Vungle

Roopa, Uros

October 12, 2017

Vungle uses 1/16 of the data to test the algorithm and they are trying to see which algorithm is the better one. The problem that arises is that the sample may be biased since it's such a small portion of the whole data set. Ideally they should have set the portion of the data to be much higher.

The key variables for managers are the conversion rates and the price at which the actual impression or click was served. Meaning that the biggest problem in the data we have is the fact that we aren't able to calculate the actual contribution per each user step.

Load the data and split it into two data frames for comparison

```
#.libPaths(c("C:/Users/Uros Randelovic/Documents/R/win-library/3.3", "C:/Program Files/R/R-3.3.2/librar

install.packages('dplyr',repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/Uros Randelovic/Documents/R/win-library/3.3'
## (as 'lib' is unspecified)

## package 'dplyr' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\Uros Randelovic\AppData\Local\Temp\RtmpI3SP5G\downloaded_packages
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.3.3

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
install.packages('xlsx',repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/Uros Randelovic/Documents/R/win-library/3.3'
## (as 'lib' is unspecified)

## package 'xlsx' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\Uros Randelovic\AppData\Local\Temp\RtmpI3SP5G\downloaded_packages
```

```
library(xlsx)
```

```
## Loading required package: rJava
```

```
## Loading required package: xlsxjars
```

```r
install.packages('Rmisc',repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/Uros Randelovic/Documents/R/win-library/3.3'
## (as 'lib' is unspecified)

## package 'Rmisc' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\Uros Randelovic\AppData\Local\Temp\RtmpI3SP5G\downloaded_packages
```

```r
library(Rmisc)
```

```
## Warning: package 'Rmisc' was built under R version 3.3.3

## Loading required package: lattice

## Loading required package: plyr

## -------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)

## -------------------------------------------------------------------------

##
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```r
data <- read.xlsx("C:/Users/Uros Randelovic/Documents/R workspace/BUS 111/Vungle/vungle data.xlsx",2
        )
library(car)
```

```
## Warning: package 'car' was built under R version 3.3.3

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode
```

```r
library(DescTools)
```

```
## Warning: package 'DescTools' was built under R version 3.3.3

##
## Attaching package: 'DescTools'

## The following object is masked from 'package:car':
##
##     Recode
```

```r
#View(data)

#split the data into different dataframes
```

```r
Data1 <- subset(data, Strategy %in% c("Vungle A"))
Data2 <- subset(data, Strategy =="Vungle B")
#View(Data1)
#View(Data2)
```

From the following piece of code We observe only absolute numbers that each strategy yielded without knowing how many customers have actually been assigned to each case. The fact that we only have absolute numbers makes it hard for us to compare the strategies. variables that we have been given are absolute numbers of users that went trough each segment of the conversion process. They are each continuous variables excluding date which we disregard in future analysis (row number is a proxy for the date).

```r
#drop uncessary empty columns to clean the data and leave only columns with data in them
Data1 <- subset(Data1, select = c(1:7))
Data2 <- subset(Data2, select = c(1:7))
#since the absoulte numbers are different calculate rations provided in the case to compare the models


summary(Data1)
```

```
##      Strategy        Date             Impressions        Completes
##  Vungle A:30   Min.   :2014-06-01   Min.   :5832627   Min.   :5193549
##  Vungle B: 0   1st Qu.:2014-06-08   1st Qu.:7773320   1st Qu.:6951338
##                Median :2014-06-15   Median :7999200   Median :7165950
##                Mean   :2014-06-15   Mean   :7881980   Mean   :7037023
##                3rd Qu.:2014-06-22   3rd Qu.:8392917   3rd Qu.:7436648
##                Max.   :2014-06-30   Max.   :9027910   Max.   :8075018
##      Clicks          Installs          eRPM
##  Min.   :291384   Min.   :23382   Min.   :2.943
##  1st Qu.:389044   1st Qu.:30585   1st Qu.:3.214
##  Median :402210   Median :31672   Median :3.326
##  Mean   :399899   Mean   :31722   Mean   :3.347
##  3rd Qu.:436483   3rd Qu.:33090   3rd Qu.:3.478
##  Max.   :478901   Max.   :38260   Max.   :3.830
```

```r
summary(Data2)
```

```
##      Strategy        Date             Impressions       Completes
##  Vungle A: 0   Min.   :2014-06-01   Min.   :420187   Min.   :373085
##  Vungle B:30   1st Qu.:2014-06-08   1st Qu.:506569   1st Qu.:452704
##                Median :2014-06-15   Median :520847   Median :464537
##                Mean   :2014-06-15   Mean   :527513   Mean   :468281
##                3rd Qu.:2014-06-22   3rd Qu.:559108   3rd Qu.:492302
##                Max.   :2014-06-30   Max.   :586702   Max.   :522522
##      Clicks         Installs         eRPM
##  Min.   :20629   Min.   :1360   Min.   :2.587
##  1st Qu.:24826   1st Qu.:1748   1st Qu.:3.324
##  Median :25343   Median :1846   Median :3.434
##  Mean   :25985   Mean   :1868   Mean   :3.459
##  3rd Qu.:27896   3rd Qu.:1946   3rd Qu.:3.674
##  Max.   :29483   Max.   :2221   Max.   :4.073
```

Summary statistics differ from each other because we have only absolute numbers thus making it impossible for us to compare them. The only variable we can compare is the eRPM where we observe that B outperforms A, as stated in the case, but we do not know which other variable affected it, if any. We now proceed to create new variables that will help us compare these two data sets.

```r
#creating a dataframe to store the calculations
compareData <- data.frame(fillRate=double(),
                 completeRate=double(),
                 clickRate=double(),
                 conversionRate=double(),
                 stringsAsFactors=FALSE)


#equalize the number of rows
compareData <- rbind(Data1, compareData)
#delete all columns
compareData <- subset(compareData, select = -c(1:28))

#dataset 1 - calculate each ratio to be used later on for comparison
Data1$completeRate <- Data1$Completes/Data1$Impressions
Data1$clickRate <- Data1$Clicks/Data1$Impressions
Data1$conversionRate <- Data1$Installs/Data1$Impressions

#we explore how each of the rates interacts with the the revenue figure
summary(lm(Data1$eRPM ~ Data1$completeRate+Data1$clickRate+Data1$conversionRate))
```

```
##
## Call:
## lm(formula = Data1$eRPM ~ Data1$completeRate + Data1$clickRate +
##     Data1$conversionRate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33227 -0.08466  0.02032  0.09305  0.29966
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.490      6.376   0.234   0.8171
## Data1$completeRate     -3.390      6.266  -0.541   0.5932
## Data1$clickRate       122.724     25.854   4.747 6.56e-05 ***
## Data1$conversionRate -331.774    149.511  -2.219   0.0354 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1525 on 26 degrees of freedom
## Multiple R-squared:  0.5556, Adjusted R-squared:  0.5044
## F-statistic: 10.84 on 3 and 26 DF,  p-value: 8.446e-05
```

From the above we see that rates explain 50% of the variance in the eRPM but also show that conversion rate is the only significant variable.

We proceed to look at how each of the created variables correlate with each other to determine if there is significant relationship between eRPM and the ratios

```r
Data1 <- subset(Data1, select = -c(Data1$Strategy) )
Data1 <- within(Data1, rm("Date"))
corrplot::corrplot(cor(Data1),method="number")
```


```r
#load to compare data frame
compareData <- cbind(Data1$completeRate,compareData)
compareData <- cbind(Data1$clickRate,compareData)
compareData <- cbind(Data1$conversionRate,compareData)
```

We repeat the same process for the strategy B

```r
#dataset 2
Data2$completeRate <- Data2$Completes/Data2$Impressions
Data2$clickRate <- Data2$Clicks/Data2$Impressions
Data2$conversionRate <- Data2$Installs/Data2$Impressions


data2LM <-lm(Data2$eRPM ~ Data2$completeRate+Data2$clickRate+Data2$conversionRate)
summary(data2LM)
```

```
##
## Call:
## lm(formula = Data2$eRPM ~ Data2$completeRate + Data2$clickRate +
##     Data2$conversionRate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.87968 -0.07639  0.04615  0.15706  0.47380
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)              5.747     11.134   0.516    0.610
## Data2$completeRate      -7.118     10.539  -0.675    0.505
## Data2$clickRate        103.630     75.507   1.372    0.182
## Data2$conversionRate -302.705    297.192  -1.019    0.318
##
## Residual standard error: 0.3356 on 26 degrees of freedom
## Multiple R-squared:  0.1489, Adjusted R-squared:  0.05073
## F-statistic: 1.517 on 3 and 26 DF,  p-value: 0.2336
```

From the above we see that rates explain only 15% of the variance in the eRPM but also show that there are no significant variables compared to algorithm A which had much better R^2. This is consistent with our hypothesis that the sample of 1/16 of the users is not a representative.

```
Data2 <- subset(Data2, select = -c(Data2$Strategy) )
Data2 <- within(Data2, rm("Date"))
```

```
## Warning in rm("Date"): object 'Date' not found
```

```
corrplot::corrplot(cor(Data1),method="number")
```

From the correlation plots for both data sets we observe that the variables correlate with each other almost exactly the same showing us that there isn't a significant difference in the user journey from impression to installs.

```
#load to compare data frame
compareData <- cbind(Data2$completeRate,compareData)
compareData <- cbind(Data2$clickRate,compareData)
compareData <- cbind(Data2$conversionRate,compareData)
summary(compareData)
```

```
## Data2$conversionRate Data2$clickRate   Data2$completeRate
## Min.   :0.003043     Min.   :0.04772   Min.   :0.8740    
## 1st Qu.:0.003387     1st Qu.:0.04866   1st Qu.:0.8834    
## Median :0.003581     Median :0.04926   Median :0.8878    
## Mean   :0.003538     Mean   :0.04924   Mean   :0.8878    
## 3rd Qu.:0.003705     3rd Qu.:0.04981   3rd Qu.:0.8930    
## Max.   :0.003936     Max.   :0.05169   Max.   :0.8987    
## Data1$conversionRate Data1$clickRate   Data1$completeRate
## Min.   :0.003590     Min.   :0.04864   Min.   :0.8798    
## 1st Qu.:0.003911     1st Qu.:0.04992   1st Qu.:0.8903    
## Median :0.004023     Median :0.05030   Median :0.8933    
## Mean   :0.004027     Mean   :0.05068   Mean   :0.8927    
## 3rd Qu.:0.004097     3rd Qu.:0.05135   3rd Qu.:0.8961    
## Max.   :0.004592     Max.   :0.05411   Max.   :0.8995
```

```r
#differences in ratios to determine the effectiveness of B over A
compareData$clickDiff <- compareData$`Data2$clickRate`- compareData$`Data1$clickRate`
compareData$completeDiff <- compareData$`Data2$completeRate`- compareData$`Data1$completeRate`
compareData$conversionDiff <- compareData$`Data2$conversionRate`- compareData$`Data1$conversionRate`

#calculate erpm difference
compareData$erpmDiff <- Data2$eRPM- Data1$eRPM
#look at the side by side comparison of A and B algorythm and their effectiveness on erpm
summary(select(.data = compareData, clickDiff, completeDiff,conversionDiff,erpmDiff))
```

```
##    clickDiff            completeDiff        conversionDiff     
## Min.   :-0.0036065   Min.   :-0.009466   Min.   :-0.0008818  
## 1st Qu.:-0.0017519   1st Qu.:-0.006078   1st Qu.:-0.0005321  
## Median :-0.0013465   Median :-0.004767   Median :-0.0004724  
## Mean   :-0.0014378   Mean   :-0.004866   Mean   :-0.0004890  
## 3rd Qu.:-0.0009752   3rd Qu.:-0.003153   3rd Qu.:-0.0003997  
## Max.   :-0.0000544   Max.   :-0.000868   Max.   :-0.0002362  
##     erpmDiff       
## Min.   :-0.37400  
## 1st Qu.: 0.04525  
## Median : 0.16200  
## Mean   : 0.11190  
## 3rd Qu.: 0.23625  
## Max.   : 0.40800
```

Summary table of these three variables shows us that algorithm B did a poorer job than the algorithm A since it yielded a lower ratio in each of the three categories except the income that was generated from the algorithm B. It seems that on average B generated .16$ more than A. Reason for this might be the fact that the ad campaigns that were served were just more expensive. We are not provided data about the cost thus we cannot make such inferences but should be aware of potential sample bias.

```r
#testing the means
#Ho: The differnece is not statistically significant at 99% confidence
#H1: The difference is statistically signifiacnt at 99% confidence
MeanCI(compareData$erpmDiff,
       conf.level=0.99)
```

```
##       mean     lwr.ci     upr.ci 
## 0.11190000 0.01913923 0.20466077
```

We conclude that difference in eRPM is statistically significant at 99% confidence interval.

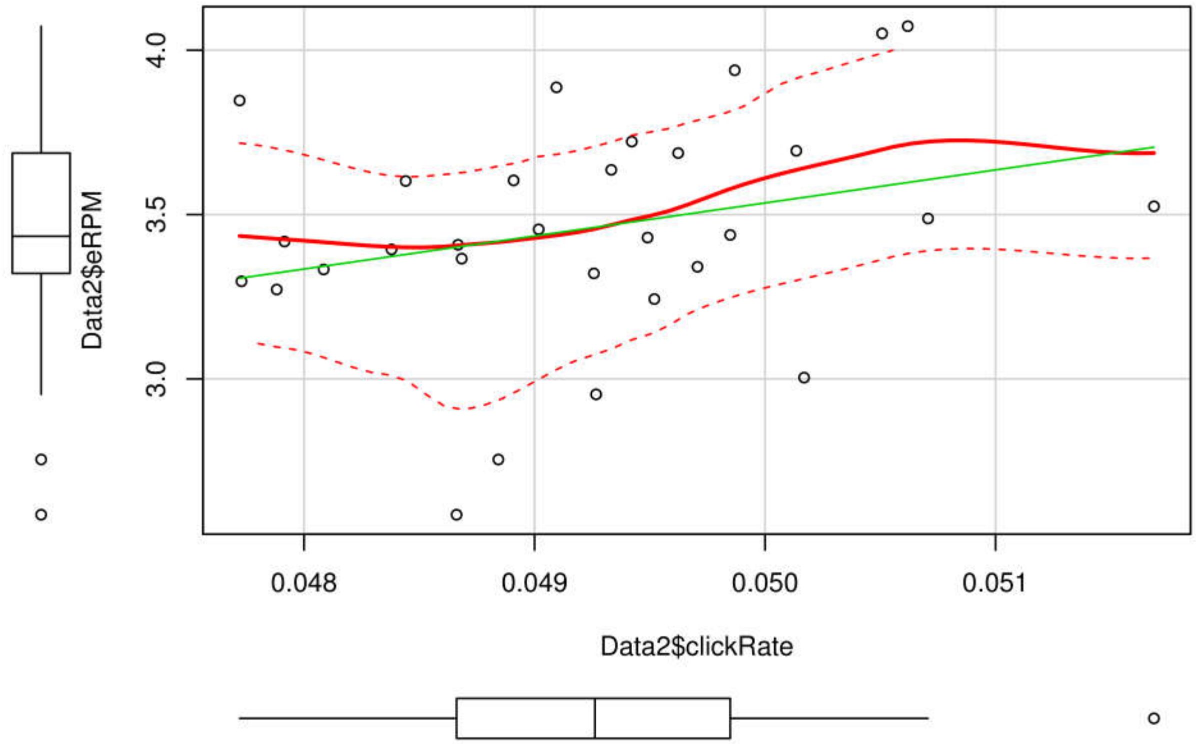
```r
summary(lm(compareData$erpmDiff~compareData$conversionDiff))
```

```
##
## Call:
## lm(formula = compareData$erpmDiff ~ compareData$conversionDiff)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31675 -0.09871  0.00155  0.10287  0.32541
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  0.4440     0.1003   4.425 0.000133 ***
## compareData$conversionDiff 679.1611   196.6224   3.454 0.001776 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1571 on 28 degrees of freedom
## Multiple R-squared:  0.2988, Adjusted R-squared:  0.2737
## F-statistic: 11.93 on 1 and 28 DF,  p-value: 0.001776
```

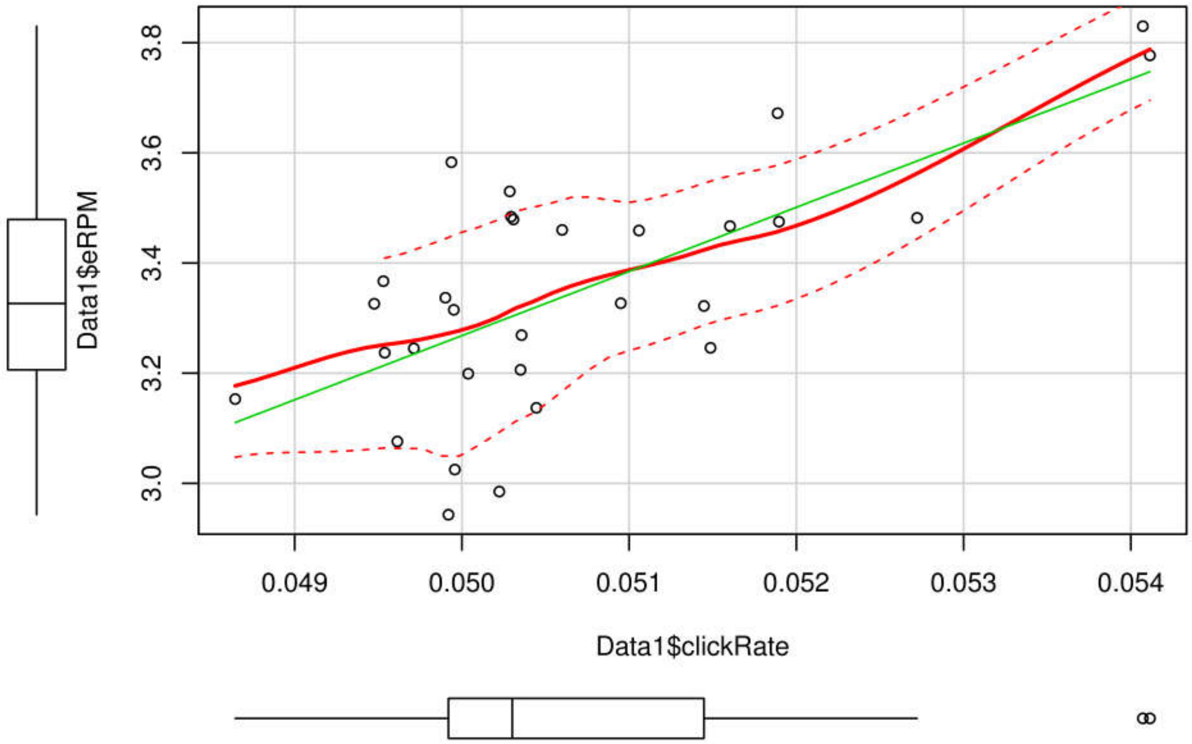
From the output above we see that conversion rate explains only 30% of the variance in data thus we conclude that the relationship in the difference between conversion rates does not really explain difference in eRPMs.

From the plots below we observe the samples and see that algorithm B has many more outliers then algorithm A thus skewing our averages.

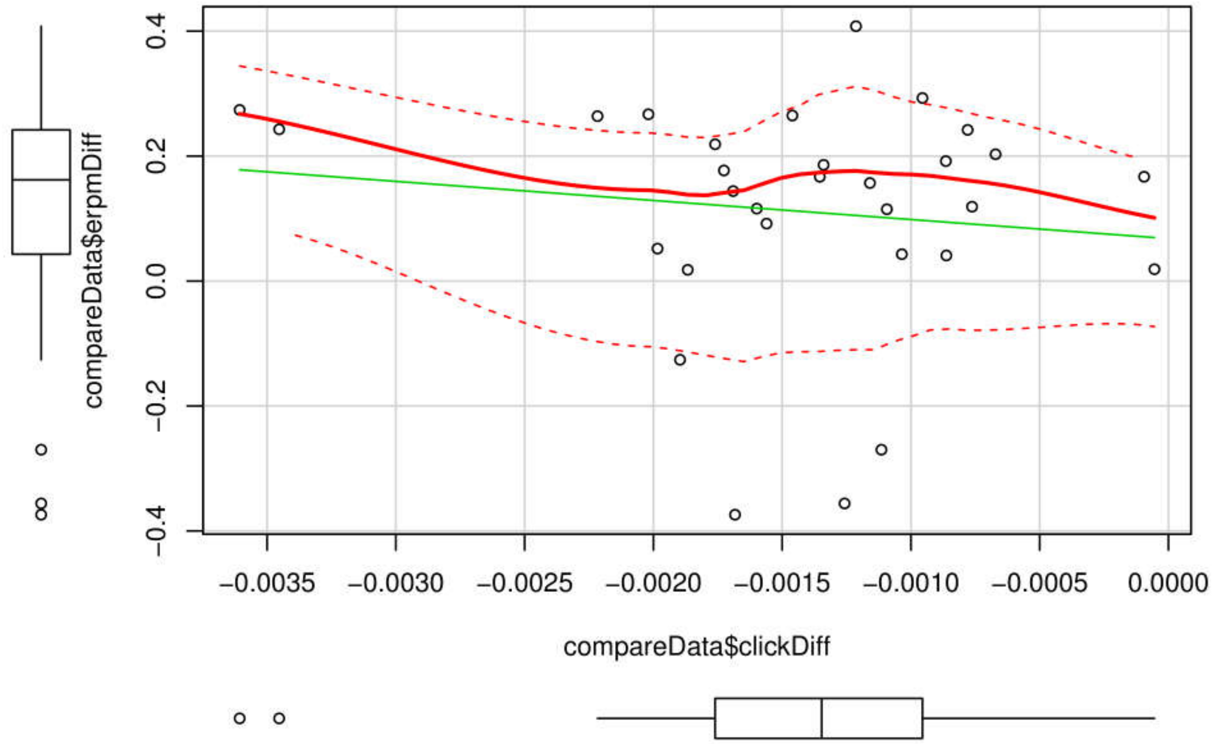
```r
scatterplot(Data2$eRPM~Data2$clickRate|Data2$clickRate, data=Data2)
```




```
scatterplot(Data1$eRPM~Data1$clickRate|Data1$clickRate, data=Data1)
```




```
scatterplot(compareData$erpmDiff~compareData$clickDiff|compareData$clickDiff, data=compareData)
```



Here we observe an odd relationship where with less clicks we have a higher eRPM.

In conclusion, we determine that we cannot conclude if algorithm B is better than A due to the sample that is not representative. Even though we show that difference in eRPMs is significant at 99% confidence level we would need more data from the manager to make a certain conclusion. Ir is recommended to repeat the test with a larger portion of customers served by B to eliminate such high variation in data points as well as recording the price of served ads because B might be really good at serving high paying ads to customers thus earning better margin despite lower ratios.