airbnb_data_challenge_solution_Ashim.R

ashimd

Sun Dec 13 23:29:40 2015

```
##Aribnb data challenge- Measure the success of product change

## Author- Ashim Datta
## email- datta.ashim2@gmail.com

## Summary:
## The test results are positive. The lift in treatment group is ~1.7x of the lift
in control. This shows that using 140 characters message is infact working.
## This test is a valid test since the treatment group and control have similar acc
eptance rate. Less than 10% variation
## However it is still not conclusive if this lift is signifcant. I would want to r
epeat this experiment for a few other groups and conclude if this feature is really
 useful


### See below for detail solution, scripts, graphs and discussion

## Extracting Data
setwd("/Users/ashimd/Documents/reashimlizairbnbdataanalyticsroles")
raw_users<-read.csv("takehome_assignments.csv")
raw_booking<- read.csv("takehome_contacts.csv")


library(sqldf)
```

```
## Loading required package: gsubfn
## Loading required package: proto
```

```
## Warning in doTryCatch(return(expr), name, parentenv, handler): unable to load sh
ared object '/Library/Frameworks/R.framework/Resources/modules//R_X11.so':
##   dlopen(/Library/Frameworks/R.framework/Resources/modules//R_X11.so, 6): Librar
y not loaded: /opt/X11/lib/libSM.6.dylib
##   Referenced from: /Library/Frameworks/R.framework/Resources/modules//R_X11.so
##   Reason: image not found
```

```
## Could not load tcltk.  Will use slower R code instead.
## Loading required package: RSQLite
## Loading required package: DBI
```

```
#Joining the 2 dataset to determine the group each guest belongs to

raw_all_data<-sqldf("select b.ab,a.* from raw_booking a join raw_users b
                    on a.id_guest=b.id_user")
# The above resulted in more rows than the booking data. There are probably duplica
te records for users. One single user is categorised both as test and control. We n
eed to make sure that each user has only one criteria
# Assigning users to the treatment group if they are both in treatment and control
raw_users_deduped<-sqldf("select distinct id_user,max(ab) as ab from raw_users grou
p by id_user")

raw_all_data_deduped<-sqldf("select b.ab,a.* from raw_booking a join raw_users_dedu
ped b
                    on a.id_guest=b.id_user")

# Since we are interested in finding the impact of using a message on the acceptanc
e rate, there is no point looking at instant book as they are automatically accepte
d by the host
# Subsetting data to get only books where the contact method was contact_me or book
_it

bookings_data<-raw_all_data_deduped[which(raw_all_data_deduped$dim_contact_channel=
=c('contact_me','book_it')),]

# creating 2 data sets- for treatment and control groups

booking_data_treatment<-bookings_data[which(bookings_data$ab=='treatment'),]
booking_data_control<-bookings_data[which(bookings_data$ab=='control'),]

#Overall percent of users who got accepted for treatment

booking_data_treatment_acceptance<-nrow(booking_data_treatment[which(booking_data_t
reatment$ts_booking_at!='NULL'),])/nrow(booking_data_treatment)

#Overall percent of users who got accepted for control

booking_data_control_acceptance<-nrow(booking_data_control[which(booking_data_contr
ol$ts_booking_at!='NULL'),])/nrow(booking_data_control)

#percent lift between the treatment and control group

lift<-(booking_data_treatment_acceptance-booking_data_control_acceptance)/booking_d
ata_control_acceptance
print(lift)
```

```
## [1] -0.07349893
```

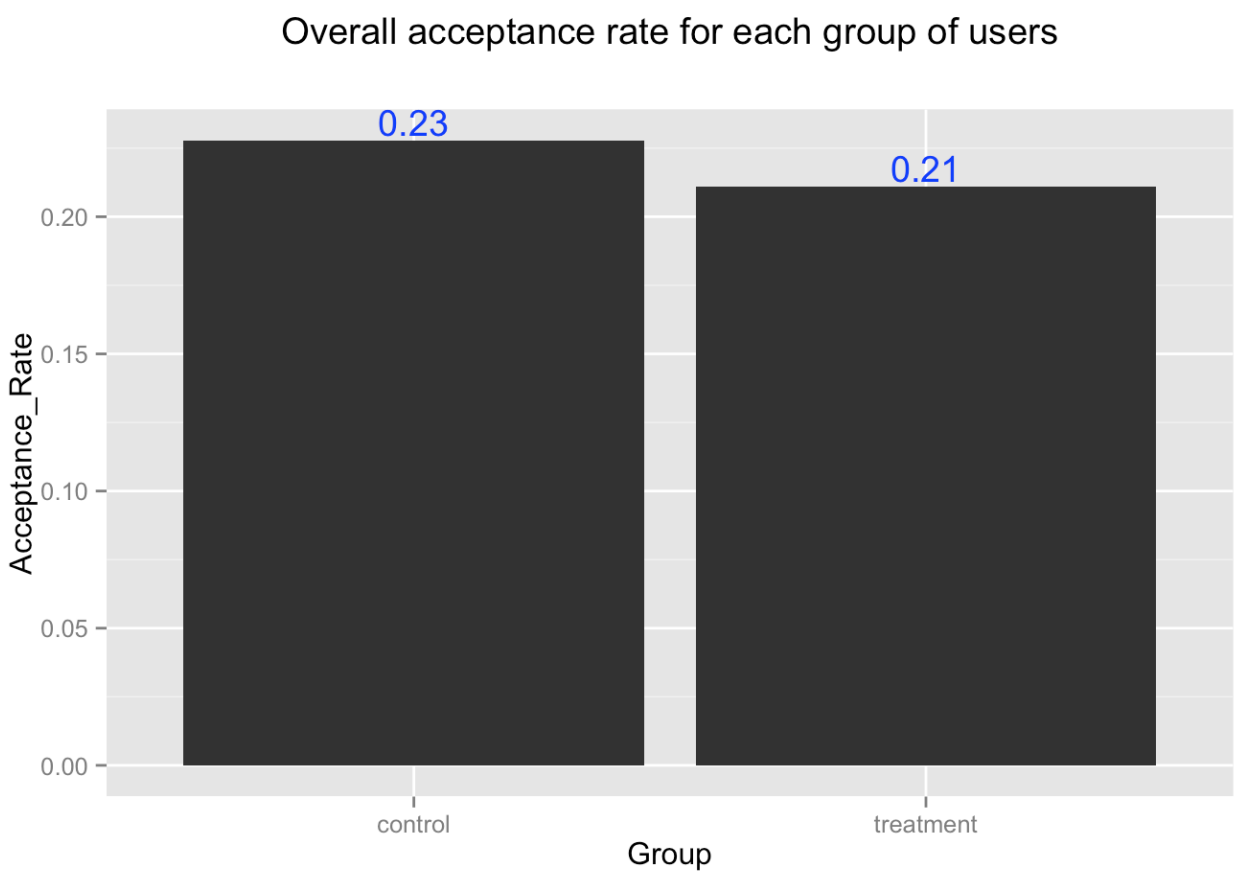
```
#This shows that the 2 groups are very similar. They have less than 10 percent diff
erence in their acceptance rates

#Plot acceptance rate:
#acceptance matrix

acceptance<-data.frame(c('treatment','control'),c(booking_data_treatment_acceptance
,booking_data_control_acceptance))

names(acceptance)<-c("ab_group","acceptance_rate")

library(ggplot2)
ggplot(acceptance, aes(x=ab_group, y=acceptance_rate,label=round(acceptance_rate,2)
)) +
  geom_bar(position=position_dodge(),stat="identity") +geom_text(vjust=-.2,size=5,
colour="blue")+ xlab("Group") +
  ylab("Acceptance_Rate") +
  ggtitle(expression(atop("Overall acceptance rate for each group of users")))
```



```
# Now let us look at lift in acceptance in both treatment and control for people wh
o wrote more than 140 characters vs the ones who did not in each group

#Percent of users who got accepted from treatment by writing over 140 charcters

booking_data_treatment_acceptance_140<-nrow(booking_data_treatment[which(booking_da
ta_treatment$ts_booking_at!='NULL'&
                                                                    as.numeric(booking_data_
treatment$m_first_message_length)>=140),])/
  nrow(booking_data_treatment)

#Percent of users who got accepted from treament by writing less than 140 character
s

booking_data_treatment_acceptance_n140<-nrow(booking_data_treatment[which(booking_d
ata_treatment$ts_booking_at!='NULL'&
                                                                          as.numer
ic(booking_data_treatment$m_first_message_length)<140),])/
  nrow(booking_data_treatment)

#lift in treatment

lift_treatment<-(booking_data_treatment_acceptance_140-booking_data_treatment_accep
tance_n140)/booking_data_treatment_acceptance_n140

#Control
#Percent of users who got accepted from control by writing more than 140 characters

booking_data_control_acceptance_140<-nrow(booking_data_control[which(booking_data_c
ontrol$ts_booking_at!='NULL'&
                                                                      as.numeric
(booking_data_control$m_first_message_length)>=140),])/
  nrow(booking_data_control)

#Percent of users who got accepted from control by writing less than 140 characters

booking_data_control_acceptance_n140<-nrow(booking_data_control[which(booking_data_
control$ts_booking_at!='NULL'&
                                                                       as.nume
ric(booking_data_control$m_first_message_length)<140),])/
  nrow(booking_data_control)

#lift in control

lift_control<-(booking_data_control_acceptance_140-booking_data_control_acceptance_
n140)/booking_data_control_acceptance_n140
```

```
#Comparing the lifts in 2 groups

lift_overall<-(lift_treatment-lift_control)/lift_control
print(lift_overall)
```

```
## [1] 1.672705
```

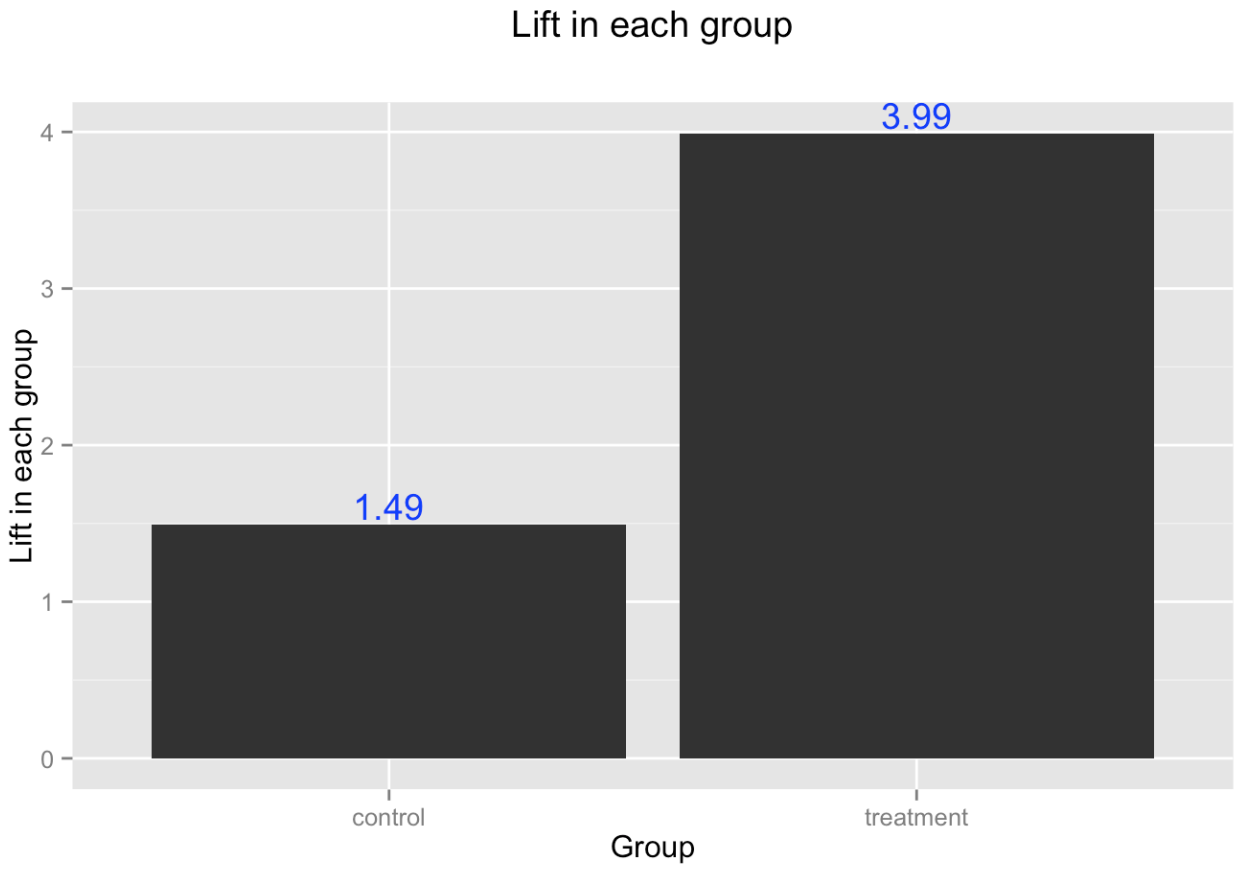
```
#The lift in treatment group is  ~1.7x of the lift in control. This shows that usin
g 140 characters message is infact working.

#Plot lift in control and treatment
#Lift matrix

lift_t_c<-data.frame(c('treatment','control'),c(lift_treatment,lift_control))

names(lift_t_c)<-c("ab_group","Lift")

library(ggplot2)
ggplot(lift_t_c, aes(x=ab_group, y=Lift,label=round(Lift,2))) +
  geom_bar(position=position_dodge(),stat="identity") +geom_text(vjust=-.2,size=5,
colour="blue")+ xlab("Group") +
  ylab("Lift in each group") +
  ggtitle(expression(atop("Lift in each group")))
```




```
#Power of the test

## Now the question is is this test significant.Airbnb has 50 million users and if
we assume each user books 2wice each year. Thus there will be 100million bookings p
er year
## Our dataset has 4984 valid observations. If we calculate the sample size with 99
% confidence level and 1 confidence interval for 100m observations then it would be
 16369. See link: http://www.surveysystem.com/sscalc.htm
## Hence I would ideally want to repeat this experiment for 4 more groups of contro
l and treatment
## Calculate the mean(expected difference between control and treatment) and sd in
lifts in control and treatment
## I will then calculate the power of the test which is probability of rejecting th
e null hypothesis(the mean difference in lift is acceptable) when it is actually tr
ue and decide on if this test is significant
```