Assignment 2

Avnee Satija - 5992424 and Harshal Sable - 5949697

2025-03-08

Star Digital: Assessing the Effectiveness of Display Advertising

Star Digital, a multichannel video service provider, conducted a controlled experiment across six websites to understand the impact of their advertisements on customer purchases. This report evaluates the effectiveness of online display advertising, analyzes whether increased ad exposure influences conversion rates, and provides insights on optimizing ad spend by comparing the performance of Site 6 against Sites 1-5. The report uses various statistical methods to help Star Digital's advertising strategy by identifying the most cost-effective and impactful approach to digital marketing.

Data Preprocessing

```
#loading the necessary libraries
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)
data <- read_excel ("~/Desktop/Semester 2/Casual Inference/Homework/hw2/M347SS-XLS-ENG.xls")
```

To assess the imbalance in group sizes, we first determine the number of users in the control and test groups. Given that 90% of users are in the test group and only 10% in the control group, we check the exact sample sizes for both groups before proceeding with data balancing. The following code extracts and counts the number of observations in each group.

```
#filtering control and treatment groups
control_population <- data %>% filter(test == 0)
test_population <- data %>% filter(test == 1)
```

```r
#population of control and treatment group
control_population_count <- nrow(control_population)
test_population_count <- nrow(test_population)

print(control_population_count)
```

```
## [1] 2656
```

```r
print(test_population_count)
```

```
## [1] 22647
```

The dataset is highly imbalanced, with 22,647 users in the test group (90%) and only 2,656 in the control group (10%), whereas an ideal experimental design would aim for a more balanced allocation. This imbalance could introduce bias, potentially inflating statistical significance and affecting the reliability of results.

To address this, we will balance the dataset by randomly selecting a sample of 2,656 users from the test group and combining it with all users in the control group. Only the sample size is adjusted; all other features remain unchanged across groups, ensuring comparability and preserving the assumptions of the experiment.

```r
#random sampling to balance the dataset
set.seed(42)
test_sample <- data %>% filter(test == 1) %>% sample_n(2656)
control_sample <- data %>% filter(test == 0)

balanced_data <- bind_rows(test_sample, control_sample)
```

Analyzing the effectiveness of advertisements

The controlled experiment was designed to ensure random assignment of users to the control and treatment groups, maintaining comparability between them. To confirm the effectiveness of this randomization and that both groups received similar exposure in terms of ad impressions, we perform a statistical test. Specifically, we conduct a t-test on total impressiosn (sum of impressions across Sites 1 to 6) to assess whether the distribution of ad exposures differs significantly between the groups. Ensuring balance in total impressions is critical to isolating the causal impact of advertising.

```r
#t-test on total_impressions and test
balanced_data <- balanced_data %>% mutate(total_impressions = imp_1 + imp_2 + imp_3 + imp_4 + imp_5 + i
t.test(total_impressions ~ test, data = balanced_data)
```

```
##
##  Welch Two Sample t-test
##
## data:  total_impressions by test
## t = 0.80331, df = 4938.2, p-value = 0.4218
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.6529791  1.5596056
## sample estimates:
## mean in group 0 mean in group 1
##        7.929217        7.475904
```

The t-test results yield a p-value of 0.4218, which exceeds the standard significance threshold of 0.05. Therefore, we fail to reject the null hypothesis, indicating that the mean total impressions do not significantly differ between the control and test groups. This confirms that the randomization process effectively balanced ad exposure between groups.

To further assess the robustness of our experimental design, we will determine the minimum sample size required for detecting a statistically significant difference if one exists. This will be done through a power analysis, ensuring that the study has sufficient sensitivity to detect meaningful effects.

```
#minimum sample reuired
power.t.test(n = NULL, delta = 0.1, sd = 1, sig.level = 0.05, power = 0.8, type = "two.sample", alterna
```

```
##
##      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 power analysis indicates that a minimum of 1,570 observations is required for the test to achieve statistical significance. Since both the control and test groups exceed this threshold, the sample size is sufficient for reliable statistical inference. Therefore, we can proceed with the analysis.

Next, we examine the effect of online advertising on purchases by estimating a logistic regression model, where purchase behavior serves as the dependent variable and test group assignment as the independent variable. This will allow us to quantify the causal impact of ad exposure on customer conversion rates.

```
#online ad and purchases effect
model_balanced <- glm(purchase ~ test, data = balanced_data, family = "binomial")
summary(model_balanced)
```

```
##
## Call:
## glm(formula = purchase ~ test, family = "binomial", data = balanced_data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.05724    0.03882  -1.474    0.140
## test         0.11449    0.05490   2.085    0.037 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7364.0  on 5311  degrees of freedom
## Residual deviance: 7359.6  on 5310  degrees of freedom
## AIC: 7363.6
##
## Number of Fisher Scoring iterations: 3
```

The estimated coefficient for the test variable in the logistic regression model is 0.11449. To interpret this in terms of odds, we exponentiate the coefficient:

```
#odds of purchase
exp(0.11449)
```

```
## [1] 1.121301
```

This result indicates that the odds of purchase for the test group are exp(0.11449) i.e.1.1213 times the odds of purchase for the control group. This translates to a 12.13% increase in the likelihood of purchase for users exposed to the online advertisements compared to those in the control group.

The p-value for the test variable is 0.037, which is below the conventional 0.05 significance threshold. This provides strong statistical evidence that online advertising has a significant impact on purchase behavior.

Based on these findings, we conclude that online advertising is effective for Star Digital, as exposure to display ads positively influences customer conversion rates.

Assessing Whether Increasing Ad Frequency Increases the Probability of Purchase

To evaluate the impact of advertising frequency on purchase probability, we will estimate a logistic regression model. In this model, purchase behavior serves as the dependent variable, while total impressions (sum of impressions across Sites 1 to 6) is the independent variable. This will allow us to determine whether higher ad exposure increases the likelihood of conversion.

```
#regression for total impressions and purchase
model2 <- glm(purchase ~ total_impressions, data = balanced_data, family = "binomial")
summary(model2)
```

```
##
## Call:
## glm(formula = purchase ~ total_impressions, family = "binomial",
##     data = balanced_data)
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -0.179440   0.031682  -5.664 1.48e-08 ***
## total_impressions  0.027569   0.002775   9.936  < 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: 7364.0  on 5311  degrees of freedom
## Residual deviance: 7197.5  on 5310  degrees of freedom
## AIC: 7201.5
##
## Number of Fisher Scoring iterations: 5
```

The p-value for the total_impressions variable is <2e-16, which is significantly below the 0.05 threshold, indicating strong statistical evidence that ad frequency influences purchase probability.

Additionally, the coefficient for total_impressions is positive (0.027569), suggesting that an increase in ad impressions is associated with a higher likelihood of purchase. This implies that greater ad exposure has a statistically significant positive effect on customer conversion rates.

```
#odds of purchse
exp(0.027569)
```

```
## [1] 1.027953
```

This result indicates that each additional ad impression increases the odds of purchase by exp(0.027569) = 1.027953 times. This translates to a 2.79% increase in the odds of purchase for every additional ad impression.

Given the statistically significant positive coefficient and the low p-value (<2e-16), there is strong evidence to suggest that increasing the frequency of advertising positively impacts the probability of purchase.

Evaluating the Effectiveness of Sites 1-5 vs. Site 6 in Driving Purchases

Star Digital's advertising campaign ran across six websites, divided into two groups: Sites 1-5, which belong to a single ad network where ads are distributed algorithmically, and Site 6, an independent platform where ad placement can be directly controlled. The key decision is determining which group delivers higher conversions at a lower cost, given that Sites 1-5 cost $25 per 1,000 impressions, while Site 6 costs $20 per 1,000 impressions.

To assess effectiveness, we will first run a logistic regression model with an interaction term to evaluate whether impressions from Sites 1-5 and Site 6 have a combined effect on purchase probability. If the interaction term is not significant, we will proceed with a logistic regression without interaction to isolate the individual effects of each site group. Finally, we will assess cost-effectiveness by calculating the cost per conversion for both site groups, ensuring that advertising budget allocation is optimized for both impact and efficiency.

```
balanced_data <- balanced_data %>% mutate(group_1_imp = imp_1 + imp_2 + imp_3 + imp_4 + imp_5)

#combined effect on purchase probability
model_sites <- glm(purchase ~ group_1_imp + imp_6 + group_1_imp * imp_6, data = balanced_data, family =
summary(model_sites)
```

```
##
## Call:
## glm(formula = purchase ~ group_1_imp + imp_6 + group_1_imp *
##     imp_6, family = "binomial", data = balanced_data)
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -0.1590598  0.0312291  -5.093 3.52e-07 ***
## group_1_imp        0.0293373  0.0034646   8.468  < 2e-16 ***
## imp_6              0.0039181  0.0042578   0.920   0.3575
## group_1_imp:imp_6  0.0013637  0.0007413   1.840   0.0658 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 7364.0  on 5311  degrees of freedom
## Residual deviance: 7181.1  on 5308  degrees of freedom
## AIC: 7189.1
##
## Number of Fisher Scoring iterations: 6
```

The p-value for group_1_imp is < 2e-16, which is well below the 0.05 significance threshold, indicating strong statistical evidence that impressions from Sites 1-5 positively impact purchase probability. Additionally, the coefficient for group_1_imp (sum of impressions from site 1 to 5) is 0.0293373, confirming that higher ad impressions on these sites are associated with an increased likelihood of purchase. In contrast, the p-value for imp_6 is 0.3575, which is greater than 0.05, suggesting that impressions from Site 6 do not have a statistically significant effect on purchases. These findings highlight that advertising on Sites 1-5 is more effective in driving conversions than Site 6.

```
#odds of purchase
exp(0.0293373)
```

```
## [1] 1.029772
```

For each additional ad impression on sites 1-5 increases the odds of purchase by exp(0.0293373) = 1.029772 times. Therefore, there is a 2.9772% increase in the odds of purchase for each additional ad impression on sites 1-5.

For site 6, this effect is not statistically significant (p = 0.108), meaning we cannot be confident that ads on Site 6 impact purchase behavior.

The interaction term (group_1_imp:imp_6) is weakly significant (p = 0.0658). This suggests a potential relationship between the number of impressions on Sites 1-5 and Site 6, but it is not strongly significant at the 5% level. Since p = 0.066, it is close to statistical significance but not conclusive.

To further validate these findings and isolate the true impact of ad impressions on purchase probability, we conduct an additional logistic regression. This model estimates the effects of ad impressions from Sites 1-5 and Site 6 separately, ensuring a more accurate assessment of their influence on conversions.

```
balanced_data <- balanced_data %>% mutate(group_1_imp = imp_1 + imp_2 + imp_3 + imp_4 + imp_5)

#isolated impact of ad impressions on purchase proba
model_sites <- glm(purchase ~ group_1_imp + imp_6, data = balanced_data, family = "binomial")
summary(model_sites)
```

```
##
## Call:
## glm(formula = purchase ~ group_1_imp + imp_6, family = "binomial",
##     data = balanced_data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.167175   0.031378  -5.328 9.94e-08 ***
## group_1_imp  0.032219   0.003253   9.904  < 2e-16 ***
## imp_6        0.007967   0.004954   1.608    0.108
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 7364.0  on 5311  degrees of freedom
## Residual deviance: 7185.7  on 5309  degrees of freedom
## AIC: 7191.7
##
## Number of Fisher Scoring iterations: 5
```

The logistic regression confirms that each additional impression on Sites 1-5 increases purchase odds by 3.28% ($p < 0.05$), making it a statistically significant driver of conversions. In contrast, Site 6 impressions remain insignificant ($p > 0.05$), indicating no strong impact on purchases. Therefore, advertising should be prioritized on Sites 1-5 for better effectiveness.

To analyze cost-effectiveness across sites, the cost per impression for Sites 1-5 was evaluated as a group, since advertisers do not have the ability to select specific sites within this network. Instead, ad placements are determined algorithmically within the group. This approach ensures a fair comparison between Sites 1-5 and Site 6.

To ensure accurate cost-per-conversion estimates, purchases from users exposed to both Sites 1-5 and Site 6 were evenly allocated between the two groups, preventing double counting and ensuring a fair comparison of advertising effectiveness.

```
# cost 1,000/- impressions
cost_1_5 <- 25
cost_6 <- 20

# purchases based on unique exposure
total_purchases_only_1_5 <- sum(balanced_data$purchase[balanced_data$group_1_imp > 0 & balanced_data$imp
total_purchases_only_6 <- sum(balanced_data$purchase[balanced_data$imp_6 > 0 & balanced_data$group_1_imp
total_purchases_both <- sum(balanced_data$purchase[balanced_data$group_1_imp > 0 & balanced_data$imp_6 >

# handling users exposed to both
total_purchases_adj_1_5 <- total_purchases_only_1_5 + (total_purchases_both / 2)
total_purchases_adj_6 <- total_purchases_only_6 + (total_purchases_both / 2)

# total ad spend
total_imps_1_5 <-sum(balanced_data$imp_1+balanced_data$imp_2+balanced_data$imp_3+balanced_data$imp_4+bal
total_imps_6 <-sum(balanced_data$imp_6)

#total ad spending
total_ad_spend_1_5 <- (total_imps_1_5/1000)*cost_1_5
total_ad_spend_6 <- (total_imps_6/1000)*cost_6

# cost per conversion
cost_per_convers_1_5 <- total_ad_spend_1_5 / total_purchases_adj_1_5
cost_per_convers_6 <- total_ad_spend_6 / total_purchases_adj_6

cost_per_convers_1_5
```

```
## [1] 0.4737951
```

```
cost_per_convers_6
```

```
## [1] 0.191853
```

Site 6 is more cost effective, the cost per conversion for Site 6 ($0.191) is significantly lower than for Sites 1-5 ($0.47). This suggests that while Sites 1-5 may have a stronger impact on purchase probability, advertising on Site 6 generates conversions at a lower cost.

There is a trade-off between effectiveness and cost. Previous logistic regression results showed that Sites 1-5 significantly increase of purchase, while Site 6 does not. However, cost per conversion favors Site 6, meaning it is more budget-friendly despite its weaker statistical significance in driving purchases.

Final Recommendations and Conclusion -

1. The analysis confirms that online advertising is effective in increasing conversions. Star Digital can continue investing in online advertising as it demonstrates a measurable impact on customer acquisition.
2. Results show that increasing ad frequency positively influences purchase behavior, particularly on Sites 1-5, where each additional impression increases the odds of purchase by 3.28% ($p < 0.05$). However, ad impressions on Site 6 do not have a statistically significant impact on conversions. Star Digital can increase ad exposure on Sites 1-5 to maximize effectiveness.
3. However even though, sites 1-5 are the strongest drivers of conversions, site 6 does not significantly influence purchase behavior, but it has a lower cost per conversion ($0.191) compared to Sites 1-5 ($0.47). It is recommended to prioritize sites 1-5 for advertising investment, given their stronger impact on conversion rates. But also allocating a portion of the budget to Site 6 to take advantage of its cost efficiency.
4. Further testing on Site 6 can also be done to determine whether improved targeting or creative adjustments could enhance its effectiveness.

Conclusion

Star Digital's online advertising strategy is effective and should remain a key component of its marketing efforts. Sites 1-5 should receive the majority of ad spend due to their strong impact on conversions, while Site 6 can be leveraged as a cost-efficient alternative. A balanced approach that maximizes return on investment while optimizing cost efficiency will ensure sustained advertising effectiveness.

This report provides data-driven insights to support budget allocation decisions and recommends further testing to refine digital marketing strategies.

Assumptions of the Study

While the experiment follows a standard A/B testing framework, certain assumptions and limitations must be considered:

1. Stable Unit Treatment Value Assumption (SUTVA): The analysis assumes that a user's purchase decision is influenced only by their own ad exposure, without external spillover effects. cross-device behavior, spillover effects and heterogeneous treatment effects.
2. Randomization & Sample Representativeness: The study assumes random assignment between the test and control groups, ensuring comparable user characteristics. Any systematic differences in unobserved factors may introduce bias.
3. Attribution of Conversions: The model assumes that an observed purchase is causally linked to ad exposure and not external factors.
4. Confidence level : The analysis assumes a 95% confidence level (alpha = 0.05), meaning there is a 5% probability of rejecting the null hypothesis when it is actually true.

5. Other Assumptions: The analysis assumes that all impressions within a site group are equally valuable, despite possible differences in ad placement, audience engagement, or visibility. Future studies could explore marginal return analysis to refine budget allocation across sites.