Value Miner: A Data Mining Environment for the Calculation of the Customer Lifetime Value with Application to the Automotive Industry

Katja Gelbrich[1] and Reza Nakhaeizadeh[2]

[1] DaimlerChrysler Research Center, Postfach 2360, D-89013 Ulm, Germany
Rheza.Nakhaeizadeh@daimlerchrysler.com
[2] DaimlerChrysler Research Center, Postfach 2360, D-89013 Ulm, Germany
gelbrich@rcs.urz.tu-dresden.de

Abstract. The acquisition of new accounts is a major task of marketers. It is often carried out rather unsystematically, though. However, by now, one has come to terms that customer acquisition is a matter of quality. An instrument to evaluate prospective accounts is the Customer Lifetime Value (CLV). This paper introduces a Data Mining environment for its calculation and demonstrates its applicability to marketing in the automotive industry. The Car Miner refers to the evaluation of prospects rather than of current customers. This and other restrictions will be discussed along with guidelines for future research.

1 Introduction

Not all customers contribute to the profit of their suppliers. According to the "pareto-rule", a minority of 20% high-valuable customers subsidizes 80% less valuable ones [12]. Therefore, the acquisition budget should be spent on the right prospects. The Customer Lifetime Value (CLV) supports this decision [4], [9]. *Chapter 2* introduces a definition of CLV and describes constraints concerning the customer acquisition in the automotive industry which leads to an adjusted CLV model. *Chapter 3* reflects the development of a Data Mining environment to calculate the CLV. *Chapter 4* introduces restrictions of the model and discusses ideas to improve the Value Miner.

2 Conceptualization of the Customer Lifetime Value

2.1 Classical Definition of CLV

Several models of CLV are introduced in the literature. They agree that CLV is the *present value of expected revenues less costs caused directly by a customer during his relationship with the seller* [2], [4], [8]. Costs include spending for acquisition (e.g. advertising, promotion) and account maintenance (e.g. post purchase marketing). Revenues include the monetary benefit from the customer, i.e. the money he spends

R. López de Mántaras, E. Plaza (Eds.): ECML 2000, LNAI 1810, pp. 154-161, 2000.

on the supplier's products / services [8]. Some authors also mention soft benefits, e.g. the customer's reference value [6]. The definition reveals several *key statements*:

1. Since the value of a customer is revenues less costs, it represents a *net* value.
2. The term "lifetime" refers to the *duration of the relationship* between buyer and seller. This requires an estimation of the prospective end of the relationship.
3. Revenues can be economic (e.g. turnover) and *non-economic* (e.g. reference value). Non-economic benefits have to be quantified.
4. The term "present" suggests to *discount* future payments. The implied devaluation of future streams of payment is due to the fact that they are uncertain and that the company could alternatively invest its money into the capital market.
5. The definition only covers revenues and costs *in the future*.

2.2 Adjustment of the Definition with Respect to the Acquisition of Car Buyers

Given the acquisition of new accounts in the automotive industry, only two of the statements above are relevant: Future payments should be discounted to the present, because they are uncertain (statement 4). Past payments from the prospect should be neglected, because – as opposed to current customers – they are not relevant to the company concerned (statement 5). However, the statements 1, 2, and 3 do not hold:

- *Statement 1 (net value).* As opposed to current customers, the individual costs of a prospect can hardly be estimated, because there is no individual historical data. Subsequently, some researchers [4] simply divide the historical overall spending for the acquisition and retention of customers by the number of accounts yielding to per capita costs. But if one does so, costs can easily be neglected at all, because they reduce the revenue of all prospects by the same amount.
- *Statement 2 (time frame).* Most authors, e.g. [10], equate the time frame of CLV with the relationship's duration. But the end of a relationship is uncertain. We argue to extend the time frame to the day when the buyer stops consuming, i.e. when he dies. This is reasonable, because one should aim at keeping customers for good.
- *Statement 4 (non-economic benefit).* Many authors point out the relevance of soft benefits, but – with only a few exceptions [3] – refrain from measuring them, because they can hardly be quantified. If soft facts are included at all, they are more easily gathered from actual customers rather than from prospects. Table 1 summarizes the adjustments of the state-of-the-art definition.

Table 1. Definition of CLV adjusted to the customer acquisition in the automobile industry

Problem	Classical Definition	Adjustment	Reason for Adjustment
Gross or net value?	Net value	Gross value	Costs are assumed to be equal among customers.
Present value?	Present value	No adjustment	–
Time frame?	Estimated duration of the relationship	Remaining lifetime	The goal is to keep the customer for ever.
Soft benefits?	Inclusion of soft benefits	No inclusion of soft benefits	Soft benefits can hardly be estimated for prospects.
Past payments?	No past payments	No adjustment	–

Based on these adjustments, the CLV is conceptualized by discounting the *price acceptance* in every year the customer purchases a car to the presence. Given the restrictions which come along with the evaluation of prospects, this conceptualization is preferred to competing definitions (e.g. CLV as net present value). Cars are not purchased frequently, though. Thus, the term y, representing the year of purchase, does not increase by one, but in smaller steps, called purchase frequency (see equation 1).

$$CLV = \sum_{y=0}^{n} \frac{PA_y}{(1+r)^y} = PA_0 + \frac{PA_{y+PF}}{(1+r)^{[PF]}} + \frac{PA_{y+2PF}}{(1+r)^{[2 \bullet PF]}} + ... + \frac{PA_{y+n \bullet PF}}{(1+r)^{[n \bullet PF]}} \; . \qquad \textbf{(1)}$$

PA ... Price acceptance r ... Rate of discount PF ... Purchase frequency
y ... Year of purchase n ... Last year of purchase (when customer dies)

According to equation (1), the following information are required: purchase frequency, price acceptance, rate of discount, age of customer, average life expectancy. The main source for the *Data Warehouse* was the "Consumer Analysis 1999" (CA) with data from 31,337 German residents. Since only purchasers of *new cars* were considered, the data bases melted down to 6,039 people. Some data (life expectancy, discount rate) were added from external sources (Federal Statistical Office, FAZ).

3 Data Mining Environment for the Calculation of the CLV

Predicting future revenues and discounting them to the presence seems to be the main Data Mining task for calculating the CLV (chapter 3.2). The calculation, however, is not as trivial as equation (1) suggests: Upcoming arguments necessitated a refinement of the model. First, the purchase frequency is not constant (chapter 3.1). Second, the discount rate is a component of market interest and price increase (chapter 3.3).

3.1 Data Mining Task 1: Prediction of the Purchase Frequency

The purchase frequency was not acquired directly. Thus, it had to be predicted by other variables. We assumed, that the purchase frequency, i.e. the *time span an individual keeps a car*, decreases with income, intensity of care usage, usage for business reasons, and a positive attitude towards brands. Moreover, it was supposed that the frequency is not constant – as equation (1) states –, but that older people purchase less often. In order to predict the purchase frequency, we had to consider *two more items*:

- People were asked how old their current car was (CAR_0).
- They were asked if they intended to buy a new car in the course of the this year (INTENTION). To the people who did, CAR_0 represents their *purchase frequency*.

We draw a *subsample* of those people who declared their upcoming purchase intention (n = 2,260) and conducted several analyses with CAR_0 (i.e. purchase frequency) as the dependent and the items stated above as *independent variables*:

- Age of the customer in t = 0 (AGE_0),
- Net income of the household (INCOME),

- Intensity of car usage, quantified by the kilometers driven per year (KILO),
- Usage for business or private reasons (PRIVATE),
- Attitude towards the consumption of brands (BRAND).

Cross-validation was used to choose the model which best predicts the purchase frequency. The idea is to split the data base into two subsamples, to calibrate the model on one part and validate it on the other. The model yielding to the highest R square on the validation subsample should be chosen [7]. *Three alternative models* were tested on the calibration sample: a multiple linear regression, a multiple non-linear regression, and a chaid analysis. The *linear regression* yielded to equation 2. As hypothesized, older people and private users purchase less often (PF increases); high income, high intensity of usage, and positive attitude towards brands reduce the time span.

In the multiple *non-linear regression*, the directions of influence were the same, but the dependencies for AGE_0 (cubic), INCOME (exponential), KILO (cubic), and PRIVATE (exponential) were non-linear (see equation 3). The *chaid analysis* searches for the independent variables yielding to the highest split of the dependent variable. Chaid stands for Chi-squared Automatic Interaction Detector pointing out that it is based on chi-square tests automatically detecting interactions between variables [1]. Figure 1 partly displays the chaid output[1].

$$\begin{aligned}\text{PF} = 5.674 &+ (0.170 \bullet \text{AGE}_0) - (0.124 \bullet \text{INCOME}) - (0.182 \bullet \text{KILO}) \\ &+ (0.847 \bullet \text{PRIVATE}) - (0.906 \bullet \text{BRAND}) \quad .\end{aligned} \tag{2}$$

$$\begin{aligned}\text{PF} &= 5.834 + (-0.075 \bullet \text{AGE}_0) + \left(0.0017 \bullet \text{AGE}_0^2\right) + \left(-0.00000096 \bullet \text{AGE}_0^3\right) \\ &+ \left(e^{-0.109 \bullet \text{INCOME}}\right) + (-1.556 \bullet \text{KILO}) + \left(0.429 \bullet \text{KILO}^2\right) + \left(-0.040 \bullet \text{KILO}^3\right) \\ &+ \left(e^{0.647 \bullet \text{PRIVATE}}\right) - (0.961 \bullet \text{BRAND}) \quad .\end{aligned} \tag{3}$$

PF	... Purchase frequency	AGE_0	... Current age of customer
INCOME	... Net household income	BRAND	... "The reputation of a brand is a crucial criteria for the purchase of a new car."
PRIVATE	... Private usage		
KILO	... km driven per year		

Applied to the validation subsample, the non-linear regression model performed best. It explained 12.3% of PF's variance, the linear regression model explaining slightly less (11.4%). The chaid analysis hardly yielded to an R square of 5%. Subsequently, the *non-linear model* was selected to calculate the purchase frequency. Using equation 3, the purchase frequency of each customer *at any age* could be predicted in the main sample. However, the first purchase frequency depends on *two cases*:

1. The car is now "younger" than a person's purchase frequency at his age usually is → $\text{CAR}_0 < \text{PF}(\text{AGE}_0)$. If there is a 40-year-old, whose predicted purchase fre-

[1] One disadvantage of the chaid analysis is, that it is limited to independent variables with a maximum of 31 categories. Therefore, the numerical variable "age" had to be categorized. Pre-analyses with different intervals (constant, non-constant) and different numbers of categories caused no better split of the purchase frequency than the one displayed in figure 1.

quency according to equation (3) is, say, 5.4 years and his car is only three years old (3 years < 5.4 years), he will purchase his next car in 2.4 years, in t = 1.

2. The car is "older" than or as old as the purchase frequency is → $CAR_0 \geq PF (AGE_0)$. For the car of our 40-year-old, which is, say, eight years old, this means: 8 years > 6.3 years. The customer's car is "overdue", he purchases now, in t = 0.

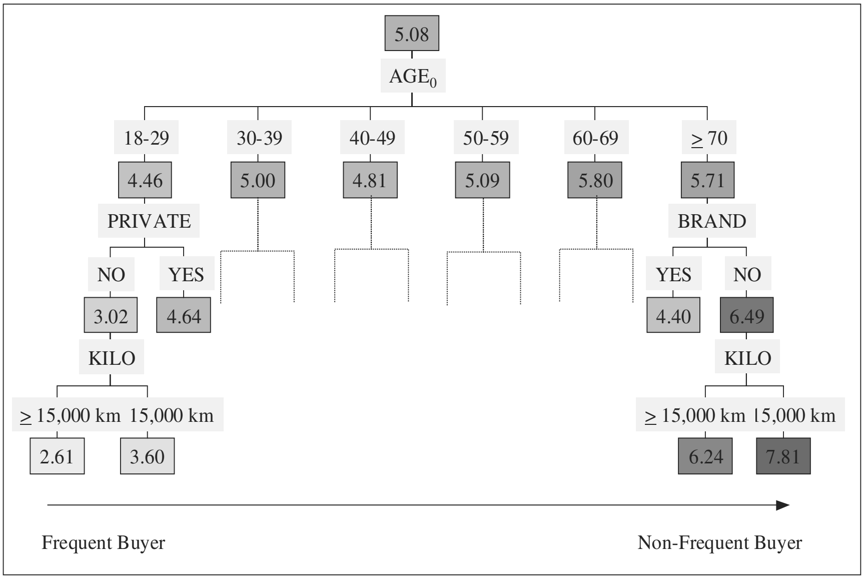

Fig. 1. PF, which is 5.08 years on average, is first split by AGE_0 yielding to 6 subgroups. Young people (18–29 years) purchase cars every 4.46 years, old people (≥ 70) less often. However, the dependency is not linear. Then, the algorithm searches for the variables causing the highest split of PF in the 6 subgroups. In level two, two different attributes split the purchase frequency: In AGE_0 = 18–29 years, PRIVATE is used while in category $AGE_0 \geq 70$ years, BRAND causes the best split. The corresponding leaves of the tree show, that people who use their cars for private reasons purchase less frequently than others. Similarly, consumers who are concerned with brand are more frequent buyers than others. On level three, KILO is used in both categories shown: The more people drive the faster they replace their old car

Using equation (3) and taking the two cases above into account, the purchase frequencies at any future purchase could now be computed by the following algorithm:

```
If $CAR_0 < PF (AGE_0)$    {case 1, first purchase in t = 1}
          then   $(Age_0 - Car_0) + PF (AGE_0) = AGE_1$
If $CAR_0 \geq PF (AGE_0)$    {case 2, first purchase in t = 0}
         then  $[AGE_0 + PF (AGE_0)] = AGE_1$
Compute
$PF(AGE_1) = 5.834 + (-0.075 \bullet AGE_1) + (0.0017 \bullet AGE_1^2) + (-0.00000096 \bullet AGE_1^3)$
$+ (e^{-0.109 \bullet INCOME}) + (-1.556 \bullet KILO) + (0.429 \bullet KILO^2) + (-0.040 \bullet KILO^3)$
$+ (e^{0.647 \bullet PRIVATE}) - (0.961 \bullet BRAND)$
Compute  $AGE_2 = AGE_1 + PF (AGE_1)$
```

```
If        sex = male
          then  continue until $AGE_m \geq 72,99$
                {average life expectancy of males}
If        sex = male
          then  continue until $AGE_m \geq 79,59$
               {average life expectancy of females}
```

(*Note:* AGE_t = Age of customer in t. t is not equivalent to y in equation (1): t increases by one, y represents the years when a car is purchased. So, for our 40-year-old whose car is overdue, t = 1 will be in 5.4 years from now (y = 5). t and y coincide at present, when both are zero.)

3.2 Data Mining Task 2: Prediction of Price Acceptance

To calculate the CLV, the price acceptance *at any year of purchase* had to be estimated. Price acceptance increases with age decreasing again when people retire. We related the individual price acceptance in t = 0 to the price acceptance a person of the same age usually got (see table 2). If our 40-year-old customer intends to pay 35 TDM for a car, he spends 27% more than people of his age (see equation 4). The price acceptance at any of the customer's future purchases (variable PA_y in equation 1) has to be multiplied by this price ratio, because one can assume that if a person spends more on a car than others today, he will do so in the future as well.

Table 2. Price acceptance (PA) in terms of age (excerpt)

Age	PA (Median)	Age	PA (Median)	Age	PA (Median)	Age	PA (Median)
22	17,500 DM	35	27,500 DM	50	27,500 DM	65	22,500 DM
25	22,500 DM	40	27,500 DM	55	27,500 DM	70	22,500 DM
30	22,500 DM	45	27,500 DM	60	27,500 DM	75	22,500 DM

Note: PA was acquired in categories. To avoid spans, we substituted the class by its mean.

$$\text{Price Ratio} = PR = \frac{35{,}000\text{DM}}{27{,}500\text{DM}} = 1.2727 \mathrel{\hat{=}} 127.27\% \quad . \tag{4}$$

3.3 Data Mining Task 3: Prediction of the Rate of Discount

In order to discount future streams of payment, we must consider both the *market interest* and the *inflation* (see equation 5). The *market interest* depends on the alternative investment. We assumed a certain investment keeping complexity minimal. The time span considered, i.e. the time until the customer dies, is rather long. The investment with the longest repayment period is a 30-years federal loan. Its interest receivables, the so-called spot interest rates, were taken from a leading German newspaper [5]. However, the time to maturity is 30 years, while our stream of payments goes much more far into the future. Taking the most extreme example, a 18 year old female who dies at almost 80, will purchase cars over the next 62 years. To calculate the market interest for the years 30– 62, we conducted a regression analysis with the time as the independent and the spot interest rates as the dependent variable. Several regression models (linear, logarithmic, cubic, exponential etc.) were tested if they could reflect the slope of the observed spot interest rates, i.e. the market interest. The logarithmic function (see equation 6) yielded to the highest explained variance (99,8%).

$$\text{Discount Rate } (r) = \text{Market Interest } (mr) - \text{Price Increase } (pi) \ . \tag{5}$$

$$\text{Spot Interest Rate} = 3.1052 + [0.9485 \bullet \ln(t)] \ . \tag{6}$$

The market interest (*mi*) for the years 30–62 were estimated using equation 6. In order to estimate the *price increase* for cars (*pi*) for the next 62 years, we extrapolated historical data from the Federal Statistical Office. According to this, the price increase fluctuated quite evidently within the last 30 years (8.2% to -0,5%). Several regressions models explained only up to 51% of the variance. Thus, we used *exponential smoothing*. One problem is to determine the smoothing factor alpha. The higher it is, the heavier the weight of recent data. Moreover, a model with a high alpha is sensitive to structural changes [11]. These arguments suggest a relatively high alpha, say 0.5, which yields to a future price increase of about 0.8%. This seemed to be too less, the most recent years (with almost zero inflation) being quite untypical. So we chose alpha = 0.1 yielding to a future price increase of 2.5%. Finally, the *discount rate* was computed by subtracting the 2.5% price increase from the market interest.

3.4 Prediction of CLV with the Car Miner

To summarize the discussion from the last chapters, equation (1) has to be refined: As equation 7 shows, the CLV is the present value of the price acceptance at any future year of purchase. Y does not increase by one, but by the purchase frequency, which is a function of AGE_t, INCOME, PRIVATE, KILO, and BRAND. AGE_t is the only variable in this function which, in turn, depends on the purchase frequency of the year before (see the algorithm in chapter 1). Applied to the given data base, we predicted the CLV for all customers. Table 3 shows the average CLV of car drivers in the upper market segment. According to this, it is desirable to acquire BMW drivers. A driver of a BMW 7, for example, will spend about 240,000 DM on cars in his remaining life.

$$\text{CLV} = \text{PA}_0 + \sum_{y=0}^{n} \frac{\text{PA}(\text{AGE}_y) \bullet \text{PR}}{(1 + \text{mr} - \text{pi})^y} \ . \tag{7}$$

$$\text{while } y = \sum_{t=0}^{m} \text{PF}_t = \sum_{t=0}^{m} \text{PF}(\text{AGE}_t, \text{INCOME, PRIVATE, KILO, BRAND})$$

PA_0	... Price acceptance in t = 0	PF	... Purchase frequency
PR	... Price ratio	y	... Year of purchase
n	... Year of last purchase	t	... Time period
m	... Last period	r	... Rate of discount
mr	... Market rate of interest	pi	... Price increase for cars

Note: PA_0 only has to be included if the car is "overdue". Furthermore, PRIVATE is not constant. We set it to YES when people reached their retirement age.

Table 3. CLV as gross present value in terms of preferred brand

Brand and Type	CLV (in TDM)	Brand and Type	CLV (in TDM)
BMW 7	240	Audi 100 / A 6	127
BMW 5	180	Mercedes S	111
BMW 3	148	Mercedes 190 / C	103
Mercedes 200 / E	145	Audi 80 / A4	97

4 Restrictions and Guidelines for Future Research

- The regression model explains only 12% of the purchase frequency. Its power could be improved by including external variables (e.g. macro-economic trends).
- The FED influences the market interest as well as the price increase. Thus, the Value Miner should be calibrated with respect to different FED policy scenarios.
- The life expectancy has been increasing for the last years and is expected to rise in the future. So people will purchase more cars than assumed. On the opposite, people don't drive until they die. So the two effects cancel each other out to some degree. However, a more precise model should take both effects into consideration.
- The composite model introduced in chapter 3.4 is subject to further evaluation. The impact of varying constituents and / or parameters (e.g. linear regression model for the estimation of PF, uncertainty when estimating the interest receivables from the alternative investment) should be shown in alternative models.
- When it comes to the retention of current customers, the CLV should be calculated as a net value. This requires a sophisticated accounting. Moreover, soft benefits such as reference potential should be included in that case.

References

1. Brosius, F. 1997. SPSS Chaid (in German). Internat. Thomson Publ., Bonn (1997)
2. Bitran, G. R., Mondschein, S.: Mailing Decisions in the Catalogue Sales Industry. Management Science 9 (1996) 1364–1281
3. Cornelsen, J.: Customer Value. Working Paper, University of Erlangen-Nürnberg (1996)
4. Dwyer, F. R.: Customer Lifetime Profitability to Support Marketing Decision Making. Journal of Direct Marketing 4 (1989) 8–15
5. FAZ Frankfurter Allgemeine Zeitung (1999) 41
6. Johnson, M. D., Herrmann, A., Huber, F., Gustafsson (eds.): Customer Retention in the Automotive Industry. Gabler Verlag, Wiesbaden (1997)
7. Homburg, C.: Cross-Validation and Information Criteria in Causal Modeling. Journal of Marketing Research, 2 (1991) 137–145
8. Jackson, D. R.: Strategic Application of Customer Lifetime Value in the Direct Marketing. Journal of Targeting, Measurement and Analysis for Marketing 1 (1994) 9–17
9. Keane, T. J., Wang, P.: Applications for the Lifetime Value Model in Modern Newspaper Publishing. Journal of Direct Marketing 2 (1995) 59–66
10. Mulhern, F. J.: Customer Profitability. Journal of Interactive Marketing 1 (1999) 25–40
11. Nieschlag, R., Dichtl, E., Hörschgen, H.: Marketing, Duncker & Humblot, Berlin (1998)
12. Reichheld, F. F., Aspinall, K.: Building High-Loyalty Business Systems. Journal of Retail Banking 4 (1993/1994) 21–29