BB and BG/BB models

2/3/2020

Joyful Voyages, Inc. (JV) provided cruise tourism services to customers worldwide. In this report, a cohort of 18402 customers who took the JV cruise for the first time in 2009 was recorded in the dataset. From 2010 to 2014, each customer in the cohort was taking zero- or one-time cruise service on a yearly basis. Various perspectives of repeat behavior for this cohort of customers were taken and presented in the following report.

Data Preparation

To make our analysis process easier, we grouped the total number of attendances for each unique ID and created two new columns to show the recency and repetition of each customer. Recency was calculated by subtracting the base year 2009 from the latest year a customer attended the cruise. Rep was the sum of the total number of attendances for each customer minus one (subtracting the attendance of 2009).

```r
#Input Data
calib <- read.table("cruises_calib.txt", header = TRUE)
hold <- read.table("cruises_hold.txt" , header = TRUE)
options(digits=3)
```

```r
#filter data
calib1 <- calib %>% group_by(ID) %>% summarize(
  recency = max(Year)-2009,
  rep = n_distinct(Year)-1
)
hold1 <- hold %>% group_by(ID) %>% summarize(
  recency = max(Year)-2015,
  rep = n_distinct(Year)-1
)
```

Model Approach

First, we tried with the Beta-Binomial model. Base on the assumptions for the BB model.Each customer is independent when choosing whether or not he/she takes the cruise across each of the five years. If the customer decides to take the cruise in a given year, then the customer's decision is treated as a "success", vice versa. The results of customer decisions for each year are also independent of each other. The probability of "success" p should be stationary for each year, and p should obey beta distribution across five years.
We found $m = 5$, $a = 0.903$, $b = 9.053$ as model results.Since the dataset accounts for observations from 2009 to 2014, there should be five independent trials across the years ($m = 5$). Using the value of a and b, we can calculate the underlying probability for cruise repeat:

$$E(p) = a/(a + b) = 0.091$$

.

Since $a < 1$ and $b > 1$, the mixing distribution of p is L-shaped. Beta Binomial distribution formula:

$$P(X = x|m, a, b) = mx(B(a + x, b + m - x))/B(a, b)$$

```r
#Question 1
##BB Model
###Define log prob. for every person for beta-binomial distribution
log_bb <- function(x,m,a,b){
  lchoose(m,x) + lbeta(a+x,b+m-x)- lbeta(a,b)
}
###Define LL_BB for the group
LL_BB <- function(pars,x,m){
  a <- exp(pars[1])
  b <- exp(pars[2])
  z <- log_bb(x,m,a,b)
  return(-sum(z))
}

###Find the optimal a,b for BB model
start <- c(0,0)
opt <- optim(start, LL_BB, x = calib1$rep, m=5)
a_bb <- exp(opt$par[1])
b_bb <- exp(opt$par[2])
p_bb = a_bb/(a_bb+b_bb)
c(a_bb,b_bb)
```

```
## [1] 0.903 9.053
```

Also, we used the Beta-Geometric/Beta-Binomial model. Again, base on the assumptions for the BG/BB model The customer behaviors are characterized as: “alive”& “death” and “purchase” & “not purchase”. In the model, p depicts the probability of purchase, theta represents the customer churn rate (alive/death). Once a customer churns, he/she will never come back. While alive. a customer’s purchase probability p, and death probability theta are heterogeneous and follow the beta distribution. Plus, each customer’s decision is independent of each other. Based on assumptions above, the results from the BG/BB model are: $a = 1.472$, $b = 10.784$, $c = 0.332$, and $d = 2.365$. Using the value of a and b, we can calculate the underlying probability for cruise repeat:

$$E(p) = a/(a + b) = 0.1249$$

The result is higher compared to the BB model. It is more realistic because the BG/BB model captures the dropout process in which “death” rates vary across customers. Using the value of c and d, we can calculate the mean θ:

$$E(\theta|c, d) = c/(c + d) = 0.156$$

The average churn probability for each customer in this dataset is 0.156 and it is stationary over time.

```
##BB/BG model
###Define LL_BB for the group
LL_BGBB <- function(pars, x, tx, m) {
  p <- exp(pars)
  k <- pmax(0,lchoose(tx-1,x-1))
  LLi <- bgbb.LL(p, x, tx, m)+k
  return(-sum(LLi))
}

###Find the optimal a,b,c,d
st <- rep(0,4)
bgbb_opt <- optim(st, LL_BGBB,
                  x=calib1$rep,
                  tx=calib1$recency,
                  m=rep(5,nrow(calib1)),
                  method="BFGS")
bgbb_pars <- exp(bgbb_opt[['par']])
bgbb_pars  ##a,b,c,d
```

```
## [1]  1.472 10.784  0.332  2.365
```

```
bgbb_LL <- -bgbb_opt[['value']]
```

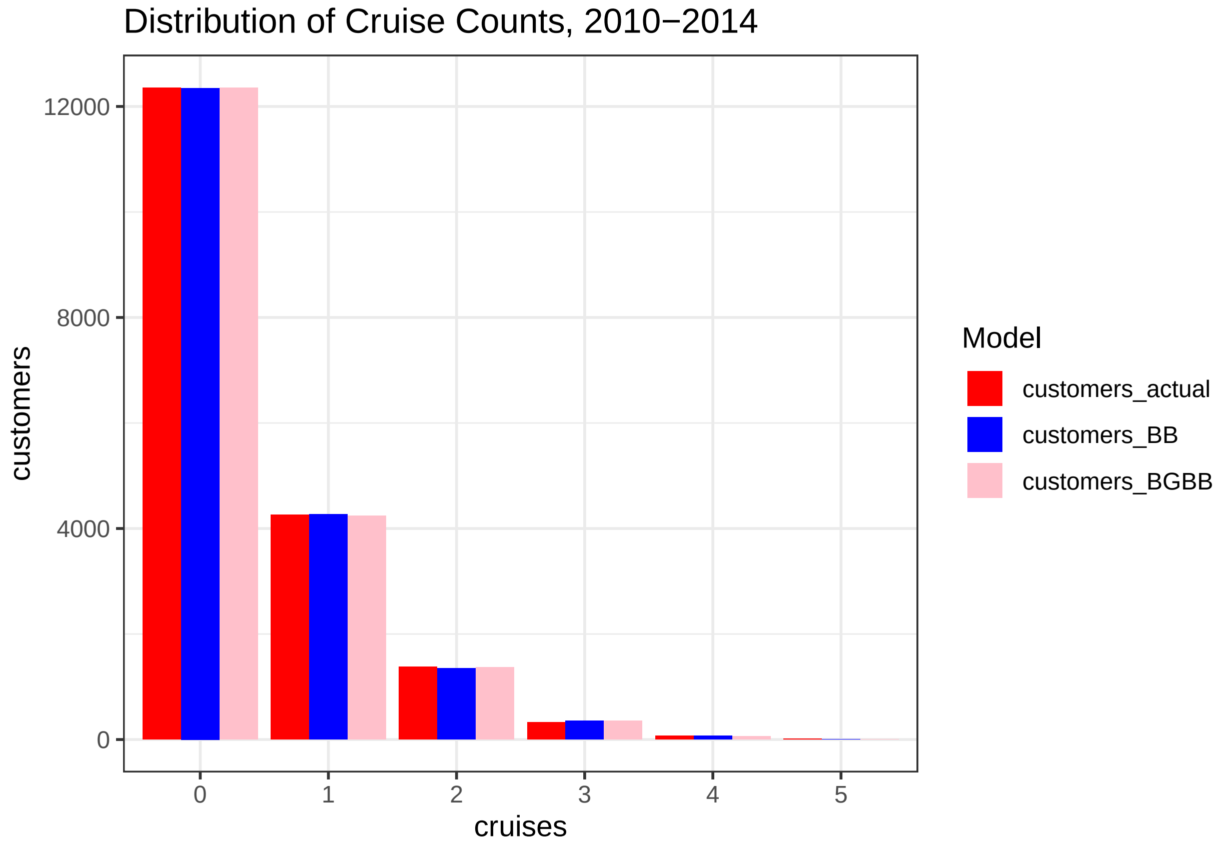
Model evaluation and comparison

We plotted table and bar charts to visualize the estimated parameters of the BB and the BB/BG models compared to the actual data.No clear discrepancies are shown in the bar chart. Refer to the numbers in the previous table, there are no large differences between the actual and the estimates from two different models. However, for customers who took cruises 5 times, the estimates from the two models were 6 and 7, which is a little off compared to the actual, 12. This may not be a problem of the model, but due to noises. Overall, both the BB and the BG/BB models fit well for the period of 2010 - 2014, but we do not know if the pattern will unceasingly fit for later periods.

```
#Question 2
##BB customers
m <- 5
prob_bb <- exp(log_bb(x = 0:m,m = m,a = a_bb,b = b_bb))
customers_bb <- round(nrow(calib1)*prob_bb,0)
table_bb <- tibble(cruises=0:m,customers_BB = customers_bb)

##BG/BB customer: get the prob. that a randomly chosen customer makes
#x=0,1,2,3,4,5 transactions in the first m=5 opportunities
m <- 5
prob_bgbb <- bgbb.pmf(bgbb_pars,m,0:m)
customers_bgbb <- round(nrow(calib1) * prob_bgbb,0)
table_bgbb <- tibble(cruises = 0:m, customers_BGBB = customers_bgbb)

##actual customers
table_actual <- calib1 %>% group_by(rep) %>% summarize(customers_actual=n_distinct(rep))
table_actual <- as.data.frame(table(calib1$rep))
colnames(table_actual)[1] <- "cruises"
colnames(table_actual)[2] <- "customers_actual"

##comprehensive table
table <- merge(table_actual,table_bb, by = 'cruises')
table <- merge(table,table_bgbb, by = 'cruises')
kable(table,digit=0,align = c('c', 'c', 'c','c'))
```

cruises	customers_actual	customers_BB	customers_BGBB
0	12351	12350	12358
1	4260	4272	4246
2	1380	1349	1370
3	330	354	355
4	69	69	66
5	12	7	7

```
  table <- table %>%
  gather(Model,customers, c(customers_actual,customers_BB,customers_BGBB))

ggplot(table,aes(x=cruises,y=customers,fill=Model)) %>%
  + geom_bar(position="dodge",stat="identity", size = 2) %>%
  + scale_fill_manual("Model",values=c(customers_actual='red', customers_BB='blue',
                                      customers_BGBB='pink')) %>%
  +theme_bw()%>%
  +ggtitle("Distribution of Cruise Counts, 2010-2014")
```



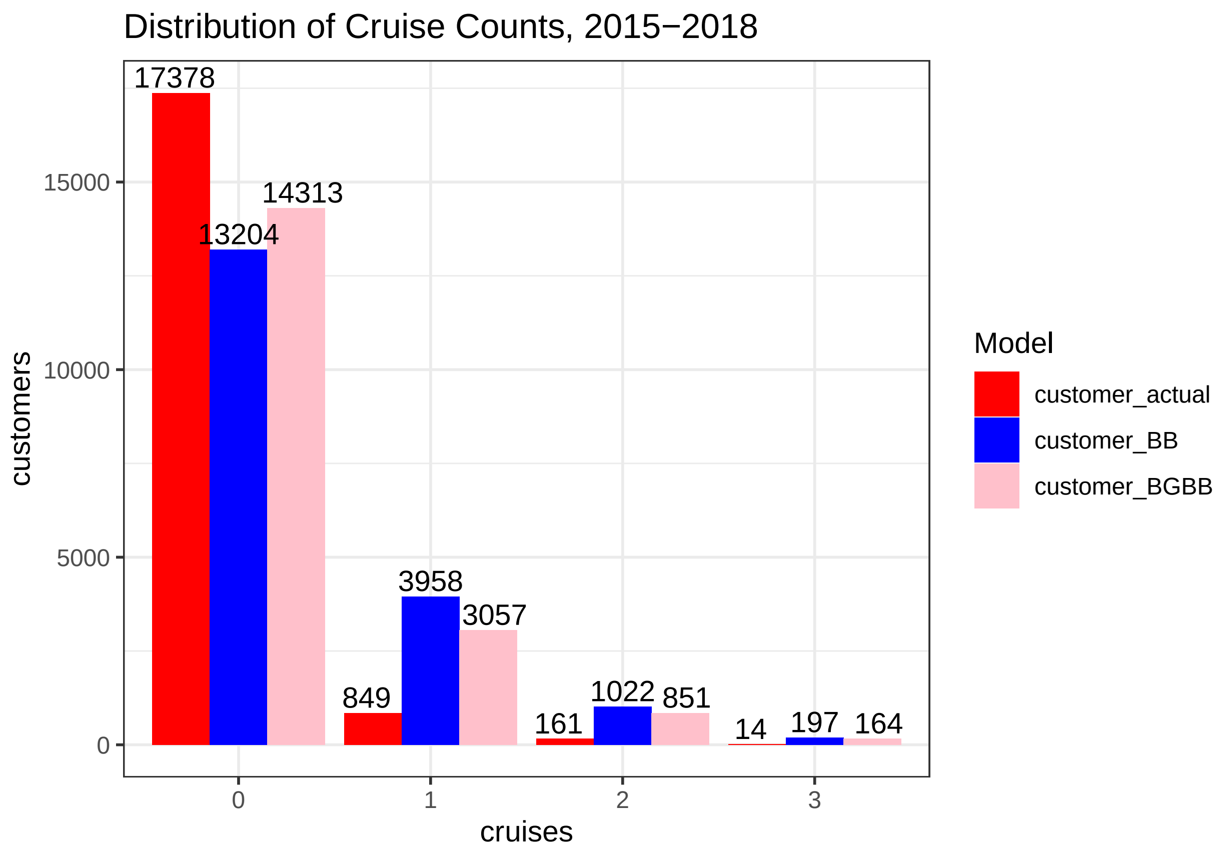
Similar visualizations were done to 2015-2018.Overall, counts for customers who took cruises 1, 2, and 3 times from 2015 to 2018 were overestimated, and count for customers did not take the cruise at all were underestimated. The BG/BB model, in general, had better predictions than the BB model because the BG/BB model took the customer churn rates into account: shown in the bar chart or the table, predictions using BG/BB model is less different from the actual cruises. The result of BG/BB model in the period of 2015-2018 was not good compared to the period of 2009 - 2014, because the estimated parameters were taken directly from the calibration data and applied to the holdout data, given that p and theta were stable from period one to period two. However, this assumption is not realistic. The unexpected and unpredicted change in p and theta may result in prediction imperfection.

```
#Question3
##Join two datasets and select the full hold dataset
colnames(calib1) <- c("ID","cal_recency","cal_rep")
colnames(hold1) <- c("ID", "hold_recency","hold_rep")
table_all =left_join(calib1,hold1,by='ID')
table_all[is.na(table_all)] <- 0
table_hold <- select(table_all,c("ID", "hold_recency","hold_rep"))

##actual customers
table_hold_actual <- as.data.frame(table(table_hold$hold_rep))
colnames(table_hold_actual) <- c("cruises", "customer_actual")

##BB customers
m.hold <- 4
```

```
prob_hold_bb <- exp(log_bb(x = 0:m.hold,m = m.hold,a = a_bb,b = b_bb))
customers_hold_bb <- round(nrow(table_hold)*prob_hold_bb,0)
table_hold_bb <- tibble(cruises = 0:m.hold, customer_BB = customers_hold_bb)

##BG/BB customer
m.hold <- 4
prob_hold_bgbb <- bgbb.pmf.General(bgbb_pars,m,m.hold,0:m.hold)
customers_hold_bgbb <- round(nrow(table_hold) * prob_hold_bgbb,0)
table_hold_bgbb <- tibble(cruises = 0:m.hold, customer_BGBB = customers_hold_bgbb)

##comprehensive customer
table_hold_all <- merge(table_hold_actual,table_hold_bb, by="cruises")
table_hold_all <- merge(table_hold_all,table_hold_bgbb,by="cruises")
kable(table_hold_all, digits = 0,align = c("c","c","c","c"))
```

cruises	customer_actual	customer_BB	customer_BGBB
0	17378	13204	14313
1	849	3958	3057
2	161	1022	851
3	14	197	164

```
##bar chart
table_tep <- table_hold_all %>%
  gather(Model,customers, c(customer_actual,customer_BB,customer_BGBB))
ggplot(table_tep,aes(x=cruises,y=customers,fill=Model)) %>%
  + geom_bar(position="dodge",stat="identity",size=.2) %>%
  + scale_fill_manual("Model",values=c(customer_actual='red', customer_BB='blue',
                                       customer_BGBB='pink')) %>%
  +theme_bw()%>%
  +geom_text(aes(label=customers), position=position_dodge(width=1), vjust=-0.25)%>%
  +ggtitle("Distribution of Cruise Counts, 2015-2018")
```



Usage

We are the company earns a margin of $241 from each cruise, and its annual cost of capital (discount rate) is 13%.Under the assumptions of the BG/BB model and conditioning on observed data in the period of 2010 to 2018, the expected number of the customer cohort from 2009 who remained “alive” at the beginning of 2019 is 10,253.

```
#Question 4
m.total <- 9
table_total <- rbind(hold,calib)
table_total1 <- table_total %>% group_by(ID) %>% summarize(
  recency = max(Year)-2009,
  rep = n_distinct(Year)-1
)

##a)
individual_PA <- bgbb.PAlive(bgbb_pars,table_total1$rep, table_total1$recency, m.total)
n_alive <- sum(individual_PA)
n_alive
```

```
## [1] 10253
```

We can also know that at the beginning of 2019, the total expected residual lifetime value of the 2009 is $1,997,407.

```
##b)
delta <- 0.13
margin <- 241
```

```
individual_NT <- bgbb.DERT(bgbb_pars,table_total1$rep,table_total1$recency,m.total,delta)
total_ERLV <- sum(individual_NT)*margin
total_ERLV <- as.numeric(format(total_ERLV,scientific=FALSE))
total_ERLV
```

```
## [1] 1997407
```

The company often acquires new customers through targeted online marketing campaigns. It believes that in the upcoming campaign for 2019, it can acquire 11,993 new customers. We tried to show the number of cruises the company can expect from this new cohort in each year from 2020 to 2024 below:

```
#Question 5
#a)
nc <- 11993
cum_t <- bgbb.Expectation(bgbb_pars, 1:5)
increm <- c(cum_t[1],diff(cum_t))
table <- data.frame(Year=2020:2024,
                    `Expected number of cruise`=increm*nc)
kable(table,digits=0, align=c('c','c'))
```

Year	Expected.number.of.cruise
2020	1263
2021	1149
2022	1068
2023	1006
2024	956

Also, for The spending on the campaign should not exceed the total margin/profit it can get from customers they acquire via the campaign in the future. With the margin of 241 bucks for one transaction, and it is believed that the number of customers it can acquire via the campaign is 11,993. Using the BB/BG model we built and tested, we got $DET = 1.59$, which means for those newly acquired customers via the campaign, we estimate that each of them will purchase 1.59 times during his/her "alive" time, based on our BB/BG model. Multiplying these numbers, we get total profits of $4,598,361, and Joyful Voyages, Inc. (JV) should not spend more than that on the campaign.

```
#b)
individual_DET <- 1+bgbb.DERT(bgbb_pars, 0, 0, 0, delta)
individual_DET
```

```
## [1] 1.59
```

```
DET <- margin*individual_DET*nc
DET
```

```
## [1] 4598361
```