Advanced Business Analytics - Power of Predictive Modeling 226161-0131

Adrianna Wołowiec

e-mail: *aw60259@sgh.waw.pl*

The simple and general retention models

Foundations

Introduction

Companies acquire customers, provide them with a product or service and make a certain amount of profit each month until they terminate the relationship forever.

Questions:

1. How much profit do you expect to make from customers during their lifetimes?

2. How would increasing retention rates affect future profit?

It is the description of situation typical for contractual service providers such as Internet service providers, health clubs and media content providers.

Example: Subscribers to Netflix pay a certain amount each month until they cancel.

Estimating the customer's value

One application is determining how much can be spent to acquire a customer.

Example:

It may cost a cellular phone company $400 to acquire a new customer and provide a handset. Even if he only generates $50 profit per month, this would be a good investment as long as the cellular phone company can retain him sufficiently long to recoup the aquisition cost.

Likewise, a company that is considering whether to invest marketing resources in retaining customers longer will need to know CLV.

Models for estimating the value of customer

1. **The customer annuity model**

- Customers sign a contract for a certain numer of periods and are not allowed to cancel

2. **The simple retention model**

- It allows customers to cancel, but assumes that the retention rate is constant over time and across customers, and that cash flows are independent of the cancelation time

3. **The general retention model**

- It assumes retention rates can change over time and payment amounts depend on the time of cancelation

The customer annuity model

The customer annuity model

Customers sign a contract to make T payments of amount m in the future and are not allowed to cancel the contract. Customers in this business situation are annuities.

Notice that the payments come at the end of every period. Suppose further that the discount rate is d. The CLV of a customer is given by the formula for the present value (PV) on an ordinary annuity:

$$PV_T = \sum_{t=1}^{T} \frac{m}{(1+d)^t} = m \frac{1 - (1+d)^{-T}}{d}$$

More generally, the payment m could be a sum of cash inflows and outflows (net contribiution). An example of negatvive cash flow is the cost of marketing to a customer each period.

Example 1

In the book-of-the-month club customers receive a series of 12 books on different topics, one each month. Those who join it agree to buy T = 12 books, one at the end of each month., each generating a gross margin of m = $10. Find the CLV of a customer, which is the present value of the 12 payments using a monthly discount rate of d = 1%.

$10	$10	$10	$10	$10	$10	$10	$10	$10	$10	$10	$10
Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Month 8	Month 9	Month 10	Month 11	Month 12

Solution

$$PV_T = m \frac{1-(1+d)^{-T}}{d} = 10 \frac{1-(1+0,01)^{-12}}{0,01} = \$112,55$$

What if the club requires payments not at the end of the month, but at the beginning?

$$PV_T = m \frac{(1+d)[1-(1+d)^{-T}]}{d} = 10 \frac{(1+0,01)[1-(1+0,01)^{-12}]}{0,01} = \$113,68$$

The simple retention model

Assumptions

The customer annuity model assumes that customers make a pre-determined number of payments and that they never stop making payments before the end of the contract.

These assumptions are usually not realistic.

The simple retention model (SRM) estimates CLV assuming the folowing:

- The percentage of customers retained each month (the retention rate) r is constant over time and across customers

- The period cash flow m is unaffected by the cancelation date

- The event that a customer cancels in time period t is independent of the event that the customer cancels in any other time period

Example 2

An Internet service provider (ISP) acquires 1 000 customers who will pay $50 at the end of each month for the service, with a gross margin of $25. The ISP retains 80% of its customers each month and discontinues service immediately to anyone who fails to make a payment. Using a spreadsheet find the CLV of this cohort, assuming a monthly discount rate of 1%.

A review of elementary probability theory

A **random variable** X assigns a number to the outcome of a random experiment.

The **probability mass function** (PMF) of a random variable that takes discrete values gives the probability $f(x) = P(X = x)$ that random variable X will take the value x. (Upper-case letters represent random variables and corresponding lower-case letters indicate a realization of the random variable.)

The **mean** or **expected value** of X, which describes the location of the middle of the distribution of X: $E(X) = \sum_x x\, P(X = x)$. It can also be interpreted as a weighted average of the x values, with the weights determined by the probabilities.

For $g(X)$ - a transformation of X, the mean of $g(X)$: $E[g(X)] = \sum_x g(x)\, P(X = x)$.

A probabilistic model for CLV

Assume that all customers in some group are retained each period with probability r (the retention rate) for all periods and that the event a customer cancels during some period is independent of the event od cancelation during any other period.

Let T be a random variable indicating the time of cancelation. Under these assumptions, T has a geometric distribution. Probabilities of a geometric distribution are given by PMF: $f(x) = P(T = t) = r^{t-1}(1 - r)$. If t is the time of customer cancelation, the customer must be retained for $t - 1$ periods. Because defaulting is assumed to be independent across time periods, the retention probabilities can be multiplied, so that r^{t-1} is the probability of retaining a customer for $t - 1$ periods and $(1 - r)$ is the probability of defaulting (in the last period).

A probabilistic model for CLV

The survival function is the probability that a customer has survived until the beginning of period t: $S(t) = P(T \geq t) = r^{t-1}$. It is also equivalent to the probability that the customer cancels at time t or later, or that the customer survives the first $t-1$ periods.

The survival function can be used to find quantiles of T. The α quantile of random variable T, call it P_α, divides a distribution so that α percent of the distribution has $T \leq P_\alpha$ and $1 - \alpha$ percent has $T \geq P_\alpha$, that is $P(T \leq P_\alpha) = \alpha$ and $P(T \geq P_\alpha) = 1 - \alpha$. We can find the α quantile of the cancelation time by solving $S(t) = P(T \geq P_\alpha) = r^{P_\alpha - 1} = 1 - \alpha$. Taking logs of both sides and solving for P_α we find that: $P_\alpha = 1 + \frac{\log(1-\alpha)}{\log(r)}$.

For example, the median time until cancelation is found by substituting $\alpha = 0{,}5$ into the equation.

The mean of a geometric distribution is given by $E(T) = \frac{1}{1-r}$

Example 3

Graph the PMF and survival function of cancelation time for the ISP having retention rate of 80%. Find and interpret the mean and median time of cancelation. Suppose the ISP implements a new retention campaign and is able to increase the retention rate to 90%. Find the mean and median under this new retention rate.

Solution

For retention rate $r = 0{,}8$:

$$E(T) = \frac{1}{1 - r} = \frac{1}{1 - 0{,}8} = 5$$

Interpretation: The expected time until attrition is 5 months. Therefore ISP expects 4 payments from each customer if the payments come at the end of a period.

$$P_{0,5} = 1 + \frac{\log(1 - \alpha)}{\log(r)} = 1 + \frac{\log(0{,}5)}{\log(0{,}8)} = 4{,}106 \approx 4$$

Interpretation: At least half of the customers will survive until period 4.

Solution

For retention rate $r = 0,9$:

$$E(T) = \frac{1}{1 - r} = \frac{1}{1 - 0,9} = 10$$

Interpretation: The expected time until attrition is 10 months. Therefore ISP expects 9 payments from each customer if the payments come at the end of a period.

$$P_{0,5} = 1 + \frac{\log(1 - \alpha)}{\log(r)} = 1 + \frac{\log(0,5)}{\log(0,9)} = 7,579 \approx 8$$

Interpretation: At least half of the customers will survive until period 8.

CLV in simple retention model

CLV is the sum of the present values of future cash flows. When a customer cancels during period t, there will be t cash flows if they occur at the beginning of a period and $t-1$ cash flows if they come at the end.

The cancelation time T is random, thus CLV will have a distribution. Those customers who have larger T have a larger CLV. We can summarize the distribution of CLV with its mean.

For cash flows at the beginning of a period: $E(CLV) = \frac{m(1+d)}{1+d-r}$

For cash flows at the end of a period: $E(CLV) = \frac{m}{1+d-r}$

Example 4

Find the expected time until attrition and expected CLV of a single ISP customer assuming a monthly discount rate of 1%, monthly cash flows of 25% at the beginning of each month and retention rates of 70%, 75%, … 95%. Plot the CLV against the retention rate.

Estimating retention rates

Previously we assumed that the retention rate r was known, but in practice it usually must be estimated from data. An organization observes when customers are acquired and can trace the history of payments. Some but not all of these customers will cancel. A customer who has not yet canceled is said to be **censored** and the organization will not have observed this customer's cancelation time yet.

We want to estimate the retention rate for some group of $n_0 + n_1$ customers that were acquired in the past, where n_1 of the customers have already canceled (so the default time t has been observed), while n_0 others are still active (censored).

Estimating retention rates

Let T denote a random variable and t be an observed cancelation time, that is, a realisation of T. If customer i has already canceled, let t_i be the observed times of defection, so that customer i has been active $t_1 - 1$ periods and canels at time t_i. For those still active, let c_i be the time of censoring so that customer i has been active c_i periods and the company knows that the time of cancelation for this customer is $T_i > c_i$.

We can estimate the retention rate as: $\hat{r} = 1 - \dfrac{n_1}{\sum t_i + \sum c_i}$

The hat over the $\hat{r}$ indicates that it is an estimation of parameter r. This equation has an intuitive interpretation. The denominator of the second term gives the total number of periods in which customers can cancel (opportunities to cancel) and numerator is number of cancelations . Thus, the second term estimates the default rate.

Example 5

Basing on the crosstab below, estimate the retention rate and both the expected and median time until cancelation.

Status	\multicolumn{12}{c}{Time of cancelation/censoring}	Total
	1	2	3	4	5	6	7	8	9	10	11	12	Total
Canceled	0	4	16	20	37	28	61	24	19	13	10	13	245
Censored	3	0	2	1	7	33	49	63	30	16	34	188	426
Total	3	4	18	21	44	61	110	87	49	29	44	201	671

The general retention model

Assumptions

An organization acquires customers who generate cash flows at time $t = 0,1,2,\dots$, until canceling during period T, which is a random variable. Let r_i be the probability of retaining a customer at time t. The general retention model relaxes the assumption that $r_t = r$ for all t. Assume further that the event a customer cancels at time t is independent of canceling at time $t' \neq t$.

The **survival function** gives the chance that the customer cancels at time t or later (that the customer is retained for the first $t-1$ periods: $S(t) = P(T \geq t) = \prod_{i=1}^{t-1} r_i$
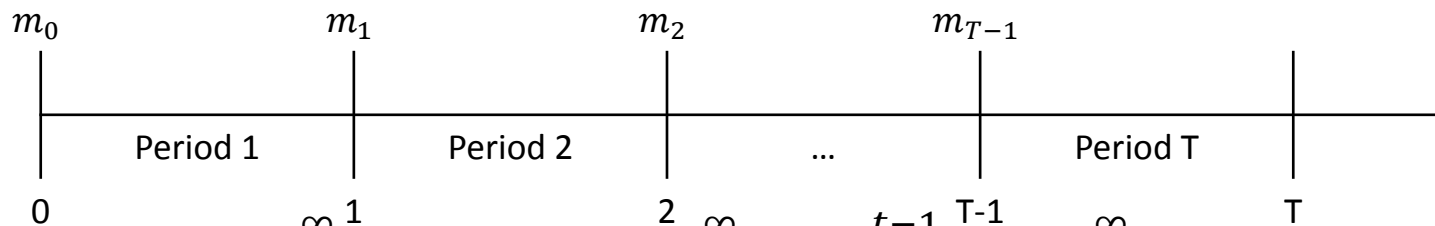
The **probability mass function** (PMF) gives the probability that a customer is retained during the first $t-1$ periods and cancels during period t: $f(t) = P(T = t) = S(t)(1 - r_i) = S(t) - S(t-1)$

For analysis we refer to $f(t)$ as a **probability density function** (PDF).

Let $\pi_t = 1 - r_t$ be the **hazard rate**, which is the conditional probability of canceling at time t given that the customer has not already canceled: $\pi_t = P(T = t | T \geq t) = \frac{P(T=t)}{S(t)} = 1 - r_t$

Advanced Business Analytics - Power of Predictive Modeling
226161-0131

CLV in general retention model

Let m_t be the discounted cash flow at time t. Time is indicated on the diagram below the tick marks. A customer is acquired at time 0, at which time cash flow m_0 occurs. Period 1 refers to the time between time 0 and 1, so if a customer cancels in period 1 the company receives only 1 payment.

$$E[CLV(T)] = \sum_{t=1}^{\infty} f(t)CLV(t) = \sum_{t=1}^{\infty} f(t) \sum_{i=0}^{t-1} m_i = \sum_{t=1}^{\infty} m_{t-1}S(t)$$

$$E(T) = \sum_{t=1}^{\infty} S(t)$$

Example 6

A hypothetical call phones provider offers only six-month contracts, where customers pay $50 at the beginning of each month. Suppose that 80% are retained in the first month, 90% in the second month and 95% thereafter except at the times of contract renewal (months 6,12,18,…), when the retention rate is 50%. Find the expected lifetime revenue, assuming a monthly discount rate of 1%. Graph the PDF, survival and hazard functions.

Introduction to survival analysis

Survival analysis is a set of methods for understanding the following questions about the time when some event such as cencelation occurs:

1. What is the distribution of the event times? The shape of the distribution is characterized by the hazard function (or equivalently the PDF or survival function)

2. How do static characteristics of customers such as acquisition source, demographics and the lenght of the initial contract affect the distribution of the event? Static covariates are variables that do not change over time such.

3. How do things that happen during the course of a relationship – time-dependent covariates – affect the hazard of T? Time-dependent covariates change over time.

Survival analysis enables us to estimate retention rates so that we can find CLV using the general retention model. These methods describe the shape of distribution without making strong assumptions about its shape, they also accomodate censoring.

Survival analysis methods

Two survival analysis methods we can use for GRM are Kaplan-Meier and the life-table method.

$$\hat{S}(T) = \prod_{i=1}^{t} (1 - \frac{d_i}{n_i}) = \hat{S}(t - 1)\left(1 - \frac{d_t}{n_t}\right)$$

For the Kaplan-Meier method: $n_{t+1} = n_t - d_t - c_t$

For the life-table method: $n_{t+1} = n_t - d_t - \frac{c_t + c_{t+1}}{2}$

where: d_t is the number of customers who canceled at time t, n_t is the number of customers at risk of cancelling at time t and c_t is the number of customers censored at time t.

Example 7

Estimate the survival, probability density and hazard functions using SAS. Use *service5yr* data from the service organization over a five-year period. Assume a monthly discount rate of 1% and payments *m* = $23.20, made at the beginning of each period. Compare the estimates with those from SRM.

Variables:

startlen – the length in months of the strating contract

bigT – time of cancelation or censorung (in months)

cancel – 1, when a customer canceled; 0, when an observation is censored

count – frequency of observations

Example 8

Using *service5yr* data, compare the survival, hazard and probability density functions accross starting contrac lengths for the service provider. Also compare the expected value estimates.