A/B Testing of Marketing Strategies

Yumi Yu

Executive Summary

Star Digital, a video service provider was curious about whether their advertisement is really effective, so it designed an experiment for advertising through internet to test it.

Main Business Goals

There are three business goals I would like to discuss here: 1. Is online advertising effective for Star Digital? 2. Is there a frequency effect of advertising on purchase? In particular, whether increasing the frequency of advertising increases the probability of purchase? 3. Which sites should Star Digital advertise on? In particular, should it put its advertising dollars in Site 6 or in Sites 1 through 5?(the cost of advertising at Sites 1 through 5 is 25 dollar per thousand impressions, while the cost of advertising at Site 6 is 20 dollar per thousand impressions.)

Analysis Process

1. Explore the data set
2. Check the sample size of the data
3. Check whether there is missing data
4. Evaluate the data set randomization efficiency
5. Apply t.test and logistic regression model to analyze the business problems accordingly
6. Analyze the results and interpret the business values

Whole Picture of the Experiment

By considering the several aspects, including baseline conversion rate, campaign reach, the minimum lift that the advertiser cares to detect and power of the experiment. There are 2,656 control group and 22,647 treatment group. The duration of the experiment was 2 months in 2012.

Potential Concerns

There may be SUTVA problems, for example, one of the member of treatment group and one of the member of control group are family. They might share information with others.

Terminology to Know

1. "t-test" is a statistic method. In this case, we will use it to check if one feature can cause any difference to control group and treatment group.
2. "linear regression" is the model that we can learn about whether one feature have any effect to our purchase. We will interpret their effects by their coefficient. Positive means positive effect,negative indicates the opposite. And the absolute size of the coefficient means the size it cause.
3. "p-value" is a statistic results, we use p-value 0.05 as threshold to prove our assumption.

Load the data

```
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(readxl)
star_dig = read_xls('/Users/yumi/Library/CloudStorage/GoogleDrive-yu015524@umn.edu/My Drive/msba6441/HW/
```

Explore the dataset

```
# In this sample, there are 2,656 people in control group(10%) and
# 22,647 in treatment group(90%)
participant_counts <-star_dig %>% group_by(test) %>% count()
```

Check the sample size of the data

```
# detect a difference of 0.1 between the means of the two groups
power.t.test(delta = 0.1, sig.level = 0.1, power = 0.8, type=c('two.sample'), alternative=c('two.sided')
```

```
##
##      Two-sample t test power calculation
##
##               n = 1237.188
##           delta = 0.1
##              sd = 1
##       sig.level = 0.1
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

The result shows that the required sample size is around 1238. The data here is larger than minimum requirement, which means it is an overpowered study.

Check whether there is missing data

```
# There is no missing data in the dataset.
missing_data <- is.na(star_dig)

col_missing <- colSums(missing_data)

# We do not filter out the outliers data in each ad site.
# Higher ad impressions possess business value as well.
# We do not normalize data because the the unit of imp_1 to imp_6 are the same.
```

Evaluate the Randomization efficiency

```
star_dig <- star_dig %>%
  mutate(total_ad_impression = imp_1 + imp_2 + imp_3 + imp_4 + imp_5 + imp_5)

t.test(total_ad_impression ~ test, star_dig) # p-value = 0.9118
```

```
##
##  Welch Two Sample t-test
##
## data:  total_ad_impression by test
## t = -0.11074, df = 3270.2, p-value = 0.9118
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.8593623  0.7674812
## sample estimates:
## mean in group 0 mean in group 1
##        6.099774        6.145715
```

The data is efficiently randomized into control and treatment groups. Because the p-value is large than our threshold(5%), we cannot reject the assumption that there is no statistical difference in two groups.

Check whether online ad is effective to purchase

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

```
##
##  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
```

According to statistical result, it shows that the mean purchase of the control group is 0.4857 and the test group is 0.5049. There is roughly .2 difference in average number of purchase the people hot in the treatment group relative to the control group.

```r
# Apply logistic regression model to check the relationship
# between online ad exposure and purchase behavior
summary(glm(purchase ~ test, star_dig, family = binomial))
```

```
##
## Call:
## glm(formula = purchase ~ test, family = binomial, data = star_dig)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.186  -1.186   1.169   1.169   1.202
##
## 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
```

It shows that the coefficient of the testing group is 0.019. That is, the treatment group has 1.9% higher probability than control group on the purchase for one unit change in treatment group. Therefore, online ad is effective to purchase.

Is there a frequency effect of advertising on purchase?

```
# We first calculate the total frequency.
star_dig = star_dig %>% mutate(total_freq = imp_1+imp_2+imp_3+imp_4+imp_5+imp_6)

# Then, run the regression again.
summary(lm(purchase ~ test*total_freq, star_dig))
```

```
##
## Call:
## lm(formula = purchase ~ test * total_freq, data = star_dig)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.89562 -0.47994 -0.05711  0.51280  0.53228
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.4651265  0.0101335  45.900  < 2e-16 ***
## test             0.0111885  0.0107209   1.044   0.2967
## total_freq       0.0025937  0.0004131   6.278 3.49e-10 ***
## test:total_freq  0.0010362  0.0004408   2.351   0.0188 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4942 on 25299 degrees of freedom
## Multiple R-squared:  0.02317,    Adjusted R-squared:  0.02306
## F-statistic:   200 on 3 and 25299 DF,  p-value: < 2.2e-16
```

In this linear regression model, the coefficient of test:total_freq(interaction of test and total_freq) shows how likely customers who expose to real ads will increase their probability of purchase if the frequency of shown ads increases. The p-value is small enough so we would say that the interaction between ads shown frequency and purchase willingness is statistically significant. That is, showing more ads to customers will increases 0.10% probability in the purchase. Therefore, we conclude that there is a frequency effect of advertising on purchase.

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

```
# The effect of posing ads on Site 1 to Site 5 on customers.
star_dig = star_dig %>% mutate(freq_15 = imp_1+imp_2+imp_3+imp_4+imp_5)
summary(glm(purchase ~ test*freq_15, star_dig, family=binomial))
```

```
##
## Call:
## glm(formula = purchase ~ test * freq_15, family = binomial, data = star_dig)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -5.1561  -1.1306   0.1256   1.2207   1.2460
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.159804   0.041934  -3.811 0.000139 ***
## test          0.014395   0.044465   0.324 0.746140
## freq_15       0.019539   0.003441   5.679 1.35e-08 ***
## test:freq_15  0.014830   0.003790   3.913 9.11e-05 ***
## ---
## 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: 34202  on 25299  degrees of freedom
## AIC: 34210
##
## Number of Fisher Scoring iterations: 5
```

```
(exp(0.014830)-1)*100
```

```
## [1] 1.494051
```

The result of the logistic regression shows that customers who expose to real ads, when the company increases one unit of ads impression on Site 1 through Site 5, the probability of the purchase will increase 1.494%. The costs of advertising at these sites are all 25 dollar per thousand impressions. So to make customer who already expose to real ads, increases 1% of probability to purchase, the cost will be 0.016 dollar.

```
# The effect of posing ads on Site 6 on customers.
summary(glm(purchase ~ test*imp_6, star_dig, family=binomial))
```

```
##
## Call:
## glm(formula = purchase ~ test * imp_6, family = binomial, data = star_dig)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6440  -1.1791   0.8694   1.1854   1.2062
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.067534   0.039733  -1.700  0.08918 .
## test         0.048705   0.042218   1.154  0.24864
## imp_6        0.005684   0.004865   1.168  0.24275
## test:imp_6   0.017085   0.005845   2.923  0.00347 **
## ---
## 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: 35004  on 25299  degrees of freedom
## AIC: 35012
##
## Number of Fisher Scoring iterations: 4
```

```
(exp(0.017085)-1)*100
```

```
## [1] 1.723178
```

The result of the logistic regression shows that customers who expose to real ads, when the company increases one unit of ads impression on Site 6, the probability of the purchase will increase 1.723%. The costs of advertising at these sites are all 20 dollar per thousand impressions. So to make customer who already expose to real ads, increases 1% of probability to purchase, the cost will be 0.012 dollar.

Conclusion

Based on the cost to increase the probability of purchase by 1%, Star Digital should advertise on Site 6.