Forecast Customer Retention | Beta-geometric Model

1/18/2020

Business overview

GetFit is a company that runs outdoor health fitness classes. The classes run on a monthly cycle that begins on the first day of each month. New customers are “booked” at the beginning of the month, and decide to renew or cancel just before the beginning of the next month. For example, a cohort that is acquired initially on January 1 will be active for all of January. Just before February 1, some members of the January cohort will renew (and will be active in February), while others will cancel. By definition, those who cancel have been active for one month, while those who renew are active for at least two months. At the end of February, some members of the January cohort will renew for March (active for at least three months), while others will cancel (active for exactly two months), and so forth.

It is now early September, and eligible customers have already made their renewal decisions for the month. The data set provided contains some information about 2,132 customers who were acquired in January, February, March, or April (i.e., four different cohorts), and may or may not have churned through August. These customers can be considered to be a random sample from the population of potential GetFit customers, and there is no substantive difference across cohorts.

Data

For each of these customers, we observe a customer ID number, the index of the month in which the customer was acquired (the first month in which the customer was active), and the index of the month after which the customer canceled service (i.e., the last month of the customer relationship). If last is missing (denoted by NA), it means the customer is still active in September (and thus must have been active in August).

For example, a customer for whom first is 3 and last is 8 was active for 6 months (March through August), but decided to not renew for September. Another customer with first=3 and a missing last value has been active for 6 months so far, but remains active in September. Put another way, the first customer churned at the 6th renewal opportunity, while the second customer has survived all 6 renewal opportunities. Note that the duration of the survival time depends on when the customer was first acquired.

```
##      cust first last
##  1: 00001     2   NA
##  2: 00002     3    3
##  3: 00003     3    3
##  4: 00004     4   NA
##  5: 00005     4    5
##  6: 00006     2    2
##  7: 00007     2    3
##  8: 00008     2    2
##  9: 00009     2    3
## 10: 00010     2    2
```

Data preprocessing

Create a new variable named dur (“duration”) to calculate the duration for each customer, and another new variable named sur (“survivor”) that takes a value of sur = 0 for churners and sur = 1 for survivors.

```
for (i in 1:nrow(getfit)) {
  if (is.na(getfit$last[i])) {
    getfit$dur[i] <- 8 - getfit$first[i] + 1
    getfit$sur[i] <- 1
  } else {
    getfit$dur[i] <- getfit$last[i] - getfit$first[i] + 1
    getfit$sur[i] <- 0
  }
}

head(getfit, 5)
```

```
##      cust first last dur sur
## 1: 00001     2   NA   7   1
## 2: 00002     3    3   1   0
## 3: 00003     3    3   1   0
## 4: 00004     4   NA   5   1
## 5: 00005     4    5   2   0
```

Build the BG model

Estimate a BG model using these data. Report the maximum likelihood estimates (MLE) of the model parameters, and well as the log likelihood at the MLE. Explain how to interpret the parameter estimates, in terms of how they describe the distribution of churn probabilities across the population of GetFit customers.

```
# define log_P and log_S
log_P <- function(t, a, b) {
  lbeta(a+1, b+t-1) - lbeta(a, b)
}
log_S <- function(t, a, b) {
  lbeta(a, b+t) - lbeta(a, b)
}

# define the log likelihood
LL_BG <- function(pars, t, s) {
  a <- exp(pars[1])
  b <- exp(pars[2])
  LL_ind <- (1-s)*log_P(t, a, b) + s*log_S(t, a, b)
  return(-sum(LL_ind))
}
```

```
# Estimating the BG model
start <- c(0,0)
opt_BG <- optim(start,fn=LL_BG, t=getfit$dur, s=getfit$sur)
a <- exp(opt_BG[["par"]][1])
b <- exp(opt_BG[["par"]][2])
LL <- -opt_BG[["value"]]
```

```r
c(a,b)
```

```
## [1] 0.8289739 1.5650389
```

```r
LL_BG(log(c(a,b)), t=getfit$dur, s=getfit$sur)
```

```
## [1] 3602.33
```

As shown above, the log likelihood function is maximized at a=0.829 and b=1.565, with LL=-3602.33.

- Each customer has a churn probability θ that does not change over time. But Customers have different θ's. The distribution of θ is estimated as a beta distribution with parameter a = 0.829 and b = 1.565.

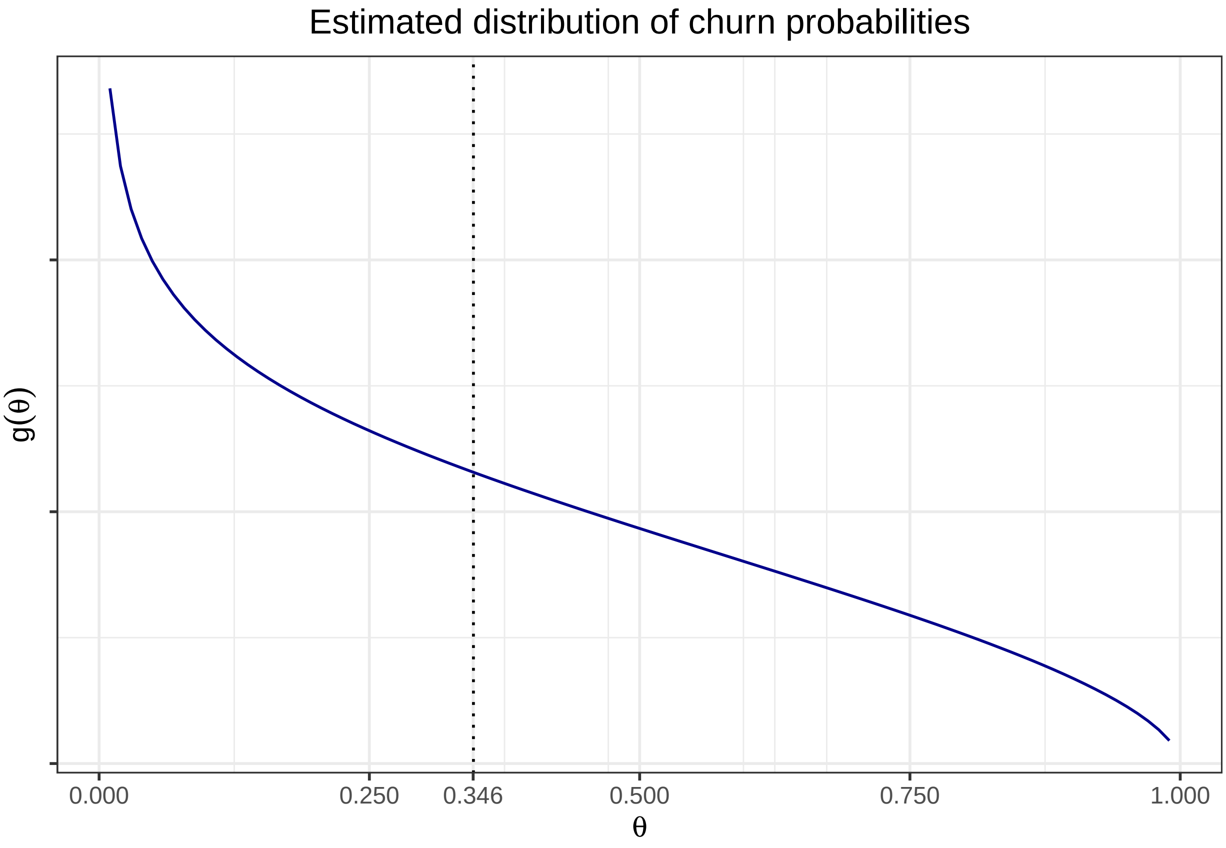
- The estimated mean churn probability is

$$E(\Theta|a,b) = \frac{a}{a+b} = \frac{0.829}{0.829+1.565} = 0.346$$

- The estimate variance of churn probability is

$$var(\Theta|a,b) = \frac{ab}{(a+b)^2(a+b+1)} = \frac{ab}{(0.829+1.565)^2(0.829+1.565+1)} = 0.067$$

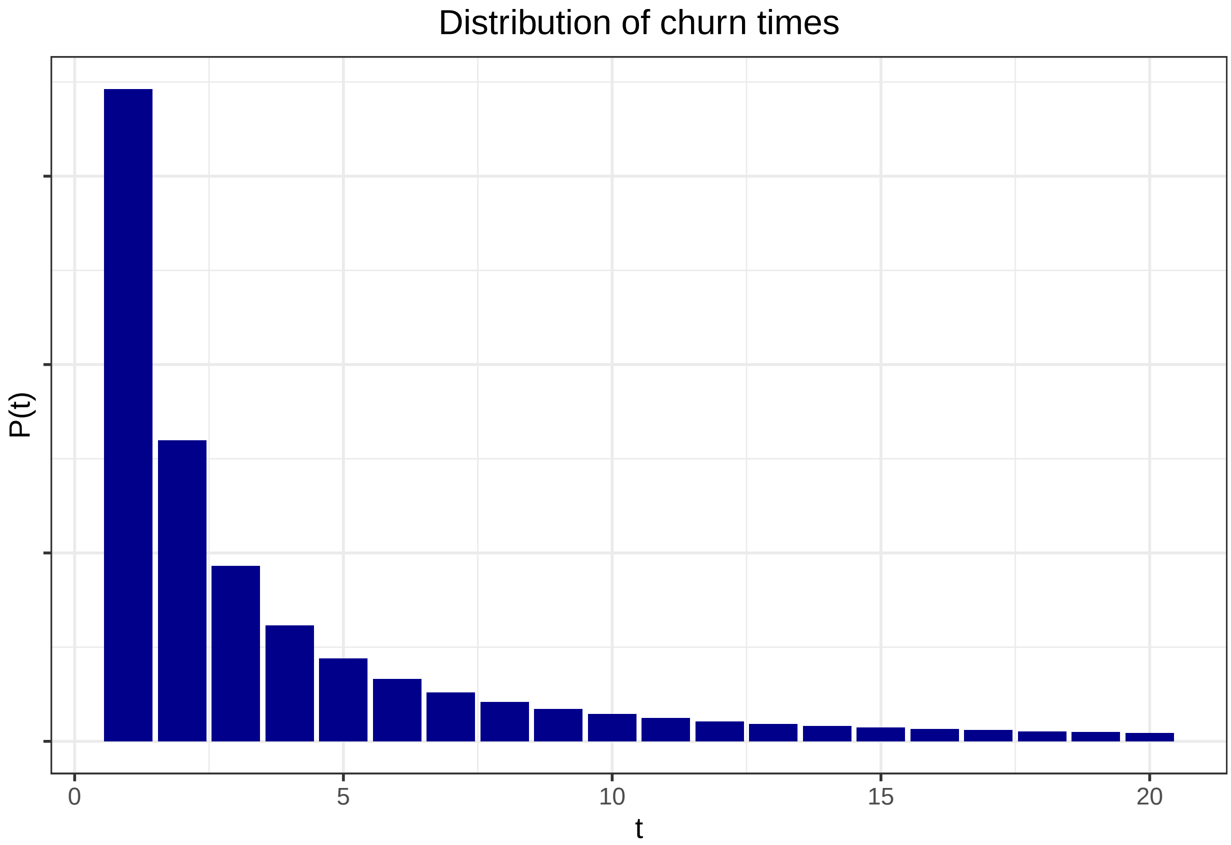
- The shape: with a < 1 and b > 1, the distribution of churn probability is L-shaped as below, which shows high heterogeneity in θ:



Question 1

Consider a randomly-chosen member of the population who was newly acquired in January.

(a) What is the estimated probability that this customer will cancel service after only one month? The probability that a newly acquired customer will churn after exactly t periods is $P(T=t|a,b)=\dfrac{B(a+1,b+t-1)}{B(a,b)}$, which could be shown as the plot below.



For t=1 (customer churn after only one month), the estimated probability is

$$P(T=1|a,b)=\frac{B(a+1,b+t-1)}{B(a,b)}=\frac{B(0.829+1,1.565+1-1)}{B(0.829,1.565)}=0.346$$

```r
beta(a+1,b+1-1)/beta(a,b)
```

```
## [1] 0.3462696
```

(b) What is the expected probability that this customer will cancel service after two months? For t=2 (customer churn after only one month), the estimated probability is

$$P(T=2|a,b)=\frac{B(a+1,b+t-1)}{B(a,b)}=\frac{B(0.829+1,1.565+2-1)}{B(0.829,1.565)}=0.16$$

```
beta(a+1,b+2-1)/beta(a,b)
```

```
## [1] 0.159671
```

(c) What is the probability this customer is still active? The probability that a newly acquired customer will retain beyond t periods is

$$S(t|a,b) = \frac{B(a,b+t)}{B(a,b)}$$

Because this customer was newly acquired in January, we expect this customer to retain beyond 8 months if this customer is still active. For t=8, the estimated probability is

$$S(8|a,b) = \frac{B(a,b+t)}{B(a,b)} = \frac{B(0.829,1.565+8)}{B(0.829,1.565)} = 0.215$$

```
beta(a,b+8)/beta(a,b)
```

```
## [1] 0.2154201
```

Question 2

Suppose there are 800 new customers acquired this month (September), and the distribution of their characteristics is the same as for customers acquired in previous months. Please predict the number of these customers who are active in each month, through next June.

Under the beta-geometric model, $S(t|\theta) = \dfrac{B(a,b+t)}{B(a,b)}$.
Predicted survivors $= 800 * S(t|a,b)$.

```
month <- c(9, 10, 11, 12, 1, 2, 3, 4, 5, 6)
NO = 800
t <- 1:length(month)
d1 <- tibble(Month = month.abb[month],
             Num_active_customer = NO*exp(lbeta(a,b+t-1)-lbeta(a,b)))

d1 %>% kable(digits=c(1,1,1), booktabs = TRUE, align=rep("c",3),
             caption = "Forecasting customer survival")
```

Table 1: Forecasting customer survival

Month	Num_active_customer
Sep	800.0
Oct	523.0
Nov	395.2
Dec	320.7
Jan	271.4
Feb	236.2
Mar	209.7
Apr	189.0
May	172.3
Jun	158.6

Question 3

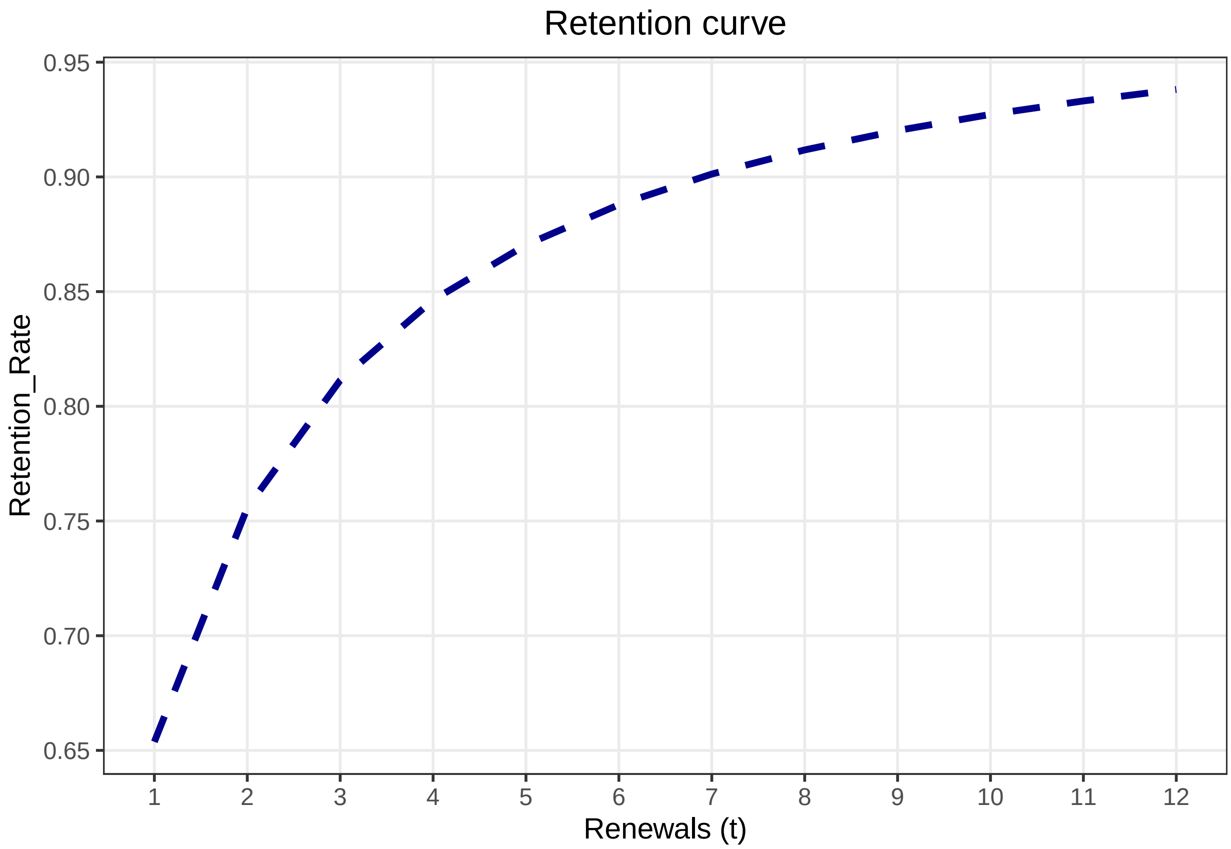
Plot the estimated retention curve for this population for 12 months and give a managerial explanation of why the retention curve has the shape that it does.

The retention rate is an expected proportion of active customers at the beginning of a period who are retained through the next period.

$$r(t|a,b) = \frac{B(a,b+t)}{B(a,b+t-1)} = \frac{b+t-1}{a+b+t-1}$$

```r
rt <- tibble(t=1:12,Retention_Rate=(b+t-1)/(a+b+t-1))

rt_plot <- ggplot(rt, aes(x=t, y=Retention_Rate)) %>%
  + geom_line(size=1.2, linetype = "dashed", color="darkblue") %>%
  + scale_y_continuous() %>%
  + labs(title = "Retention curve") %>%
  + theme_bw() %>%
  + theme(panel.grid.minor=element_blank(), plot.title = element_text(hjust = 0.5))
rt_plot%>%
  + scale_x_continuous("Renewals (t)", limits=c(1,12),
                       breaks=1:12)
```



- The plot shown above is the estimated retention curve for this population for 12 months.

- The curve is concave down, which shows increasing retention rate while slope is decreasing.
- The reason why retention curve has such a shape is because fitness industry is a heterogeneous market, "risky" (high θ) customers are more likely to churn at each renewal opportunity.
- But as high θ customers churn, the proportion of low θ customers among the survivors goes up. Since low-θ customers are more likely to be retained, the retention curve increases.