Evaluation and Optimal Calibration of Purchase Time Recommendation Services

Benjamin Buchwitz
Catholic University of Eichstätt-Ingolstadt
benjamin.buchwitz@ku.de

Abstract

Price Comparison Sites enable customers to make better – more informed, less costly – buying decisions through providing price information and offering buying advice in the form of prediction services. While these services differ to some extent, they are comparable regarding their prediction target and usually monitor every arbitrarily small price decrease. We use a large data set of daily minimum prices for 272 smartphones consisting of 198,560 daily price movements from a Price Comparison Site to show that the standard prediction setting is not optimal. A custom evaluation framework allows the maximization of the achievable savings by altering the calibration of the forecasting service to monitor changes that exceed a certain threshold. Additionally, we show that time series features calculated in a calibration period can be used to obtain precise out of sample estimates of the saving optimal forecasting setting.

1. Introduction

Price Comparison Sites (PCS) offer clients useful insights into the pricing of consumer goods throughout the world wide web. The importance of Price Comparison Sites has grown quickly within the last few years. Customers use PCS either to gain knowledge of the spectrum of available retailers and prices or to improve their overview over the range of available items in a specific product category. Thus, PCS provide a reference price as well as comparability of products and vendors, influencing tremendously the consumers' offline price evaluation and therefore traditional stationary points of sales [1]. However, the information PCS provide are not tailored and enable clients only to extract information for a buying decision at a certain point in time. This shortcoming of the current decision support has been identified by the Price Comparison Sites themselves, which is why they provide more and more additional information like customizable price alarms or historic price information on their websites. Both features not only try to retain customers respectively force clients to revisit the site, but they also allow the customer to make better – more informed, less costly – buying decisions [2].

It is well known that historic price information and their visualization influence buyers in their decision [3], [4]. Price development expectations are influenced by historic prices and are directly linked to purchase time decisions [5]. As [6] points out, customers get easily overstrained with too many information and – in contrast to companies – are usually not equipped to predict future developments. Therefore, online shoppers are seldomly able to efficiently use and value the additional information presented by PCS.

For that reason, PCS recently began to explicitly advise their customers on buying decisions. Vivid examples of the attempt to provide customers with more tangible and detailed support systems can be seen when looking at Hopper.com, Kayak.com or AirHint.com, all focusing on airfares. Recommendations from PCS are usually given in the form of a dichotomous signal that suggests either to buy immediately or to wait for the price to drop. Kayak.com augments their advice with an explanation that "prices are unlikely to decrease within 7 days" [7], while AirHint.com states that "it is unlikely for the price to drop" and additionally presents the probability estimate for a price drop between today and the departure date of the monitored offer [8]. The largest German PCS Idealo, with over 42 million monthly visitors focusses mainly on consumer electronics. The reason for this is that electronic goods are homogenous and therefore easily comparable. Besides [9] showed that over 72% of buyers of electronic goods purchase online. While buying recommendations for electronic goods and airfares are comparable, the underlying data generation processes and thus the resulting time series properties differ. Flights between destinations reoccur and are managed by a revenue management engine, thus offer more features for prediction [10], [11].

PCS' for electronic goods themselves have yet failed to enhance their service by a recommendation system for optimal purchase time decisions. An approach to fill

URI: https://hdl.handle.net/10125/59627
ISBN: 978-0-9981331-2-6

this gap can be found in [12]. Due to the relevance of electronic goods for the e-commerce market, the developed methodology in this paper is tested using a data set from a PCS for consumer electronic goods.

1.1. Research Question and Approach

General information about the architecture and operating principles used in corporate purchase time recommendation systems is rather rare and cannot be extracted directly from the literature. The discussed examples reveal some of the underlying mechanics of the prediction generating systems while simultaneously showing possible parameterization options. On the one hand, at least some of the recommendation services differ in the period considered in the issued prediction and thus assume different decision and action horizons. On the other hand, all recommendation services resemble the same characteristics, when it comes to the actual prediction target. Each inspected service focusses on predicting the same price event – a simple price decrease – meaning all aim to monitor every, even arbitrarily small price changes. However, there are multiple reasons why defining and focusing on a different event definition makes sense:

First, setting economically noticeable price decrease values as prediction target increases the relevance of the prediction to the customer. It is well known and comprehensible that customers show varying sensitivity to differently sized price reductions as well as that this sensitivity differs between customer segments [13]. [14] additionally shows that consumers react differently to discount levels for various product segments and it is therefore necessary to model segment sensitive discounts. [15] go even further and are able to demonstrate that temporal discounts tailored to consumers' purchase timing significantly increase companies' profits. Therefore, it can be expected that a customer is willing to postpone her/his acquisition in case there is a sufficiently large price decrease. Additionally, it can be stated that the size of discounts is directly associated with its relevance and that specific e.g. product-specific characteristics need to be considered when determining relevant price decreases.

Second, introducing such thresholds for relevant or sufficient price decreases would also raise the average potential savings as well as the maximum potential savings for the customer, when following the recommendation. The underlying reason for this is that even though the number of customers that witness price decreases is lower when setting a threshold, while at the same time customers that experience a satisfactory price decline profit even more.

Changing the prediction target by explicitly defining a price decrease threshold can therefore not only increase customer satisfaction but also maximizes the savings PCS are able to generate for their customer base.

Thus, in this paper we aim to answer the question what the saving optimal price decrease threshold for calibrating a recommendation system is and how such a threshold can be estimated. We discuss whether it makes sense for a PCS to recommend waiting for small thresholds or if purchase time recommendation services should focus on larger price decreases.

While the information about the evaluation of time series forecasting approaches is quite large and well understood, the evaluation of decision recommendation scenarios is more complex and highly contextual. Additionally, even though decision-based evaluation approaches exist, they mainly focus on evaluating decisions conditional on singular horizons [16]. Contrary, decision recommendations in the context of purchase time recommendation services aggregate expectations about a single event over multiple horizons, while being conditional on the previous price level. Therefore, the economic perspectives are unclear and standard evaluation toolsets for assessments of classifiers cannot be applied as shown by [12].

We therefore introduce a specialized decision evaluation framework that allows the assessment of the economic potential of general prediction settings as well as the individual economic performance of externally given prediction services and show that such a framework can be used to identify optimal forecast settings independent of the employed prediction methodology.

1.2. Calibration Objectives and Restrictions

One can argue that a price decrease threshold is a buyer specific characteristic and has to be specified for each consumer individually [17]. Given explicit and excessive knowledge about the target audience, one may be able to identify a concrete value that reflects the significance of discounts for each specific group of interest. If sufficient information is available, one may even determine factors for each customer individually that accounts for the user-specific economic situation and individual valuation of discounts.

While this scenario would be ideal, given the excessive number of products carried by PCS and the large active user base, it seems unlikely to develop a method to generate precise and meaningful thresholds for each consumer individually. Besides, even though one might be able to calculate a user-specific price decrease threshold, this value cannot be transferred from one product to another due to different user preferences or significances of the product.

Additionally, this threshold is not time constant due to the urgency of product successors, personal schedules

and so on. Consider for example a customer who lost his smartphone and has to replace it. Such a customer has arguably a lower price sensitivity and her/his relevant price threshold is much higher, because the discount in prospect that is necessary for the customer to delay the purchase is larger.

Furthermore, for new customers the PCS must provide a default value to calibrate the recommendation service independent of a users' characteristics.

Consequently, one can and should set an optimization criterion for determining a sufficient saving for their customers. This could be done either by maximizing the number of customers that save money when using the recommendation system or by maximizing the savings of the entire client base.

The size of empirically observed price changes differs, depending on the type of product, price level, product properties and market conditions. Yet, small price changes are more likely to occur than larger ones, so that the total amount of situations, where a specific price change occurs, declines when the economic significance of the price event increases. Therefore, it is obvious that maximizing the number of clients experiencing a price decline can be achieved by setting the threshold as low as possible, meaning that each price fall is taken into account. But, this goes at the expense of the overall saving each customer could generate.

Considering the urge to generate relevant savings for their customers, there are two options to calibrate a PCS' recommendation service – either to maximize the average saving per recommendation or to optimize the maximum potential saving. However, the maximization of the potential savings for the whole user base is the only reasonable choice, because aiming to maximize the average saving per issued recommendation ignores the amount of issued predictions, thus the number of profiting clients. Optimizing the average saving would misguide the calibration objective and would lead to only predicting the single most valuable event when no auxiliary constraints are set. Therefore, we focus on setting a price decrease threshold that maximizes the sum of savings for all consumers.

When specifying a price decrease threshold, price changes smaller than the specified change are ignored; the customer is directed to only buy the product if the price is expected to drop more than the threshold. In exchange, this means that a customer is advised to buy directly even though the price might decline within the decision horizon, but the magnitude of the predicted price decrease is not sufficient. If the aforementioned reduction in the number of affected customers is offset by the increase in savings by leaping over the small changes, the total amount of possible savings increases, and customer satisfaction grows.

Customer satisfaction resembles a mixed effect of the actual performance and the accuracy of the used prediction procedure and is expected to be highly correlated with the generated savings. Specifying a target price level that is expected to be met within the decision horizon therefore has a direct effect on customer satisfaction as well as the potential savings that can be harvested through the prediction process. However, keep in mind that the actual savings of the customers heavily depend on the quality of the recommendation service itself and by that the usage rate of the system compared to the maximum potential saving. Setting a price decrease threshold influences first of all the maximum potential savings and is dependent on the way savings are calculated. This paper aims to advise on optimal calibrations for purchase time recommendation systems and outlines a conservative evaluation approach for the calculation of savings.

2. Decision Evaluation Framework

The goal of a purchase time recommendation methodology is to derive a binary buying recommendation for a specific and homogeneous consumer good. Here, we assume that the decision to buy the product is fix and that the consumer has decided which product s/he wants to buy. Additionally, we presume that the customer intends to buy the product within a decision horizon of H days. In this setting, the client has the option either to buy the product immediately or to delay the purchase to one of the later points in the decision horizon. If s/he decides to wait with the purchase, s/he will have the chance to obtain the product to a potentially lower price but will also have the risk of a price increase. The urge to save money when buying the product therefore creates uncertainty about the time, when to buy the product. We assume further that a purchase time recommendation service is in place to give each customer a dichotomous advice either to buy the product immediately or to wait because the recommendation system predicts that a "price drop occurs within the next H days". The advice as well as the observed reality can be used subsequently to evaluate each given recommendation using the confusion table explained in chapter 2.1 based on [12].

The primary source of information, which we use to develop and evaluate our calibration setting, is the knowledge about the historic price development $\{y_1, \dots, y_T\}$ up to the current point in time T. The recommendation given by the PCS's service reflects the expectation that at least one of the prices in the decision horizon $\{y_T, \dots, y_{T+H}\}$ will be smaller than the current price. When applying a price decrease threshold τ, the recommendation system adjusts its expectations to the

event and advises to wait with the purchase if it predicts $\min(y_{T+1}, \dots, y_{T+H}) < y_T - \tau$.

We explicitly used the "less than" condition to employ this general form also in the base case, where $\tau = 0$ respects every price decrease regardless of its magnitude. Reformulating the event definition so that the event includes phases with constant prices, by using the "less than or equal" condition, would lead to classifying price changes as valuable and therefore issuing a waiting recommendation, despite the fact that customers can not generate any saving. To keep the following explanations, compact and lucid, we focus on an *absolute* price decrease threshold τ instead of a scale factor to model a *relative* threshold. However, results for both approaches are consistent.

To analyze the impact of variations of τ on savings, first an evaluation framework needs to be defined. The case where every price change is monitored ($\tau = 0$) will be referred to as the base case, while $\tau > 0$ resembles the extended evaluation scenario. In the first case, the evaluation of savings is straightforward as the statistical evaluation arises directly out of the evaluation of binary classifiers. Contrary to this, the extended evaluation requires a more detailed breakdown of events and hence an extension of the evaluation of the base scenario.

2.1. Base Scenario

As the generated recommendation as well as the observed event are dichotomous, the match between predictions and observations form the confusion table, shown in figure 1a where four situations occur.

True Positive (TP): In case that the purchase decision has been delayed by the recommendation system and the price drops, the advice was correct and a saving is obtained. If multiple price drops occur within the decision horizon, savings are calculated as the difference between the price when the recommendation is issued and the first price decrease. This represents a conservative and consistent way for the economic evaluation and resembles the lifelike situation of a customer in the buying process. It can be assumed that a customer waits for the price to fall (if the recommendation tells him to) and buys at the first occasion when s/he observes a price decrease.

False Positive (FP): When the purchase of the product has been delayed due to the recommendation, but the price does not decrease, the advice was inaccurate. It is assumed that the customer will wait with her/his acquisition until the end of the decision horizon, hoping that the price will fall. If this does not happen, s/he will purchase at the end of the decision horizon. The difference between the price at the beginning and the price at the end of the decision horizon yields the customer's additional cost, thus a loss.

False Negative (FN): A price decrease occurs within the decision horizon, but the recommendation was to buy the product immediately. Therefore, the decision was wrong, and the corresponding loss is calculated as the difference between the price at the beginning (for which the product was bought) and the first price drop (for which the product would have been bought otherwise).

True Negative (TN): The product has been bought for the price at the beginning and the price stays constant or increases during the decision horizon, thus the decision was exact. The saving is calculated as the difference between the price when the recommendation was issued and one of the higher or equal prices within the decision horizon. To select one of these prices a convention is needed as the first price increase as well as the last price are no reasonable benchmarks, because it is unlikely that the customer would not have bought the product for this price. To solve this issue and to settle on a course of action, we select the minimum price within the decision horizon. This leads to the smallest possible saving and thus is the most conservative valuation.

The resulting gain g_T for a specific forecasting origin T dependent on the outcoming state $s_T \in \{TP, FN, FP, TN\}$ is given by

$$g_T = \begin{cases} y_T - y_{T+h'} > 0, & if\ s_T = TP \\ y_{T+h'} - y_T < 0, & if\ s_T = FN \\ y_T - y_{T+H} \leq 0, & if\ s_T = FP \\ \min\{y_{T+h}\} - y_T \geq 0, & if\ s_T = TN. \end{cases}$$

Here, $T + h$ with $h \in \{1, \dots, H\}$ denotes the periods of the decision horizon, whereby $T + h'$ represents the first period in the decision horizon, where the defined event is reached and therefore the respective price fulfills $y_{T+h'} < y_T - \tau$. The scenarios where a gain is generated are shaded in figure 1.

It is important to note, that only cases classified as TP and FP generate a real gain respective loss for the customer. The TN and FN scenarios constitute just hypothetical outcomes because the customer already bought the product at the beginning of the decision horizon and therefore did not actually lose money. Besides, it can be argued that the customer's initial situation is that s/he intends to buy a product and has no additional information about the future price development and would therefore buy the product immediately for the best price advertised on the PCS. Recommendations in the described spirit interfere with this standard behavior and suggest that the user can save money if s/he follows the advice to delay the purchase. For the following analysis, we therefore primarily consider the cases where the modus operandi has been changed, and the purchase was delayed (TP and FP).

The total economic impact g_R now results from summing up every real gain over all observations for which a prediction was or could have been issued and is therefore given by

$$g_R = \sum_{t=1}^{N} g_t^{(s_t)} \; with \; s_t \in \{TP, FP\},$$

with N denoting the last time point in the price time series for which a prediction can be evaluated. Given a time series $\{y_1, \dots, y_M\}$ with M consecutive observations and a recommendation setting with a decision horizon of H days, it follows that $N = M - H$. $t = 1$ corresponds to the first decision made after the calibration phase of C days (if applicable). Calculating the economic impact in this manner implies that in every time window of H days, exactly one customer receives and follows a recommendation from the PCS' service. Thus, for the entire time series $N - C$ recommendations are issued. Because we assume that a prediction system is in place, C can be neglected for now. The simplification of one buyer per product per decision horizon allows isolating the effect of calibrating price decrease threshold clearly.

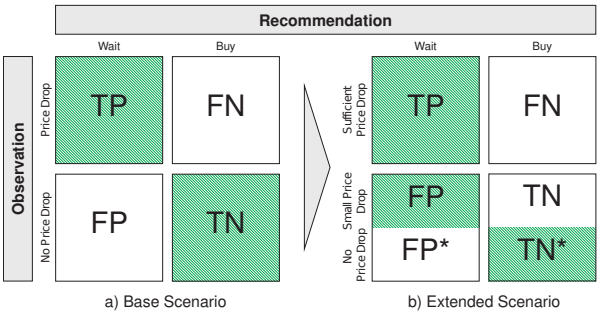

Figure 1. Confusion table for base and extended case.

2.2. Extended Scenario

While the base case considers every price decrease, a more relevant calibration can be achieved by setting $\tau > 0$, meaning only price declines larger than τ units are considered.

For the performance analysis of true positive and false negative scenarios no adjustments are necessary compared to the base scenario. When delaying the buying decision and the expected price drop occurs (TP) the resulting saving can be calculated as specified in the base case. The same holds if a price decrease occurs and the customer bought immediately (TN).

Consequences for the performance evaluation arise when the resulting price drops but not to the magnitude expected. The adjusted evaluation framework is illustrated in figure 1b. Imagine the situation, where a customer specified and thus expects a price reduction of τ units. If the recommendation service calculates a high chance that a compliant price drop occurs, it will instruct the customer to delay her/his purchase. As the recommendation expresses the belief that the specified price drop will occur, the customer would ignore smaller price decreases and waits until the observed price undercuts her/his modified reference price $y_T - \tau$. If the calibrated price decrease does not appear (FP) and consequently, the customer waits and buys the product at the end of her/his decision horizon, while at this point in time the price is lower than the price when the prediction was made (y_T) but bigger than $y_T - \tau$, evaluation problems arise.

Technically, the issued recommendation is erroneous, and the evaluation does not deviate from the base case, so that $s_T = FP$ and $g_T = y_T - y_{T+H}$. However, compared to the price at period T at which the prediction was generated the customer still saves money ($g_T > 0$), so that the economic and statistical evaluation views fall apart. While the economic impact is calculated on the premise of an occurring purchase and can therefore be categorized as saving or as loss, the statistical evaluation cannot be neglected. The reason for this, is that the customer trusting the recommendation expects a price decrease of τ units. Even though s/he still saves money, her/his expectations are not met and therefore s/he will be dissatisfied, especially if one of the prices within the decision horizon was lower than the price at the end. To align with the transaction-based view of the evaluation framework, we classify the economic outcome as a small, not sufficient but real saving, but distinguish this state from the ones defined in the base case, so that $s_T = FP^*$ for false positive scenarios with resulting savings.

A comparable evaluation dilemma occurs in the case of true negative recommendations, where the purchasing decision has not been delayed and the specified price drop does not occur. However, the price drops below y_T within the decision horizon but not by τ units. Technically, the recommendation is correct and would result in a positive loss (saving) when applying the evaluation framework from the base scenario. This approach leads to an unrealistic overestimation of savings in case of multiple (small) price drops within the decision horizon. To be more conservative and realistic, we suggest referring to the first price decrease as reference for the calculation of a small, not sufficient and hypothetical loss when $s_T = TN^*$. Following the same notation as in the base case the economic implications of the two additionally emerging states are given by

$$g_T = \begin{cases} y_T - y_{T+H} > 0, & if \; s_T = FP^* \\ y_{T+h'} - y_T < 0, & if \; s_T = TN^*. \end{cases}$$

2.3. Application Possibilities

The developed framework serves multiple distinct purposes. First, assuming the availability of multiple prediction techniques, the presented evaluation scheme can be used to compare and decide between methodological alternatives.

Second the approach can be used to evaluate the general forecasting setting. Assuming a flawless recommendation system allows to calculate an upper limit of gains $\overline{g_R}$ that can be achieved under the current parameterization. Contrary, if all recommendations are erroneous, the economic impact reaches its lowest value and (assuming fluctuating prices) is expected to be negative. Usually, these boundaries for savings and losses are not equal in size, so that the user is exposed to and confronted with an asymmetric risk function.

Because $\overline{g_R}$ is a function of τ, all possible thresholds can be compared by assessing the gain potential that they unleash. By evaluating this setting for all possible values of τ an optimal threshold can be identified.

3. Empirical Analysis

The following chapter focusses on the analysis of potential gains and their dependency to the price decrease threshold. Besides, it will be shown how to calibrate a gain optimal threshold ex-post and ex-ante using the example of a large data set from a German PCS for consumer electronic goods.

3.1. Data set and Descriptive Statistics

Basis for the following analysis is a sample from a PCS for the German consumer electronic market. The data set consists of 272 smartphone price time series from well-known and established brands. The products stem from different market segments and have different initial prices. Additionally, the time series show different concentrations of non-changing price phases and thus differ in their intensities of volatility clustering. To make the results comparable, we focus on the length of a typical product life cycle of an electronic consumer product and analyze the first two years of data, resulting in 730 observations for each item. In total, this yields a sample of 198,560 daily minimum prices. Products with less than 730 observations, not enough price movements (extremely high zero-inflation of price changes) or obvious data errors have not been considered and were removed beforehand. All time series represent a specific entity with completely homogeneous properties and features, meaning that phones with different memory sizes, brands or colors constitute different time series.

Figure 2 shows a characteristic example of a price time series of a smartphone. The price time series displays typical features of a technological consumer good. First, the level y_t exhibits the expected price deterioration and the descent of the price constantly declines until the level reaches approximately half of the starting value. Second, calm and active market phases are visible in the price graph, so that the product shows fluctuating prices. Third, periods with constant prices, meaning that the day-to-day price does not change, make up for a significant proportion of the series. These constant phases represent roughly 45% of the 730 observations and are indicated by tick marks on the abscissa in figure 2.

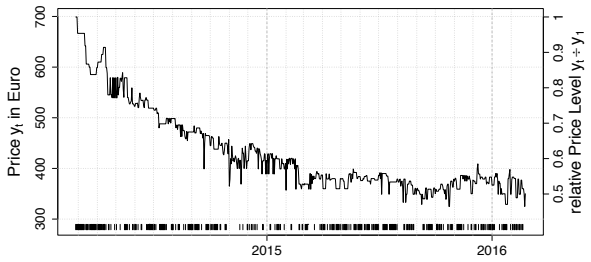

Figure 2. Price time series of a smartphone between 25 Feb. 2014 and 24 Feb. 2016; units in €.

The initial price of the product displayed is 699.00€. The average over all price changes amounts to -0.48€, resulting in an average price over 2 years of 430.31€. The descriptive statistics for the whole sample, aggregated and averaged over all products of a brand are shown in table 1.

Table 1. Descriptive Statistics.

Brand	N	Avg. Initial Price in €	Average Price in €	Avg. Daily Price Change in €
Apple	46	834.98	661.68	-0.40
HTC	19	536.56	358.21	-0.33
Nokia	20	350.62	192.40	-0.25
Samsung	76	542.58	339.78	-0.34
Sony	34	488.55	312.09	-0.28
Others	77	324.15	212.36	-0.20
All	272	508.91	345.14	-0.29

The brand with the largest portfolio in the sample is Samsung with 76 different phones, followed by Apple with 46 phones. Brands that offer less than 15 products, such as LG, Huawei, Lenovo-Motorola, Microsoft or Google have been bundled together and are represented by the "Others" group. The average initial price as shown in table 1 is calculated as the arithmetic mean of the first observed price y_1 for all products of the respective brand. It coincides often, but not in all cases, with the manufacturer's suggested retail price. The brand with the most affordable average initial price is Nokia with 350.56€, while Apple offers the lineup with the highest average starting price. The average price shown in the second column displays the same pattern. As all products show strong price deterioration the

average day-to-day price movement is negative and peaks with an average decrease of 0.40€ in the Apple lineup. As the "Others" group consist of a considerably large amount of budget phones that have low initial prices as well as low average prices, their average daily price reduction of 0.20€ is the lowest among the brands.

3.2. Effects of Calibration

Based on the presented dynamics in the prices the resulting effect of setting and deviating a price decrease threshold can now be examined. Because customers have usually quite precise ideas about when they need a product [18], we decide to use a decision horizon of seven days ($H = 7$). While we focus on the presentation of results for fixed H the main findings and conclusions do not change if the decision horizon is altered.

Figure 3 shows the effects of setting varying thresholds on the resulting saving potential $\overline{g_R}$ for the representative product introduced in figure 2. The solid black line is associated with the left ordinate and displays $\overline{g_R}(\tau)$, the maximum saving potential as function of the price decrease threshold, which in turn is shown on the abscissa. The blue, circled line corresponds to the right ordinate and shows the relative amount of price decrease events that were observed in the price time series given a specific price decrease threshold. A relative event occurrence count of 0.5 denotes that within the 723 periods in half of the time windows the specified event occurs and thus represents a perfectly balanced sample.

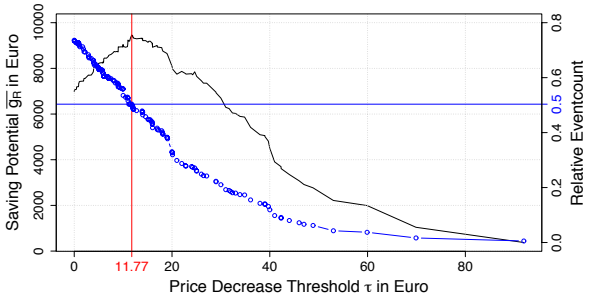

Figure 3. Effect of variating the price decrease threshold on saving potential and event occurrences.

The initial calibration of $\tau = 0$ leads to possible savings of 6979.42€ generated through 532 of 723 (73.58%) event occurrences. With rising τ the gain potential also increases and peaks at $\tau = 11.77$€ with a possible saving of 9465.71€; an increase of 35.62% compared to the default threshold value. The gain optimal price decrease threshold τ_t^* corresponds to a relative event count of 50.35% and thus counters the natural imbalance of the prediction setting.

The variation of thresholds can also be observed among different brands. The boxplots for the saving optimal thresholds by brand are shown in figure 4. The highest median threshold is found in the Apple lineup, whereas the smallest saving optimal threshold can be observed for Nokia phones. Besides, the variances of the optimal thresholds clearly differ by brand. HTC phones show a comparably small deviation of only 1.88€ contrary to the standard deviation of Samsung phones of 4.99€.

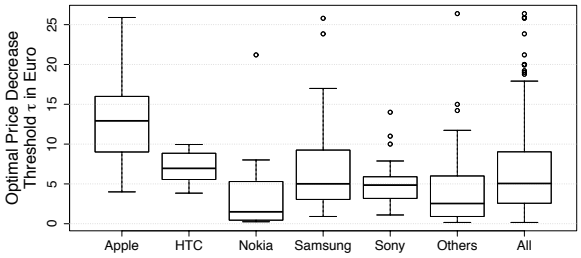

Figure 4. Distribution of optimal saving potential price decrease thresholds by brand.

3.3. Optimal Calibration

Searching for a universal, global threshold to optimize the saving potential is not promising because as figure 4 showed thresholds heavily depend on the brand affiliation. Some of the saving potential values quickly overshoot the optimum, while others do not receive a ramp up that is worth mentioning.

In a practical calibration setting, a defined amount of price observations to forecast the contained time series dynamic and to form an expectation about the future price development is needed. This in return means that a considerable amount of price observations has to be available, when issuing the first recommendation. This first fraction of the time series can then also be used to calculate a gain optimal price decline threshold. Given a calibration time of $C = 90$ days, the saving optimal threshold τ_{90}^* emerges. This threshold can be interpreted as an early estimate for the ex-post gain optimal threshold τ_{730}^* at the end of a products' life cycle.

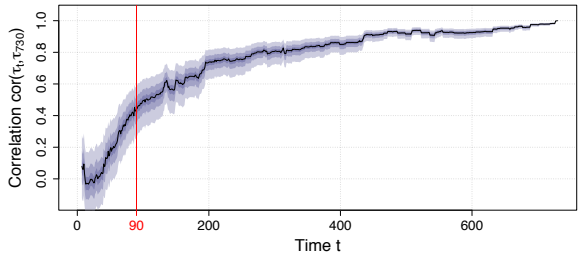

Figure 5. Correlation between ex-post saving potential optimal threshold τ_{730}^* and previous thresholds τ_t^*.

Figure 5 shows the correlation of the threshold τ_t^* after t days with τ_{730}^*, which is indicated by the black solid line. The shaded areas around the line correspond

to the 99%, 90% and 80% confidence intervals of the correlation coefficient, which become narrower over time. The correlation coefficient was calculated utilizing all 272 price time series in the sample. The correlation at $t = 90$ turns out to be 0.44 so that τ_{90}^* seems to be weakly informative with respect to the final threshold at this early stage. Therefore, using τ_{90}^* as calibration threshold is not promising.

However, an early stage τ_t^* can be included in a broader setting to predict τ_{730}^*. The price decrease threshold at any point in time is always a function of the realized price path and captures certain characteristics that represent a share of the information present in the time series. To extract additional characteristics from the analyzed time series, we revert to a collection of time series features that are calculated on the shortened time series $\{y_1, \dots, y_C\}$ with $C \in \{90,180,360\}$. The idea of characterizing time series by their features has been previously used for classification and identification of outliers or anomalous series in time series databases and collections [19], [20]. In the following analysis, we consider five features in addition to the gain optimal threshold τ_C^* that we found to be informative.

Price Level y_1: As indicator for the price level and the corresponding market segment of the product, the price level in form of the initial price observation of the time series is extracted. One would expect a positive effect of the price level on the threshold because when relative discounts are granted, they result in a bigger absolute price reduction. Besides, generally for higher prices small rebates are easier to achieve.

Autocorrelation ACF10: To obtain a general appeal of the correlation structure that also respects the price deterioration; we take the sum of the squared first ten autocorrelation coefficients [21]. A high value shows that future values of the series are more dependent on past values, whereas a low value indicates a higher amount of random noise in the series.

Curvature: The curvature measures the degree of non-linearity in the price deterioration of the series. This feature is obtained through fitting a second order polynomial regression on the trend component obtained by the Season-Trend-Level decomposition of the price time series [22], [23]. Curvature is the second order coefficient that is extracted from the regression results.

Spectral Entropy: Assessing the role of uncertainty or noise and the complexity of a signal using entropy based measures is an approach frequently employed in information theory and actively used in the literature [24]–[27]. Entropy is also used as a measure for "forecastability" of series in the context of classical time series forecasting [24], [28]–[30]. For the calculation we use an estimator of the Shannon entropy of the spectral density of the series, which represents the importance of different frequencies in the data. The measure is described in more detail in [31]. A relatively high value suggests more uncertainty and therefore a hard to forecast series. Small values of the Shannon entropy are found when the series contains more signal and therefore is richer in information and easier to forecast. The effect on the gain optimal threshold is expected to be positive because the more price movement the higher the chances of capturing a big discount.

Brand Indicator: As seen before the analyzed products and their respective saving optimal thresholds at the end of the life cycle are dependent on the brand affiliation. We therefore include an indicator for the brand capturing all brand-specific effects aside from different levels of initial prices.

Table 2. Regression Results.

	Dependent variable:		
	Threshold τ_{730}^*		
	$C = 90$	$C = 180$	$C = 360$
Threshold τ_C^*	0.086***	0.133***	0.529***
	(0.030)	(0.035)	(0.044)
y_1	0.011***	0.011***	0.004***
	(0.001)	(0.001)	(0.001)
ACF10	0.494***	0.838***	0.378
	(0.185)	(0.208)	(0.262)
Curvature	-0.232***	-0.149**	-0.018
	(0.075)	(0.060)	(0.043)
Entropy	13.265***	17.399***	5.224
	(3.465)	(4.212)	(4.549)
Brand: HTC	-2.112**	-1.041	-2.883***
	(0.947)	(0.946)	(0.815)
Brand: Nokia	-2.481**	-1.295	-3.083***
	(0.970)	(1.003)	(0.836)
Brand: Samsung	-2.445***	-1.301*	-2.874***
	(0.708)	(0.750)	(0.649)
Brand: Sony	-3.116***	-2.079**	-3.206***
	(0.814)	(0.846)	(0.757)
Brand: Others	-2.562***	-1.361*	-2.619***
	(0.749)	(0.753)	(0.664)
Constant	-7.883**	-13.421***	-2.303
	(3.114)	(3.900)	(4.477)
Observations	272	272	272
Adjusted R^2	0.661	0.671	0.772
Residual Std. Error	3.091	3.046	2.535
F Statistic (df = 10; 261)	53.775***	56.149***	92.649***
Note:		*p<0.1; **p<0.05; ***p<0.01	

Table 2 shows the results from regressing the gain optimal threshold τ_{730}^* on the described features calculated on the first 90, 180 and 360 observations of the respective time series. The values in parentheses represent standard errors of the respective estimates.

The regression results show that the selected features have a significant effect on the threshold. Price level, autocorrelation structure, curvature as well as the entropy are highly significant in the first 90 days. All mentioned coefficients have a positive sign, meaning higher feature values, lead to higher thresholds. The coefficient of the curvature is the only exception; an increase in curvature decreases the saving optimal threshold. The constant of the regression line refers to

the Apple brand as cornering effect. As expected, the differentiation by brand is significant even at an early stage. All brand indicators exhibit a negative sign, due to the cornering effect. Signs and effects shown in table 2 are consistent throughout all models. Additionally, the effect τ_C^* on the ex-post optimal threshold expectedly grows and therefore follows the pattern shown in the time dependent correlation structure in figure 5. Consequently, the influence of the features decreases and mostly vanishes in the middle of the life cycle, where essentially the time dependent optimal threshold τ_{360}^* shows increasing correlation with the dependent variable. Only the price level and brand indicator are of value when using a 360-day calibration period. As expected the explained variance of the model increases, as the adjusted R^2 grows from 66.1% to 77.2% explained variation of the ex-post gain potential optimal threshold. Also, the residual standard error decreases, which is consistent with the applied reasoning of shrinking discrepancy from the threshold at 730 days with growing price history.

To receive reliable out of sample forecasts the results shown in tables 3 have been computed on the basis of regression equations obtained through leave-one-out cross-validation (LOOCV) before aggregating them by brand. The results are robust in the sense that 10-fold cross-validation shows comparable out of sample estimates and yields the same findings. Because only positive values for price decrease thresholds are meaningful, negative thresholds are set to zero.

It can be shown that the brand-averaged thresholds at an early stage are already very close to the optimal ex-post values. In general, the difference of the time point-specific thresholds to the optimal value decreases when expanding the information base. After a year of data and therefore in the middle of the product life cycle, the shown estimates are almost identical to the ones at the end of the life cycle.

To evaluate the economic quality of the predicted thresholds, we calculate the resulting saving potentials applying estimated thresholds as calibration parameters. The first column of table 3 displays the base value for $\tau = 0$. The last column shows the maximum possible saving if τ_{730}^* is set. When using estimates conditional on the information base the resulting potential averaged by brand is shown in the three remaining columns.

Table 3. Saving Potential based on LOOCV; units in €.

Brand	Base	90 days	180 days	360 days	Max.
Apple	9959.44	12188.68	12177.11	12222.81	12457.76
HTC	5309.90	6364.78	6374.56	6380.75	6477.05
Nokia	2718.99	3305.39	3283.44	3281.42	3381.22
Samsung	4859.22	6014.29	6036.81	6043.84	6192.63
Sony	4134.32	5067.91	5062.34	5050.60	5190.44
Others	3070.75	3607.21	3620.78	3664.51	3740.29
All	4998.96	6061.55	6068.16	6088.96	6245.82

All estimated values are close to the optimal value and in general grow with the amount of observations included in the estimation of the threshold. Interestingly, the values resulting from the estimation are comparable and do not differ much, which indicates a successful calibration. Employing the thresholds that were estimated for the HTC products, results in a gain potential of 6364.78€ after 90 price observations, which accounts for 98.27% of the maximum potential. Even in the heterogeneous "Others" group, the same early estimate results in 96.44% potential generation. Disturbances in the gain potential over time are minimal and only occur in the case of Nokia and Sony. Compared to a global threshold for all products, our customized estimations yield considerably better prediction settings for all brands as well as higher overall gain potential.

The single-sided t-Test indicates that the saving potential calculated with an adjusted threshold, in all cases outperforms the base case. All differences are highly significant at the $p<0.001$ level, also when using the Wilcox-Test. The F-Test in the analysis of variance provides strong evidence that the emerging gain potential, when using our estimation and forecasting approach, do not significantly differ from each other or from the ex-post gain optimal values. Therefore, the estimation procedure delivers reasonable performance and allows for economically meaningful estimation of thresholds already at an early stage.

4. Conclusions and Future Work

In this study, we showed the need for calibrating purchase time decision services used on PCS and demonstrated that the common prediction setting is not optimal. We evaluated different options of calibrating recommendation services by applying an evaluation framework for binary classifiers adapted from [11], described gain optimal scenarios and showed that the resulting gain potential significantly improves when following our approach. Additionally, we provide options to estimate and calibrate the gain optimal threshold under incomplete information. Time series features offer a great value added and allow predicting gain optimal thresholds at an early stage of the product life cycle. Their influence on the time series decreases with growing price history and the informative value of the gain optimal threshold increases, which also allows improving economic significance over time.

The findings of this paper raise several starting points for future research contributions. In addition to the presented features it may be possible that additional features enable the improvement of the estimation process. The huge variety of feature generation methods allow using and developing regularization or decomposition techniques that automatically select the

best feature set with respect to the asymmetric risk function to further improve results. For practical purposes the exemplary assumption of a one-unit demand per decision can also be exchanged with more complex and contextual demand patterns. By choosing a specific prediction engine and evaluating their performance using the evaluation framework, one can analyze the dependencies between the presented gain potential and the method specific gain exploitation. This can support PCS by making better recommendation services available to a broad range of customers.

5. References

[1] H. O. Bodur, N. M. Klein, and N. Arora, "Online Price Search: Impact of Price Comparison Sites on Offline Price Evaluations," *J. Retail.*, vol. 91, no. 1, pp. 125–139, 2015.

[2] M. Smith, "The impact of shop bots on electronic markets," *J. Acad. Mark. Sci.*, vol. 30, no. 4, pp. 442–450, 2002.

[3] W. Drechsler and M. Natter, "Do Price Charts Provided by Online Shopbots Influence Price Expectations and Purchase Timing Decisions?," *J. Interact. Mark.*, vol. 25, no. 2, pp. 95–109, 2011.

[4] N. H. Lurie and C. H. Mason, "Visual Representation: Implications for Decision Making," *J. Mark.*, vol. 71, no. 1, pp. 160–177, 2007.

[5] S. Danziger and R. Segev, "The Effects of Informative and Non-Informative Price Patterns on Consumer Price Judgments," *Psychol. Mark.*, vol. 23, no. 6, pp. 535–553, 2006.

[6] D. Bawden and L. Robinson, "The dark side of information: Overload, anxiety and other paradoxes and pathologies," *J. Inf. Sci.*, vol. 35, no. 2, pp. 180–191, 2009.

[7] P. Sarkar, *Data as a Service - A Framework for Providing Reusable Enterprise Data Services*. Hoboken, NJ: Wiley, 2015.

[8] A. Udachny, "When to buy Ryanair tickets? - Presenting AirHint.com," 2015. [Online]. Available: http://www.airhint.com/about.

[9] P. Schrader, *E-Commerce Trends 2018: Wünsche und Ängste von Online-Shoppern [Desires and Fears of Online Shoppers]*. Berlin, Germany: Idealo, 2018.

[10] W. Groves and M. Gini, "A regression model for predicting optimal purchase timing for airline tickets," *Dep. Comput. Sci. Eng. Univ. Minnesota - Tech. Rep.*, pp. 1–17, 2011.

[11] O. Etzioni, R. Tuchinda, C. A. Knoblock, and A. Yates, "To buy or not to buy: Mining airfare data to minimize ticket purchase price," in *Proceedings of the ninth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 2003, pp. 119–128.

[12] B. Buchwitz and U. Küsters, "A Time Series based Monitoring Methodology to optimize Purchase Timing Decisions," *SSRN Work. Pap. 3179987*, pp. 1–32, 2018.

[13] C. Narasimhan, "A Price Discrimination Theory of Coupons," *Mark. Sci.*, vol. 3, no. 2, pp. 128–147, 1984.

[14] A. Greenstein-Messica, L. Rokach, and A. Shabtai, "Personal-discount sensitivity prediction for mobile coupon conversion optimization," *J. Assoc. Inf. Sci. Technol.*, vol. 68, no. 8, pp. 1940–1952, 2017.

[15] J.-P. Dubé, Z. Fang, N. Fong, and X. Luo, "Competitive Price Targeting with Smartphone Coupons," *Mark. Sci.*, vol. 36, no. 6, pp. 944–975, 2017.

[16] M. P. Clements, *Evaluating Econometric Forecasts of Economic and Financial Variables*. New York, NY: Palgrave Macmillan, 2005.

[17] J. Johnson, G. J. Tellis, and E. H. Ip, "To Whom, When, and How Much to Discount? A Constrained Optimization of Customized Temporal Discounts," *J. Retail.*, vol. 89, no. 4, pp. 361–373, 2013.

[18] E. Meierhoff, *Data based Overview of the German E-Commerce Market*. Berlin, Germany: Idealo, 2018.

[19] F. Mörchen, "Time series feature extraction for data mining using DWT and DFT," *Dep. Math. Comput. Sci. Univ. Marburg, Ger. - Tech. Rep.*, vol. 33, pp. 1–31, 2003.

[20] A. Nanopoulos, "Feature-based Classification of Time-series Data," *Int. J. Comput. Res.*, vol. 10, no. 3, pp. 49–61, 2001.

[21] R. J. Hyndman, E. Wang, Y. Kang, and T. Talagala, "tsfeatures: Time Series Feature Extraction. R package version 0.1." 2018.

[22] R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. Terpenning, "STL: A seasonal-trend decomposition procedure based on loess," *Journal of Official Statistics*, vol. 6, no. 1. pp. 3–73, 1990.

[23] R. J. Hyndman and Y. Khandakar, "Automatic time series forecasting: The forecast package for R," *J. Stat. Softw.*, vol. 27, no. 3, pp. 1–22, 2008.

[24] Y. Kang, R. J. Hyndman, and K. Smith-Miles, "Visualising forecasting algorithm performance using time series instance spaces," *Int. J. Forecast.*, vol. 33, no. 2, pp. 345–358, 2017.

[25] R. J. Hyndman, E. Wang, and N. Laptev, "Large-Scale Unusual Time Series Detection," in *Proceedings of the 15th IEEE Int. Conf. on Data Mining Workshop*, 2015, pp. 1616–1619.

[26] C. Bandt and B. Pompe, "Permutation Entropy: A Natural Complexity Measure for Time Series," *Phys. Rev. Lett.*, vol. 88, no. 17, pp. 1–4, 2002.

[27] E. Maasoumi and J. Racine, "Entropy and predictability of stock market returns," *J. Econom.*, vol. 107, no. 1–2, pp. 291–312, 2002.

[28] G. M. Goerg, "Forecastable Component Analysis," in *JMLR Workshop and Conference Proceedings*, 2013, vol. 28, pp. 1–9.

[29] B. D. Fulcher, M. A. Little, and N. S. Jones, "Highly comparative time-series analysis: the empirical structure of time series and their methods," *J. R. Soc. Interface*, vol. 10, no. 83, pp. 1–12, 2013.

[30] J. Garland, R. James, and E. Bradley, "Model-free quantification of time-series predictability," *Phys. Rev. E*, vol. 90, no. 5, pp. 1–15, 2014.

[31] G. M. Goerg, "ForeCA: An R package for Forecastable Component Analysis. R package version 0.2.4." 2016.