HW3

Tamsir, Su, Chu, Hisham

2/19/2019

Question 1. Set Key.

```
trains = fread("train_subset.csv")
is.data.table(trains)
```

```
## [1] TRUE
```

```
trainsrand = subset(trains, random_bool == 1) #This is the random dataset
trainsnorm = subset(trains, random_bool == 0) #This is the ranking generated by Expedias algorithm.
uniqueN(trains[, .(srch_id, prop_id)] ) / trains[,.N]
```

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

```
#Set Key
#Question 1

setkey(trains,srch_id, prop_id)
setkey(trainsnorm,srch_id, prop_id)
setkey(trainsrand,srch_id, prop_id)
```

The key We found is the combination of srch_id, which is searching ID, and prop_id, which is the hotel ID. Each observations represent a consequent click on the search results of accommodations appearing on Expedia's websites. The srch_id is a index that records the search, and the search ID might occur more than once with different hotel ID since each search can come with multiple matched accommodations and visitors might click more than one of them.

Question 2a. Chunk 1.

```
trainsnorm[, prop_starrating := as.numeric(prop_starrating) ]
lm1 = lm(position~log(price_usd + 1) + prop_starrating, data=trainsnorm)

#Site_ID
lm2 = lm(position~log(price_usd + 1) + prop_starrating + site_id, data=trainsnorm)
#visitor_location_country_id

lm3 = lm(position~log(price_usd + 1) + prop_starrating + site_id + visitor_location_country_id, data=tra

#visitor_hist_starrating
trainsnorm[, visitor_hist_starrating := as.numeric(visitor_hist_starrating) ]
lm4 = lm(position~log(price_usd + 1) + prop_starrating + site_id + visitor_location_country_id + visito

#visitor_hist_adr_usd
trainsnorm[, visitor_hist_adr_usd := as.numeric(visitor_hist_adr_usd) ]
lm5 = lm(position~log(price_usd + 1) + prop_starrating + site_id + visitor_location_country_id + visito

#Remove site_id and visitor location id.
lm6 = lm(position~log(price_usd + 1) + prop_starrating + visitor_hist_starrating + visitor_hist_adr_usd

stargazer(lm1, lm2, lm3, lm4, lm5, lm6,
          title="Position", type="text",
```

```
          column.labels=c( "Random", "Not-Random"),
          df=FALSE, digits=2, star.cutoffs = c(0.05,0.01,0.001))
```

```
##
## Position
## ===========================================================================================
##                                                   Dependent variable:
##                                 -----------------------------------------------------------
##                                                         position
##                                 Random    Not-Random
##                                   (1)        (2)         (3)        (4)      (5)       (6)
## -------------------------------------------------------------------------------------------
## log(price_usd + 1)              0.24***    0.25***     0.24***    0.21***  0.21***   0.21***
##                                 (0.01)     (0.01)      (0.01)     (0.03)   (0.03)    (0.03)
##
## prop_starrating                -0.34***   -0.35***    -0.35***   -0.36*** -0.36*** -0.36***
##                                (0.004)    (0.004)     (0.004)    (0.02)   (0.02)    (0.02)
##
## site_id                                    0.01***     0.01***    0.002    0.001
##                                           (0.0005)    (0.0005)   (0.002)  (0.002)
##
## visitor_location_country_id                          -0.0003***  -0.0001  -0.0001
##                                                       (0.0001)   (0.0002) (0.0002)
##
## visitor_hist_starrating                                           0.12***  0.13***   0.14***
##                                                                   (0.02)   (0.02)    (0.02)
##
## visitor_hist_adr_usd                                                      -0.0002   -0.0002
##                                                                          (0.0001)  (0.0001)
##
## Constant                        5.41***    5.35***     5.41***    5.24***  5.20***   5.18***
##                                 (0.03)     (0.03)      (0.03)     (0.13)   (0.14)    (0.13)
##
## -------------------------------------------------------------------------------------------
## Observations                    745,709    745,709     745,709    47,791   47,773    47,773
## R2                               0.01       0.01        0.01       0.01     0.01      0.01
## Adjusted R2                      0.01       0.01        0.01       0.01     0.01      0.01
## Residual Std. Error              2.97       2.97        2.97       2.97     2.97      2.97
## F Statistic                   3,807.45*** 2,621.50*** 1,972.12*** 107.76*** 90.14*** 134.99***
## ===========================================================================================
## Note:                                                             *p<0.05; **p<0.01; ***p<0.001
```

Going forward, we will not include site_id, visitor_location_id, or visitor_hist_adr_usd due to the insignificant coefficients and conceptually.

Question 2a Chunk 2.

```
#Prop_country_Id
lm7 = lm(position~log(price_usd + 1) + prop_starrating + visitor_hist_starrating + prop_country_id, data

#Prop Review Score
trainsnorm[, prop_review_score := as.numeric(prop_review_score) ]
lm8 = lm(position~log(price_usd + 1) + prop_starrating + visitor_hist_starrating + prop_country_id + pro
#Prop Brand Bool
lm9 = lm(position~log(price_usd + 1) + prop_starrating + visitor_hist_starrating + prop_review_score + p
```

```r
#Prop Location Score 1
lm10 = lm(position~log(price_usd + 1) + prop_starrating + visitor_hist_starrating + prop_review_score +
#Prop Location Score 2
trainsnorm[, prop_location_score2 := as.numeric(prop_location_score2) ]
lm11 = lm(position~log(price_usd + 1) + prop_starrating + visitor_hist_starrating + prop_review_score +
#promotion_flag
lm12 = lm(position~log(price_usd + 1) + prop_starrating + visitor_hist_starrating + prop_review_score +

#srch_destination_id, INSIGNIFICANT
lm13 = lm(position~log(price_usd + 1) + prop_starrating + visitor_hist_starrating + prop_review_score +
lm14 = lm(position~log(price_usd + 1) + prop_starrating + visitor_hist_starrating + prop_review_score +
lm15 = lm(position~log(price_usd + 1) + prop_starrating + visitor_hist_starrating + prop_review_score +
#srch_adults_count
lm16 = lm(position~log(price_usd + 1) + prop_starrating + visitor_hist_starrating + prop_review_score +

stargazer(lm7, lm8, lm9, lm10,  lm11, lm12, lm13, lm14, lm15, lm16,
          title="Position", type="text",
          column.labels=c( "Random", "Not-Random"),
          df=FALSE, digits=2, star.cutoffs = c(0.05,0.01,0.001))
```

```
##
## Position
## =====================================================================================================
##                                                              Dependent variable:
##                          ----------------------------------------------------------------------------
##                                                                    position
##                           Random   Not-Random
##                             (1)       (2)         (3)        (4)        (5)        (6)        (7)        (8)
## -----------------------------------------------------------------------------------------------------
## log(price_usd + 1)        0.21***   0.25***    0.25***    0.23***    0.22***    0.09**     0.09**     0.09
##                           (0.03)    (0.03)     (0.03)     (0.03)     (0.03)     (0.03)     (0.03)     (0.0
##
## prop_starrating          -0.37***  -0.33***   -0.33***   -0.34***   -0.39***   -0.32***   -0.32***   -0.32
##                           (0.02)    (0.02)     (0.02)     (0.02)     (0.02)     (0.02)     (0.02)     (0.0
##
## visitor_hist_starrating   0.11***   0.10***    0.11***    0.10***    0.08***    0.09***    0.09***    0.09
##                           (0.02)    (0.02)     (0.02)     (0.02)     (0.02)     (0.02)     (0.02)     (0.0
##
## prop_country_id          -0.001*   -0.0003
##                          (0.0002)  (0.0002)
##
## prop_review_score                  -0.14***   -0.15***   -0.15***   -0.16***   -0.14***   -0.14***   -0.14
##                                     (0.02)     (0.02)     (0.02)     (0.02)     (0.02)     (0.02)     (0.0
##
## prop_brand_bool                                0.06*      0.08**     0.08*      0.01       0.01       0.0
##                                               (0.03)     (0.03)     (0.03)     (0.03)     (0.03)     (0.0
##
## prop_location_score1                                      0.05***    0.22***    0.28***    0.28***    0.27
##                                                           (0.01)     (0.01)     (0.01)     (0.01)     (0.0
##
## prop_location_score2                                                -2.74***   -2.84***   -2.84***   -2.83
##                                                                      (0.09)     (0.09)     (0.09)     (0.0
##
## promotion_flag                                                                 -0.85***   -0.85***   -0.86
```

```
##                                                                              (0.03)    (0.03)    (0.0
##
## srch_destination_id                                                                   0.0000
##                                                                                      (0.0000)
##
## srch_length_of_stay                                                                              0.0
##                                                                                                 (0.0
##
## srch_booking_window
##
##
## srch_adults_count
##
##
## Constant                    5.36***   5.57***    5.47***   5.55***   5.59***   6.06***   6.04***   6.06
##                             (0.13)     (0.14)    (0.13)    (0.13)    (0.14)    (0.14)    (0.14)    (0.1
##
## ------------------------------------------------------------------------------------------------------
## Observations                47,791     47,738    47,738    47,738    40,719    40,719    40,719    40,7
## R2                           0.01       0.01      0.01      0.01      0.04      0.05      0.05      0.0
## Adjusted R2                  0.01       0.01      0.01      0.01      0.04      0.05      0.05      0.0
## Residual Std. Error          2.97       2.96      2.96      2.96      2.93      2.91      2.91      2.9
## F Statistic                136.07*** 125.87***  126.35*** 109.94*** 226.77*** 281.71*** 250.50*** 251.0
## ======================================================================================================
## Note:                                                                                          *p<0.
```

When adding variables to this regression, we decided to exclude prop_country_id, prop_brand_bool, search_destination_id, and search_booking_window.

Question 2a Chunk 3

```r
#Question 2 Chunk 3
#srch_children_count
lm17 = lm(position~log(price_usd + 1) + prop_starrating + visitor_hist_starrating + prop_review_score +

#srch_room_count
lm18 = lm(position~log(price_usd + 1) + prop_starrating + visitor_hist_starrating + prop_review_score +

#srch_saturday_night_bool
lm19 = lm(position~log(price_usd + 1) + prop_starrating + visitor_hist_starrating + prop_review_score +

#srch_query_affinity_score
trainsnorm[, srch_query_affinity_score := as.numeric(srch_query_affinity_score) ]
lm20 = lm(position~log(price_usd + 1) + prop_starrating + visitor_hist_starrating + prop_review_score +

#orig_destination_distance
trainsnorm[, orig_destination_distance := as.numeric(orig_destination_distance) ]
lm21 = lm(position~log(price_usd + 1) + prop_starrating + visitor_hist_starrating + prop_review_score +

finalmodel = lm(position~log(price_usd + 1) + prop_starrating + visitor_hist_starrating + prop_review_s

stargazer(lm17, finalmodel, lm19, lm20, lm21,
          title="Position", type="text",
          column.labels=c( "lm17", "Final Model"),
          df=FALSE, digits=2, star.cutoffs = c(0.05,0.01,0.001))
```

```
##
## Position
## ============================================================================
##                                            Dependent variable:
##                           --------------------------------------------------
##                                                 position
##                             lm17    Final Model
##                              (1)        (2)        (3)      (4)       (5)
## ----------------------------------------------------------------------------
## log(price_usd + 1)         0.10**     0.09**     0.09**    0.17     0.11**
##                           (0.03)     (0.03)     (0.03)    (0.37)   (0.04)
##
## prop_starrating          -0.33***   -0.33***   -0.33***   -0.06   -0.37***
##                           (0.02)     (0.02)     (0.02)    (0.23)   (0.02)
##
## visitor_hist_starrating   0.09***    0.09***    0.09***   -0.14    0.11***
##                           (0.02)     (0.02)     (0.02)    (0.26)   (0.03)
##
## prop_review_score        -0.14***   -0.14***   -0.14***   -0.43   -0.11***
##                           (0.02)     (0.02)     (0.02)    (0.29)   (0.03)
##
## prop_location_score1      0.27***    0.28***    0.28***    0.26    0.31***
##                           (0.01)     (0.01)     (0.01)    (0.15)   (0.02)
##
## prop_location_score2     -2.89***   -2.95***   -2.95***  -3.92*** -3.17***
##                           (0.09)     (0.09)     (0.09)    (0.90)   (0.11)
##
## promotion_flag           -0.86***   -0.87***   -0.87***   -0.80*  -0.97***
##                           (0.03)     (0.03)     (0.03)    (0.39)   (0.04)
##
## srch_length_of_stay        0.02*      0.02*      0.02*     0.01     0.03**
##                           (0.01)     (0.01)     (0.01)    (0.10)   (0.01)
##
## srch_adults_count         0.04**     0.09***    0.09***    0.12    0.07***
##                           (0.02)     (0.02)     (0.02)    (0.18)   (0.02)
##
## srch_children_count      -0.17***   -0.15***   -0.15***   -0.13   -0.14***
##                           (0.02)     (0.02)     (0.02)    (0.25)   (0.03)
##
## srch_room_count                     -0.18***   -0.17***   -0.33   -0.17***
##                                      (0.03)     (0.03)    (0.41)   (0.03)
##
## srch_saturday_night_bool                        0.03
##                                                 (0.03)
##
## srch_query_affinity_score                                  0.01
##                                                           (0.01)
##
## orig_destination_distance                                          0.0000
##                                                                   (0.0000)
##
## Constant                  5.99***    6.16***    6.15***  7.20***   5.96***
##                           (0.14)     (0.14)     (0.14)    (1.73)   (0.19)
##
```

```
## ------------------------------------------------------------------------
## Observations               40,719     40,719     40,719     370     24,545
## R2                          0.05       0.05       0.05      0.08     0.06
## Adjusted R2                 0.05       0.05       0.05      0.05     0.06
## Residual Std. Error         2.90       2.90       2.90      2.95     2.89
## F Statistic               231.60***  215.23***  197.36***  2.64**  136.51***
## ========================================================================
## Note:                                        *p<0.05; **p<0.01; ***p<0.001
```

Question 2b. Comparison.

```
finalmodel = lm(position~log(price_usd + 1) + prop_starrating + visitor_hist_starrating + prop_review_s

#Need to conver all to numeric for the randomly selected dataset since I subsetted the data beforehand.
trainsrand[, prop_starrating := as.numeric(prop_starrating) ]
trainsrand[, visitor_hist_starrating := as.numeric(visitor_hist_starrating) ]
trainsrand[, visitor_hist_adr_usd := as.numeric(visitor_hist_adr_usd) ]
trainsrand[, prop_review_score := as.numeric(prop_review_score) ]
trainsrand[, prop_location_score2 := as.numeric(prop_location_score2) ]
trainsrand[, orig_destination_distance := as.numeric(orig_destination_distance) ]


finalmodelrandom = lm(position~log(price_usd + 1) + prop_starrating + visitor_hist_starrating + prop_re

stargazer(finalmodel, finalmodelrandom,
          title="Position", type="text",
          column.labels=c( "Final Model Expedia", "Final Model Randomly Selected"),
          df=FALSE, digits=2, star.cutoffs = c(0.05,0.01,0.001))
```

```
##
## Position
## ========================================================================
##                                      Dependent variable:
##                          ---------------------------------------------
##                                            position
##                          Final Model Expedia Final Model Randomly Selected
##                                  (1)                       (2)
## ------------------------------------------------------------------------
## log(price_usd + 1)              0.09**                    0.02
##                                 (0.03)                   (0.06)
##
## prop_starrating                -0.33***                  -0.14***
##                                 (0.02)                   (0.04)
##
## visitor_hist_starrating         0.09***                   -0.01
##                                 (0.02)                   (0.05)
##
## prop_review_score              -0.14***                   0.02
##                                 (0.02)                   (0.04)
##
## prop_location_score1            0.28***                  0.20***
##                                 (0.01)                   (0.02)
##
## prop_location_score2           -2.95***                  -1.82***
##                                 (0.09)                   (0.19)
```

```
##
## promotion_flag               -0.87***                  -0.001
##                               (0.03)                   (0.09)
##
## srch_length_of_stay           0.02*                    -0.01
##                               (0.01)                   (0.02)
##
## srch_adults_count             0.09***                   0.08*
##                               (0.02)                   (0.04)
##
## srch_children_count          -0.15***                  -0.12*
##                               (0.02)                   (0.05)
##
## srch_room_count              -0.18***                  -0.19**
##                               (0.03)                   (0.06)
##
## Constant                      6.16***                  5.41***
##                               (0.14)                   (0.30)
##
## ------------------------------------------------------------------------
## Observations                  40,719                    7,641
## R2                             0.05                     0.02
## Adjusted R2                    0.05                     0.02
## Residual Std. Error            2.90                     2.96
## F Statistic                  215.23***                12.26***
## ========================================================================
## Note:                                        *p<0.05; **p<0.01; ***p<0.001
```

Question 2c. Randomly Generated Position.

```
trainsnorm[, rPosition := ceiling(10*runif(.N))] #Expedia ranked
trainsrand[, rPosition := ceiling(10*runif(.N))] #Randomly generated

finalmodelz = lm(rPosition~log(price_usd + 1) + prop_starrating + visitor_hist_starrating + prop_review

finalmodelrandomz = lm(rPosition~log(price_usd + 1) + prop_starrating + visitor_hist_starrating + prop_

stargazer(finalmodel, finalmodelrandom, finalmodelz, finalmodelrandomz,
          title="Position vs rPosition", type="text",
          column.labels=c( "Pos Expedia", "Pos Random", "rPos Expedia", "rPos Random"),
          df=FALSE, digits=2, star.cutoffs = c(0.05,0.01,0.001))
```

```
##
## Position vs rPosition
## ========================================================================
##                                      Dependent variable:
##                          ---------------------------------------------
##                                 position               rPosition
##                          Pos Expedia Pos Random rPos Expedia rPos Random
##                              (1)         (2)          (3)          (4)
## ------------------------------------------------------------------------
## log(price_usd + 1)          0.09**      0.02        -0.06        0.03
##                             (0.03)      (0.06)      (0.03)       (0.06)
##
## prop_starrating            -0.33***    -0.14***     -0.01        0.01
```

```
##                            (0.02)      (0.04)      (0.02)      (0.04)
##
## visitor_hist_starrating   0.09***     -0.01        0.01        -0.05
##                            (0.02)      (0.05)      (0.02)      (0.05)
##
## prop_review_score        -0.14***      0.02        0.01        -0.01
##                            (0.02)      (0.04)      (0.02)      (0.04)
##
## prop_location_score1      0.28***    0.20***      0.0005      0.0003
##                            (0.01)      (0.02)      (0.01)      (0.02)
##
## prop_location_score2     -2.95***    -1.82***      0.07        0.18
##                            (0.09)      (0.19)      (0.09)      (0.18)
##
## promotion_flag           -0.87***     -0.001       0.05       -0.21*
##                            (0.03)      (0.09)      (0.03)      (0.08)
##
## srch_length_of_stay        0.02*      -0.01        0.01       -0.002
##                            (0.01)      (0.02)      (0.01)      (0.01)
##
## srch_adults_count          0.09***     0.08*        0.01        -0.03
##                            (0.02)      (0.04)      (0.02)      (0.03)
##
## srch_children_count      -0.15***     -0.12*      -0.03        0.10*
##                            (0.02)      (0.05)      (0.02)      (0.05)
##
## srch_room_count          -0.18***    -0.19**      -0.01        -0.01
##                            (0.03)      (0.06)      (0.03)      (0.06)
##
## Constant                   6.16***    5.41***     5.74***      5.54***
##                            (0.14)      (0.30)      (0.14)      (0.30)
##
## ------------------------------------------------------------------------
## Observations               40,719      7,641       40,719       7,641
## R2                          0.05        0.02        0.0003       0.002
## Adjusted R2                 0.05        0.02        0.0001      0.0005
## Residual Std. Error         2.90        2.96         2.88        2.89
## F Statistic              215.23***   12.26***       1.21        1.34
## ========================================================================
## Note:                                        *p<0.05; **p<0.01; ***p<0.001
```

For the interpretation, we see that once we randomize the position ourselves (rPosition), our model becomes totally insignificant and so if Expedia truely randomized their data, the randomized subset will be all insignificant. When we use this model on the original randomized position, we see that there are some variables that are insignificant. This details that Expedia did not properly randomize their samples as well as they could have.

Question 3a.

```
randompos1 = subset(trainsrand, position == 1)
randompos10 = subset(trainsrand, position == 10) #Ask him about clickthrough rate

#Clickthrough Rate
clickthroughpos1 = mean(randompos1$click_bool) #.143
clickthroughpos10 = mean(randompos10$click_bool)
```

```r
clickthroughanswer= clickthroughpos1 - clickthroughpos10
#There is a 9.5 percent increase for clickthrough rate in position 1 versus postion 10.

#Booking Rate
bookingpos1 = mean(randompos1$booking_bool) #.01906
bookingpos10 = mean(randompos10$booking_bool)
bookinganswer = bookingpos1 - bookingpos10
#There is a 1.3% increase for bookings in position 1 versus position 10.
```

Question 3b. Yes, it is more beneficial to be at the top of the randomized list because we can still observe a positive increase in both click through and booking rate.

Question 4

```r
table <- data.frame("position"=NULL, "a"=NULL, "b"=NULL, "e"=NULL, "c"=NULL,"d"=NULL, "f"=NULL)
for (x in 1:10){
   table = rbind(table,data.frame("position" = x,"a" = mean(trains[position == x & random_bool == 0, cl
}

names(table) = c("position", "click rate: algorithm","click rate: random"," difference of click rate",
```

4A. From positions 1 to 4, the clickthrough rate of algorithms are significantly higher than the randomized dataset. From position 6 - 10, the clickthrough rate doesn't change too much and so we conclude that the algorithm is beneficial for the higher positions compared to lower positions.

4B. We observe similar effects for positions 1 to 4 where we see a huge difference in booking rate between algorithm and randomized. For positions 6-10, we see there is still positive difference but not as much as the top positions.

5. From our analysis, top ranking positions results in higher clickthrough and booking rate. But since Expedia's "randomized" dataset was not properly randomized, we cannot conclude the causal estimates of their algorithm since they don't have a proper control group. We assume that the Expedia algorithm is attempting to maximize bookings, but since the randomization is incorrect, the effectiveness of the algorithm is ambigous.