Digital Marketing Group Assignment 2

Team 9

2024-03-14

1. Exploratory Analysis (0.5 PT)

a. How many users are in test and control groups?

```
control <- df[df$test==0,]
test <- df[df$test==1,]

n_test <- nrow(test)
n_control <- nrow(control)

print(paste("There are", n_test, "users in test group."))
```

```
## [1] "There are 564577 users in test group."
```

```
print(paste("There are", n_control, "users in control group."))
```

```
## [1] "There are 23524 users in control group."
```

b. What is the conversion rate of each group?

```
# calculate total impressions and convererted for each groups
total_test_impr <- sum(test$tot_impr)
total_test_converted <- sum(test$converted)

total_control_impr <- sum(control$tot_impr)
total_control_converted <- sum(control$converted)

cvr_test <- round(total_test_converted / total_test_impr * 100, 3)
cvr_control <- round(total_control_converted / total_control_impr * 100, 3)

print(paste("The conversion rate of test group is ", cvr_test, "%.",
            sep = ""))
```

```
## [1] "The conversion rate of test group is 0.103%."
```

```
print(paste("The conversion rate of control group is ", cvr_control, "%.",
            sep = ""))
```

```
## [1] "The conversion rate of control group is 0.072%."
```

2. Randomization Checks (1 PT)

a. What is the total number of impressions for an average user it test versus control? Are they statistically different?

```r
imp_test <- t.test(control$tot_impr, test$tot_impr)
imp_test
```

```
##
##  Welch Two Sample t-test
##
## data:  control$tot_impr and test$tot_impr
## t = -0.218, df = 25608, p-value = 0.8274
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.6217286  0.4972735
## sample estimates:
## mean of x mean of y
##  24.76114  24.82337
```

The null hypothesis for this test is that there is no difference in the means of impressions between the control and test groups. The alternative hypothesis is that there is a difference.

Since the p-value (0.8274) is greater than the typical significance level of 0.05, we fail to reject the null hypothesis. This means that we do not have sufficient evidence to conclude that there is a statistically significant difference in the average number of impressions per user between the test and control groups. So they are not statistically different.

b. What do you conclude?

This minimal difference suggests that users in both groups received a similar number of impressions on average,which could indicate that the randomization process was effectively neutral in terms of exposure to impressions. Since we cannot identify significant differences between the test and control groups, we should continue to monitor the performance of both groups to determine if differences emerge over time.

3. Treatment Effect (1 PT)

a. Use a t-test to see if the conversion rate of the test group is higher than the control? What is the difference (this is usually called the "lift")?

```r
cvr_test <- t.test(test$converted, control$converted, alternative = "greater")
cvr_test
```

```
##
##  Welch Two Sample t-test
##
## data:  test$converted and control$converted
## t = 8.6572, df = 26384, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.006230842         Inf
## sample estimates:
##  mean of x  mean of y
## 0.02554656 0.01785411
```

```r
lift <- mean(test$converted) - mean(control$converted)
print(paste("Lift:", lift))
```

```
## [1] "Lift: 0.00769245319220152"
```

The p-value is smaller than the common significance level of 0.05, so there are enough evidences can support the hypothesis that the test group has a higher conversion rate than the control group with a high degree of confidence, which means we are able to reject null hypothesis.

b. Use a logistic regression to calculate the lift. Is your answer the same as in 3.a?

```r
# two models for test and control
model1 <- glm(formula = converted ~ tot_impr, data=test, family='binomial')
summary(model1)
```

```
##
## Call:
## glm(formula = converted ~ tot_impr, family = "binomial", data = test)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.0421580  0.0101991  -396.3   <2e-16 ***
## tot_impr     0.0101626  0.0001005   101.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 134260  on 564576  degrees of freedom
## Residual deviance: 124271  on 564575  degrees of freedom
## AIC: 124275
##
## Number of Fisher Scoring iterations: 6
```

```r
model2 <- glm(formula = converted ~ tot_impr, data=control, family='binomial')
summary(model2)
```

```
##
## Call:
## glm(formula = converted ~ tot_impr, family = "binomial", data = control)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.4034654  0.0590085  -74.62   <2e-16 ***
## tot_impr     0.0096955  0.0005093   19.04   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4213.9  on 23523  degrees of freedom
## Residual deviance: 3918.0  on 23522  degrees of freedom
## AIC: 3922
##
## Number of Fisher Scoring iterations: 7
```

```
# predict value
test_prob <- predict(model1, newdata=data.frame(tot_impr = test$tot_impr),
                     type="response")
control_prob <- predict(model2, newdata=data.frame(tot_impr = control$tot_impr),
                        type="response")

# calculate lift
log_lift <- mean(test_prob) - mean(control_prob)
print(paste("Lift based on logtistic regression:", log_lift))
```

```
## [1] "Lift based on logtistic regression: 0.00769245326361128"
```

Yes, the lift is the same as in pervious question.

4. Costs and ROI (1.5 PT)

a. How much more money did TaskaBella make by running the campaign (excluding advertising costs)?

```
profit_margin <- 40
rev_test <- round(lift * profit_margin * n_test, 3)

print(paste("TaskaBella made", rev_test, "dollars by running the campaign."))
```

```
## [1] "TaskaBella made 173719.286 dollars by running the campaign."
```

b. What was the cost of the ad campaign?

```
cpc <- 9/1000
cost_test <- round(cpc * sum(df$tot_impr), 3)

print(paste("TaskaBella spent", cost_test, "dollars by running the campaign."))
```

```
## [1] "TaskaBella spent 131374.638 dollars by running the campaign."
```

c. Calculate the ROI of the campaign?

```r
ROI <- round((rev_test - cost_test) / cost_test, 3)

print(paste("The ROI is ", ROI, ".", sep = ""))
```

```
## [1] "The ROI is 0.322."
```

d. What was the opportunity cost of including a control group? i.e., how much more could TaskaBella have made with a smaller control group or not having a control group at all?

```r
oppprtunity_cost <- round(lift * profit_margin * n_control, 3)

print(paste("TaskaBella would make",  oppprtunity_cost,
            "dollars if the company dropped the control group entirely."))
```

```
## [1] "TaskaBella would make 7238.291 dollars if the company dropped the control group entirely."
```