is subtracted in full of the absolute margin. As we do not consider multiple treatments in our paper, d is a constant.

$$\mathrm{E}\left(r_i \mid X_i, \tau_i = 1\right) * m - \mathrm{E}\left(r_i \mid X_i, \tau_i = 1\right) * d \geq \mathrm{E}\left(r_i \mid X_i, \tau_i = 0\right) * m \tag{5}$$

The profit zone is reached when the quotient between sales exceeds a constant value.

$$\frac{\mathrm{E}\left(r_i \mid X_i, \tau_i = 1\right)}{\mathrm{E}\left(r_i \mid X_i, \tau_i = 0\right)} = \frac{1}{\left(1 - \frac{d}{m}\right)} \tag{6}$$

If we imagine an example with a 50% margin and a 20% discount, turnover within a marketing campaign must be at least 66.6% higher than without treatment. Thus, if a discount is applied, it is not the difference, but the quotient appears the correct measure.

2.1.3 The influence of the redemption behavior

Regarding response-depending costs like vouchers and rebates, we can consider another specialty (Figure 1). Being assigned to the treatment group does not necessarily entail costs. Even further, whether costs arise is no longer in the campaign manager's control but in the customers' redeem behavior. Why is this distinction so important? One could easily say that a non-redeemer is equal to a customer not treated. However, this is only valid for the costs. The contact, even if it is for free, can alter the purchase behavior favorably. Consequently, the individual redemption behavior directly influences the costs incurred and, thus, on the ROI.

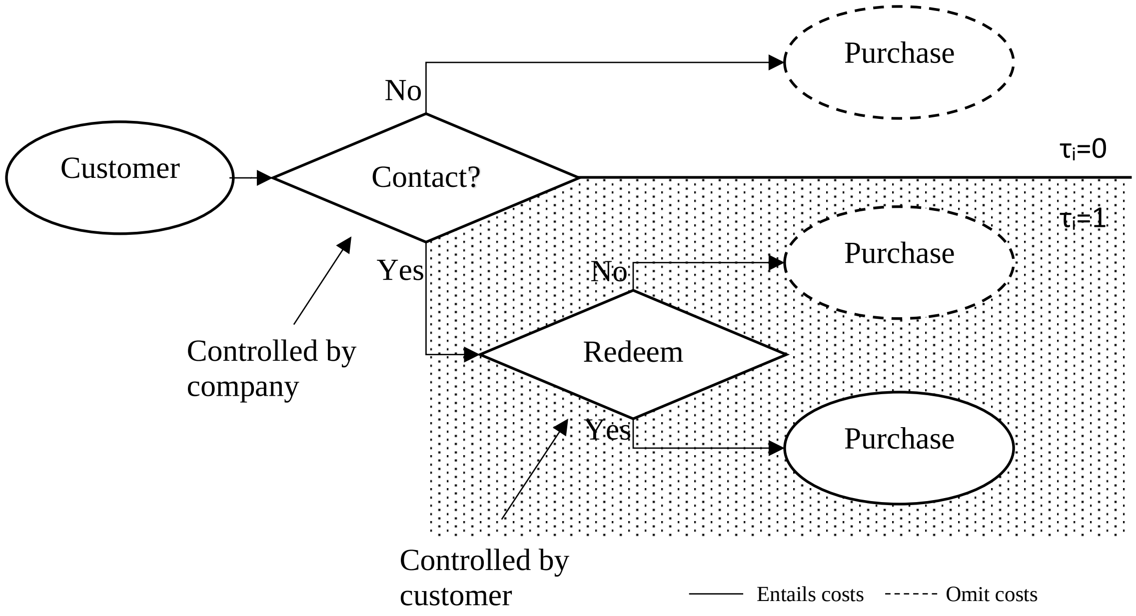


Fig.1 Occurrence of response-related campaign costs

Bawa and Shoemaker (1989) addressed this field at a general campaign level. They investigated how the redemption behavior of coupons affects incremental sales. Among other findings, they proved that customers who had received a coupon promotion but did not redeem it also showed incremental sales growth. We could not find an application to individual customers in the literature.

To achieve ideal sorting, not only the estimated causal effect but also the redemption behavior is critical. If, again, only response-dependent costs are present, a look at the redemption behavior can be useful. Once a customer is assigned to a campaign but does not redeem the coupon, no costs occur. In deriving the causal relationships between the return and the different cost structures, we implicitly assumed a 100% redemption rate, representing the extreme case. Suppose a customer does not redeem a coupon with a sale. In that case, a notable effect arises: Customers who respond positively to a campaign, but do not redeem a coupon, add to the contribution margin with their entire turnover. Therefore, all non-redeeming customers with a positive estimated causal effect should receive the highest score and be selected first.

The considerations are valid from the classification view, too; see Figure 2. First of all, the customers with the highest difference in the estimated causal effect in return and the lowest probability of redemption should be selected—lastly, the customers with the lowest causal effect and the highest redemption probability.

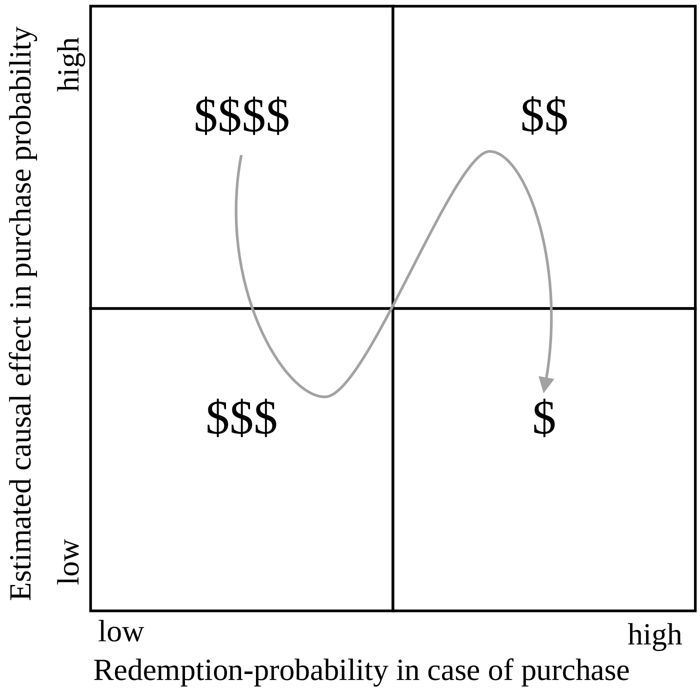


Fig. 2 Ideal scoring order considering the redemption probability if only binary returns are available

2.2 Estimating the Causal Effect Under Consideration of the Different Cost Types

Determining the causal effect is a challenge for us. The methods widely used in statistics, such as classification and regressions, are best suited to predict a customer's behavior. Response modeling predicts which value can be expected under given conditions. For example, when considering the customer lifetime value, also called CLV, predictions are made about how a customer will develop in the future. This prediction provides ceteris paribus, an approximation of future customer development. To determine the causal effect, we do not want to make a prediction ceteris paribus, but rather "what if." A customer cannot be given treatment at the same time and not in order to develop a model afterward.

At this place, Rubins' causal model (Holland 1986) comes in. Rubin describes that by finding statistical twins, a transfer of the prediction becomes possible. In the best case, a customer base is divided into two equal parts by a randomized experiment. One group, we call it the test group, receives a marketing campaign; the other control group receives nothing. The pending variable is now observed in both groups (Table 2).

Table 2 Data sets obtained from a randomized test

ID	X_i	$yi \mid \tau_i=0$	$yi \mid \tau_i=1$	*Estimated causal effect*
1	1,2,3,4	3	?	? –3
2	4,6,7,8,9	7	?	? –7
3	1,2,3,4	?	3	3–?
4	4,6,7,8,9	?	10	10–?

We now get a data set with different customers, their characteristics (also called predictor vector), and the measured reaction. In our very simplified model, customers 1 and 3 and 2 and 4 show the same characteristic values. Rubin now argues that the gaps (marked with "?") can be filled with the other statistical twin (Table 3).

Table 3 Imputation of the missing values and calculation of the causal effect

ID	*Xi*	$y_i \mid \tau_i=0$	$y_i \mid \tau_i=1$	*Estimated causal effect*
1	1,2,3,4	3	3*	0
2	4,6,7,8,9	7	10*	3
3	1,2,3,4	3*	3	0
4	4,6,7,8,9	7*	10	3

* statistical estimation

Two statistical models are developed, one prediction for $\tau_i=1$ and one for $\tau_i=0$ with two statistically equally distributed samples (Rubin 1974, 1977; Holland 1986). The two samples are observed in their behavior at the same time. Thus, a statistical prediction model is developed for both samples, the treatment (τ_i=1), and the control group (τ_i=0).

Let the individual estimated causal effect (u_i) be the conditional propensity to respond r_i, given the predictors X_i.

(7) $$u_i^{dichotomous} = \Pr\left(r_i \mid X_i, \tau_i = 1\right) - \Pr\left(r_i \mid X_i, \tau_i = 0\right)$$

And in the case of a continuous outcome:

(8) $$u_i^{continuous} = \mathrm{E}\left(r_i \mid X_i, \tau_i = 1\right) - \mathrm{E}\left(r_i \mid X_i, \tau_i = 0\right)$$

As in this case, two models, one for τ_i=1 and one for τ_i=0, are developed; this approach is also called the two-model approach.

In his work, Rubin describes the causal effect as a difference. In our example, no causal effect can be determined for customers 1 and 3, and a causal effect for customers 2 and 4. By considering the different relationships in the campaign costs, we supplement Rubins' determination of the causal effect with a quotient.

2.3 Related Work

As already described in our theoretical framework, Rubin and other authors (for the historical origins see Rubin 1990, 2005 or Imbens and Rubin 2006 for economics) have already provided the first approaches to modeling the causal effect. A marketing-specific application followed later.

The literature uses various terms to describe the causal effect in the field of marketing: "uplift modeling" (Devriendt et al. 2018): "differential response analysis" (Radcliffe and Surry 1999), "true-lift modeling" (Lo 2002; Kane 2014), "true response modeling" (Radcliffe and Surry 1999), "net lift modeling" (Larsen 2010), "differential marketing" (Radcliffe and Surry 1999), "incremental value modeling” (Hansotia and Rukstales 2002), “incremental impact modeling” (Hansotia and Rukstales 2002) and “personalized treatment selection” (Zhao et al. 2017a).

For this study, we consider papers related to marketing, modeling strategies of continuous outcomes, and the possible considerations of ROI. Also, we focus on approaches that indirectly calculate the causal effect (Guelman 2014), i.e., calculate two predictions in order to be able to show the different cause-effect relationships in the costs. Direct methods either show the membership in a class (Jaskowski and Jaroszewicz 2012; Lai et al. 2006; Kane 2014; Su et al. 2012; Tian et al. 2014) or the expected continuous causal effect (Tian et al. 2014; Su et al. 2012) and, therefore, cannot be applied here.

The literature on causal effects in marketing is rare, especially in the early years. One reason may be that only a few data sets are available for research, containing data from an A/B test scenario. Therefore, many papers rely on the marketing data set from the so-called Hillstrom Challenge (Hillstrom 2008) or data from medical trials (UCI Repository, Dua, Dheeru, and Graff, Casey 2017).

The first application goes back to Radcliffe and Surry (1999). They use the causal effect estimation as the difference between two purchase probabilities (Pr) or continuous response (E) for the presence of treatment and not. Therefore, the estimated causal effect expresses how much the purchase probability or continuous response increases or decreases in absolute terms through a marketing campaign. Therefore, the goal is to increase efficiency in winning additional responses, conversions, or orders at a given cost per contact. Radcliffe (2007a) classifies customers into four groups based on how likely they are to buy in the event τ_i=1 and τ_i=0 and proposes to contact customers who change their behavior because of the treatment of the so-called „persuadables“ (Figure 3).

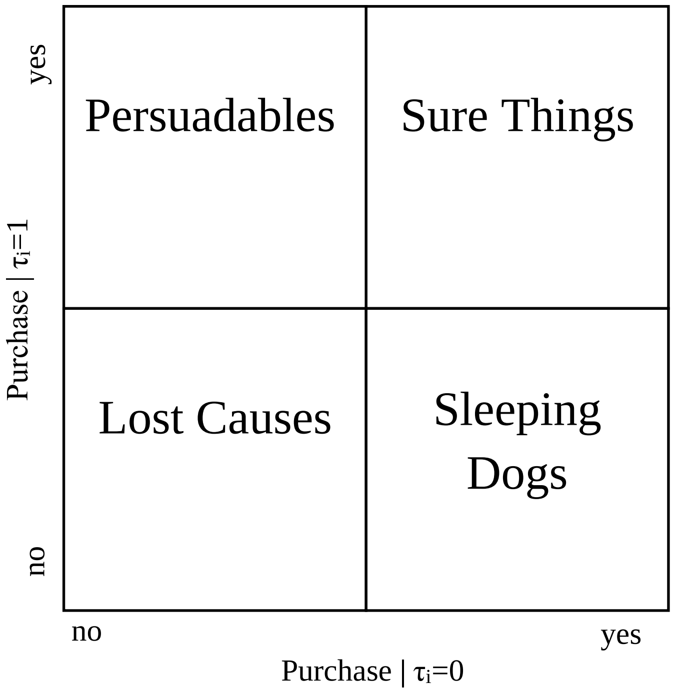


Fig. 3 Which customers should ideally be addressed? (Radcliffe 2007a)

Lo (2002) later suggests using an interaction term instead of the two-model approach. Instead of two models, one model is developed in which a treatment indicator (τ_i) occurs as an interaction term for all independent variables.

$$\mathrm{E}\left(r_i \mid X_i\right) = f\left(X_i, \tau_i, X_i * \tau_i\right) \quad (9)$$

The causal effect is estimated by calculating the function twice, with τ_i=1 and with τ_i=0. The causal effect is again the difference between the two outcomes.

Another noticeable fact is that many researchers in the context of the marketing application approach the estimation of the causal effect with classification methods, but only a few papers focus on continuous outcomes (Baier, Stöcker 11/14/2019; Gubela et al. 2020; Radcliffe and Surry 2011). Another difficulty in direct marketing campaigns is usually the excess of zeros. Gubela et al. (2020) tackle this with a two-stage model and Baier, Stöcker (11/14/2019) with Random Forest, Heckman selection model, and zero-inflated negative binomial regression model.

Although economic reasons are used in many marketing-related papers to justify the causal effect method's employment, the interaction between the emergence of the causal effect (additional buyers vs. more sales) and different cost structures in marketing campaigns is not sufficiently investigated. A discussion on return on investment can be found at Hansotia and Rukstales 2002; Sołtys et al. 2015; Gubela et al. 2020. Chickering and Heckerman (2000), Hansotia and Rukstales (2002), and Baier, Stöcker (11/14/2019) approach the ROI decision by introducing the expected lift in profit, respectively, the incremental break-even decision rule. A customer should be contacted if the expected causal effect on profit is higher than the cost per contact. The term profit is not defined more precisely at this point. Gubela et al. (2020) generalize the consideration of marketing costs of Lessmann et al. (2019), whereas profit is only used to evaluate the models. Again, there is no discussion about the appropriate modeling strategy depending on different cost structures.

Beyond this, some work addresses multiple treatments (Zhao et al. 2017a, 2017b; Lo and Pachamanova 2015; Rzepakowski and Jaroszewicz 2012), which is not the focus of this paper.

The detailed understanding of the estimated causal effect provides valuable information on the optimization approaches. An effect can arise from additional purchases (Lai et al. 2006) and more significant turnovers (Rubin and Waterman 2006). Even if, e.g., the sales do not increase, the average turnover by a customer can also increase through cross and up-sell. This specialty is a critical point to the standard approach because the gain in average turnover and not in additional sales would lead to no causal effect in conversion. If a customer buys for 50 € in τ_i=0 but exceeds it to 100 € when treated, we assume a causal effect (Rubin 2005).

A suitable campaign for activating additional buyers, respectively, responses must be optimized differently than one that hardly activates any new responses but influences their average turnover.

In the case of „only additional buyers," an algorithm that predicts the purchase probability may have a better chance. So if the turnover fades into the background, the customers who will only order based on the measure should be addressed. In the case of „only higher value per buyer," where it is crucial to predict the turnover as accurately as possible, the purchase probability hardly differs. It becomes essential to find customers who show the most considerable increase in turnover. Depending on the marketing campaign, either the optimization of the purchase probability or the turnover may be the right choice.

2.4 Performance Measuring

A widely used valuation method for customer scorings is the Lift Chart (Ling and Li 1998) Figure 4. Lift Charts visualize the prediction quality of a model. For this purpose, the records are sorted by their score. The x-axis shows the population targeted, and the y-axis shows the percentage of predicted events; the reference line illustrates the ratio of the events in a random sorting. An ideal model sorts all, e.g., buyers to the front, so with a response rate of 30%, the curve would rise to 30% response within the first 30% of customers and remain constant after that.

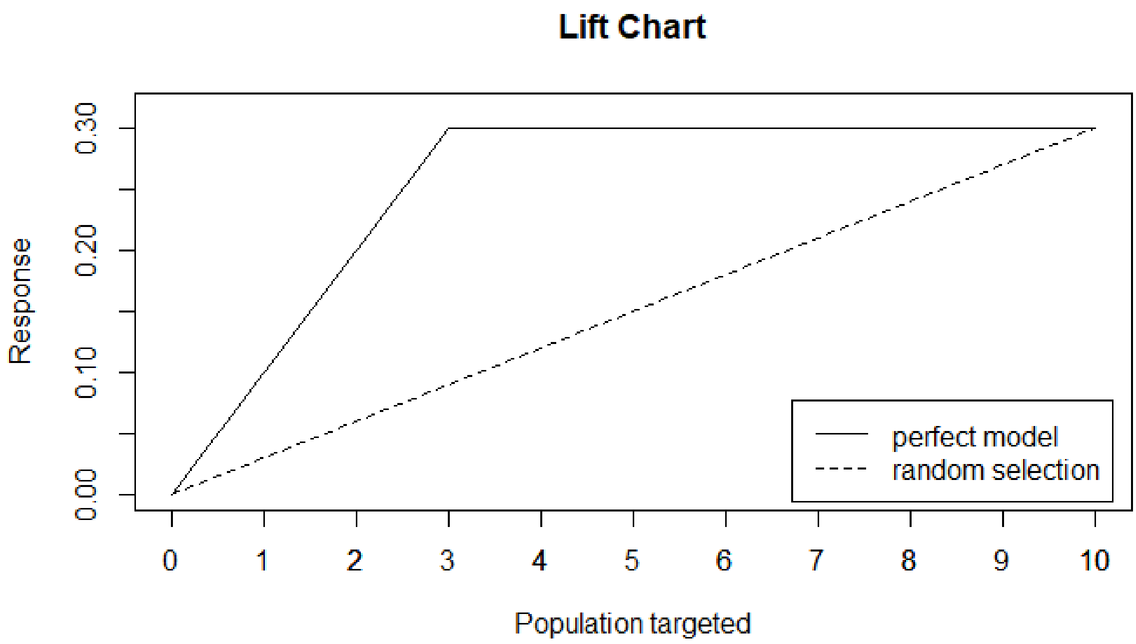


Fig.4 Lift chart (Ling and Li 1998)

Radcliffe (2007b; Radcliffe and Surry 2011) proposes a new metric: the Qini coefficient combined with the Qini curve Figure 5. The Qini curve plots the customers in $\tau_i=1$ on the x-axis. However, on the y-axis this time, the incremental purchases, i.e., the additional purchases or the total incremental value compared to the control group. To stay with the above example: 30% of the addressed customers buy, but 10% also buy when not addressed (causal effect of 20%). In an ideal model, 10% of customers appear to the very end. The curve would then drop from 30% to 20%.

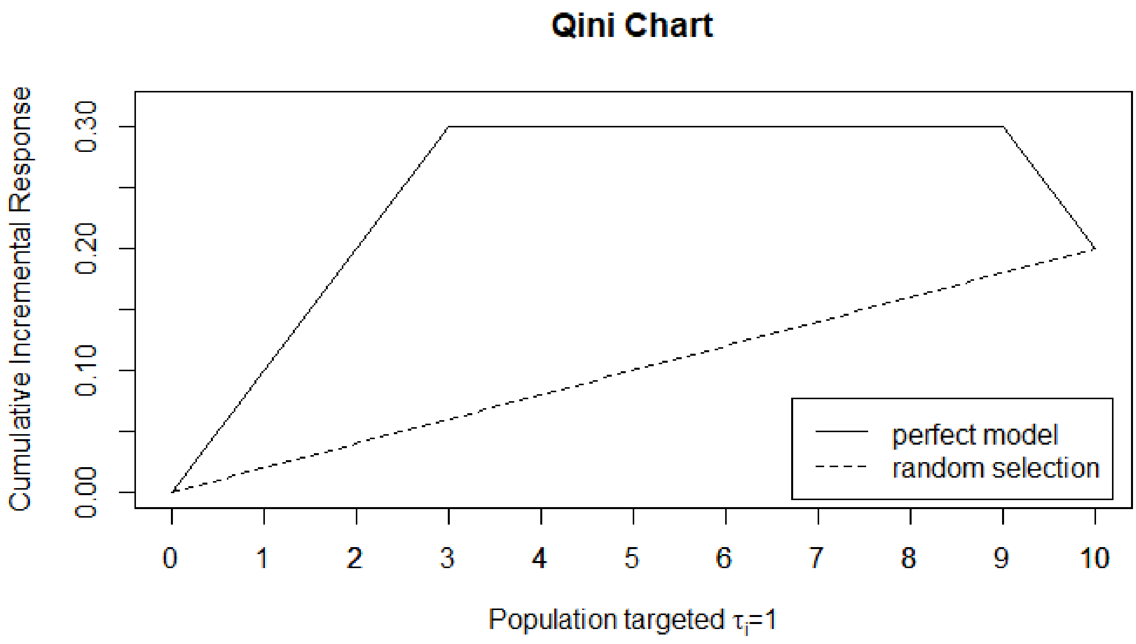


Fig. 5 Qini chart (Radcliffe 2007a)

The Qini bar-chart (Figure 6) shows the causal effect for each quantile. To follow the example above, each of the first three quantiles has the same causal effect, and the last shows the down lift.

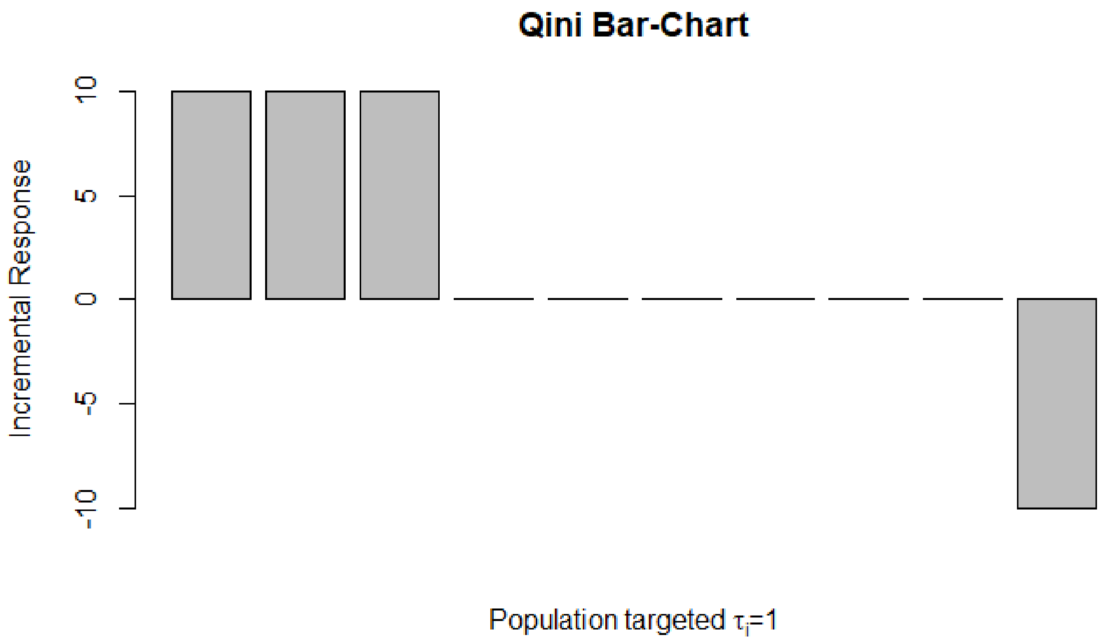


Fig. 6 Qini bar-chart

Table 4 Qini table

Decile	*Customers τ_i=1 and τ_i=0*	*Orders*			*Turnover*		
		τ_i=1	τ_i=0	Causal Effect	τ_i=1	τ_i=0	Causal Effect
1	10,000	10,000	0	10,000	360,000	0	360,000
2	10,000	10,000	0	10,000	240,000	0	240,000
3	10,000	10,000	0	10,000	120,000	0	120,000
4	10,000	0	0	0	0	0	0
5	10,000	0	0	0	0	0	0
6	10,000	0	0	0	0	0	0
7	10,000	0	0	0	0	0	0
8	10,000	0	0	0	0	0	0
9	10,000	0	0	0	0	0	0
10	10,000	0	10,000	-10,000	0	120,000	-120,000
Total	100,000	30,000	10,000	20,000	720,000	120,000	600,000

Figures 5 and 6 are based on an evaluation table, the so-called Qini table (Table 4). According to the commonly used Qini metrics, a Qini area of 1.25 (15,000 absolute) is obtained to measure purchases and 1.05 (330,000 absolute) for the turnover, where 0.5 corresponds to a random draw.

We are still in the environment of cost per contact, as described above. Let us now transfer the Qini curve to the cost per response (Table 5). With a conversion rate of 30%, the optimization will also only affect 30% of the customers, 70% of customers do not buy anything and generate neither turnover nor costs. This inequality can be seen quite well in the comparison of the individual cost structures in Table 5. To better compare the structures, the marketing costs in all three cases add up to 30,000 €, each only related to the group τ_i=1. With the fixed costs, the meaningfulness of the Qini metrics does not change. Each decile is just as significant to the marketing costs used. A fixed amount is incurred for each completed order in response-fixed costs, in the example 1 €. Although the first three deciles are equally important from a cost perspective, the remaining 70% do not affect costs. Regarding response-variable costs, the importance shifts even further into the first deciles (we have also assumed different order values). Here the marketing costs correspond to 4.2% of turnover.

Alternatively, in other real-world examples: in the data set we have at our hand, only 8% of customers buy; in e-commerce, even lower conversion rates are real (~3% conversion rate of online shoppers in the U.S., statista (2020)). In these cases, the educational value of the Qini curve diminishes. Firstly, in the context of 8% conversion, 4% estimated causal effect, and voucher, an ideal model, would have 4% of customers in the front and 4% in the end. The groups would have to be plotted even more finely than in the usual 10% steps to visualize the edges' optimization. The large area, formed between >4% and <96% of the customers, significantly influences the area under the curve and does not affect ROI. Because as already mentioned, these non-buyers do not generate any turnover or costs either.

Table 5 Qini table supplemented with the three different cost structures

Decile	*Customers τ_i=1 and τ_i=0*	*Orders Causal Effect*	*Turnover Causal Effect*	*Fixed Costs*	*Response-fixed Costs*	*Response-variable Costs*
1	10,000	10,000	360,000	3,000	10,000	15,000
2	10,000	10,000	240,000	3,000	10,000	10,000
3	10,000	10,000	120,000	3,000	10,000	5,000
4	10,000	0	0	3,000	0	0
5	10,000	0	0	3,000	0	0
6	10,000	0	0	3,000	0	0
7	10,000	0	0	3,000	0	0
8	10,000	0	0	3,000	0	0
9	10,000	0	0	3,000	0	0
10	10,000	-10,000	-120,000	3,000	0	0
Total	100,000	20,000	600,000	30,000	30,000	30,000

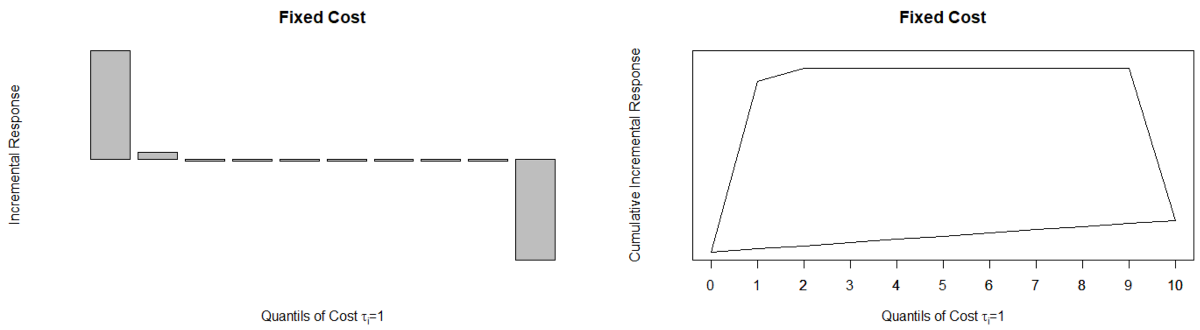
At this point, we suggest a new metric. As the basis for the scoring valuation, the independent variable should now show the costs instead of customers. The customers are again sorted based on the score, but the grouping bases on cost quantiles of equal size. This new approach has the advantage that each quantile is now equally important in its costs and ROI. Also, this procedure is very variable regarding the cost structures that can arise. The dependent variable is then the causal effect in return in terms of, e.g., sales. Therefore, a good model sorts the customers with the highest estimated causal effect in return to the front. Since the grouping centers on equal size cost groups, the model performance shows the incremental causal effect in return. In reference to Qini and ROI, we call the new metric Rini. The Qini is, therefore, a particular case of Rini if just costs per contact occur.

A given point g(Φ) is defined as the estimated causal effect in return of the best-scored customers $\Phi\in\{0,1\}$ in $\tau_i=1$.

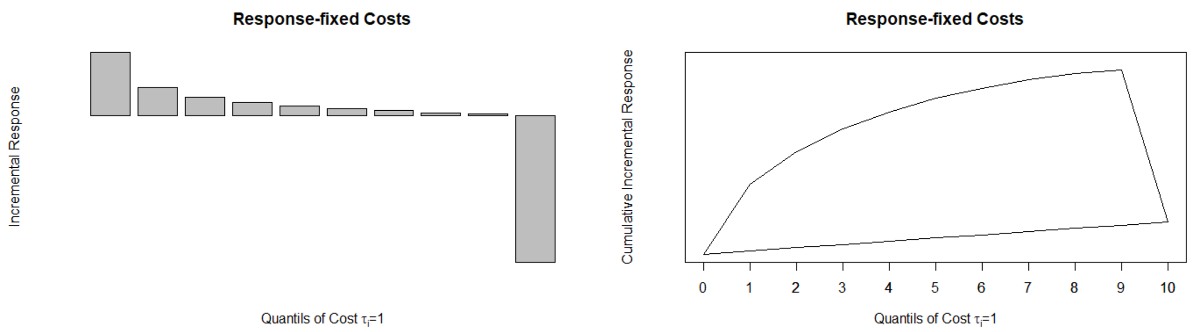
(10) $$g(\phi)=\sum_{i\in N_\phi}\left(r_i^{\tau_i=1}-r_i^{\tau_i=0}\right)$$

The Rini area represents the model quality using the area under the curve. A random selection leads to a Rini area of 0.5. Since the Rini chart quantiles are derived from individual customers' campaign costs, and since these rarely lead to equal-size groups, the Rini chart's return is weighted with the average costs.

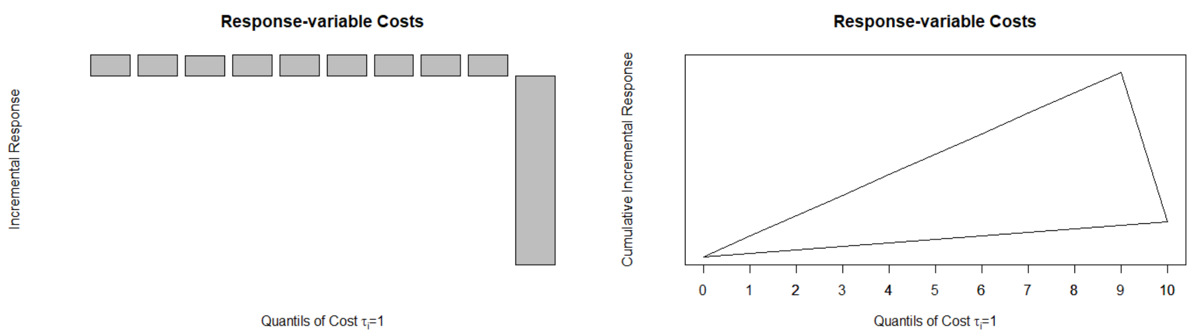
The three different cost structures also show patterns related to the Rini curve unique for each cost type (Figure 7). In the case of fixed costs (Panel A of Figure 7), the curve shows a linear increase at the beginning. On the x-axis, groups of costs of the same size are mapped. Since each contact is equally expensive, the Rini curve ideally (100% of contacts buy because of the measure) also shows the same return. For the first 10% of the costs, a maximum of 100 contracts can be generated, for example. All contract conclusions in $\tau_i=0$ are ideally sorted into the last group. Therefore the ideal Rini curve shows the characteristic downward trend for the last group, unlike the Qini curve. Thus, if it is possible, for example, to consider the conclusion of contracts or purchases not only in binary but also in metric terms, the Rini curve will also change. We may assume that the shopping baskets differ in height. The ideal model now sorts the customers with the highest estimated causal effect per contact to the front. The effect in the groups is now non-linear. The bar chart shows a significant gap without returns, larger or smaller, depending on the conversion rate. Hence, if fixed costs per order are present (Panel B of Figure 7), this gap no longer exists because a return can be determined whenever costs occur. Since the costs are constant, but the basket height fluctuates, a non-linear increase is expected here. In the last case of variable costs per order (Panel C of Figure 7), we make the following assumption: The discount offered refers, for example, to the entire order and not only to particular items. Thus the Rini curve rises linearly again. With an assumed discount of 20%, the ideal model's profit corresponds to a 100% turnover. The bar-chart shows idyllically a constant causal effect in spending for all but the least groups.



A: Rini bar-chart and Rini chart for fixed costs



B: Rini bar-chart and Rini chart for response-fixed costs



C: Rini bar-chart and Rini chart for response-variable costs

Fig. 7 Rini diagrams of the different cost structures

3. Empirical Investigation

After we could already show in our theoretical framework and literature review that different cost types have a considerable influence on customers' optimal sorting, we would like to test this in practice. For this purpose, we are introducing a new data set from an e-commerce retailer in Germany. In this data set, we simulate different cost types based on the shown ordering behavior and apply the different sorting options. To measure the models' performance, we discuss the method used so far and generalize it to cope with different cost structures. Finally, in a Monte Carlo setting, we calculate the results on one hundred randomly selected data splits and discuss the results.

3.1 Description of the Data Set and Data Preparation

The data set consists of 295,040 unique customer data from two print mailing campaigns. The customers were statistically equally assigned to τ_i=1 (147,520) and τ_i=0 (147,520). For each data set, purchase (1.045 variables, last 24 months) are available. The data includes ordering behavior in terms of recency frequency, monetary value, and coupon usage (Table 6). The outcomes include orders, turnover, and discount costs in the two following weeks.

We randomly split the dataset into training (70%) and validation (30%) samples to validate our models. This dataset's causal effect arises mostly from driving orders (+109%) rather than the turnover (+16%). To assess our assumptions on a dataset with opposite drivers, we copied the dataset and upsampled the τ_i=0 group by doubling the records containing a purchase. After that, we divided the spending on τ_i=0 by two. So the constraints between purchases in τ_i=1 and the redemption stay untouched (thru orders +4%, thru turnover +131%).

Table 6 Structure of the predictors

	Domain	*Example*
	Customer (16)	Customer since, lifetime value, age, sex, place of residence, status newsletter permission
For each period* (147)	Order value and quantity whole period (16)	Turnover, returns, install payments, mailings received
	Value and quantity for 30 assortment-clusters (60)	Clothing, garden, shoes, kids, furniture, lingerie
	Price brackets for fashion (low, medium, high) (36)	Women's clothing, men's clothing, kids clothing
	Usage of coupons (15)	For rebate, voucher, assortment+rebate, assortment+voucher
	Order value and quantity for each device (10)	App, desktop, mobile, tablet
	Order value and quantity for online marketing (4)	Brand, performance
	Order value and quantity for order channel (6)	Telephone, web

* four periods in sum, one period lasts 6 months, and combinations, e.g., last 12, 18, or 24 months

The preprocessing of the predictors can determine whether a good statistical model can be found at all. Also, clever preprocessing has a direct influence on the performance of the models. In the context of ROI, the dependent variable is the turnover depending on the treatment. As described, we do not directly model the causal effect, but the expected value of turnover as a function of the coefficients and the treatment. Usually, many variables show little variance and have a high correlation. Procedures such as linear regression have special requirements for the predictor variables. The input variables should/must have the same dispersion within the data (homoscedasticity). Especially in econometrics, this requirement is usually challenging to meet. The data available here is highly skewed and does not show a normal distribution around zero (just one example: the number of sales can never be negative). To improve the model's quality, we filtered the predictor variables for low variance and high correlation, then transformed using Box-Cox to mitigate the distributions' skewness, centered and scaled. To reduce the remaining variables to more meaningful linear combinations, we used the Principal Component Analysis for the predictors without considering the dependent variable (unsupervised learning method).

3.2 Different Approaches to the Calculation of the Causal Effect

To bring the customers in the right order for campaign selections, an order, the so-called score, is necessary. We now combine our findings regarding return and costs to describe ideal strategies (Table 7). Of course, in practice, marketing campaigns can entail varied forms. The existing literature well covers the ideal strategies in a cost per contact setting, the traditional approach for the case of additional orders (Devriendt et al. 2018; Gubela et al. 2019; Pierre Gutierrez and Jean-Yves Gérardy 2017), and the revenue (Gubela et al. 2020) for an increase of the response value.

The prediction of the turnover can be understood as an extension of the classification problem (buy or not buy), wherein the case of purchase additional information is available. This additional information can then be used to fine-tune the customers in the scoring and, consequently, the ROI. Thus, this dimension's classic approach can be extended by assuming a continuous variable for the return.

Table 7 Idealized cost and causal effect constellations to derive a reasonable approach

	Contact	*Response**	
	Fix	Fix	Variable
Additional sales	I	II	III
More response value per customer	IV	V	VI

* can be between 0 and 100% for the population respectively ∈ {0,1} for each customer

When considering the costs per contact, it is essential to weigh whether it makes sense to contact the customer. In case I), a difference in purchase probability can be determined for $\tau_i=0$ and $\tau_i=1$. Therefore, it makes sense to address customers with the highest positive causal effect on conversion (ideally already finely sorted with their expected shopping basket). This consideration also applies to cases II) and III). Here too, the additional buyers must be identified. In contrast to I), however, no costs are incurred if the customer does not respond. Therefore, we assume that estimating the causal effect as a difference (probability of purchase or shopping basket value) is best suited here.

In case IV), however, we imagine the estimated causal effect towards conversion is always zero because no additional buyers appear. Instead, the causal effect appearing in the turnover becomes more critical. Now customers with the highest distinction in turnover are to be addressed, and customers whose turnover has hardly changed are sorted out. The considerations regarding the origin of the causal effect and campaign costs structures now apply to V) and VI). For response-fixed costs such as vouchers, the difference should be decisive, and for response-variable costs such as discounts, the quotient should be decisive.

For response-related costs (II, III, V, VI), another thought is worth considering. We assume that the purchase probability is also influenced by other, unobserved variables, such as a purchase already made by a competitor. Since a retailer can only train our models on their data, these missed opportunities are not captured. With response-dependent costs, it is theoretically possible to contact all customers at no cost. We now take the prediction from the count model of the zero-inflated model (which can also be 0 in contrast to the hurdle regression) and apply it to all customers, regardless of how high their response probability is. This approach could lead to better results, mostly if the unobserved variables play a significant role in the purchase probability.

3.3 Modeling

A glance at the histogram for turnover shows a significant problem for modeling continuous outcomes in marketing campaigns with low response rates: The zero is by far the most common value. A transformation of the distribution through, e.g., the logarithm does not help, so that we speak here of an excess of zeros. A solution can be mixed models such as hurdle regression or zero-inflated regression models. They model a probability for the occurrence of zero and a second model for the prediction of the count data.

Lambert (1992) first described zero-inflated models. The original model referred to the Poisson distribution and described a mixed model consisting of a count data model with Poisson distribution and an additional model that describes the occurrence of zero, later extended to include the negative-binomial distribution (Ridout et al. 2001). The zero-inflated count-data model determines the probability of belonging to one of the two latent classes. Secondly, a zero can stem from both the zero and the count model.

In the first part of the equation, we calculate the probability of belonging to one of the two latent classes. The two classes refer to turnover (j) equals zero and are greater than zero and can thus be estimated with a binary logistic regression. Theoretically and practically, the influencing variables for this regression can differ from those of the count model. It may feel slightly unusual that we are modeling the probability of the occurrence of zero (i.e., the non-purchase) here and not the purchase probability.

$$\Pr\left(r_i = j\right) = \begin{cases} \pi_i + \left(1-\pi_i\right) g\left(y_i = 0\right) & if\ j = 0 \\ \left(1-\pi_i\right) g\left(y_i\right) & if\ j > 0 \end{cases} \tag{11}$$

Where π_i is the logistic regression with probit link function for the occurrence of zero with the predictors Z_i and the corresponding ß factors from the logistic regression, all parameters are estimated by maximum likelihood. We chose the probit link over the logit because of the slightly better results in the validation. In the probit model, the binary outcome depends on a hidden Gaussian variable.

$$\pi_i = \Phi\left(Z_i^{'}\beta\right) \tag{12}$$

The negative-binomial distribution is given by g with the shape parameter θ and the mean μ_i.

$$g\left(y_i\right) = \Pr(Y = y_i \mid \mu_i, \Theta) = \frac{\Gamma\left(y_i + \Theta^{-1}\right)}{\Gamma\left(\Theta^{-1}\right)\Gamma\left(y_i + 1\right)} * \left(\frac{1}{1+\Theta\mu_i}\right)^{\Theta^{-1}} * \left(\frac{\Theta\mu_i}{1+\Theta\mu_i}\right)^{y_i} \tag{13}$$

3.4 Application of the Current Approach

Using a selected real data example, we would like to show how to calculate the Qini metrics. After the data set has been sorted according to the score, we summarize the data on cost-deciles and form different sums (Table 8). This example table is based on fixed costs, so the cost deciles are equally distributed.

Table 8 Exemplary Qini table

Decile	*Costs per group*		*Turnover per group*			
	τ_i=0	τ_i=1	τ_i=0	τ_i=1	Est. causal effect targeted	Est. causal effect random selection
0.1	0	16,460	17,683,354	52,716,349	35,032,995	7,821,170
0.2	0	16,460	7,429,908	20,191,487	12,761,579	7,821,170
0.3	0	16,460	5,648,185	12,968,962	7,320,777	7,821,170
0.4	0	16,460	3,274,953	10,683,175	7,408,222	7,821,170
0.5	0	16,460	4,003,650	8,877,335	4,873,685	7,821,170
0.6	0	16,460	3,750,799	6,774,247	3,023,448	7,821,170
0.7	0	16,460	2,595,724	4,759,497	2,163,773	7,821,170
0.8	0	16,460	2,171,362	4,899,204	2,727,842	7,821,170
0.9	0	16,460	1,849,125	3,700,226	1,851,101	7,821,170
1.0	0	16,460	7,386,810	8,435,091	1,048,281	7,821,170

The Qini charts (Fig. 8) show a well-performing model. In the Qini bar-chart, we can see that the first two deciles perform above average and the latter below. We can also tell that each decile is ranked in descending order from high positive to low negative, which means no discontinuities within the ranking. The model can reproduce a good sorting. The Qini chart shows how much the increase in profit changes entirely if someone adds the next best 10% to the best 20% of customers. In this example, however, the typical bend to the last decile is missing. The Qini area is 0.770, and since these are fixed costs, this is equivalent to the Rini area.

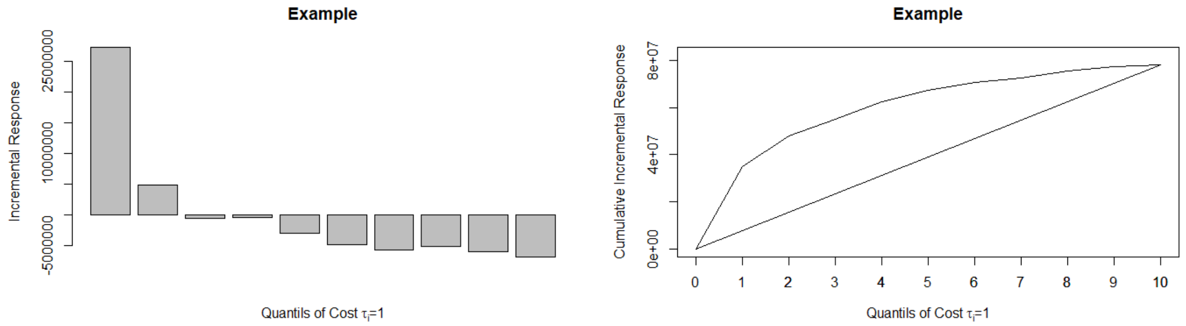


Fig. 8 Exemplary Qini bar-chart and Qini chart

In the next section, we will compare the current approaches with our new approaches. The current approaches are denoted by "pr_d" and "e_d." The nomenclature is explained in the following.

3.5 Application of the new Approach and Discussion

As described, there are different ways of calculating the causal effect. Only the Two-model and the Interaction-model can calculate a continuous outcome such as turnover. From the considerations regarding different natures of the causal effect and cost settings, six different approaches for the evaluation seem promising: the consideration of the purchase probability (pr), the turnover considering the purchase probability (e), and the turnover in case of purchase (ep), each calculated as difference or quotient (denoted by the addendum "d" respectively "q"). We have modeled the voucher and the rebate with a 100% redemption rate to see more evident differences for the different cost types. We now combine all six possible scorings with the three cost types on two datasets.

We now form two zero-inflated models on each of the preprocessed training data sets since we have chosen the indirect method. The first model includes only data from the treatment group, the second model only from the control group; the treatment information itself is not an input variable. The dependent variable is the amount of the shopping basket, i.e., the turnover made during the campaign, which is available as an integer in euro cents. The respective initial models with all input variables were then gradually reduced. Therefore only input variables were included, whose p-value was smaller than 0.2, to calculate the model again with these reduced input variables. We also calculated models for validation with stricter p-values; this did not significantly affect the Rini area. Based on the zero-inflated model, three predictions are now calculated, in which, in addition to the entire model (e), the two components probability of occurrence of zero (pr) and the amount of the shopping basket, if a purchase was present (ep), are also calculated separately. The scoring variable now results either from the difference or the quotient of the prediction pairs. Also, we extract the estimate from the model τ_i=1, which represents a response model, and can thus conclude the benefit of using the causal model (denoted by the addendum "r").

In this case, the highest values represent the most worthwhile customers, so we sort in descending order here. For validation, we apply the models to the new validation data set, and the Rini area is determined individually on the different cost structures.

The computations were performed on 100 randomly generated data-splits into training and validation data sets. The preprocessing was also carried out again in each case. Since there is no optimization procedure for zero-inflated models, we calculated the analysis on different maximum thresholds for input variables' significance.

3.5.1 Cost per Contact

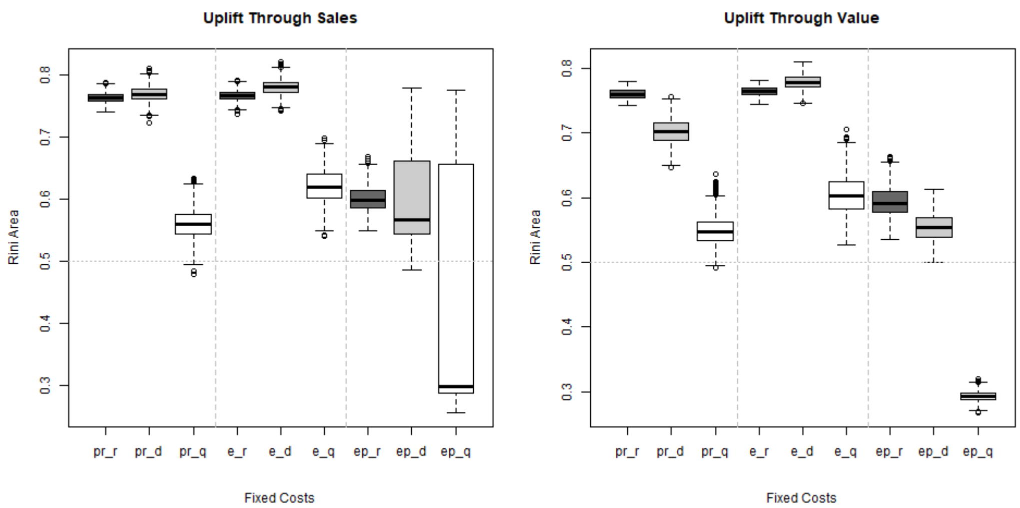
In Panel A of Figure 9, we see the classic model with a difference (pr_d) as having an advantage in the sales-driven and value-driven causal effect. A fine-tuning by predicting the turnover (e_d) brings about a significant improvement, especially in the value-driven causal effect. A prediction that does not consider the probability of purchase (ep_d, ep_q) cannot keep up with when costs per contact are present. As assumed, in our data set, which shows a low response rate, even a response score can provide excellent results.

3.5.2 Cost per Order

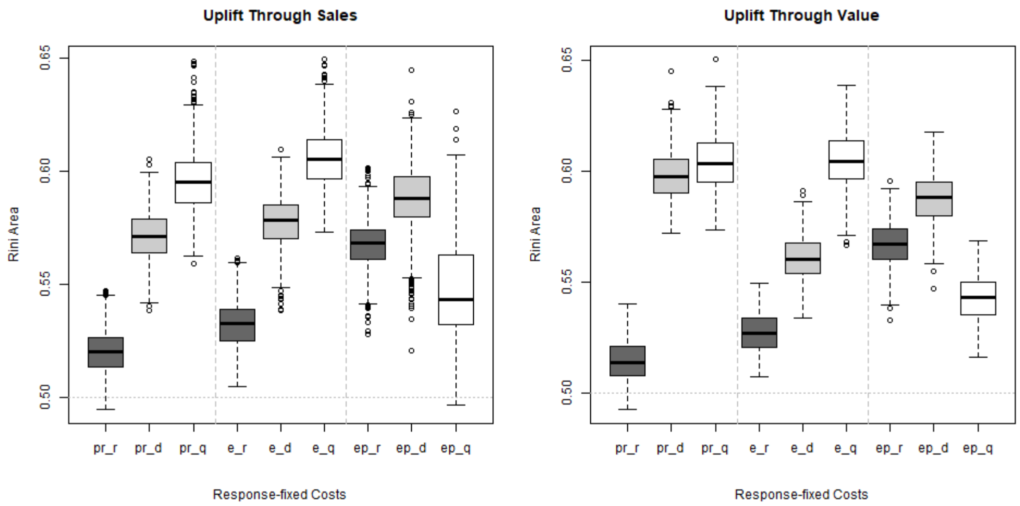
Although theoretically, the calculation of the difference should be in advantage here, it shows that the quotient seems to be slightly superior (Panel B of Figure 9). Here, too, fine-tuning could be demonstrated by predicting the turnover. Our model can also show decent results for response-variable costs without considering the purchase probability when calculating the difference (ep_d).

3.5.3 Cost per Turnover

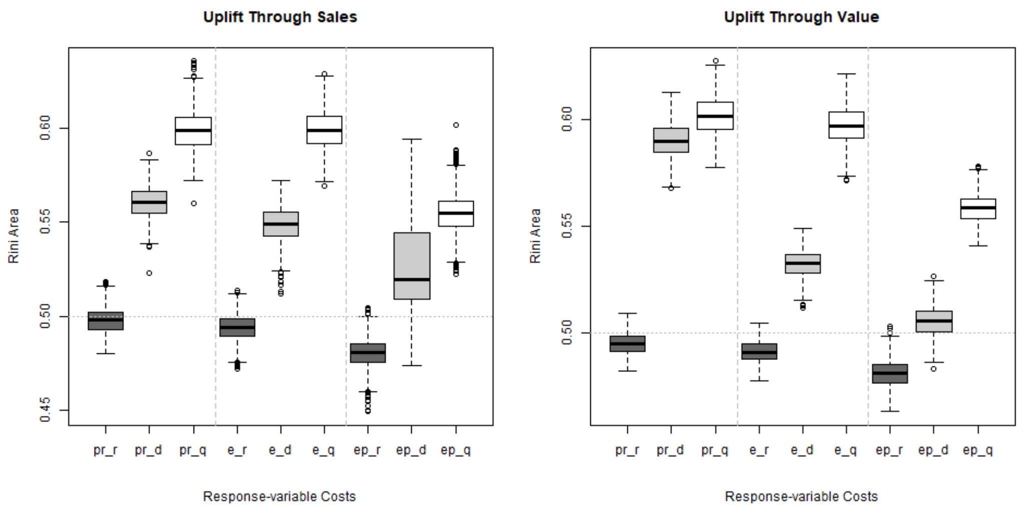
As expected, the quotient determination leads to the best results with this cost structure (Panel C of Figure 9). However, it is also surprising that the difference in the probability of purchase (pr_d) can also achieve good results, at least for value-driven causal effects. Models that do not take the purchase probability into account perform significantly worse here. Here the response score is even worse than a random sort.



A: Results in the context of fixed costs



B: Results in the context of response fixed costs



C: Results in the context of response variable costs

Fig. 9 Box-plots of the Rini areas of the different approaches from 100 randomly selected data splits

4. Conclusion and Limitations

We wanted to investigate whether and how different marketing costs influence determining the causal effect and, thus, on the ROI. We succeeded in proving this effect theoretically and practically and in making a recommendation.

The effects of the different cost types on the ROI are evident. It is essential to closely examine the costs before modeling to choose the optimal solution for calculating the score. For example, if one chooses the wrong approach, initial results could be so devastating that the vital issue is not pursued further within the company. It is not beneficial for the scientific discussion that the different cost types are not mentioned so far and that fixed costs per contact are implicitly assumed. This shortcoming leads to the point that the Qini metrics used so far also need to be revised.

Also, how the causal effect occurs, i.e., through more purchases or higher turnover, receives too little attention. Again, at least theoretically, we could show a significant difference in approaching the given dataset.

Our research has confirmed the previously published work seen in the context of cost per contact. Furthermore, we have extended the application of causal effect in marketing to other cost types that are also important and frequently used in marketing. We found another way to describe the causal effect, which leads to significantly better results in other cost types not yet considered in research. In this context, we also have generalized the measurement of model quality to be used flexibly.

We have discussed the implications of redemption behavior in the theory section. Costs that only arise when actively initiated by the customer, e.g., activating a voucher before placing the order, represent a considerable challenge. Two independent statistical models have already been developed to calculate the causal effect. In order to now also still consider redemption behavior, a third one is needed. We could not find any improvement in the practical application by adding the third model. We assume that the errors resulting from the combination of three models become too large so that a valid prediction is no longer possible. We would like to encourage researchers to investigate this problem more closely since the impact on ROI is evident.

For calculating the different cost and causal effect simulations, the same data set was used, created by a discount campaign. The simulation could be problematic because experience shows that different marketing campaigns with either cost per contact, vouchers, or discounts lead to different responses and turnover results. In the simulations, we also assumed a 100% redemption rate of vouchers and discounts, which is not always the case in reality.

The application and computation of the causal effect bring some problems to read up in the extensive literature, e.g., Holland (1986) discusses this in detail. A statistical model is an approximation of observed behavior at a given time. We have found that it is useful to develop and optimize the two models independently. The test and control groups should differ in their behavior. Thus, different input variables enter the models, which leads to different predictions.

References

Baier, Daniel; Stöcker, Björn (2019): Maximizing Return on Investment from Direct Marketing Campaigns: A New Uplift Modeling Approach for Online Shops. „Advanced Data Analysis Techniques with Marketing Applications“. Karlsruhe, Germany, November 14, 2019. Available online at http://www.gfkl.org/working-groups/data-analyses-and-classification-in-marketing/.

Bawa, Kapil; Shoemaker, Robert W. (1989): Analyzing Incremental Sales from a Direct Mail Coupon Promotion. In *Journal of Marketing* 53 (3), pp. 66–78. DOI: 10.1177/002224298905300308.

Chickering, David Maxwell; Heckerman, David (2000): A Decision Theoretic Approach to Targeted Advertising. Available online at http://arxiv.org/pdf/1301.3842v1.

Devriendt, Floris; Moldovan, Darie; Verbeke, Wouter (2018): A Literature Survey and Experimental Evaluation of the State-of-the-Art in Uplift Modeling: A Stepping Stone Toward the Development of Prescriptive Analytics. In *Big Data* 6 (1), pp. 13–41. DOI: 10.1089/big.2017.0104.

Dua, Dheeru and Graff, Casey (2017): UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences. Available online at http://archive.ics.uci.edu/ml.

Gubela, Robin; Bequé, Artem; Lessmann, Stefan; Gebert, Fabian (2019): Conversion Uplift in E-Commerce: A Systematic Benchmark of Modeling Strategies. In *International Journal of Information Technology and Decision Making*. 18 (03), pp. 747–791. DOI: 10.1142/S0219622019500172.

Gubela, Robin M.; Lessmann, Stefan; Jaroszewicz, Szymon (2020): Response transformation and profit decomposition for revenue uplift modeling. In *European Journal of Operational Research* 283 (2), pp. 647–661. DOI: 10.1016/j.ejor.2019.11.030.

Guelman, Leo (2014): Optimal Personalized Treatment Learning Models with Insurance Applications. PhD in Economics: Universitat de Barcelona.

Hansotia, Behram; Rukstales, Brad (2002): Incremental value modeling. In *Journal of Interactive Marketing* 16 (3), pp. 35–46. DOI: 10.1002/dir.10035.

Hillstrom, Kevin (2008): The MineThatData e-mail analytics and data mining challenge. In *MineThatData blog*. Available online at https://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.html, checked on October 27, 2020.

Holland, Paul W. (1986): Statistics and Causal Inference. In *Journal of the American Statistical Association* 81 (396), p. 945. DOI: 10.2307/2289064.

Imbens, G. W.; Rubin, D. B. (2006): Rubin causal model. In *The New Palgrave Dictionary of Economics* 2nd ed.

Jaskowski, Maciej; Jaroszewicz, Szymon (2012): Uplift modeling for clinical trial data. In *ICML 2012 Workshop on Clinical Data*, Edinburgh, Scotland.

Kane, Kathleen; Lo, Victor S.Y.; Zheng, Jane (2014): Mining for the truly responsive customers and prospects using true-lift modeling: Comparison of new and existing methods. In: *Journal of Marketing Analytics* 2 (4), S. 218–238. DOI: 10.1057/jma.2014.18.

Lai, Yi-ting; Wang, Ke; Ling, Daymond; Shi, Hua; Zhang, Jason (2006): Sixth IEEE International Conference on Data Mining. ICDM 2006: proceedings: December 18 - 22, 2006, Hong Kong. Los Alamitos, Calif: IEEE Computer Society. Available online at http://ieeexplore.ieee.org/servlet/opac?punumber=4053012.

Lambert, Diane (1992): Zero-inflated Poisson regression, with an application to defects in manufacturing. In *Technometrics* 34 (1), pp. 1–14.

Larsen, Kim (2010): Net Lift Models. In: Slides of a talk given at the. A2010-Analytics Conference, September. Copenhagen, Denmark.

Lessmann, Stefan; Haupt, Johannes; Coussement, Kristof; Bock, Koen W. de (2019): Targeting customers for profit: An ensemble learning framework to support marketing decision-making. In *Information Sciences*.

Ling, Charles X.; Li, Chenghui (1998): Data Mining for Direct Marketing: Problems and Solutions. In Rakesh Agrawal (Ed.): Proceedings / The Fourth International Conference on Knowledge Discovery and Data Mining, August 27 - 31, 1998, New York, New York. Menlo Park, Calif.: AAAI Press, pp. 73–79.

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

Lo, Victor S. Y.; Pachamanova, Dessislava A. (2015): From predictive uplift modeling to prescriptive uplift analytics: A practical approach to treatment optimization while accounting for estimation risk. In *Journal of Marketing Analytics* 3 (2), pp. 79–95.

Pierre Gutierrez; Jean-Yves Gérardy (2017): Causal Inference and Uplift Modelling: A Review of the Literature. In Claire Hardgrove, Louis Dorard, Keiran Thompson, Florian Douetteau (Eds.): Proceedings of The 3rd International Conference on Predictive Applications and APIs, vol. 67. Microsoft NERD, Boston, USA: PMLR (Proceedings of Machine Learning Research), pp. 1–13. Available online at http://proceedings.mlr.press/v67/gutierrez17a.html.

Radcliffe, Nicholas (2007a): Generating Incremental Sales. Maximizing the incremental impact of cross-selling, up-selling and deep-selling through uplift modelling. In *White Paper Stochastic Solutions Limited*, pp. 1–10.

Radcliffe, Nicholas J. (2007b): Using control groups to target on predicted lift: Building and assessing uplift models. In *Direct Marketing Analytics Journal* 1, p. 1421.

Radcliffe, Nicholas J.; Surry, Patrick D. (1999): Differential Response Analysis: Modeling True Responses by Isolating the Effect of a Single Action. In *Proceedings of Credit Scoring and Credit Control IV*, Edinburgh, Scotland.

Radcliffe, Nicholas J.; Surry, Patrick D. (2011): Real-world uplift modelling with significance-based uplift trees. In *White Paper TR-2011-1, Stochastic Solutions*.

Ridout, M.; Hinde, J.; Demétrio, C. G. (2001): A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. In *Biometrics* 57 (1), pp. 219–223. DOI: 10.1111/j.0006-341X.2001.00219.x.

Rubin, Donald B. (1974): Estimating causal effects of treatments in randomized and nonrandomized studies. In *Journal of Educational Statistics* 66 (5), p. 688.

Rubin, Donald B. (1977): Assignment to treatment group on the basis of a covariate. In *Journal of Educational Statistics* 2 (1), pp. 1–26.

Rubin, Donald B. (1990): [On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.] Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies. In *Statistical Science* 5 (4), pp. 472–480. Available online at http://www.jstor.org/stable/2245383.

Rubin, Donald B. (2005): Causal inference using potential outcomes: Design, modeling, decisions. In *Journal of the American Statistical Association* 100 (469), pp. 322–331.

Rubin, Donald B.; Waterman, Richard P. (2006): Estimating the Causal Effects of Marketing Interventions Using Propensity Score Methodology. In *Statistical Science* 21 (2), pp. 206–222. DOI: 10.1214/088342306000000259.

Rzepakowski, Piotr; Jaroszewicz, Szymon (2012): Decision trees for uplift modeling with single and multiple treatments. In *Knowledge and Information Systems* 32 (2), pp. 303–327. DOI: 10.1007/s10115-011-0434-0.

Sołtys, Michał; Jaroszewicz, Szymon; Rzepakowski, Piotr (2015): Ensemble methods for uplift modeling. In *Data Mining and Knowledge Discovery* 29 (6), pp. 1531–1559. DOI: 10.1007/s10618-014-0383-9.

statista (2020): U.S. online shopper conversion rate 2018-2019. Available online at https://www.statista.com/statistics/439558/us-online-shopper-conversion-rate/, checked on November 6, 2020.

Su, Xiaogang; Kang, Joseph; Fan, Juanjuan; Levine, Richard A.; Yan, Xin (2012): Facilitating score and causal inference trees for large observational studies. In *Journal of Machine Learning Research* 13 (Oct), pp. 2955–2994.

Tian, Lu; Alizadeh, Ash A.; Gentles, Andrew J.; Tibshirani, Robert (2014): A Simple Method for Estimating Interactions between a Treatment and a Large Number of Covariates. In *Journal of the American Statistical Association* 109 (508), pp. 1517–1532. DOI: 10.1080/01621459.2014.951443.

Zhao, Yan; Fang, Xiao; Simchi-Levi, David (2017a): Uplift Modeling with Multiple Treatments and General Response Types. In Nitesh Chawla, Wei Wang (Eds.): Proceedings of the 2017 SIAM International Conference on Data Mining. Philadelphia, PA: Society for Industrial and Applied Mathematics, pp. 588–596.

Zhao, Yan; Fang, Xiao; Simchi-Levi, David (2017b): A Practically Competitive and Provably Consistent Algorithm for Uplift Modeling. Available online at http://arxiv.org/pdf/1709.03683v1.

Chapter 6

Conclusion

This thesis should open new approaches to CRM for fashion online retailers. Two broad research topics were addressed based on current and relevant problems from practice, and two research papers were methodically developed for each of these two topics.

For Part A, attention was given to the frontstage, the area where customers and company interact:

RQ 1 addressed how skewed response behavior in customer satisfaction studies affects the validity of PRCA and how this can be avoided. Our two studies have shown that the PRCA works incorrectly with skewed distributions, which can possibly lead to diametrically different conclusions. A new method developed by us, using cubic regression, can meet this challenge. In direct comparison, it became clear how much the categorizations occasionally varied depending on which of the three different calculation logics of the PRCA was followed. With cubic regression, this choice was omitted, and the results were simpler to understand. The consequences of an incorrect categorization should not be underestimated if this leads to incorrect strategies for managing touchpoints. Therefore, the application of cubic regression represents an essential new method for categorizing the different cause-effect relationships in the measurement of customer satisfaction.

RQ 2 dealt with the question of which measures in return management would have the most substantial influence on customer satisfaction from the customer's point of view. Further insights can be gained by examining the returns management process holistically, that is, from presales to returning. It could be determined for all examined measures that they are still at the beginning of their life cycles. For many respondents, these measures are appealing but are not expected to be universally implemented. The highest effect on customer satisfaction was measured in monetary incentives in the avoidance of returns, a second cluster that was important as well and related to the improved presentation. This hierarchy was interesting because we suspected that returns would additionally represent a significant customer burden, which the consumer would have liked to avoid. However, since measures in the last purchase phase have been clearly successful, it can be assumed that customers who are willing to shop in the mail-order business have at least not excluded returns in advance of purchase and are therefore not deterred from making a purchase. The future trends will be interesting to observe when other customers have to switch to the mail-order business due to retail stores being affected by the structural change. Here, avoiding returns in the presales phase could have a further positive impact.

Part B addressed the backstage and examined customers' optimal selection for a direct marketing campaign:

RQ 3 focused on comparing methods for the optimal selection of customers in direct marketing campaigns, with a particular focus on the prediction of continuous values, namely profit and the treatment of the excesses of zeros in prediction models. We extended well-known procedures in uplift modeling with a new perspective, namely profit. Direct marketing campaigns in the mail-order business generally have a low response rate, leading to the "no buy" result being the most common one of the effort. Therefore, three statistical methods were chosen to cope with these special conditions: the Heckman sample selection model, the zero-inflated negative binomial regression model, and random forest-based regression. The research in Paper #3 shows that all three approaches are well suited to handling uplift modeling concerning profit. Additionally, the handling of continuous values in uplift modeling continues to be minimally studied; thus, our results are an essential contribution to further theoretical development.

RQ 4 examined the implications for the ROI-optimal selections of direct marketing campaigns, which result from different campaign cost structures. Uplift modeling or causal effect modeling has continued to be mainly based on a cost per contact approach. However, we were able to show on a real data set how strongly the cost structure affected the successful application and thus the ROI. Each cost structure must be optimized differently in order to exploit the full potential of this method.

For this purpose, we additionally generalized the common validation methods, known as Qini metrics, which could then be applied to all cost types. We further pointed out other important influencing factors that were hardly or not considered in the literature, such as what the expression of the additional sales in a test group is, and what influence the redemption behavior of coupons and discounts have. Through this extensive work, for the first time in the literature, the application of causal effects has been integrated into a holistic campaign cost framework and draws attention to significant shortcomings in the current approach. We hope that this will provide new impulses for further, more in-depth research.

The answers to the four research questions represent significant methodological advancements in CRM and thus make a significant contribution to current research. I hope that these impulses will be well received by researchers and practitioners and lead to further research questions and practical improvements.

Appendix

Academic output of research papers and individual contributions

Appendix A: Academic Output of Research Papers

This dissertation is cumulative in nature; this means that chapters two to five are based on individual papers. These papers have been published in or are under review at academic journals. The following list summarizes the included papers and their respective academic output.

Research Paper #1 (Chapter 2). Stöcker, B., Nasseri, A. (2020). Penalty Reward Contrast Analysis (PRCA) for Categorizing Service Components: A New Approach
This paper has been published in the *Archives of Data Science*, Series A (Online First), Volume 6, Issue 2. (VHB JOURQUAL 3: Category: none)

Research Paper #2 (Chapter 3). Stöcker, B., Baier, D., Brandt, B. (2020). New Insights in Online Fashion Retail Returns From a Customers' Perspective and Their Dynamics.
This paper is under review and considered for publication in the *Journal of Business Economics*. (VHB JOURQUAL 3: Category B)

Research Paper #3 (Chapter 4). Baier, D., Stöcker, B. (2020). Maximizing Profit from Direct Marketing Campaigns: Profit Uplift Modeling Approaches for Online Shops.
This paper is under review and considered for publication in the *Journal of Business Economics*. (VHB JOURQUAL 3: Category B)

Research Paper #4 (Chapter 5). Stöcker, B. Baier, D. (2020). A Better Understanding of Cost-related Dependencies in the Estimation of the Causal Effects in Direct Marketing Campaigns.
This paper is under review at the *International Journal of Research in Marketing*.
(VHB JOURQUAL 3: Category A)

Appendix B: Individual Contributions to the Included Research Papers

All included research papers were written in settings with multiple authors. In the following, I detail the settings and my individual contribution to each of the four papers included in this thesis.

Research Paper #1, which is presented in Chapter 2, was written by two researchers. The conception, design, and analysis were carried out by Björn Stöcker. Aydin Nasseri was the scientific support, assisted in the preparation of the presentation at the European Conference on Data Analysis (ECDA), and reviewed the paper.

Research Paper #2, which is presented in Chapter 3, was written by three researchers. The conception, design, and analysis were carried out by Björn Stöcker and Daniel Baier. The theoretical background and review were completed by Benedikt Brand. The first draft of the manuscript was written by Björn Stöcker, and all authors commented on previous versions of the manuscript

Research Paper #3, which is presented in Chapter 4, was written by two researchers, with Daniel Baier serving as the corresponding author and being mainly responsible for writing the article and analyzing the data. I was substantially involved in developing the overall motivation and preprocessing of the data as well as the discussion of the results. Furthermore, I contributed to the paper by reviewing it.

Research Paper #4, which is presented in Chapter 5, was written by two researchers. I served as the corresponding author. I conceived of the presented idea, developed the theory, performed the computations and wrote the paper. Daniel Baier verified the analytical methods and supervised the findings of this work. All authors discussed the results and contributed to the final manuscript.