Star Digital online advertisements A/B testing experiment

Claire Tsao

2023-09-30

Background

Star Digital, a multimedia video service provider with an annual advertising budget exceeding $100 million, has been increasingly focusing on online advertising. To assess the effectiveness of their digital display advertising, Star Digital conducted a controlled experiment.

Business Goal

The primary business goal is to assess the effectiveness of online advertising and measure the causal impact of displaying Star Digital's advertisements on sales conversions. The specific objectives are if the ads will increase the number of package sales or not.

We conducted the A/B testing experiment to run logistic regression test on the Star Digital advertising campaign on six websites.

Experiment Design

To overcome challenges in measuring digital campaigns, Star Digital implemented a controlled experiment for their marketing campaign. The campaign was executed on six different websites, with the primary goal of boosting subscription sales. The experiment comprised two main treatment variations:

1.Customer Group Split:

Star Digital divided its customer base into treatment and control groups, with a 90%-10% split.

- *Treatment Group*: This group will be exposed to Star Digital's online advertisements.
- *Control Group*: This group will not be exposed to Star Digital's advertisements.(but charity organizations)

2.Key questions:

- Is online advertising effective for Star Digital?
- Does the frequency of advertising impact purchase behavior?
- Which websites should Star Digital prioritize for advertising, including whether to invest in Site 6 compared to Sites 1 through 5?

Analysis Performed

Read the data set into the file and find out the summary statistic of this data set - star

```
library(dplyr)
```

```
##
## 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
```

```
library(ggplot2)
library(readxl)
star <- read_excel("Star_Digital_Dataset.xls")
# sum the impression
star <- star %>% rowwise() %>% mutate(tot_imp = sum(imp_1,imp_2, imp_3, imp_4, imp_5, imp_6))
# sum the first five impression
star <- star %>% rowwise() %>% mutate(imp1to5 = sum(imp_1,imp_2, imp_3, imp_4, imp_5))

summary(star)
```

```
##        id             purchase           test           imp_1
##  Min.   :     27   Min.   :0.0000   Min.   :0.000   Min.   :  0.0000
##  1st Qu.: 353880   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:  0.0000
##  Median : 708344   Median :1.0000   Median :1.000   Median :  0.0000
##  Mean   : 708953   Mean   :0.5029   Mean   :0.895   Mean   :  0.9309
##  3rd Qu.:1062738   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:  0.0000
##  Max.   :1413367   Max.   :1.0000   Max.   :1.000   Max.   :296.0000
##      imp_2             imp_3               imp_4            imp_5
##  Min.   :  0.000   Min.   :  0.00000   Min.   :  0.00   Min.   : 0.00000
##  1st Qu.:  0.000   1st Qu.:  0.00000   1st Qu.:  0.00   1st Qu.: 0.00000
##  Median :  0.000   Median :  0.00000   Median :  0.00   Median : 0.00000
##  Mean   :  3.428   Mean   :  0.09477   Mean   :  1.59   Mean   : 0.04897
##  3rd Qu.:  2.000   3rd Qu.:  0.00000   3rd Qu.:  0.00   3rd Qu.: 0.00000
##  Max.   :373.000   Max.   :148.00000   Max.   :225.00   Max.   :51.00000
##      imp_6            tot_imp           imp1to5
##  Min.   :  0.000   Min.   :  1.000   Min.   :  0.000
##  1st Qu.:  0.000   1st Qu.:  1.000   1st Qu.:  0.000
##  Median :  1.000   Median :  2.000   Median :  1.000
##  Mean   :  1.784   Mean   :  7.875   Mean   :  6.092
##  3rd Qu.:  2.000   3rd Qu.:  6.000   3rd Qu.:  4.000
##  Max.   :404.000   Max.   :521.000   Max.   :397.000
```

Explore the star data

1. Find if there is any missing value

```r
missing_value <- colSums(is.na(star))
missing_value
```

```
##       id purchase     test    imp_1    imp_2    imp_3    imp_4    imp_5
##        0        0        0        0        0        0        0        0
##    imp_6  tot_imp  imp1to5
##        0        0        0
```

- Based on the output result, there is no missing value in this data set.

2. Randomization Check

```r
t.test(star$tot_imp ~ star$test)
```

```
##
##  Welch Two Sample t-test
##
## data:  star$tot_imp by star$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
```

- The results indicate the average impressions for both test and control are similar and they are not significantly different from each other.

3. Sample size determination:

Use "power.t.test" function helps business design experiments by calculating the minimum sample size. - We set delta equal to 0.2, meaning that we want to observe at least 20% increase in the likelihood of purchase or impression. - Set the H0: beta1 = 0 and test H0 at 5% level of significance that we allow 5% of type 1 error(Falsely reject hypothesis when it is true)

By using "power.t.test", we can calculate the minimum sample size to conclude the effective experiment is 393 in each groups(control and treatment).

```r
power.t.test(n= NULL, delta = 0.2, sd= 1, sig.level=0.05, power=0.8,
type = "two.sample", alternative = "two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 393.4067
##           delta = 0.2
##              sd = 1
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Concerns that need to be address

1. Selection Bias: Star Digital mitigated selection bias by randomly assigning treatment and control groups from the entire population and randomly selecting the dataset. This approach minimizes bias, and we confirmed it through randomization checks.
2. Omitted Variable Bias: The analysis only includes impression-related data, potentially omitting external variables correlated with impressions and purchase decisions.
3. Simultaneity Bias: There is no evidence to suggest that subscribing to the package causes an increase in ad views, eliminating simultaneity bias.
4. Measurement Error: We assume no measurement error, as tracking impressions at the user level is straightforward.

Q1: Is online advertising effective for Star Digital?

Goal: Who be exposed to Star Digital's advertisements would increase the number of package sales or not Use t.test function: a powerful tool used to compare two groups of data and determine if there is a significant difference between them.

```
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
```

The p-value helps assess the statistical significance of the observed difference. Here, the p-value(0.06139) is slightly greater than the significance level of 0.05, indicating that there were some difference between two groups, but the evidence is not very strong. Showing ads increases the probability of purchase of each customer by **3.95%**.[(0.504-0.485)/0.485*100]

```
model_1 = glm(purchase~test, data=star, family = "binomial")
summary(model_1)
```

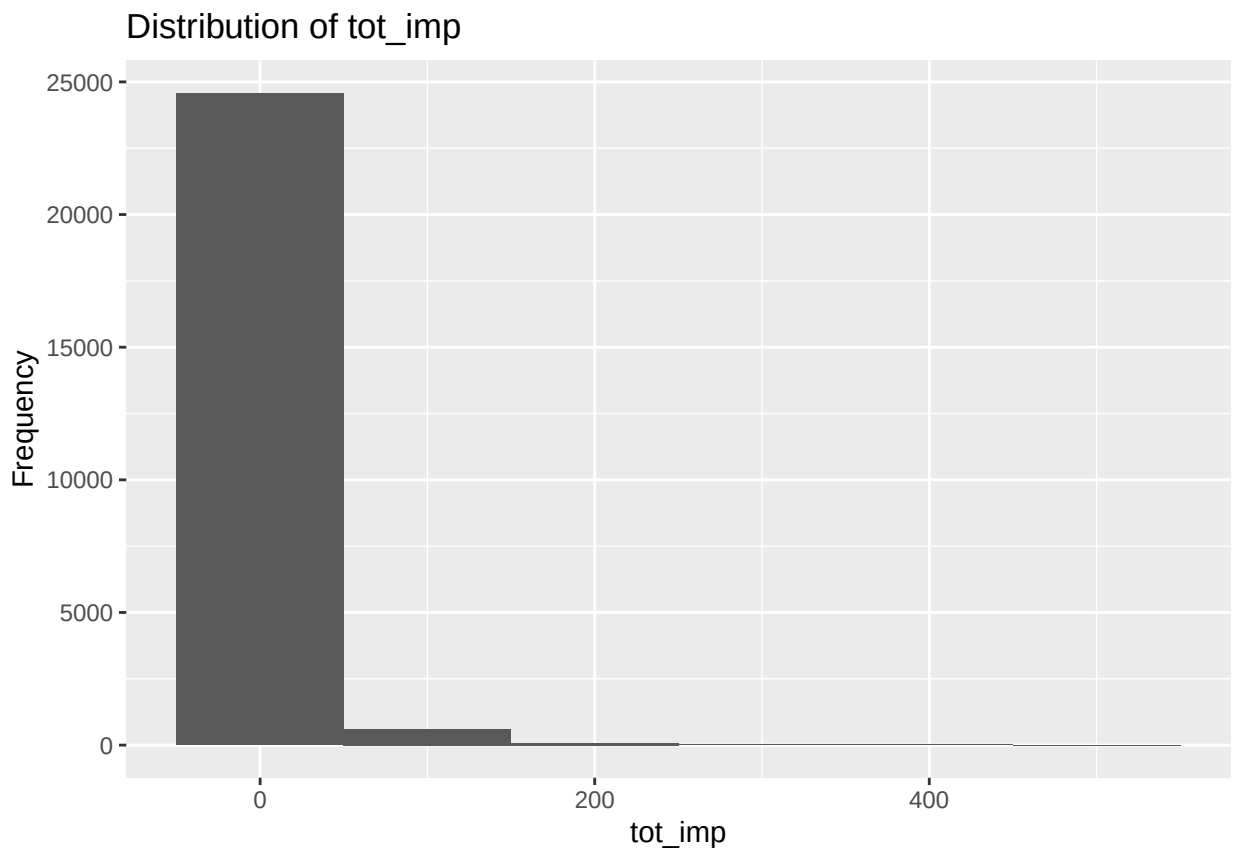
```
##
## Call:
## glm(formula = purchase ~ test, family = "binomial", data = star)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.05724    0.03882  -1.474   0.1404
## test         0.07676    0.04104   1.871   0.0614 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 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
```

In the regression, the coefficient for 'test' is 0.0767, meaning that being part of the test group would increase the odds of purchasing Star Digital by **7.98%**.

Q2: 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?

```
ggplot(star, aes(x = tot_imp)) +
  geom_histogram(binwidth = 100) +
  labs(title = "Distribution of tot_imp",
       x = "tot_imp",
       y = "Frequency")
```



Since the distribution of tot_imp is skewed, in the following logistic regression model we will use log transformation.

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

```
##
## Call:
## glm(formula = purchase ~ test * log(tot_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(tot_imp)       0.385238   0.036110  10.668   <2e-16 ***
## test:log(tot_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
```

When the frequency of Star Digital advertising (impressions) increases, there is a 7.35% increase in the odds of making a purchase for customers in the treatment group (those who saw the Star Digital's ads) compared to customers in the control group (those who did not see the ads). However, with p-value = 0.055, slightly above the typical 0.05 significant level, this suggests that the treatment effect may depend on the level of total impressions, but the evidence is not very strong.

Q3: Which sites should Star Digital advertise on? In particular, should it put its advertising dollars in Site 6 or in Sites 1 through 5?

- Background: The cost of advertising at Sites 1 through 5 is $25 per thousand impressions, and Sites 6 is $20. A purchase results in a lifetime contribution of $1,200 for Star Digital. Option 1 is to put ads dollars in Sites 1 through 5; Option 2 is to put ads dollars in Site 6.
- Build logistic regression model to understand the relationship between the "purchase" and the "imp" for treatment group.
- Next, we calculate the cost and revenue of two options and compared them to make the final decision.

```r
star$cost_1to5 <- star$imp1to5*(25/1000)
star$cost_6 <- star$imp_6*(20/1000)
model_3 = glm(purchase ~ test+log(cost_1to5+0.0001)+log(cost_6+0.0001)+test*log(cost_1to5+0.0001)
              +test*log(cost_6 +0.0001),family="binomial" ,data = star)
summary(model_3)
```

```
##
## Call:
## glm(formula = purchase ~ test + log(cost_1to5 + 1e-04) + log(cost_6 +
##     1e-04) + test * log(cost_1to5 + 1e-04) + test * log(cost_6 +
##     1e-04), family = "binomial", data = star)
```

```
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   0.75412    0.15232   4.951 7.38e-07 ***
## test                          0.35114    0.16124   2.178   0.0294 *
## log(cost_1to5 + 1e-04)        0.13185    0.01461   9.023  < 2e-16 ***
## log(cost_6 + 1e-04)           0.03025    0.01579   1.916   0.0554 .
## test:log(cost_1to5 + 1e-04)   0.02590    0.01545   1.676   0.0936 .
## test:log(cost_6 + 1e-04)      0.02376    0.01672   1.421   0.1553
## ---
## 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: 33835  on 25297  degrees of freedom
## AIC: 33847
##
## Number of Fisher Scoring iterations: 4
```

Option 1(Choose website 1 to 5):

- Cost: $25 per thousand impressions
- Every percentage increase in dollar invested: (0.13+0.025)*100% = 16% #### Option 2(Choose website 6):
- Cost: $20 per thousand impressions
- Every percentage increase in dollar invested: (0.03+0.024)*100% = 5%

We can see that investing in option 1(website 1 to 5) is more effective as it results in 16% compared to option 2(website 6) only increases by 5.4%. Hence, we would recommend to choose option 1: put its advertisting dollars in Sites 1 through 5.

Conclusion and Recommendation

- We definitely recommend Star Digit Company to do online advertising. From the experiment, we can tell that their ads indeed increase the sales of purchases.
- If the budget is not tight, we recommend to increase the frequency of ads in all websites, since there is a 7.35% increase in the odds of making a purchase based on our testing result.
- If the budget is tight, option 1,which is to invest in website 1 to website 5, is preferred compared to website 6.