The Analysis of the effectiveness of Display Advertising for Star Digital

Chunwei Pan

2023-10-07

Online ads has become one of the most effective way of advertising. It provides a platform that is far better in its potential to measure ad effectiveness by enabling connection between ad impressions and conversions at the costomer level.

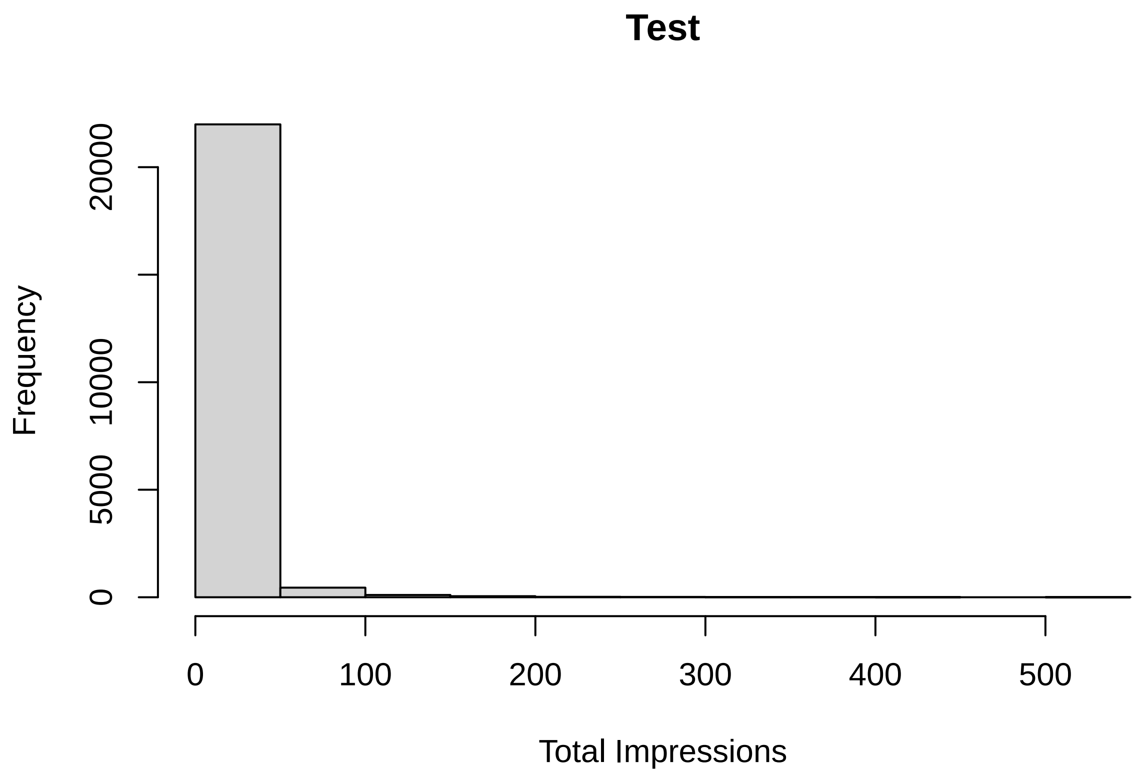
In this report, we are evaluating the result of Star Digital's online display advertising campaign and analyzing if it is effective or not.
In the experiment, Star Digital assigned 9/10 of customers into test group, who will be exposed to the ads for Star Digital, while the remaining 1/10 can only see ads for a charity organization. By doing the experiment, Star Digital wanted to understand if their ads actually help improving the conversion rate.

The data we use contains 25K of customers, with their group(control or test group), purchase or not, and the number of ad impressions they saw on different websites.

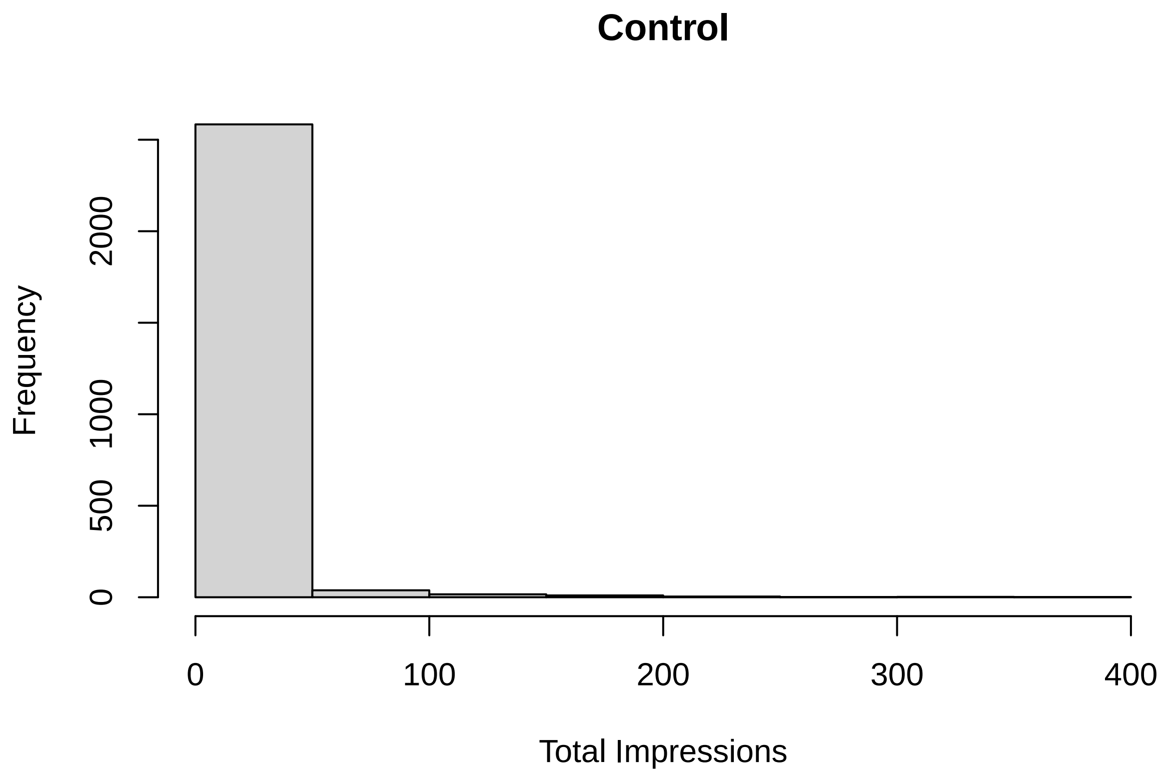
Randomization Check

First, before we implement any test on the data, we need to check if the randomization of the test is efficient. Before using t test, we first checked the distribution of total impressions for both groups.

```r
star_test <- star %>% filter(test == 1)
star_control <- star %>% filter(test == 0)
hist(star_test$total_imp,xlab = 'Total Impressions', main = "Test" )
```




```r
hist(star_control$total_imp,xlab = 'Total Impressions', main = "Control")
```



We can see that both groups have similar right skewed distribution.
Then we use t test to see if the test and control groups are similar.

```
t.test(total_imp ~ test, star)
```

```
##
##  Welch Two Sample t-test
##
## data:  total_imp by test
## t = 0.12734, df = 3204.4, p-value = 0.8987
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.8658621  0.9861407
## sample estimates:
## mean in group 0 mean in group 1
##        7.929217        7.869078
```

From the last line of the result, the control group saw 7.93 ad impressions in average, and the test group saw 7.87 ad impressions in average. Though there's a small different between the number, we can look at the p-value, which is 0.8987, and this number implies that the averages are not statistically different. Therefore, we can conclude that the randomization of the two groups is efficient.

Power Test

In the data, we got 22647 data of test group and 2656 data of control group. Thus we need to run a power test to check the minimum effect size that can be observed from this sample.

```r
power.t.test(delta = 0.1, sd = 1, sig.level = 0.05, power = 0.8,
             type = "two.sample", alternative = "two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 1570.737
##           delta = 0.1
##              sd = 1
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

By the result, the sample size is enough for the test.

Effectiveness

Then, we have the following question:
Is online advertising effective for Star Digital?
To answer this question, we do a t test to understand if the average number of purchases in the test group is higher than that in the control group.

```r
t.test(purchase ~ test, star)
```

```
##
##  Welch Two Sample t-test
##
## data:  purchase by test
## t = -1.8713, df = 3309.2, p-value = 0.06139
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.039289257  0.000916332
## sample estimates:
## mean in group 0 mean in group 1
##       0.4856928       0.5048792
```

From the result, we see that the mean purchase of the control group is 48.6% and the mean purchase of the test group is 50.5%, which is higher. Then we look into the p-value, which is a small number 0.06, though it is slightly higher than 5%, we can still conclude that the difference between the mean of the two groups is statistically significant.
Therefore, we can conclude that the ads are effective for Star Digital.
As for how effective, we can do a logistic regression. The following is the result of the model.

```r
summary(glm(purchase ~ test, star, family = "binomial"))
```

```
##
## Call:
## glm(formula = purchase ~ test, family = "binomial", data = star)
```

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.05724    0.03882  -1.474   0.1404
## test         0.07676    0.04104   1.871   0.0614 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 35077  on 25302  degrees of freedom
## Residual deviance: 35073  on 25301  degrees of freedom
## AIC: 35077
##
## Number of Fisher Scoring iterations: 3
```

From the result, we can also see that the coefficient of the test is positive and with a relatively small p-value(0.0614), we can conclude that the ads are effective.
If we look at the coefficient of test, it shows that seeing Star Digital ads increases the odds of 7.676% to purchase.

Frequency effect

Secondly, we are interested in the next question.
Is there a frequency effect of advertising on purchase? In particular, the question is whether increasing the frequency of advertising increases the probability of purchase?
To answer this question, we need to look into the relationship between the number of ads and purchase. We use the following logistic regression model to see the relationship.

```r
model_1 = glm(purchase ~ test + log(total_imp) + test*log(total_imp), data=star ,family="binomial")
```

In this model, we want to calculate the relationship of purchase and the total number of ads, along with the influence of being in the test group or not.

```r
summary(model_1)
```

```
##
## Call:
## glm(formula = purchase ~ test + log(total_imp) + test * log(total_imp),
##     family = "binomial", data = star)
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -0.448864   0.053301  -8.421   <2e-16 ***
## test                -0.008307   0.056453  -0.147    0.883
## log(total_imp)       0.385238   0.036110  10.668   <2e-16 ***
## test:log(total_imp)  0.073458   0.038278   1.919    0.055 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 35077  on 25302  degrees of freedom
## Residual deviance: 33485  on 25299  degrees of freedom
## AIC: 33493
##
## Number of Fisher Scoring iterations: 4
```

The coefficient of total impressions is 0.38, and that of the interaction between test and total_imp is 0.07, which means that the change in odds by adding the influence of being in the test group is about 7%, which made the sum of coefficient:
`0.38 + 0.07 = 0.45`
Because it's in log, so the actual number should be:
`e^(0.45) - 1 = 0.56 = 56%`
The number indicates that for customers being in the test group, every 1% more of real ads(ads for Star Digital) shown increase 56% of the probability of purchase.

Which website is more effective?

After knowing that these real ads do have positive influence on the decision of customers, we want to further understand if the budget is limited, which website is more effective, which leads to the following question: **Which sites should Star Digital advertise on? In particular, should it invest in Site 6 or Sites 1 through 5?**

First, we create 2 new variables showing the cost of each class.

```
star = star %>% mutate(cost_1to5 = imp_1to5 * (25/1000),
                       cost_6 = imp_6 *(20/1000))
```

Then, we build a logistic regression model that shows the relationship between purchase and the influence of each class(website 1 t0 5/ website 6) and the interaction with test.

```
model_2 = glm(purchase ~ test + log(cost_1to5 + 1e-3)+log(cost_6 + 1e-3)+test*log(cost_1to5 + 1e-3)
                 +test*log(cost_6 + 1e-3),family="binomial" ,data = star)
```

```
summary(model_2)
```

```
##
## Call:
## glm(formula = purchase ~ test + log(cost_1to5 + 0.001) + log(cost_6 +
##     0.001) + test * log(cost_1to5 + 0.001) + test * log(cost_6 +
##     0.001), family = "binomial", data = star)
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  0.99096    0.17745   5.584 2.34e-08 ***
## test                         0.41524    0.18798   2.209   0.0272 *
## log(cost_1to5 + 0.001)       0.20205    0.02016  10.023  < 2e-16 ***
## log(cost_6 + 0.001)          0.04842    0.02374   2.039   0.0414 *
## test:log(cost_1to5 + 0.001)  0.03756    0.02133   1.761   0.0783 .
## test:log(cost_6 + 0.001)     0.03680    0.02515   1.463   0.1434
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 35077  on 25302  degrees of freedom
## Residual deviance: 33619  on 25297  degrees of freedom
## AIC: 33631
##
## Number of Fisher Scoring iterations: 4
```

From the model result, we can see that:
1. For website 1-5, the sum of coefficient = 0.2 + 0.04 = 0.24
`e^(0.24) - 1 = 0.27 = 27%`
2. For website 6, the sum of coefficient = 0.048 + 0.037 = 0.085
`e^(0.085) - 1 = 0.09 = 9%`
From the calculation above, we found that investing in website 1 to 5 could increase the odds of purchases by more percentage than investing in website 6.
Thus investing more advertising on website 1 to 5 should be a better choice for Star Digital.