Samba TV - Data Challenge

Posted on January 8, 2017

Contents

Problem Description

The company XYZ ran A/B testing for its Spanish users by localizing the Spanish language based on the country the user was visiting from.

After running the experiment, the analytics team concluded that the localization did not help with the conversion rate and that the non-localized content still performed better.

We are asked to verify this and provide suitable solution/algorithm to avoid a false conclusin to if was indeed erroneous.

Exploring the data

I will be using R to analyze the pings data. The following packages are loaded :

- *data.table* : I prefer to use data.table over data.frame for performance efficiency.

- *ggplot2* : To visualize data
- *dplyr* : To perform data operations
- *rpart* : To perform regression using decision trees
- *rpart.plot* : To visualize the decision tree

```
> if (!require("pacman")) install.packages("pacman")
Loading required package: pacman
> pacman::p_load(data.table, ggplot2, dplyr, rpart, rpart.plot)
```

I import the data sets and store them in data.table format

```
> setwd("~/Projects/data-analysis/samba/")
> test_data <- read.csv("test_table.csv")
> test_table <- data.table(test_data)
>
> user_data <- read.csv("user_table.csv")
> user_table <- data.table(user_data)
>
> summary(test_table)
    user_id                     date              source            device      
 Min.   :      1   2015-11-30: 71025   Ads    :181877   Mobile:201756  
 1st Qu.: 249816   2015-12-01: 70991   Direct: 90834   Web   :251565  
 Median : 500019   2015-12-02: 70649   SEO    :180610                  
 Mean   : 499938   2015-12-03: 99493                                   
 3rd Qu.: 749522   2015-12-04:141163                                   
 Max.   :1000000                                                       

        browser          conversion            test       
 Android_App:155135   Min.   :0.00000   Min.   :0.0000  
 Chrome     :101929   1st Qu.:0.00000   1st Qu.:0.0000  
 FireFox    : 40766   Median :0.00000   Median :0.0000  
 IE         : 61715   Mean   :0.04958   Mean   :0.4764  
 Iphone_App : 46621   3rd Qu.:0.00000   3rd Qu.:1.0000  
 Opera      :  6090   Max.   :1.00000   Max.   :1.0000  
 Safari     : 41065                                     
>
> nrow(test_table)
[1] 453321
```

We have data for 5 days of activity from 30th November, 2015 to 4th of December 2015 containing about 450K rows of data.

I verify that each row has a unique user id and also make sure that there are no duplicated rows.

```
> nrow(test_table) ==  length(unique(test_table$user_id))
[1] TRUE
>
> sum(duplicated(test_table))
[1] 0
> sum(duplicated(user_table))
[1] 0
>
```

Next, I verify that we have information for all the users in the test_table.

```
> unknown_users <- filter(test_table, !(user_id %in% user_table$user_i
> nrow(unknown_users)
[1] 454
```

We don't have information for 454 users. Ideally, I would check with the data infrastructure/analytics team to get this information. For the purposes of this data challenge, I choose to ignore these users since the number is relatively small compared to the total number of users. I now merge the two tables for further analysis.

```
> translation_table <- merge(test_table, user_table, by = "user_id")
>
```

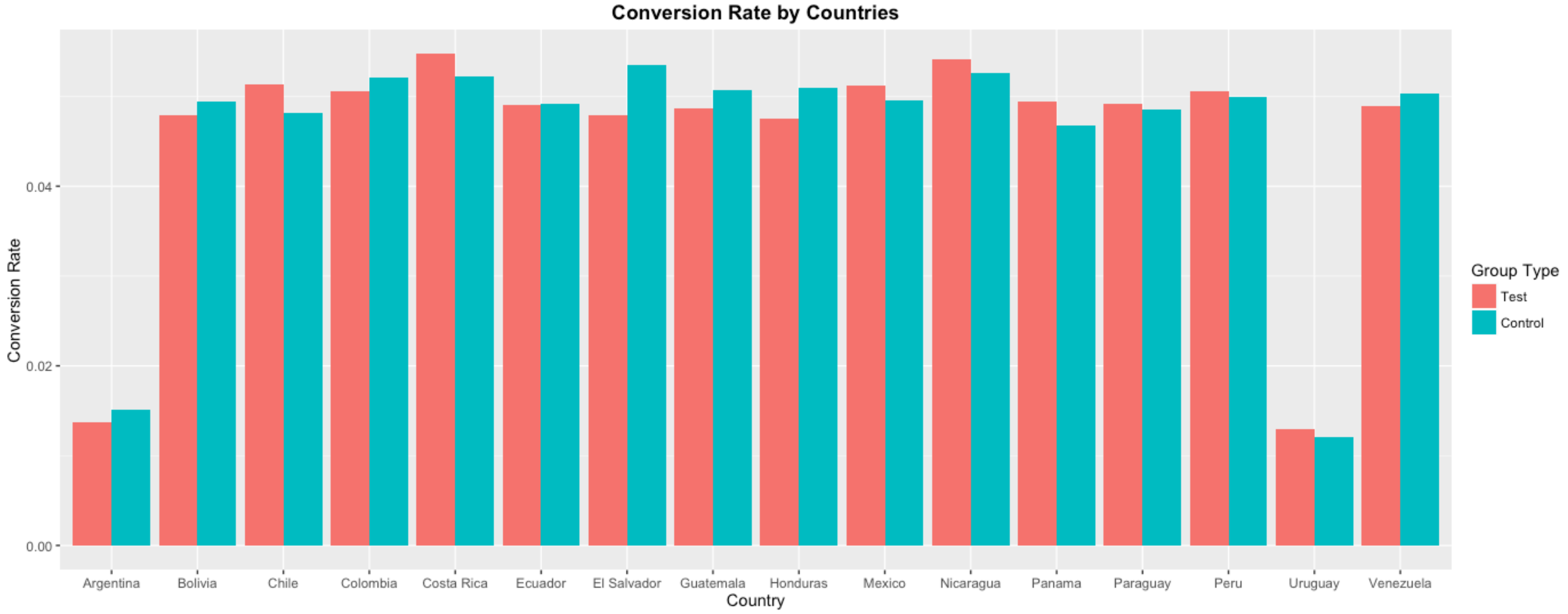
I verify that the test users are not from Spain since their translations remain the same. And then create separate data tables for test and control users.

```
> nrow(filter(translation_table, test==1, country=="Spain"))
[1] 0
>
> test_users_table <- filter(translation_table, test==1)
> control_users_table <- filter(translation_table, test==0)
> nrow(test_users_table)
[1] 215774
> nrow(control_users_table)
[1] 237093
>
```

The split percentage between test and control users is 47% to 53% which is almost even.

Now, I want to compare the conversion rate between the test and control users for each country. I will plot this information on a bar chart.

```
> test_users_grouped_by_country = test_users_table %>%
                                    group_by(country) %>%
                                    summarise(test_conversion_rate = m
>
> control_users_grouped_by_country = control_users_table %>%
                                        group_by(country)
                                        %>% summarise(control_conversi
>
> conversion_rate_by_country <- merge(test_users_grouped_by_country, c
>
> ggplot(melt(conversion_rate_by_country), aes(x=country, y=value)) +
Using country as id variables
```



The conversion rate is not always worse for test users over control users. In countries like Chile, Costa Rica, Nicaragua, Panama the conversion rate is slightly better for test users.

In the next section, I will evaluate if these differences are actually significant.

Analyzing the data

While there are differences in the conversion rate for each country, the sample sizes of users for each country is different.

```
> num_of_users_in_each_country <- translation_table %>% group_by(count
>
> num_of_users_in_each_country
# A tibble: 17 × 2
       country num_of_users
        <fctr>        <int>
1    Argentina        46733
2      Bolivia        11124
3        Chile        19737
4     Colombia        54060
5   Costa Rica         5309
6      Ecuador        15895
7  El Salvador         8175
8    Guatemala        15125
9     Honduras         8568
10      Mexico       128484
11   Nicaragua         6723
12      Panama         3951
13    Paraguay         7347
14        Peru        33666
15       Spain        51782
16     Uruguay         4134
17   Venezuela        32054
```

So a direct comparion of conversion rates is insufficient. I will use the t-test to evaluate if these differences are actually significant.

```
> translation_table_excluding_spain <- filter(translation_table, count
> t_test_results <- translation_table_excluding_spain %>% group_by(cou
> t_test_results
# A tibble: 16 × 4
       country   p_value test_conversion_rate control_conversion_rate
        <fctr>     <dbl>                <dbl>                   <dbl>
1       Mexico 0.1655437           0.05118631              0.04949462
2  El Salvador 0.2481267           0.04794689              0.05355404
3        Chile 0.3028476           0.05129502              0.04810718
4    Argentina 0.3351465           0.01372502              0.01507054
5     Colombia 0.4237191           0.05057096              0.05208949
6     Honduras 0.4714629           0.04753981              0.05090576
7    Guatemala 0.5721072           0.04864721              0.05064288
8    Venezuela 0.5737015           0.04897831              0.05034367
9   Costa Rica 0.6878764           0.05473764              0.05225564
10      Panama 0.7053268           0.04937028              0.04679552
11     Bolivia 0.7188852           0.04790097              0.04936937
12        Peru 0.7719530           0.05060427              0.04991404
13   Nicaragua 0.7804004           0.05417676              0.05264697
14     Uruguay 0.8797640           0.01290670              0.01204819
15    Paraguay 0.8836965           0.04922910              0.04849315
16     Ecuador 0.9615117           0.04898842              0.04915381
>
```

The p-values indicates that the differences are not significant to draw any conclusions.

If this was the case then why did the team conclude that the conversion rates were worse for localized translations? I will now compare the overall conversion rates for the test and the control group.

```
> test_vs_control_t_test_result <- t.test(translation_table_excluding_
>
> test_vs_control_t_test_result

        Welch Two Sample t-test

data:  translation_table_excluding_spain$conversion[translation_table_
t = -7.3539, df = 385260, p-value = 1.929e-13
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.006181421 -0.003579837
sample estimates:
 mean of x  mean of y
0.04341116 0.04829179


>
```

Here, the p-value is signigicant enough that we cannot ignore the differences in means. This is possibly why the team concluded that the conversion rate for the test group was worse (0.43 compared to 0.48 for control group). This indicates a bias in the selection process for the control group. Given that when we look at the data grouped by country we don't observe a sginificant difference , there is likely a bias in the selection by country.

I will now perform a logisitic regression to see the affect of country on the test group.

```
> glm.model <- glm(test~country,data = translation_table_excluding_spa
>
> glm.model

Call:  glm(formula = test ~ country, family = binomial, data = transla

Coefficients:
      (Intercept)      countryBolivia        countryChile     country
           1.3850             -1.3807             -1.3819
    countryEcuador  countryEl Salvador    countryGuatemala     country
          -1.4073             -1.3951             -1.4008
  countryNicaragua       countryPanama     countryParaguay         cou
          -1.4193             -1.3754             -1.3722
  countryVenezuela
          -1.4003

Degrees of Freedom: 401084 Total (i.e. Null);  401069 Residual
Null Deviance:      553700
Residual Deviance: 535000      AIC: 535000
>
```

I notice that Uruguay has a positive coefficient where as all the other countries have negative coefficients. Argentina is used as the reference level.

I will change the reference level to another country to get an estimate for Argentina.

```
> releveled_translation_table <- within(translation_table_excluding_sp
>
> glm.model <- glm(test~country,data = releveled_translation_table, fa
>
> glm.model

Call:  glm(formula = test ~ country, family = binomial, data = relevel

Coefficients:
      (Intercept)    countryArgentina      countryBolivia         coun
       -0.0041439           1.3891816           0.0084589            0
    countryEcuador  countryEl Salvador    countryGuatemala      country
       -0.0181282          -0.0058868          -0.0115919           -0
  countryNicaragua       countryPanama     countryParaguay          cou
       -0.0300703           0.0137618           0.0169384           -0
  countryVenezuela
       -0.0110807

Degrees of Freedom: 401084 Total (i.e. Null);  401069 Residual
Null Deviance:      553700
Residual Deviance: 535000      AIC: 535000
```

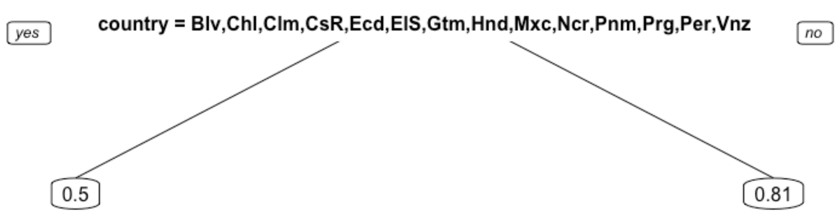
I notice that the coeffcients for Argentina and Uruguay are larger than compared to other countries. This possibly means that users from Argentina and Uruguay were more likely to be selected for the test group.

I will now construct a decision tree to verify this further.

```
> tree.model <- rpart(test~., translation_table_excluding_spain)
> tree.model
n= 401085

node), split, n, deviance, yval
      * denotes terminal node

1) root 401085 99692.820 0.5379757
  2) country=Bolivia,Chile,Colombia,Costa Rica,Ecuador,El Salvador,Gua
  3) country=Argentina,Uruguay 50867  7894.097 0.8079108 *
> prp(tree.model)
>
```



The decision tree verifies our earlier observation. Users from Argentina and Uruguay were 81% more likely to be selected for the test group. If the split between test and control group were truly random then our decision tree shouldn't show any splits. This decision tree cane be used to avoid making this bias again in future.

Conclusion

The data science incorrectly concluded that the non-localized translations weren't doing better. They need to run this experiment again with a truly random sample.

(https://github.com/sabman83) (mailto:saby83@gmail.com)
(https://linkedin.com/in/sebastin-kolman)