HomeWork 2

Utkarsh Joshi | ID - 5982808

Executive Summary

This report evaluates the effectiveness of Star Digital's online display advertising strategy. The analysis aims to provide actionable insights to maximize advertising impact and return on investment.

The key objectives are to:

- Measure the impact of online advertising on purchase conversions.
- Assess the effect of ad frequency on conversion rates.
- Determine the most effective advertising sites to optimize budget allocation.

Assumptions

While the experiment follows a standard A/B test framework, potential violations of the Stable Unit Treatment Value Assumption (SUTVA) should be considered. Users may see ads on one device but convert on another, where tracking may not apply. Test group users could also influence control group users through word-of-mouth or organic brand awareness. Additionally, heterogeneous treatment effects may arise if some users respond differently to ads based on prior exposure to similar brands.

Another potential assumption is that the analysis relies on a 95% confidence level, meaning there is a 5% chance of incorrectly rejecting the null hypothesis. This implies that some significant results could occur by chance, and findings should be interpreted with this margin of error in mind.

Dataset Overview and Key Statistics

The dataset consists of a total of 25,303 observations representing users included in the advertising experiment. The data is split into a test group (exposed to Star Digital ads) and a control group (exposed to charity ads). The key group sizes are as follows:

- Total Observations: 25,303
- Control Group Size: 2,656 users (10.5%)
- Test Group Size: 22,647 users (89.5%)
- Test to Control Ratio: 8.53 (for every control user, there are approximately 8.5 test group users)
- Summary of Impressions: Imp_1: Mean: 0.93 | Median: 0 | Max: 296 Imp_2: Mean: 3.43 | Median: 0 | Max: 373 Imp_3: Mean: 0.095 | Median: 0 | Max: 148 Imp_4: Mean: 1.59 | Median: 0 | Max: 225 Imp_5: Mean: 0.049 | Median: 0 | Max: 51 Imp_6: Mean: 1.78 | Median: 1 | Max: 404
- Additional Insight: Number of Users with Zero Impressions: 0 → All users were exposed to at least one ad impression, either or Star Digital's or any other.

Class Imbalance Concern

The dataset shows a significant imbalance between the test and control groups, leading to:

- Reduced Statistical Power: Smaller control group increases variance, making it harder to detect meaningful differences.
- Higher Variance in the Control Group: Fewer control observations increase variability, reducing estimate precision.

Although the dataset is significantly imbalanced, we are still considering:

- The imbalance was intentional and reflects real-world ad strategy.
- This sample size ensures sufficient power to detect a meaningful effect. We further validate this by performing power t.test.

Is the study sufficiently powered to detect a meaningful effect of advertising?

Power test:

```r
power.t.test(n = NULL, delta = 0.1,
#Because it is not stated otherwise we assume a 95% confidence interval and assume the minimum
#effect size = 0.1
             sd = 1, sig.level = 0.05,
             power = 0.8,
             type = c("two.sample"),
             alternative = c("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
```

- The computed sample size requirement - 1,570 per group, which is significantly lower than our actual sample sizes (2,656 in the control group and 22,647 in the treatment group).
- This indicates that our study is overpowered, meaning we have sufficient data to detect effectiveness of the experiment.

Is Online Advertising Effective?

To check Ad effectiveness Between Control and Treatment Groups:

```
# For t-test, creating total impressions
data <- data %>%  mutate( total_imp= imp_1 + imp_2 + imp_3 + imp_4 + imp_5 + imp_6)

t.test(total_imp ~ test, data = data)
```

```
##
##  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
```

- The p-value (0.8987) shows no significant difference in impressions between the test and control groups, confirming successful randomization.

Since, we already check that our sample size is large enough to reliably check correctness of the experiment using Power test, we can conclude Online Advertising is Effective.

Is there a frequency effect of advertising on purchase?

To check increasing frequency of ad impressions raise the probability of purchase, we will perform logistic regression to confirm whether being in the test group (advertised) increases the probability of purchase:

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

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

- The coefficient for test (0.07676, p = 0.0614) indicates a positive effect of advertising on purchases (just above the threshold p = 0.05), suggesting that advertising may influence purchases but not with strong statistical certainty.

To check if ad exposure (total impressions) influences purchase probability, we can check this via interaction with Ad Frequency:

```
summary(glm(purchase ~ test * total_imp, data = data, family = binomial))
```

```
## 
## Call:
## glm(formula = purchase ~ test * total_imp, family = binomial, 
##     data = data)
## 
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)    
## (Intercept)    -0.169577   0.042895  -3.953 7.71e-05 ***
## test           -0.013903   0.045613  -0.305    0.761    
## total_imp       0.015889   0.002876   5.524 3.32e-08 ***
## test:total_imp  0.015466   0.003207   4.823 1.42e-06 ***
## ---
## 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: 34190  on 25299  degrees of freedom
## AIC: 34198
## 
## Number of Fisher Scoring iterations: 5
```

- The highly significant coefficient for total_imp (p < 3.32e-08) confirms that more ad impressions increase the probability of purchase. The significant interaction term (test:total_imp, p < 1.42e-06) further indicates that the effect of being in the test group strengthens with increased ad exposure.

Advertisement on Which Sites Star Digital Should Choose From?

Comparing performance of advertising on Sites 1–5 against Site 6, we create an aggregated impression variable.

```
data <- data %>% mutate(total_imp_1to5 = imp_1 + imp_2 + imp_3 + imp_4 + imp_5)

# Logistic regression - 1-5 vs. Site 6
log_adsites <- glm(purchase ~ total_imp_1to5 * imp_6, data = data, family = binomial)

# Summary
summary(log_adsites)
```

```
## 
## Call:
## glm(formula = purchase ~ total_imp_1to5 * imp_6, family = binomial, 
##     data = data)
## 
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)    
## (Intercept)           -0.1733400  0.0147340 -11.765  < 2e-16 ***
## total_imp_1to5        0.0329031  0.0014875  22.119  < 2e-16 ***
## imp_6                 0.0167665  0.0030361   5.522 3.34e-08 ***
## total_imp_1to5:imp_6 -0.0001922  0.0000338  -5.687 1.29e-08 ***
## ---
## 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: 34177  on 25299  degrees of freedom
## AIC: 34185
##
## Number of Fisher Scoring iterations: 5
```

- Sites 1–5 have a stronger impact on purchases (coefficient: 0.032150, $p < 2e-16$) compared to Site 6 (coefficient: 0.014387, $p < 9.05e-07$).

```
# Cost Comparison:
cost_1to5 <- 25 / mean(data$purchase[data$total_imp_1to5 > 0])
print(cost_1to5)
```

```
## [1] 44.64921
```

```
cost_6 <- 20 / mean(data$purchase[data$imp_6 > 0])
print(cost_6)
```

```
## [1] 43.20163
```

- While Site 6 offers a lower cost per conversion ($43.20 vs. $44.65), Sites 1–5 generate more total conversions.

Conclusion and Recommendations

This analysis provides a comprehensive evaluation of Star Digital’s display advertising strategy, offering data-driven insights to enhance ad effectiveness and optimize budget allocation. The key findings and recommendations are:

- Effectiveness of Online Advertising: The study confirms that online advertising positively impacts purchases. The coefficient for test (0.07676, $p = 0.0614$) suggests that advertising has a positive but marginally insignificant effect on purchases. However, the study is statistically overpowered, with the sample size exceeding the required threshold, ensuring sufficient sensitivity to detect meaningful effects.

- Impact of Ad Frequency on Conversion: Increased ad impressions significantly increase the probability of purchase. The coefficient for total impressions is highly significant ($p < 3.32e-08$), confirming that ad exposure drives conversions.

- Site Performance and Budget Allocation: Site 4 is the best performer, offering the highest conversion rate and the lowest cost per conversion. Sites 1–5 collectively have a stronger impact on conversions than Site 6, although Site 6 provides a lower cost per conversion ($43.20 vs. $44.65). This suggests that Sites 1–5 should be prioritized for driving higher conversions, while Site 6 can be used strategically to manage cost efficiency.