

LTV prediction using probabilistic models

Jonatan Piñol and Peter Weber

CONTENTS

1	Introduction	2
2	Customer Basis Analysis	2
2.1	Probabilistic models for Customer Basis Analysis	2
2.2	Possible scenarios:	2
2.3	Our case: Model LTV prediction	3
3	Dataset description	3
4	Model: Shifted Geometrical Beta distribution for retention prediction	4
4.1	Model Assumptions	4
4.2	Prior and Likelihood	5
4.3	Maximum Likelihood	7
5	The experiment	8
5.1	Parameter estimation	8
5.2	Quality of the model	9
6	Conclusions and Future Work	10

1 INTRODUCTION

In this project we want to perform an estimation of customer life-time value (LTV) as a case of a Customer-Base Analysis using what literature in this domain calls probabilistic models, which is in reality no other thing than using a Bayesian approach. We will be using two datasets from the company where we work.

This work will be divided in the following parts: First, a discussion of Customer-Basis Analysis and possible scenarios and characterization of our scenario. Second, a description of the model we will use (assumptions etc...), and, third, a discussion about implementation of the model. Finally, we will present the results obtained and a short evaluation of the model.

We found that this would make a proper project for a "Bayesian Statistics" course, as it is in between a Data-oriented and an Application-oriented project type.

2 CUSTOMER BASIS ANALYSIS

Customer Basis Analysis is a concept that is being more and more critical as companies have access (or generate) to more customer's transaction data (increase in complexity and volume). The availability of these datasets has led to a general change from transaction based strategies to more customer-centric strategies. In other words, the first strategies lead typically to descriptive analysis (for example summary statistic like rates, average number of subscriptions ...), while the customer-centric approach is more focused on analysis on the user behaviour which forms the basis for a more predictive approach. Using characteristics and past behaviour (found in those databases) a forecast can be made. This is why a Bayesian Probabilistic approach can be convenient.

2.1 PROBABILISTIC MODELS FOR CUSTOMER BASIS ANALYSIS

As usual for a probabilistic model, the problem will be modeled in a way that observed behavior has an underlying random process. Thus, past behavior is not a true mirror of future actions. With this interpretation, the past, and the "future" are function of latent characteristics denoted by θ .

While this approach may not be new (in other fields), the novelty resides in the fact that classically for Customer Basis Analysis, the approaches have been regression-like (which also happened in the company we work) and the use of probabilistic approaches is relatively new.

2.2 POSSIBLE SCENARIOS:

Commonly in literature we will find the following classification for scenarios in Customer-Base Analysis, which is shown in Fig. 2.1. bservable and therefore must be inferred. For example, in a contractual scenario, a user might pay for a service only when needed, so that if during a long time this user didn't make any transaction either it should be considered that he not any more a client or it's just a hiatus, we don't know it directly.

The company in which we work, has a few subscription based websites, where users are subscribed in order to have access to contents and services. After each cycle (30 days) the user is

Opportunities for Transactions	Continuous	Grocery purchases Doctor visits Hotel stays	Credit card Student mealplan Mobile phone usage
	Discrete	Event attendance Prescription refills Charity fund drives	Magazine subs Insurance policy Health club m'ship
		Noncontractual	Contractual
		Type of Relationship With Customers	

Figure 2.1: Examples for the different scenarios. From [2]

automatically rebilled and in order to drop out, the user has to cancel his subscription. This scenario corresponds to a contractual scenario where transactions are done periodically. Concretely, we will work with data from two products, which underly the same type of setting but slightly different user behaviour.

2.3 OUR CASE: MODEL LTV PREDICTION

For the scope of this project, we will model the Customer Life Time Value as a particular case of Customer Basis Analysis, using a Bayesian Probabilistic model.

In marketing, customer lifetime value (CLV or often CLTV), or life-time value (LTV) is a prediction of the net revenue attributed to the entire future relationship with a customer. The prediction model can have varying levels of sophistication and accuracy, ranging from a crude heuristic to the use of complex predictive analytics techniques. Many businesses strongly depend on this metric as it is their only projection of revenue for the near future.

For this, we will implement the proposed model for Customer Retention Prediction described in [1], on our datasets.

3 DATASET DESCRIPTION

Our case is of a Discrete Contractual scenario (as we see in 2.1) , with a 30 days period cycle of rebill. The users can cancel their relation with the company at any time. The data we use has been collected over the last two years and it is composed of number of clients that rebill after each cycle. We have added two additional columns that will be useful for modeling, the number of clients as a percentage and the number of users that cancels during each cycle. In particular, we start off with about 47 thousand users that initially sign up, and then on every cycle we loose some of these users. We consider 12 rebill cycles in total. The second dataset

looks similar. Both datasets will be provided as .rds files, which are read in by the provided R file.

numRebill	Count	CountNorm	Diff
0	41747	100.00000	0
1	16026	38.38839	25721
2	10960	26.25338	5066
3	8578	20.54758	2382
4	7027	16.83235	1551
5	5878	14.08005	1149

Figure 3.1: Number of user rebilling at each cycle

4 MODEL: SHIFTED GEOMETRICAL BETA DISTRIBUTION FOR RETENTION PREDICTION

Our first idea to approach this problem was using a Negative Binomial with a Beta distribution as prior. We first tried this implementation with an MCMC implementation in order to estimate the probability of cancel of the users. For this model to be applicable we always need data of the full customer lifetime, as we only model the cancellation of users. In this approach all users should be cancelled after the last rebilling cycle that is considered.

However, as we want to use this model to predict the full customer lifetime, but having only a few rebill cycles for the user we need a model that takes into account the users that are still active at the end of our observation period. This is exactly what the model presented in [1] considers. Therefore, we will implement their approach and apply it to our data.

The model we want to test is also based on using a negative binomial distribution (more precisely a geometric distribution as the failure number r is set to 1), with an additional term that enables to consider the surviving users. Let's see this model in more detail.

4.1 MODEL ASSUMPTIONS

The assumptions of the model are:

- At the end of each cycle, the user renews its subscription or cancels with some probability
- That probability remains constant for every user, but
- is heterogeneous accross users.

Those assumptions are translated in the following probabilities. The probability of duration for a given user(we will call it retention function), in terms of counts of rebills (success) before cancel the contract (failure), which is a classical Negative Binomial Distributio with $r = 1$

$$P(T = t|\theta) = \theta(1 - \theta)^{t-1} \quad (4.1)$$

The second term is the probability of remaining as client in for a given cycle (survival function), which is a Negative Binomial with $r=0$

$$S(t|\theta) = (1 - \theta)^t, \quad (4.2)$$

where T is the number of cycles that the user has rebilled (the number of counts), and θ the probability of cancel.

So in this model, the parameter that is not directly measurable and has to be estimated is θ .

4.2 PRIOR AND LIKELIHOOD

For this setting it is not a bad assumption to consider as a prior distribution for θ a Beta distribution, since it is a conjugate pair of the negative binomial distribution, with a proper choice of α and β parameters. Prior Distribution:

$$f(\theta; \alpha, \beta) = \frac{\theta^{\alpha-1}(1 - \theta)^{\beta-1}}{B(\alpha, \beta)} \quad (4.3)$$

As for the Likelihood function, we should not forget what we want to model. In a classic scenario of counting successes/failures, we would only take into consideration the retention probability part i.e. the users that cancels after a number of rebills (thus with a Negative Binomial distribution). In our case, we want to model either the behavior of the users that cancelled after some cycles and the users that are still active. The idea on doing this approach is that with data of few cycles we can predict behavior of the following cycles. That's the reason a survival (equation 4.2) function is considered.

Let's assume that we have data for a limited number of cycles, that we call T_{max} . Assuming independence between users, we will have a likelihood distribution

$$\prod_{t=1}^{T_{max}} P(T = t|\theta)^{n_t} S(T_{max}|\theta)^{N - \sum_{t=1}^{T_{max}} n_t}, \quad (4.4)$$

where

- n_t is the number of users that cancels on cycle t
- N is the number of total subscriptions (i.e. users that made transaction rebill 0)
- T_{max} is the maximum number of cycles considered

Note that this equation has two distinguishable parts. The first term, $\prod_{t=1}^{T_{max}} P(T = t|\theta)^{n_t}$ models the probability of users that had cancelled before cycle T_{max} and the second term $S(T_{max}|\theta)^{N - \sum_{t=1}^{T_{max}} n_t}$ the probability of the users still "alive".

As this likelihood term doesn't correspond exactly to any standard distribution, we opted for a Posterior Maximization/Maximum likelihood strategy to estimate the parameter. Note, that we use prior values for α and β close to zero. Therefore, posterior maximization and likelihood maximization are equivalent. For this, the approach that is used at [1] is to compute the likelihood function as a function of α and β parameters by averaging the likelihood with the prior (thus in reality computing the prior/posterior predictive) and maximizing the likelihood for α and β .

We present the derivations to obtain these expressions (adapted from [1]) :

For the retention part :

$$\begin{aligned} P(T = t|\alpha, \beta) &= \int_0^1 P(T = t|\Theta = \theta) f(\theta; \alpha, \beta) d\theta = \int_0^1 \theta(1-\theta)^{t-1} \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)} d\theta \\ &= \frac{1}{B(\alpha, \beta)} \int_0^1 \theta^{\alpha}(1-\theta)^{\beta+t-2} d\theta = \frac{B(\alpha+1, \beta+t-1)}{B(\alpha, \beta)} \end{aligned}$$

For the survival part :

$$\begin{aligned} S(t|\alpha, \beta) &= \int_0^1 S(t|\Theta = \theta) f(\theta; \alpha, \beta) d\theta = \int_0^1 (1-\theta)^t \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)} d\theta \\ &= \frac{1}{B(\alpha, \beta)} \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta+t-1} d\theta = \frac{B(\alpha, \beta+t)}{\alpha, \beta} \end{aligned}$$

The retention part has been implemented in a recursive way, using the expression proposed in the article:

First we should note that the Γ function fulfills the following identity

$$\frac{\Gamma(x+1)}{\Gamma(x)} = x$$

and

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

These identities are used to derive the recursive expression

$$P(T = t|\alpha, \beta) = \frac{P(T = t|\alpha, \beta)}{P(T = t-1|\alpha, \beta)} P(T = t-1|\alpha, \beta)$$

First we have to compute the initial expression in which we use the above expressions of the Gamma function and the definition of Beta function in terms of Gamma functions

$$P(T = 1|\alpha, \beta) = \frac{B(\alpha+1, \beta)}{B(\alpha, \beta)} = \frac{\Gamma(\alpha+1)/\Gamma(\alpha)}{\Gamma(\alpha+\beta+1)/\Gamma(\alpha+\beta)} = \frac{\alpha}{\alpha+\beta}$$

Then we can compute the first term of the recursive expression

$$\begin{aligned}\frac{P(T = t|\alpha, \beta)}{P(T = t-1|\alpha, \beta)} &= \frac{B(\alpha + 1, \beta + t - 1)}{B(\alpha + 1, \beta + t - 2)} = \frac{\Gamma(\beta + t - 1)/\Gamma(\beta + t - 2)}{\Gamma(\alpha + \beta + t)/\Gamma(\alpha + \beta + t - 1)} \\ &= \frac{\beta + t - 2}{\alpha + \beta + t - 1}\end{aligned}$$

So at the end the recursive expression is :

$$P(T = t|\alpha, \beta) = \begin{cases} \frac{\alpha}{\alpha + \beta} & \text{if } t = 1 \\ \frac{\beta + t - 2}{\alpha + \beta + t - 1} P(T = t|\alpha, \beta) & \text{if } t > 1 \end{cases} \quad (4.5)$$

4.3 MAXIMUM LIKELIHOOD

In order to find the optimal parameters for α and β we use a maximum likelihood approach, in which we assume that the dropout of one user is independent of the dropout of another user. As discussed in previous chapters, the likelihood is a product of all the users that have dropped out before rebill cycle T_{max} , and all the users who are still active at the end of rebill cycle T_{max} . Concretely, we quantify all the dropouts at cycle t by $P(T = t|\alpha, \beta)^{n_t}$, while all the users still active after cycle T_{max} is $S(T_{max}|\alpha, \beta)^{N - \sum_{t=1}^{T_{max}} n_t}$, where N is the number of users at $t = 0$. This yields the log likelihood function (the logarithm of equation 4.4)

$$\log(LL(\alpha, \beta|data)) = \sum_{t=1}^{T_{max}} n_t \log(P(T = t|\alpha, \beta)) + \left(N - \sum_{t=1}^{T_{max}} n_t \right) \log(S(T_{max}|\alpha, \beta)). \quad (4.6)$$

We do a grid search to find the parameters α and β that maximize the log likelihood function given the data observed. Once we obtain the optimal parameters we can then fit either the posterior predictive for the survival probability

$$S(t|\alpha, \beta) = \frac{B(\alpha, \beta + t)}{B(\alpha, \beta)}, \quad t = 1, 2, \dots \quad (4.7)$$

or the posterior predictive for the probability that a user drops out

$$P(T = t|\alpha, \beta) = \frac{B(\alpha + 1, \beta + t - 1)}{B(\alpha, \beta)}, \quad t = 1, 2, \dots \quad (4.8)$$

Here, we will focus on the survival probability, since this is the quantity of interest for modeling the expected revenue per user.

5 THE EXPERIMENT

We use two datasets related to two products that our company sells. We will see that the first dataset is more regular in terms of following the shifted Beta Geometric distribution, in particular, for the first rebills, and it is actually a product that has higher retention of the users. We use in total 12 rebill cycles for both products, where we will give a comparison of the model being trained on all the 12 cycles, and being trained on only four cycles. The latter can be used to predict long term customer survival and therefore the lifetime value of a user.

5.1 PARAMETER ESTIMATION

A plot of the log likelihood function for our first data set can be found in figure 5.1. Here we use the full 12 cycles for the parameter estimation. In the plot, the parameters range between zero and one, because this is the region where we found the maximum of the function. We then repeat the same procedure for a reduced dataset where we consider only 4 cycles. In both cases we then plot the full dataset first where the first is overlayed by a fit obtained from the full parameter estimation, and second overlayed by a fit from the partial parameter estimation (4 cycles). The fits are done using Eq. 4.7.

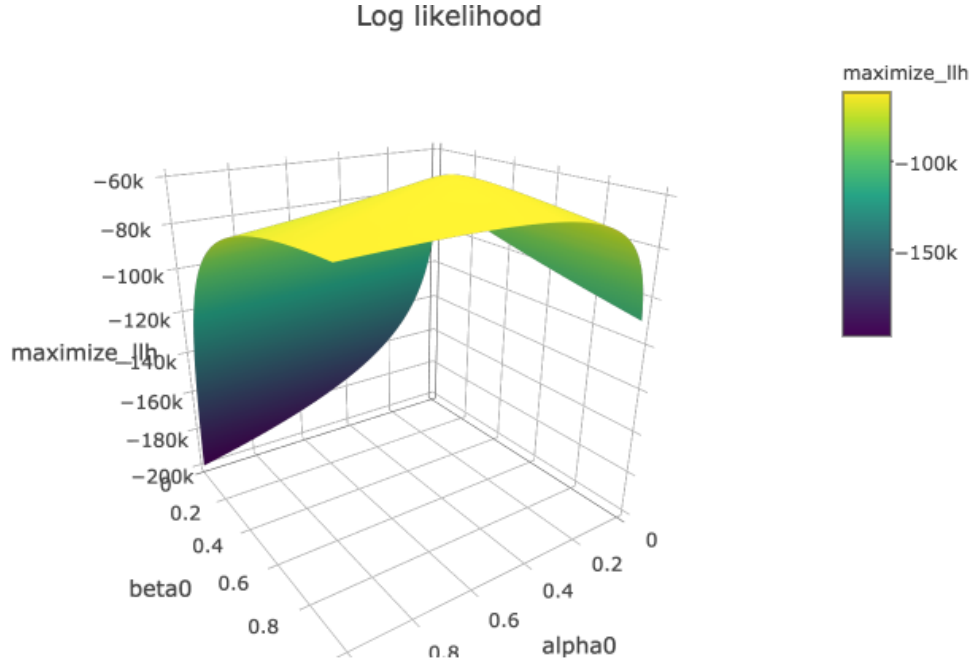


Figure 5.1: Log likelihood function of the dataset related to the first product. The maximum is found at $\alpha = 0.52$ and $\beta = 0.82$

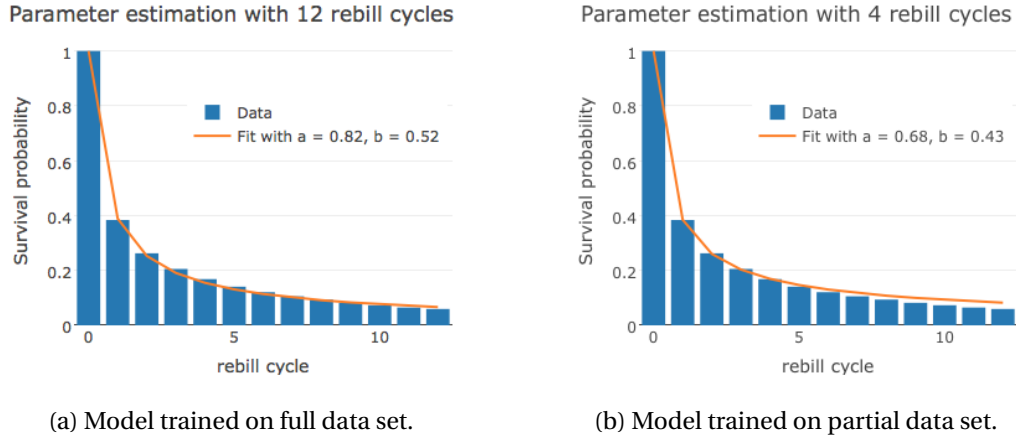


Figure 5.2: Data and modelling for the first product considered.

5.2 QUALITY OF THE MODEL

In order to judge the quality of the model, in the table below, we present the error between the model and data as a percentage. We observe that the model trained on the full data agrees better with the data than the model used for prediction, with a maximum deviation of 10% vs 40%. This is intuitively obvious. Overall, the predictive model overestimates the survival probability, which, when known, could be manually corrected by using knowledge of historical data.

We repeat the same type of analysis for a second product, see Fig. ?? and table 5.2. We observe, that the model does not fit very well the second and the third rebill cycle, but there is no obvious difference between the full model (trained on all data) and the predictive model (trained on 4 cycles). However, when checking the last table below we find that the discrepancies for the predictive model are by about 50% larger. Again, the predictive model overestimates towards large rebill cycles.

12 cycles for parameter estimation		4 cycles for parameter estimation	
Rebill cycle	Error (%)	Rebill cycle	Error (%)
0	0.00	0	0.00
1	1.31	1	0.24
2	-3.70	2	-0.85
3	-7.13	3	-1.12
4	-8.02	4	0.66
5	-6.90	5	4.26
6	-5.47	6	8.00
7	-4.13	7	11.45
8	-2.43	8	15.19
9	1.70	9	21.74
10	5.13	10	27.44
11	10.05	11	34.96
12	11.67	12	38.41

Table 5.1: Errors between data and model for first product, where two left most column quantify error for model trained on full dataset, and two left most column on partial dataset.

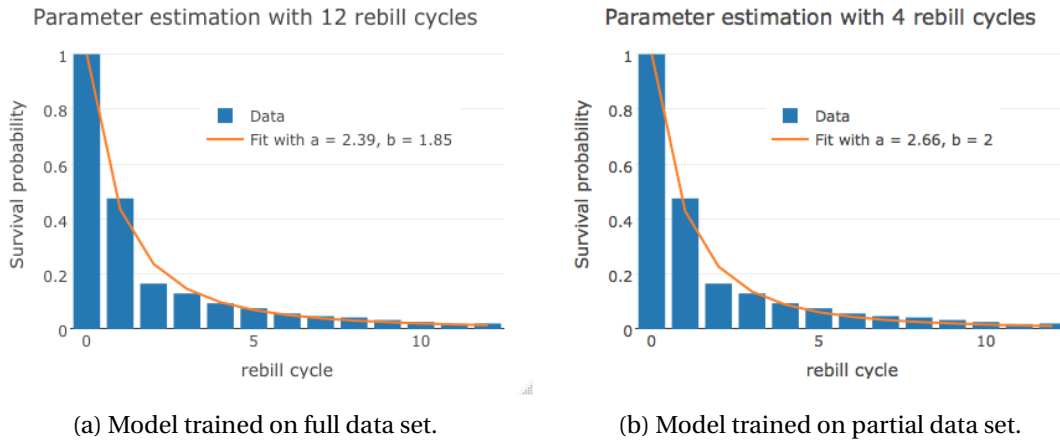


Figure 5.3: Data and modelling for the second product considered.

6 CONCLUSIONS AND FUTURE WORK

In this work, we have studied a probability model for dropout prediction of clients, which we use in the future to estimate long term customer survival and customer life time. By using it on different datasets we have seen it's adaptability for different cases. Our further goal is to implement a MCMC method for this type of model.

Future Work:

12 cycles for parameter estimation		4 cycles for parameter estimation	
Rebill cycle	Error (%)	Rebill cycle	Error (%)
0	0.00	0	0.00
1	-8.28	1	-9.78
2	43.01	2	37.09
3	12.51	3	4.99
4	4.02	4	-5.41
5	-8.46	5	-18.77
6	-9.78	6	-21.74
7	-16.68	7	-29.24
8	-26.92	8	-39.16
9	-25.16	9	-38.85
10	-21.58	10	-37.03
11	-9.49	11	-28.51
12	-30.43	12	-45.90

Table 5.2: Errors between data and model for second product, where two left most column quantify error for model trained on full dataset, and two left most column on partial dataset.

This model assumes that the probability of cancel is constant after each cycle which may not be the case for certain products in which factors like "boredom" has hight impact. A possible upgrade that can be explored is the characterization of the cancel probability as cycle dependent. A possible approach to explore is the use of states for each rebill cycle, with different transition probabilities each state (use of a Hidden Markov Model could be a possibility).

REFERENCES

- [1] Peter S. Fader and Bruce G. S. Hardie *How To Project Customer Retention*, 2007.
- [2] Peter S. Fader and Bruce G. S. Hardie *Probability Models for Customer-Base Analysis*, 2009.