MKT 500T HW 1

Sami Cheong

Jan 27 2018

Projecting customer retention rates

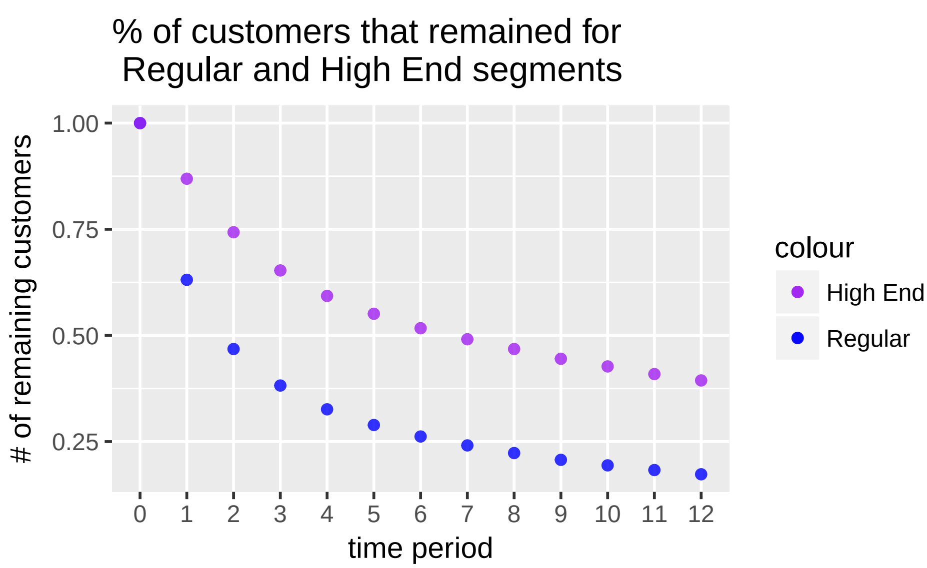
We are given the following customer retention data,

Year	Regular	High_End
0	1000	1000
1	631	869
2	468	743
3	382	653
4	326	593

and the goal is to predict the following retention rates in the following periods:

Year	Regular	High_End
5	289	551
6	262	517
7	241	491
8	223	468
9	207	445
10	194	427
11	183	409
12	173	394

Visualizing the customer population in temrs of % of total inital number of customers:



First, let's take a look at the churn rates and retention rates of each customer segment:

```
data[,'Churn_R']<- c(0, -diff(data$Regular))
data[,'Churn_H']<- c(0, -diff(data$High_End))
data[,'Ret_R']<- data$Regular/lag(data$Regular,1)
data[,'Ret_H']<- data$High_End/lag(data$High_End,1)
insamp.data<-data[1:5,]
kable(insamp.data)
```

Year	Regular	High_End	Churn_R	Churn_H	Ret_R	Ret_H
0	1000	1000	0	0	NA	NA
1	631	869	369	131	0.6310000	0.8690000
2	468	743	163	126	0.7416799	0.8550058
3	382	653	86	90	0.8162393	0.8788694
4	326	593	56	60	0.8534031	0.9081164

Assuming we only have 4 seasons of data, there are 326 survivors and 593 survivors in the *Regular* and *High End* segment respectively. Based on this training data, we are interested in knowing the period t in which a customer churns $P(T = t)$, with the following assumption:

- θ is the probability that a customer witll 'flip' and switch brands.
- $P(T = t|\theta)$ is the probability that a customer 'flip' at time t given θ, which follows the geometric distribution with density $f(t|\theta) = \theta(1 - \theta)^t$.
- θ follows a density function $p(\theta)$ characterized by $\Gamma(\alpha, \beta)$.

Now, since $P(T = t) = \int_0^1 f(t|\theta)p(\theta)d\theta$, after some algebra we can deduce that $P(T = t)$ follows the shifted Beta Geometric Distribution, which is defined by the following: $P(T = t) = \frac{\alpha}{\alpha+\beta}$ for t = 1 and $\frac{\beta+t-1}{\alpha+\beta+t-1} P(T = t)$ for t = 2,3,4...

```
## [1] "For high end segment, alpha is  1.2809761376149  beta is 7.79038228304711"
```

In here, α and β are obtained by maximizing the log-likelihood function defined by

$$\sum_{t=1}^{4} c_t \ln(P(T = t|\theta) + n_4 ln(P > 4|\theta),$$

where c_i is the number of customers who churns at time t and n_4 is the number of 'survivors' at the end of the 4^{th} season.

```
par.final<-rbind(par.final.reg,par.final.high)
colnames(par.final)<-c('alpha','beta')
row.names(par.final)<-c('Regular','High End')
kable(par.final,col.names =colnames(par.final) )
```

	alpha	beta
Regular	0.4079333	0.9816958
High End	1.2809761	7.7903823

```
data[,'Forecast_R']<-c(1,1-cumsum(pmf.reg))
data[,'Forecast_H']<-c(1,1-cumsum(pmf.high))
```

Putting it all together: