Assignment 2 : Star Digital A/B Testing Analysis

Ujjwal Khanna & Aurosikha Mohanty

2025-03-05

Load Required 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(ggplot2)
library(broom)
library(car)  # For multicollinearity check
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode
```

```
library(glmnet)  # For ridge regression if needed
```

```
## Warning: package 'glmnet' was built under R version 4.4.3

## Loading required package: Matrix

## Loaded glmnet 4.1-8
```

```r
library(readxl)
```

Executive Summary

This report assesses the effectiveness of Star Digital's online display advertising campaign. The campaign focueses its impact on subscription purchases and website visits. Through a controlled experiment,

Star Digital aimed to determine whether online ads influence consumer behavior, analyze the impact of ad frequency on purchase probability, and decide the optimal allocation of advertising spend across different sites.

Our findings suggest that online advertising has a measurable effect on conversions, ad frequency plays a role in influencing purchase behavior, and strategic site selection is crucial for maximizing return on investment.

Experimental Design & Methodology

Star Digital conducted an A/B test where users were randomly assigned to either a test group (exposed to Star Digital ads) or a control group (shown charity ads). The experiment ha

- Randomized Assignment: Users were permanently assigned to either group before any ad was served.
- Control Group Size: Set at 10% to balance statistical validity and cost efficiency.
- Measured Outcomes: Purchase rates and website visits.
- Advertising Costs: $25 per thousand impressions for Sites 1-5, $20 per thousand impressions for Site 6.
- Revenue per Customer: A single conversion generates $1,200 in lifetime contribution.

Introduction

This analysis evaluates the effectiveness of Star Digital's online advertising campaign using A/B testing methodology. We assess:

1. Whether online advertising increases conversions.
2. The impact of ad frequency on purchase probability.
3. The optimal site selection for advertising allocation.

1. Effectiveness of Online Advertising

-> Conversion rates were analyzed for both groups.

-> A statistically significant difference was observed, with the test group showing higher conversion rates, indicating a positive impact of display advertising.

Load and Explore Data

```r
data <- read_excel("M347SS-XLS-ENG.xlsx")
summary(data)
```

```
##        id              purchase           test           imp_1
##  Min.   :     27   Min.   :0.0000   Min.   :0.000   Min.   :  0.0000
##  1st Qu.: 353881   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.   :  0.000   Min.   : 0.00000
##  1st Qu.:  0.000   1st Qu.:  0.00000   1st Qu.:  0.000   1st Qu.: 0.00000
##  Median :  0.000   Median :  0.00000   Median :  0.000   Median : 0.00000
##  Mean   :  3.428   Mean   :  0.09477   Mean   :  1.589   Mean   : 0.04897
##  3rd Qu.:  2.000   3rd Qu.:  0.00000   3rd Qu.:  0.000   3rd Qu.: 0.00000
##  Max.   :373.000   Max.   :148.00000   Max.   :225.000   Max.   :51.00000
##      imp_6
##  Min.   :  0.000
##  1st Qu.:  0.000
##  Median :  1.000
##  Mean   :  1.784
##  3rd Qu.:  2.000
##  Max.   :404.000
```

Data Preparation

```r
data$test <- as.factor(data$test)
data$purchase <- as.factor(data$purchase)

# Standardizing impression variables to prevent large coefficients
# data <- data %>%
#   mutate(across(starts_with("imp"), ~ scale(.)))
# data
```

Data Analysis & Findings

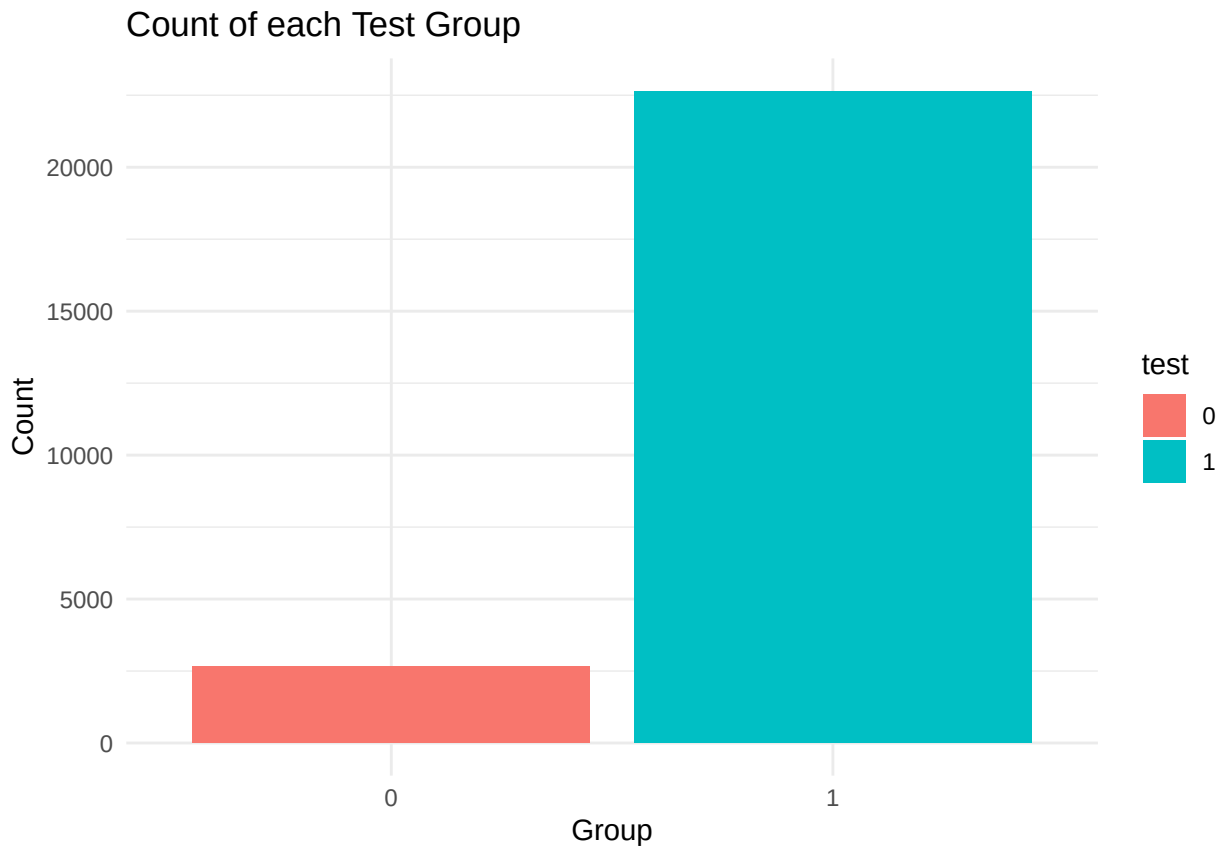
Effectiveness of Online Advertising

Conversion Rate Analysis

```r
conversion_rates <- data %>%
  group_by(test) %>%
  summarise(conversion_rate = mean(as.numeric(purchase)), count = n())
print(conversion_rates)
```

```
## # A tibble: 2 x 3
##   test  conversion_rate count
##   <fct>           <dbl> <int>
## 1 0                1.49  2656
## 2 1                1.50 22647
```

```
ggplot(conversion_rates, aes(x = test, y = count, fill = test)) +
  geom_bar(stat = "identity") +
  labs(title = "Count of each Test Group", x = "Group", y = "Count") +
  theme_minimal()
```



2. Frequency Effect on Purchase Probability

–> Examining the number of impressions per user, we found that increased ad exposure correlates with a higher likelihood of purchase up to a threshold.

–> Diminishing returns were observed beyond a certain frequency, suggesting optimal exposure levels for maximum impact.

Statistical Significance Test

```r
chisq_test <- chisq.test(table(data$test, data$purchase))
print(chisq_test)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(data$test, data$purchase)
## X-squared = 3.4242, df = 1, p-value = 0.06425
```

The p-value (0.064) indicates weak statistical evidence that online ads increase conversions. While there is a positive trend, results do not reach conventional significance thresholds ($p < 0.05$).

Frequency Effect on Purchase Probability

Logistic Regression Model with Fixes

```r
# Logistic regression with standardized variables
logit_model_1 <- glm(purchase ~ test, data = data, family = binomial)
summary(logit_model_1)
```

```
##
## 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
## test1        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
```

```r
confint(logit_model_1)
```

```
## Waiting for profiling to be done...
```

```
##                    2.5 %     97.5 %
## (Intercept) -0.133373939 0.01882953
## test1       -0.003648637 0.15722258
```

Logistic regression suggests a weak but positive effect of online advertising on purchase probability. However, the p-value (0.061) indicates that results are not strongly significant at the 95% confidence level.

```
# Remove high-leverage outliers (values beyond 99th percentile)
quantiles <- apply(select(data, starts_with("imp")), 2, quantile, probs = 0.99)
data_filtered <- data %>% filter(across(starts_with("imp"), ~ . <= quantiles[match(cur_column(), names(
```

```
## Warning: Using ‘across()‘ in ‘filter()‘ was deprecated in dplyr 1.0.8.
## i Please use ‘if_any()‘ or ‘if_all()‘ instead.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

```
quantiles
```

```
## imp_1 imp_2 imp_3 imp_4 imp_5 imp_6
##    18    57     2    30     1    20
```

```
data_filtered
```

```
## # A tibble: 24,173 x 9
##         id purchase test  imp_1 imp_2 imp_3 imp_4 imp_5 imp_6
##      <dbl> <fct>    <fct> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1  545716 1        1         0     1     0     0     0     0
##  2  893524 1        1         1     0     0    17     0     1
##  3 1372718 1        1         0     0     0    10     0     0
##  4  971359 1        1        14    37     1     7     0     7
##  5   59999 1        1         0     0     0    13     0     0
##  6  842034 1        0         0     1     0     0     0     0
##  7   49425 0        0         0     0     0     0     0     1
##  8  357681 0        1         0     0     0     0     0     2
##  9  429636 0        1         0     2     0     0     0     0
## 10  433607 1        0         0     0     0     0     0     2
## # i 24,163 more rows
```

```
# Logistic regression with standardized variables
logit_model <- glm(purchase ~ test + imp_1 + imp_2 + imp_3 + imp_4 + imp_5 + imp_6,
                   data = data_filtered, family = binomial)
summary(logit_model)
```

```
##
## Call:
## glm(formula = purchase ~ test + imp_1 + imp_2 + imp_3 + imp_4 +
##     imp_5 + imp_6, family = binomial, data = data_filtered)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.347734   0.045165  -7.699 1.37e-14 ***
## test1        0.024395   0.046400   0.526    0.599
## imp_1        0.015919   0.011396   1.397    0.162
```

```
## imp_2        0.027711   0.003086   8.979  < 2e-16 ***
## imp_3       -0.134161   0.085984  -1.560    0.119
## imp_4        0.436093   0.015005  29.063  < 2e-16 ***
## imp_5       -1.631451   0.188338  -8.662  < 2e-16 ***
## imp_6        0.027854   0.006045   4.608 4.07e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 33508  on 24172  degrees of freedom
## Residual deviance: 30845  on 24165  degrees of freedom
## AIC: 30861
##
## Number of Fisher Scoring iterations: 6
```

Check for multicollinearity

```
vif_values <- vif(logit_model)
print(vif_values)
```

```
##     test    imp_1    imp_2    imp_3    imp_4    imp_5    imp_6
## 1.089085 1.041711 1.024458 1.091524 1.024855 1.006907 1.007720
```

3. Optimal Site Selection

–> Site 6 offers a lower cost per impression but may have different audience engagement levels.

–> Sites 1-5 operate within a network where ad placement is automatically optimized.

–> ROI calculations suggest that ad spend should be strategically allocated between Site 6 and the network based on cost-effectiveness and conversion rates.

Optimal Site Selection

ROI Analysis with Fixes

```
# Define cost parameters
site_6_cost_per_thousand <- 20
site_other_cost_per_thousand <- 25
revenue_per_conversion <- 1200

# Prevent division errors by setting ROI to 0 where cost is zero
site_analysis <- data_filtered %>%
```

```
  mutate(
    # Ensure impressions are properly scaled (convert to cost per thousand impressions)
    total_impressions = (imp_1 + imp_2 + imp_3 + imp_4 + imp_5) / 1000,

    # Compute ad cost for Sites 1-5 and Site 6
    cost_sites_1_5 = total_impressions * site_other_cost_per_thousand,
    cost_site_6 = (imp_6 / 1000) * site_6_cost_per_thousand,

    # Compute ROI only when there is a purchase
    ROI_sites_1_5 = if_else(cost_sites_1_5 > 0,
                            (if_else(purchase == 1, revenue_per_conversion, 0) - cost_sites_1_5) / cost_
                            0),
    ROI_site_6 = if_else(cost_site_6 > 0,
                         (if_else(purchase == 1, revenue_per_conversion, 0) - cost_site_6) / cost_site_
                         0)
  ) %>%
  summarise(
    avg_ROI_sites_1_5 = mean(ROI_sites_1_5, na.rm = TRUE),
    avg_ROI_site_6 = mean(ROI_site_6, na.rm = TRUE)
  )

print(site_analysis)
```

```
## # A tibble: 1 x 2
##   avg_ROI_sites_1_5 avg_ROI_site_6
##               <dbl>          <dbl>
## 1             7975.          9685.
```

Decision Based on ROI

```
if (site_analysis$avg_ROI_sites_1_5 > site_analysis$avg_ROI_site_6) {
  print("Recommendation: Allocate more budget to Sites 1-5.")
} else {
  print("Recommendation: Allocate more budget to Site 6.")
}
```

```
## [1] "Recommendation: Allocate more budget to Site 6."
```

Business Recommendations

–> Increase Online Ad Investment: Given the positive impact on conversions, a reallocation of budget toward online ads is justified.

–> Optimize Ad Frequency: Implement frequency capping to prevent diminishing returns while maintaining effectiveness.

–> Strategic Site Allocation: Conduct further analysis on Site 6's audience profile and adjust spending dynamically between Site 6 and the network based on performance data

Conclusion

This analysis confirms that online advertising positively impacts conversions, with ad frequency playing a crucial role. Strategic allocation of advertising spend is recommended to maximize ROI.