Nr. 28
23. November 2016

Leopold-Franzens-Universität Innsbruck

Preprint-Series: Department of Mathematics - Applied Mathematics

A Machine Learning Framework for Customer Purchase Prediction in the Non-Contractual Setting

A. Martínez, C. Schmuck, S. Pereverzyev Jr., C. Pirker, M. Haltmeier

Technikerstraße 13 - 6020 Innsbruck - Austria
Tel.: +43 512 507 53803 Fax: +43 512 507 53898
https://applied-math.uibk.ac.at

A Machine Learning Framework for Customer Purchase Prediction in the Non-Contractual Setting

Andrés Martínez[a], Claudia Schmuck[b], Sergiy Pereverzyev Jr.[b], Clemens Pirker[c], Markus Haltmeier[b,*]

[a] *Coolblue BV, Weena 664, 3012 CN Rotterdam, The Netherlands*
[b] *Department of Mathematics, University of Innsbruck, Technikerstraße 13, 6020 Innsbruck, Austria*
[c] *Department of Strategic Management, Marketing and Tourism, University of Innsbruck, 6020 Innsbruck, Austria*

Abstract

Predicting future customer behavior provides key information for efficiently directing resources at sales and marketing departments. Such information supports planning the inventory at the warehouse and point of sales, as well strategic decisions during manufacturing processes. In this paper, we develop an advanced analytics tool that predicts future customer behavior in the non-contractual setting. We establish a new dynamic and data driven framework for predicting whether a customer is going to make purchase at the company within a certain time frame in the near future. For that purpose, we propose a new set of customer relevant features that derives from times and values of previous purchases. These customer features are updated every month, and state of the art machine learning algorithms are applied for purchase prediction. In our studies, the gradient tree boosting method turns out to be the best performing method. Using a data set containing more than 10 000 customers and a total number of 200 000 purchases we obtain an accuracy score of 89 % and an AUC value of 0.95 for predicting next moth purchases on the test data set.

Keywords: analytics, purchase prediction, sales forecast, non-contractual setting, machine learning

1. Introduction

Customer management requires that firms make a careful assessment of the costs and benefits of alternative expenditures and investments, and identify the optimal allocation of resources to marketing and sales actions over time. Decision makers will benefit from decision support models that relate costs and customer purchase behavior, and forecast the value of the customer portfolio (Berger et al., 2002). Thus, knowing who is likely to purchase within the next months is one of key drivers to allocate efficiently resources at the sales and marketing departments (see e.g. (Allenby et al., 1999)). This information is also needed when planning the inventory at the warehouse and/or point of sales, as well for deciding quantities at the manufacturing processes. Thereby, the non-contractual distinction is of fundamental importance for developing models for customer-base analysis. One of the main challenges in the non-contractual settings is how to differentiate customers who have ended their relationship with the firm from those who are simply in the midst of a pause between transactions.

It is widely accepted by business wisdom and research literature that it costs five to ten times more to acquire a new customer than to retain an existing customer (Daly, 2002; Bhattacharya, 1998). While the factor itself may vary substantially depending on the business context, re-

*Corresponding author
Email addresses: jamartinez@gmail.com (Andrés Martínez), claudia.schmuck@hotmail.com (Claudia Schmuck), sergiy.pereverzyev@uibk.ac.at (Sergiy Pereverzyev Jr.), clemens.pirker@uibk.ac.at (Clemens Pirker), markus.haltmeier@uibk.ac.at (Markus Haltmeier)

taining customers has received strong attention from both academia and practitioners (see Van den Poel and Lariviere (2004) for an overview). Thereby, it has been well established that appropriate retention strategies have strong benefits over acquisition approaches (see Ganesh et al. (2000)). However, it has to be noted that retention activities are not necessarily desirable in an unconditional way, since targeting profitable customers can make marketing spending more efficient (Mulhern, 1999; Zeithaml et al., 2001; Kumar et al., 2008), even more if this profitability can be predicted (Reinartz and Kumar, 2003). Among practitioners, it is quite desirable to consider customers' future profitability and responsiveness, specifically in terms of purchase actions, to marketing when allocating resources (Rust et al., 2011; Venkatesan and Kumar, 2004).

Firms are encouraged to develop models to predict which customers are more likely to defect (Keaveney and Parthasarathy, 2001; Neslin et al., 2006). Once identified, these likely defectors should be targeted with appropriate incentives to convince them to stay (Hadden et al., 2007).

1.1. Customer purchase prediction

While purchase prediction has received attention for a long time in consumer research (see e.g., (Herniter, 1971), the rise of customer analytics by marketing analysts has revived such issues in the recent years (Winer, 2001). As outlined in (Platzer and Reutterer, 2016), one of the most challenging areas remains the prediction of customer purchases in the non-contractual settings: The current status of the customer is not directly observable at a time T and the available historical record is censored at point T while customer data tends to vary substantially.

During the last years, large improvements in the information technology domain have resulted in the increased availability of customer transaction data (Fader and Hardie, 2009). Initial analyses of these transaction databases are usually descriptive in form of basic summary statistics such as the average number of orders or the average order size, and information on the distribution of behaviors across the customer portfolio. Further analyses of the customer base may use multivariate statistical methods and data mining tools to identify characteristics of, for instance, heavy buyers, or to determine which groups of products tend to be purchased together (i.e., performing a market-basket analysis).

The next step in terms of analysis is to undertake customer-base analysis activities that are more predictive in nature (Fader and Hardie, 2009). In this work we develop a machine learning framework for forecasting future purchasing by the firm's customers from a given customer transaction database. For that purpose we first compute a large number of customer features, that characterizes the customer at a given month. We then apply machine learning algorithms including logistic Lasso regression (Friedman et al., 2010; Tibshirani, 1996), the extreme learning machine (Huang et al., 2006) and gradient tree boosting (Chen and Guestrin, 2016; Friedman, 2001) for predicting whether the customer makes a purchase in the upcoming month. Our approach avoids prohibitive customer inquires that may be costly to be acquired in the non-contractual setting, and nevertheless shows performance comparable to state of the art approaches.

The application of machine learning or data mining techniques for predictive purposes on the customer-base is often analyzed in the customer relationship management and expert systems domain, and customer churn prediction is the most popular objective in this fields. The concept of churn and associated statistical implementations have been well studied in B2C business models (see, e.g., Verbeke et al. (2012); Neslin et al. (2006); Burez and Van den Poel (2007); Xia and Jin (2008); Xie et al. (2009)), especially in a contractual setting. Main industries include retail markets, subscription management, financial services and electronics commerce (see e.g. Chen et al. (2012) for an overview. This is in line with the general

trend of a stronger focus of academia and intelligence approaches on B2C applications (Wiersema, 2013). However churn prediction is also important in the B2B context, where has been studied much less; see (Jahromi et al., 2014). In particular, the development of business relationships remains central to B2B companies (Eriksson and Vaghult, 2000). The importance of retention for suppliers becomes even clearer in the B2B context where customers make larger and more frequent purchases with far higher transactional values (Rauyruen and Miller, 2007; Boles et al., 1997).

1.2. Relations to previous work

In the last decade, various machine learning methods for predicting customer retention and profitability have been analyzed in the academia field and some of them often used by practitioners. In most cases those approaches are based on extracting customer's latent characteristics from its past purchase behavior with the mindset that observed behavior is the outcome of an underlying stochastic process (Fader and Hardie, 2009). This approach to the customer's purchase prediction can be named as *characteristics approach*. Previous studies have analyzed the use of random forest techniques in order to predict customer retention and profitability (Larivière and Van den Poel, 2005) in a financial services and B2C context. Three major predictor categories that encompass potential explanatory variables were considered and those three categories were: past customer behavior, observed customer heterogeneity and variables related to intermediaries. In that research, it was found evidences that past customer behavior is more important to generate repeat purchasing and favorable profitability evolution, while the intermediary's role has a greater impact on the customers defection proneness. Literature on effective B2B promotions suggests to incorporate an enhanced, in depth view on the complex decision making setup such as buying center analysis (Hellman, 2005). Decision making is also more complex with B2B customers, as company's purchase decision is usually the consequence of a complex decision process, an alignment among stakeholders and business goals, and comparing the decision process done as an end consumer in the B2C domain. Also is important to highlight a very strong influence coming from the industry dynamic and other activities like product launches and campaigns.

Closely related to our work is (Jahromi et al., 2014), where the authors develop a model for predicting wether a customer performs a purchase in some prescribed future time frame based on purchase informations from the past. They propose customer characteristics such as the number of transactions observed in past time frames, time of the last transaction, and the relative change in total spending of a customer. They found an adaptive boosting method (Freund et al., 1996) to perform best on the tested data with an AUC value of 0.92. In contrast, in our study we compute a richer set of customer characteristics than the one in (Jahromi et al., 2014). These features are listed in Table 2.2 and described in detail in Section 2.3. For our framework, the best performing method (gradient tree boosting) shows an AUC value of 0.95. We point out that we obtain a higher AUC score even we use a time frame of only one month within purchases are predicted. This is much smaller than the 6 months time frame used in (Jahromi et al., 2014). A smaller time frame in beneficial in terms of actionability for a company; however it also makes the prediction much more complicated. These results demonstrate that our method provides valuable and reliable information for supporting sales and marketing departments also in the short term.

1.3. Outline

The remainder of this article is organized as follows. In Section 2 we formally describe the considered purchase prediction activity, see Problem 2.2. In that section we also describe the features characterizing the customer at a specific time. In Section 3 we describe how to solve Prob-

lem 2.2 using machine learning tools. In particular we use the logistic Lasso, the extreme learning machine and gradient tree boosting for model selection. Our framework for purchase prediction is applied in Section 4 to transactional B2B data of 10 000 customers and a total number of 200 000 transactions. The gradient tree boosting turns out to be the performing model showing an accuracy score of 88.98 % and an AUC value of 0.949. The paper ends with a short discussion in Section 5.

2. Formal problem definition and description of our customer features

In this section we establish a mathematical framework that formally describes the customer's purchase task. We give particular emphasis on the description of the features characterizing the customer at certain time instance.

customer ID	order	purchase time	purchase value
k	i	$t_{k,i}$	$V_{k,i}$
1	1	$t_{1,1}$	$V_{1,1}$
⋮	⋮	⋮	⋮
1	N_1	t_{1,N_1}	V_{1,N_1}
⋮	⋮	⋮	⋮
K	1	$t_{K,1}$	$V_{k,1}$
⋮	⋮	⋮	⋮
K	N_K	t_{K,N_K}	V_{K,N_K}

Table 2.1: *Original transactional data.* For every transaction of customer k one stores time $t_{k,i}$ and value $V_{k,i}$ of its i-th purchase made between month A and month B.

2.1. Problem description

Suppose that certain customers purchase products or services at a given company. We suppose that the company has a total number of K customers. Any costumer is represented by its ID $k \in \mathcal{K} \triangleq \{1, \dots, K\}$. Here and below $\triangleq$ means *equal by definition*. The purchases of costumer

Characteristics related to purchase time	
Number of total purchases	$x[1]$
Mean time between purchases	$x[2]$
Standard deviation of purchase frequency	$x[3]$
Maximal time without purchase	$x[4]$
Time since last purchase	$x[5]$
Thresholds for classification	$x[6]$, $x[7]$, $x[8]$
Frequency classification	$x[9]$
Characteristics related to purchase value	
Moving averages	$x[10]$, $x[11]$, $x[12]$
Maximum values of purchase	$x[13]$, $x[14]$
Mean values of purchase	$x[15]$, $x[16]$, $x[17]$
Median values of purchase	$x[18]$, $x[19]$, $x[20]$
Time frame variations	$x[21]$, $x[22]$
Purchase trend	$x[23]$
Further customer information	
Country of customer	$x[24]$
Creation of additional variables	
Pairwise products	$x[25], \dots, x[214]$
Powers of two and three	$x[215], \dots, x[254]$
Logarithms	$x[255], \dots, x[274]$

Table 2.2: Characteristic customer features derived from the transactional raw data.

k are characterized by purchase times $t_{k,i}$ and purchase values $V_{k,i}$ for $i = 1, \dots, N_k$, with N_k denoting the total number of purchases of customer k. The whole transaction data made between month A and month B can be arranged in a list as shown in Table 2.1. Here and below months are identified with elements in the set $\mathbb{Z}$ of all integer numbers. Without loss of generality we assume that the purchases of any customer are ordered chronologically, that is $t_{k,i_1} \leq t_{k,i_2}$ for all $i_1 \leq i_2$.

With the above notions, the problem under consideration can be stated as follows:

Problem 2.1 (Prediction of customer purchases)**.** *Given purchase data* $\mathcal{P}_k \triangleq \{(t_{k,i}, V_{k,i}) \mid i = 1, \dots, N_k\}$ *for every*

customer $k \in \mathcal{K}$ between month A and month B (as illustrated in Table 2.1), predict whether a given customer makes a transaction in the month following to B.

We address Problem 2.1 using machine learning algorithms. For that purpose we introduce some further notation. We define the binary variable $y_{k,\tau}$ that characterizes the purchase of the customer k in month $\tau \in [A,B] \triangleq \{A, A+1, \ldots, B\}$ by

$$y_{k,\tau} \triangleq \begin{cases} 1 & \text{if customer } k \text{ makes purchase in month } \tau \\ 0 & \text{otherwise}\,. \end{cases}$$

Further, for pairs $(k,\tau) \in \mathcal{K} \times [A,B]$ we construct a feature vector $x_{k,\tau}$ that uses characterizes the state of customer k at time τ based on purchase information of customer k up to month τ. Notice that the values $y_{k,\tau}$ are only known for $\tau \leq B$ and that we aim at estimating $y_{k,B+1}$ for the upcoming month $B+1$. Therefore, Problem 2.1 can be reformulated as follows.

Problem 2.2 (Reformulation as supervised learning problem)**.** *Estimate the values of $y_{k,B+1}$ from the feature vector $x_{k,B+1}$ representing the behavior of customer k until month B, and known input-output pairs $(x_{k,\tau}, y_{k,\tau})$ for all $k \in \mathcal{K}$ and certain $\tau \in [A,B]$.*

The efficient solution of Problem 2.2 requires the computation of significant features $x_{k,\tau}[1], \ldots, x_{k,\tau}[M]$. In this work we propose a certain set of $M = 274$ characteristic features that are listed in Table 2.2. A detailed description of these features and its computation is given in Subsection 2.3.

2.2. Data binning and smoothing

The customer features will be extracted from the original purchase information, as well as smoothed versions obtained by moving averages and a polynomial fit. For that purpose we first define binned purchase data as the sum of all purchases of a customer within a given month,

$$v_{k,\tau} \triangleq \sum_{i\colon t_{k,i}=\tau} V_{k,i} \quad \text{for } k \in \mathcal{K} \text{ and } \tau \in [A,B]\,.$$

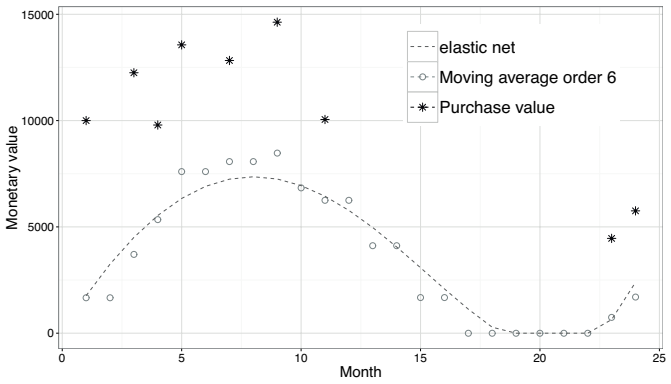

Figure 2.1: An example of the customer's purchase data $(t_{k,i}, V_{k,i})$ that are visualized by the asterisks. The corresponding 6th order moving averages $\bar{v}_{k,\tau}$ are visualized by the circles. The fitted purchase value function $\hat{v}_{k,+}(t)$ is visualized by the dashed curve.

In particular, we have $v_{k,\tau} = 0$ if there is no purchase from the customer k in month τ. Otherwise $v_{k,\tau}$ is equal to the corresponding purchase value.

We use the moving average of order 6 of the binned purchase values, defined by

$$\bar{v}_{k,\tau} \triangleq \frac{1}{6} \sum_{\Delta\tau=0}^{5} v_{k,\tau-\Delta\tau} \quad \text{for } k \in \mathcal{K} \text{ and } \tau \in [A-5,B]\,.$$

Finally, we consider a polynomial regression approximation of order 7,

$$\hat{v}_k(t) = \theta_0 + \theta_1\, t + \theta_2\, t^2 + \cdots + \theta_7\, t^7\,.$$

Here t represents a continuous time variable and $\hat{v}_k$ is constructed such that $\hat{v}_k(j) \simeq \bar{v}_{k,j}$ for $j = \tau-1, \tau-2, \ldots, \tau-T$. The coefficients θ_i are determined using the elastic-net method. Actually, we will use $\hat{v}_{k,+}(t) \triangleq \max\{1, \hat{v}_k(t)\}$.

Figure 2.1 shows an example of a customer's binned purchase data $v_{k,\tau}$ together with the corresponding 6th order moving average $\bar{v}_{k,\tau}$ and the polynomial fit $\hat{v}_{k,+}(t)$.

2.3. Customer features

We are now ready to formally define the characterizing features $x_{k,\tau}[1], \ldots, x_{k,\tau}[274]$ listed in Table 2.2. The features of any customer dynamically depend on time τ. For its computation we use subsets of the purchase data (original as well as smoothed) that containing purchases

made between months $\tau - T$ and $\tau - 1$, where T is some fixed time period. Formally, we define this past purchase data for $\tau \in [A + T, B + 1]$ as

$$\mathcal{P}_{k,\tau} \triangleq \Big\{ (t_{k,i}, V_{k,i}) \in \mathcal{P}_k \;\Big|\; \tau - T \le t_{k,i} < \tau - 1 \Big\}.$$

We denote the number of customer's purchases in the time frame $[\tau - T, \tau - 1]$ by $N_{k,\tau}$ and the corresponding purchase data by

$$\begin{aligned} t^{\tau}_{k,i} &\triangleq t_{k,M_{k,\tau}+i} && \text{for } i = 1, \dots, N_{k,\tau}\,, \\ V^{\tau}_{k,i} &\triangleq V_{k,M_{k,\tau}+i} && \text{for } i = 1, \dots, N_{k,\tau}\,, \end{aligned}$$

where $M_{k,\tau}$ is the total number of purchases made prior to month $\tau - T$.

2.3.1. Characteristics related to the purchase time

We first describe the characteristics related to the times of purchases.

- *Number of purchases.*

  The first considered feature is the number of customer's purchases in the time frame $[\tau - T, \tau - 1]$. We take

  $$x_{k,\tau}[1] \triangleq N_{k,\tau} \triangleq |\mathcal{P}_{k,\tau}|\,,$$

  as the number of elements in $\mathcal{P}_{k,\tau}$.

- *Mean time between purchases.*

  Next we consider the weighted average of the number of time units between purchases in $\mathcal{P}_{k,\tau}$ (or purchases in the time frame $[\tau - T, \tau - 1]$),

  $$x_{k,\tau}[2] \triangleq \sum_{i=2}^{N_{k,\tau}} w_i\, \Delta t^{\tau}_{k,i}\,.$$

  Here $\Delta t^{\tau}_{k,i} \triangleq t^{\tau}_{k,i} - t^{\tau}_{k,i-1}$ is the number of time units between the i-th and $(i-1)$-th purchase in $\mathcal{P}_{k,\tau}$. In this work we propose to choose the weights as $w_i \triangleq (i-1)^2 \,/\, \sum_{i=2}^{N_{k,\tau}} (i-1)^2$.

- *Standard deviation of times between purchases.*

  Using the same weights w_i as above we define the weighed standard deviation of the number of time units between purchases in $\mathcal{P}_{k,\tau}$ as

  $$x_{k,\tau}[3] \triangleq \sqrt{\sum_{i=2}^{N_{k,\tau}} w_i \,\Big(\Delta t^{\tau}_{k,i} - x_{k,\tau}[2]\Big)^2}\,.$$

- *Maximal time without purchase.*

  Here we consider the maximum number of time units between purchases in $\mathcal{P}_{k,\tau}$,

  $$x_{k,\tau}[4] \triangleq \max\big\{ \Delta t^{\tau}_{k,i} \;\big|\; i_1 = 2, \dots, N_{k,\tau} \big\}.$$

- *Time since last purchase.*

  The next feature measures the number of months since the last purchase in the time frame $[\tau - T, \tau - 1]$ has been performed,

  $$x_{k,\tau}[5] \triangleq \begin{cases} \tau - 1 - t^{\tau}_{k,N_{k,\tau}} & \text{if } N_{k,\tau} \neq 0 \\ T & \text{otherwise}\,. \end{cases}$$

- *Thresholds for classification.*

  We consider certain thresholds for the number of time units between purchases

  $$\begin{aligned} x_{k,\tau}[6] &\triangleq x_{k,\tau}[2] + h_1\, x_{k,\tau}[3]\,, \\ x_{k,\tau}[7] &\triangleq x_{k,\tau}[2] + h_2\, x_{k,\tau}[3]\,, \\ x_{k,\tau}[8] &\triangleq x_{k,\tau}[2] + h_3\, x_{k,\tau}[3]\,, \end{aligned}$$

  where h_1, h_2, h_3 are some positive numbers. We propose to take $h_1 = 2$, $h_2 = 4$ and $h_3 = 8$.

- *Frequency classification.*

  The next characteristic is a categorial feature that characterizes the customer k according on the purchase frequency. It is defined by

  $$x_9[k, \tau] \triangleq \begin{cases} \text{normal} & \text{if } x_{k,\tau}[5] \le x_{k,\tau}[6]\,, \\ \text{attrition} & \text{if } x_{k,\tau}[6] < x_{k,\tau}[5] \le x_{k,\tau}[7]\,, \\ \text{at-risk} & \text{if } x_{k,\tau}[7] < x_{k,\tau}[5] \le x_{k,\tau}[8]\,, \\ \text{lost} & \text{if } x_{k,\tau}[8] < x_{k,\tau}[5]\,. \end{cases}$$

2.3.2. Features related to the purchase value

We further consider the following customer characteristics that are related to the purchase values of costumers.

- *Moving averages.*

  We consider the moving averages of order 6 and 3 of the binned purchase values and the polynomial approximation,

$$x_{k,\tau}[10] \triangleq \bar{v}_{k,\tau-1}\,,$$
$$x_{k,\tau}[11] \triangleq \hat{v}_{k,+}(\tau-1)\,,$$
$$x_{k,\tau}[12] \triangleq \frac{1}{3}\sum_{\Delta\tau=0}^{2} v^k_{\tau-1-\Delta\tau}.$$

- *Maximum values of purchase.*

  Here we take two different characteristics defined as the maxima over the actual purchases and the polynomial fit, respectively,

$$x_{k,\tau}[13] \triangleq \max\left\{V^\tau_{k,i} \,\middle|\, i=1,\dots,N_{k,\tau}\right\},$$
$$x_{k,\tau}[14] \triangleq \max\{\hat{v}_{k,+}(t) \mid t\in[\tau-1,\tau-T]\}\,.$$

- *Mean values of purchase.*

  Here we consider the mean values of the actual, the binned and the fitted purchase values defined by

$$x_{k,\tau}[15] \triangleq \frac{1}{N_{k,\tau}}\sum_{i=1}^{N_{k,\tau}} V^\tau_{k,i}\,,$$
$$x_{k,\tau}[16] \triangleq \frac{1}{T}\sum_{t=\tau-T}^{\tau-1} v_{k,t}\,,$$
$$x_{k,\tau}[17] \triangleq \frac{1}{T}\sum_{t=\tau-T}^{\tau-1} \hat{v}_{k,+}(t).$$

- *Median values of purchase.*

  Similar, we consider the medians of the actual, the binned and the fitted purchase values,

$$x_{k,\tau}[18] \triangleq \mathrm{med}\left\{V^\tau_{k,i} \,\middle|\, i=1,\dots,N_{k,\tau}\right\},$$
$$x_{k,\tau}[19] \triangleq \mathrm{med}\{v_{k,t} \mid t\in[\tau-1,\tau-T]\}\,,$$
$$x_{k,\tau}[20] \triangleq \mathrm{med}\{\hat{v}_{k,+}(t) \mid t\in[\tau-1,\tau-T]\}\,.$$

- *Time frame variations.*

  The time frame variation is characterized as the relative change in purchase values,

$$x_{k,\tau}[21] \triangleq \begin{cases} 0 & \text{if } N_{k,\tau}\le 1 \\ \frac{V^\tau_{k,N_{k,\tau}} - V^\tau_{k,N_{k,\tau}-\Delta i}}{V^\tau_{k,N_{k,\tau}-\Delta i}} & \text{otherwise}\,, \end{cases}$$

  where $\Delta i \triangleq \min\{5, N_{k,\tau}-1\}$.

  We also consider a categorical characteristic for time frame variation that we define as follows:

$$x_{k,\tau}[22] \triangleq \begin{cases} \text{steady} & \text{if } x_{k,\tau}[21] < -\mu\,, \\ \text{within-limits} & \text{if } |x_{k,\tau}[21]| \le \mu\,, \\ \text{alternating} & \text{if } x_{k,\tau}[21] > \mu\,. \end{cases}$$

  Here μ is some positive value; we propose $\mu = 0.3$.

- *Purchase trend.*

  We characterize the purchase trend as a categorial variable depending on the relative change

$$d_{k,\tau} \triangleq \frac{\hat{v}_{k,+}(\tau-1) - \hat{v}_{k,+}(\tau-6)}{\hat{v}_{k,+}(\tau-6)}$$

  of the fitted purchase values. More precisely, we define the purchase trend by

$$x_{k,\tau}[23] \triangleq \begin{cases} \text{decreasing}_{--} & \text{if } d_{k,\tau} \le -a_3 \\ \text{decreasing}_{-} & \text{if } -a_3 < d_{k,\tau} \le -a_2 \\ \text{decreasing} & \text{if } -a_2 < d_{k,\tau} \le -a_1 \\ \text{stable} & \text{if } -a_1 < d_{k,\tau} \le a_1 \\ \text{increasing} & \text{if } a_1 < d_{k,\tau} \le a_2 \\ \text{increasing}_{+} & \text{if } a_2 < d_{k,\tau} \le a_3 \\ \text{increasing}_{++} & d_{k,\tau} > a_3\,. \end{cases}$$

  Here a_1, a_2 and a_3 are some positive values; we propose to take $a_1 = 0.15$, $a_2 = 0.225$, $a_3 = 0.3$.

2.3.3. Additional characteristics

Beside the characteristics described above we also use the categorical characteristic $x_{k,\tau}[24]$ denoting the country of the customer k. In order to further increase the

prediction accuracy we compute auxiliary variables from the variables $x_{k,\tau}[m]$ excluding the four categorical characteristics.

The auxiliary variables are created by applying the following mathematical operations to the original variables:

- *Pairwise products.*

  Here we form all products $x_{k,\tau}[m] \cdot x_{k,\tau}[m']$ of the non-categorial features with $m \neq m'$. This yields $19 + 18 + \cdots + 1 = 190$ additional variables.

- *Powers of two and three.*

  We further consider powers $x_{k,\tau}[m]^2$ and $x_{k,\tau}[m]^3$ of all non-categorial features. This yields $20 + 20 = 40$ additional variables.

- *Taking Logarithm.*

  Finally, we add the logarithms of the non-categorial features $\log(x_{k,\tau}[m])$. This yields 20 additional variables.

In summary we have $M = 24 + 190 + 40 + 20 = 274$ variables $x_{k,\tau}[m]$ characterizing the customer k at time τ. Using these variable we will train a classifier that predicts the binary purchase variable y. Although the creation of the artificial variables in principle does not increase the information content of the data, it puts the data into a higher dimensional space and significantly improves results of the machine learning algorithms. For example, the powers contain interactions between the variables which otherwise would difficult to be found by the algorithms.

3. Application of machine learning algorithms

In this section we solve Problem 2.2 (the formally described purchase prediction issue) by various machine learning algorithms for binary classification.

3.1. Binary classification

A binary classification algorithm constructs a function $\Phi\colon \mathbb{R}^M \to \{0,1\}$ in such a way that $\Phi(x) = y$ with high probability for pairs (x, y) from the so-called training data set. Due to data imperfections, not all of the training examples will be predicted exactly by the classifier. In our case, the training data set takes the form

$$\mathcal{D} \triangleq \{(x_{k,\tau}, y_{k,\tau}) \mid k \in \mathcal{K}\,, A + T \leq \tau \leq B\}\,. \tag{3.1}$$

Actually, all of our considered methods output a regression function mapping to the real numbers,

$$\phi\colon \mathbb{R}^M \to [0,1] \subseteq \mathbb{R}\colon x \mapsto \phi(x)\,. \tag{3.2}$$

The output value $\phi(x)$ of the regression function can be interpreted as the probability that a feature vector x corresponds to a next month purchase. From the estimated purchase probabilities one constructs the classifier $\Phi = \Phi_\lambda$ by taking $\Phi_\lambda(x) = 1$ for $\phi(x) > \lambda$ and zero otherwise. Here $\lambda \in [0,1]$ is a certain threshold that is selected as tradeoff between sensitivity and specificity. In this work we use a threshold of 0.5 for the final classification.

For constructing the regression function in (3.2), we apply the following state-of-the art machine learning algorithms:

- Logistic Lasso regression;
- Extreme learning machine;
- Gradient tree boosting.

These methods are briefly reviewed in the following subsections. Any of these methods will be used in combination with 10-fold cross validation for estimating optimal values of the parameters these methods depend on. In particular, applying cross validation avoids overfitting on the training data set and therefore allows to generalize the trained models to predicting customer purchases where the next-month purchase is not known.

We decided on the above classification methods because they are totally different from one another, and further are known to yield high accuracy with reasonable computational effort.

3.2. Logistic Lasso regression

For Lasso regression we use the logistic model which is one of the most common models used in the context of classification (Hastie et al., 2009). We estimated the coefficients (β_j) in the logistic model by adding the ℓ^1-penalty term

$$R(\beta) \triangleq \sum_{j=1}^{d} |\beta_j| \ ,$$

which is known as the Lasso (Tibshirani, 1996). The Lasso penalty $\sum_{j=1}^{d} |\beta_j|$ results in variable selection and shrinkage. The purpose of this penalty is to retain a subset of the characteristics and to discard the rest. This subset selection produces a model that is interpretable and has possibly a lower prediction error than the full model.

For numerically computing the coefficients we use the algorithm for logistic Lasso regression provided in the package `glmnet`; see (Friedman et al., 2010, Chapter 3).

3.3. Extreme learning machine

Another model that we consider is the single-hidden layer feedforward neural network (SLFN). We use the extreme learning machine algorithm (Huang et al., 2006) for building the SLFN on our training data. The extreme learning machine algorithm became a very popular research subject in the past years (Huang, 2015). Unlike other algorithms for building neural networks, the extreme learning machine randomly chooses hidden nodes and analytically determines the output weights of the SLFN. The extreme learning machine provides a good theoretical performance at a very fast learning speed.

For our results we use implementation of this algorithm provided in the package `elmNN`; see (Gosso and Martinez-de-Pison, 2012).

3.4. Gradient tree boosting

Among the machine learning methods, boosting (Freund et al., 1999), and specifically gradient tree boosting, frequently shows the best performance for many applications. The term gradient boosting has been invented in (Friedman, 2001). Boosting is a procedure that combines the outputs of several weak classifiers in order to produce a powerful classifier. In this way, boosting has a similarity to bagging (Breiman, 1996) and other ensemble-based machine learning methods.

Boosting and bagging both form a set of simpler classifiers that are combined by voting. In bagging the ensembles are generated by repeated bootstrap sampling of the data, and in boosting by adjusting appropriate weights of training data. The purpose of boosting is to sequentially apply the weak classification algorithm to repeatedly modified versions of the data, thereby producing a sequence of weak classifiers. The predictions from the weak classifiers are combined through a weighted majority vote to produce the final prediction; see (Hastie et al., 2009, Chapter 10) for details.

In our work we use the implementation of the gradient tree boosting algorithm provided by the package `XGBoost`; see (Chen and Guestrin, 2016).

4. Results

We apply the developed framework for prediction of customer purchase on transactional data provided by a large manufacturer located in central Europe. The data have been gathered from transactions of the B2B unit, which have been recorded from January 2009 until May 2015. We only consider transactions of customers whose first purchase is at least six months ago due to the lack of sufficient information in the other cases.

The transactions belong to $K = 10136$ different customers from 125 different countries. As the time unit we consider a month as it is not very common in the considered data to have a customer with more than one purchase per month. If a customer has more than one purchase in a month, then for the actual purchase values V_i we take sum of the purchase values in the considered month. After this monthly aggregation, the data set contains in total 192470 orders for all customers. We take January 2009 as month

$A = 1$ such that March 2015 corresponds to month 77. The time period for computing the feature values is taken as $T = 24$.

4.1. Comparison of the machine learning algorithms

We first evaluate the predictive performance of the constructed estimators on the training set (3.1) using the classifiers described in Section 3. We trained the models on the trainings data (3.1), with $A+T = 25$ and $B = 77$. The use of 10-fold cross validation avoids overfitting, thus the performance on the training data will be representative for the actual performance on data with unknown output during the prediction phase. As assessment criteria, we use the total prediction accuracy (percentage of correctly classified purchases), and the area under the receiver operating characteristic curve (AUC). Additionally, we consider the confusion matrices.

	AUC	accuracy
Lasso	0.9263	85.74 %
Extreme learning machine	0.9223	85.58 %
Gradient tree boosting	0.9340	86.68 %

Table 4.1: AUC and total prediction accuracy for the Lasso, the extreme learning machine and the gradient tree boosting method, evaluated on the training set with 10-fold cross validation.

The results on the whole training set are presented in Tables 4.1 and 4.2. All considered methods give an excellent performance in terms of the AUC and prediction accuracy. In particular, the estimator constructed by the gradient tree boosting has the highest AUC score, and it also has the best performance in the confusion matrix. The differences in the AUC can be considered as very significant for the total customer portfolio classification in the prediction phase. Therefore, we recommend the gradient tree boosting for its actual use in practice.

All computations have been performed on a virtual machine on ESX Cluster with 12 cores and 60 GB RAM. The operation system is SUSE Linux Enterprise Server, and we have run the scripts using RStudio Server with R version 3.1.2 underneath. The computation times for training a single model are 6 minutes for the Lasso, about one minute for the extreme learning machine, and 2.5 minutes for gradient tree boosting.

Actual purchase	Yes	No
Lasso		
Yes	23.42 %	07.60 %
No	06.66 %	62.32 %
Extreme learning machine		
Yes	23.09 %	07.93 %
No	06.49 %	62.49 %
Gradient tree boosting		
Yes	23.22 %	07.80 %
No	05.52 %	63.46 %

Table 4.2: Confusion matrices for the Lasso (top), the extreme learning machine (middle), and gradient tree boosting (bottom), evaluated on the training set with 10-fold cross validation.

4.2. Example for purchase prediction

The best performing model for the given data is the gradient tree boosting method. We therefore apply this classification method to actually predict the customer purchases on the test set that is not used for constructing the classifier. For that purpose we trained the gradient tree boosting classifier using the training data (3.1) with $A + T = 25$ and $B = 75$. Hence the the model is trained only using data until March 2015. The model is then applied to the test data

$$(x_{k,B+1}, y_{k,B+1}) \quad \text{for } k \in \mathcal{K}\,, \tag{4.1}$$

which corresponds to predictions of purchases in April 2015.

The resulting total prediction accuracy computed on the test data is given by 88.98 % and the AUC value is

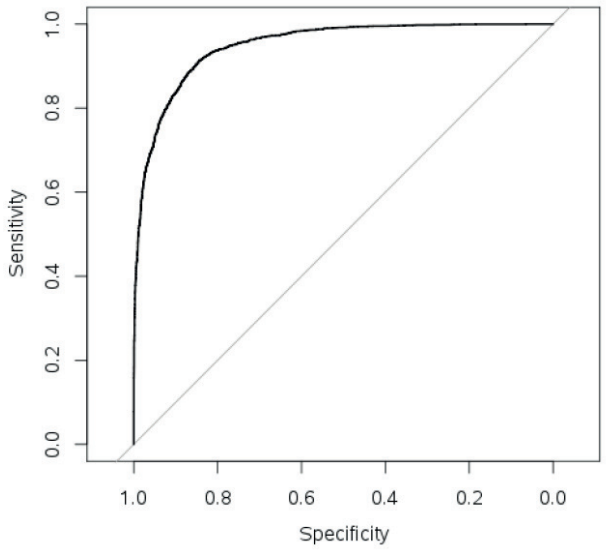

Figure 4.1: ROC curve for the gradient tree boosting estimator computed for evaluated on the independent test data for prediction of purchases in April 2015.

0.949; see Table 4.3. In Figure 4.1, we present also the ROC curve for the gradient tree boosting estimator on the test data (4.1).

Actual purchase	Yes	No
Gradient tree boosting		
Yes	21.70 %	06.37 %
No	04.66 %	67.28 %

Table 4.3: Confusion matrix for the gradient tree boosting method evaluated on the independent test data for prediction of purchases in April 2015. The total prediction accuracy computes to 88.98% an the AUC equals 0.949.

5. Discussion and conclusion

Predicting a future customer behavior provides many benefits for a company, reaching from supporting of planning the inventory at the warehouse to the identification of customer churn. While data analytic tools for such purposes are well investigated in a contractual setting (see for example, Chen et al. (2012); Miguéis et al. (2013); Verbeke et al. (2012)), they are much less developed in the non-contactual setting; see (Jahromi et al., 2014). In this work we developed a framework for dynamic and fully data driven prediction of next month customer purchase in a non-contactual setting. For that purpose, we constructed a specific set of 274 feature variables characterizing the customer at specific months. We applied several state of the art machine learning algorithms, including the logistic Lasso, the extreme learning machine and gradient tree boosting, for computing next month purchase probabilities. Our results show that the gradient tree boosting outperformes the Lasso and the extreme learning machine. Applied to transactional data from to 10136 different customers it provides a total accuracy of 88.98 % and an AUC value of 0.949 evaluated on the test data. These results are in accordance with the contractual setting, where boosted decision trees have ben reported to outperform other machine learning algorithms (Verbeke et al., 2012).

Our algorithmic framework can be extended in a straight forward way to compute probabilities for customer purchases for any month in the near future. For example, we also tested the gradient tree boosting for the prediction of purchases in the second month after the month B. For that purpose we considered the modified training data $(x_{k,\tau}, y_{k,\tau+1})$ for predicting $y_{k,B+1}$ from $x_{k,B}$. In this case we obtained a total prediction accuracy of 88 % and an AUC of 0.941 on the test data. The presented method could be extended to even later months ahead or larger time frames as in (Jahromi et al., 2014), where theoretically we should obtain even better results.

Predicting purchase probabilities highly support the forecast of the sales budget. A next step for enriching the information provided by our framework is the prediction of the actual values of future purchases. The presented approach could be modified for predicting purchase values, by training a prediction model based on the same input variables, but using the purchase values as response variable. An accurate estimation of time and value for purchases, is an interesting line of research. It would be

a powerful decision support tool for operative as well as strategic activities.

References

Allenby, G.M., Leone, R.P., Jen, L., 1999. A dynamic model of purchase timing with application to direct marketing. Journal of the American Statistical Association 94, 365–374.

Berger, P.D., Bolton, R.N., Bowman, D., Briggs, E., Kumar, V., Parasuraman, A., Terry, C., 2002. Marketing actions and the value of customer assets: A framework for customer asset management. Journal of Service Research 5, 39–54.

Bhattacharya, C.B., 1998. When customers are members: Customer retention in paid membership contexts. Journal of the academy of marketing science 26, 31–44.

Boles, J.S., Barksdale, H.C., Johnson, J.T., 1997. Business relationships: an examination of the effects of buyer-salesperson relationships on customer retention and willingness to refer and recommend. Journal of Business & Industrial Marketing 12, 253–264.

Breiman, L., 1996. Bagging predictors. Machine Learning 24, 123–140.

Burez, J., Van den Poel, D., 2007. CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services. Expert Systems with Applications 32, 277–288.

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. arXiv:1603.02754.

Chen, Z., Fan, Z., Sun, M., 2012. A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data. European Journal of operational research 223, 461–472.

Daly, J.L., 2002. Pricing for profitability: activity-based pricing for competitive advantage. volume 11. John Wiley & Sons.

Eriksson, K., Vaghult, A.L., 2000. Customer retention, purchasing behavior and relationship substance in professional services. Industrial Marketing Management 29, 363–372.

Fader, P.S., Hardie, B.G., 2009. Probability models for customer-base analysis. Journal of Interactive Marketing 23, 61–69.

Freund, Y., Schapire, R., Abe, N., 1999. A short introduction to boosting. Journal of Japanese Society For Artificial Intelligence 14, 771–780.

Freund, Y., Schapire, R.E., et al., 1996. Experiments with a new boosting algorithm, in: 13th International Conference on Machine Learning, Bari, Italy., pp. 148–156.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33, 1–22.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Annals of Statistics 29, 1189–1232.

Ganesh, J., Arnold, M.J., Reynolds, K.E., 2000. Understanding the customer base of service providers: an examination of the differences between switchers and stayers. Journal of Marketing 64, 65–87.

Gosso, A., Martinez-de-Pison, F., 2012. elmNN: Implementation of Extreme Learning Machine algorithm for single hidden layer feedforward neural networks. R package version 1.

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

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Hellman, K., 2005. Strategy-driven B2B promotions. Journal of Business & Industrial Marketing 20, 4–11.

Herniter, J., 1971. A probabilistic market model of purchase timing and brand selection. Management Science 18, 102–113.

Huang, G., 2015. What are extreme learning machines? Filling the gap between Frank Rosenblatt's dream and John von Neumann's puzzle. Cognitive Computation 7, 263–278.

Huang, G., Zhu, Q., Siew, C., 2006. Extreme learning machine: theory and applications. Neurocomputing 70, 489–501.

Jahromi, A.T., Stakhovych, S., Ewing, M., 2014. Managing B2B customer churn, retention and profitability. Industrial Marketing Management 43, 1258 – 1268.

Keaveney, S.M., Parthasarathy, M., 2001. Customer switching behavior in online services: An exploratory study of the role of selected attitudinal, behavioral, and demographic factors. Journal of the Academy of Marketing Science 29, 374–390.

Kumar, V., Venkatesan, R., Reinartz, W., 2008. Performance implications of adopting a customer-focused sales campaign. Journal of Marketing 72, 50–68.

Larivière, B., Van den Poel, D., 2005. Predicting customer retention and profitability by using random forests and regression forests techniques. Expert Systems with Applications 29, 472–484.

Miguéis, V.L., Camanho, A., e Cunha, J.F., 2013. Customer attrition in retailing: an application of multivariate adaptive regression splines. Expert Systems with Applications 40, 6225–6232.

Mulhern, F.J., 1999. Customer profitability analysis: Measurement, concentration, and research directions. Journal of Interactive Marketing 13, 25–40.

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

Platzer, M., Reutterer, T., 2016. Ticking away the moments: Timing

regularity helps to better predict customer activity. Marketing Science , 799.

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

Rauyruen, P., Miller, K.E., 2007. Relationship quality as a predictor of B2B customer loyalty. Journal of Business Research 60, 21–31.

Reinartz, W.J., Kumar, V., 2003. The impact of customer relationship characteristics on profitable lifetime duration. Journal of Marketing 67, 77–99.

Rust, R.T., Kumar, V., Venkatesan, R., 2011. Will the frog change into a prince? Predicting future customer profitability. International Journal of Research in Marketing 28, 281–294.

Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological) 58, 267–288.

Venkatesan, R., Kumar, V., 2004. A customer lifetime value framework for customer selection and resource allocation strategy. Journal of Marketing 68, 106–125.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., Baesens, B., 2012. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European Journal of Operational Research 218, 211–229.

Wiersema, F., 2013. The B2B agenda: The current state of B2B marketing and a look ahead. Industrial Marketing Management 42, 470–488.

Winer, R.S., 2001. A framework for customer relationship management. California management review 43, 89–105.

Xia, G., Jin, W., 2008. Model of customer churn prediction on support vector machine. Systems Engineering-Theory & Practice 28, 71–77.

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

Zeithaml, V.A., Rust, R.T., Lemon, K.N., 2001. The customer pyramid: creating and serving profitable customers. California Management Review 43, 118–142.