QMB: 6304

Midterm Project: The Advertising Campaign Analysis

**--- Lakshmi Priya Chintada.**

**1 Business Justification**

1. Explain why retargeting customers who initially didn’t buy a package makes business sense.

* **It makes sense to retarget customers who did not purchase a vacation package at first since it makes use of previous interactions and data insights, which provide several benefits. It lowers acquisition costs, boosts conversion rates, gives personalized information, and gives a chance to reconnect with potential customers. Retargeting can be more successful and economical if it considers the objections and preferences of the consumer. In the end, it maximizes the return on investment in marketing initiatives by maintaining a competitive edge, boosting brand recognition, and taking advantage of timing.**

1. Analyze the test/control division. Does it seem well-executed?

* **In a retargeting campaign, the primary goal of the control and test groups is to assess the campaign's efficacy.**

**Control Group:** The portion of customers who are not subjected to the retargeting campaign is known as the control group. They sort of get "untouched" by the campaign and act as a kind of reference or baseline.

**Test Group:** Customers who are exposed to the retargeting campaign make up the test group, in contrast. These are the people, or prospects, that the campaign is trying to reach again or win over. Retargeting advertisements, emails, and other marketing campaigns aimed at persuading the test group to buy something or do something else are sent to them.

A screenshot of a computer

Description automatically generated

**Conclusion:** The test/control division seems reasonably executed as from the above theory, the "test" group has 4266 individuals, while the "control" group has 4176. This indicates that the test group is slightly larger in size than the control group. The test/control division seems reasonably executed.

1. Compute summary statistics for the test variable, segmenting by available State data.

**The Test\_Control column of Abandoned.csv has the values ‘test’ and ‘control’ , the summary statistics for the ‘test’ variable in the ‘Test\_Control’ column has been calculated using the below code and it is segmented by available State data( The ‘State data’ is present in the column of ‘Address’ in the given dataset).**

**Code:**

# The below statement shows the count of 'test' by different state

abd[abd==""] <- NA

states <- abd[complete.cases(abd['Address']),]

table(states$Test\_Control)

# Filter the data to include only rows with 'test' in the 'Test\_Control' column

filtered\_abd <- abd %>% filter(Test\_Control == "test")

# Group the filtered data by 'Address' and 'Test\_Control', and compute summary statistics (overall dataset)

summary\_stats <- filtered\_abd %>%

group\_by(Address, Test\_Control) %>%

summarize(

Count = n(), # Count of observations

)

# Print the summary statistics

print(summary\_stats)

**Output:**

> # The below statement shows the count of 'test' by different state

> abd[abd==""] <- NA

> states <- abd[complete.cases(abd['Address']),]

> table(states$Test\_Control)

control test

1855 1957

>

> # Filter the data to include only rows with 'test' in the 'Test\_Control' column

> filtered\_abd <- abd %>% filter(Test\_Control == "test")

>

> # Group the filtered data by 'Address' and 'Test\_Control', and compute summary statistics (overall dataset)

> summary\_stats <- filtered\_abd %>%

+ group\_by(Address, Test\_Control) %>%

+ summarize(

+ Count = n(), # Count of observations

+ )

`summarise()` has grouped output by 'Address'. You can override using the `.groups` argument.

>

> # Print the summary statistics

> print(summary\_stats)

# A tibble: 51 × 3

# Groups: Address [51]

Address Test\_Control Count

*<chr>* *<chr>* *<int>*

1 AK test 29

2 AL test 38

3 AR test 38

4 AZ test 54

5 CA test 48

6 CO test 40

7 CT test 42

8 DE test 34

9 FL test 38

10 GA test 47

# ℹ 41 more rows

Screenshot:

A screenshot of a computer

Description automatically generated

**2 Data Alignment**

1. From your examination of both files, propose potential data keys to match customers.

**Code:**

# Merge the datasets based on the Caller\_ID

matched\_customers <- merge(abd, rs, by = "Caller\_ID", all = FALSE)

# The "all = FALSE" option ensures that only common Caller\_IDs are included in the merged dataset.

# Print the result

print(matched\_customers)

**Result:**

A close-up of a computer screen

Description automatically generated

* Selecting **'Caller\_ID'** as the potential data key to match customers because of its uniqueness. The **'Caller\_ID'** is unique for each customer in both datasets, and it can serve as an excellent key for matching. Unique identifiers help ensure that each record in one dataset corresponds to a distinct customer in the other dataset.
* The message **"<0 rows> (or 0-length row.names)"** at the end suggests that the merging process did not yield any common records between the two datasets based on the "**Caller\_ID**" column.
* Alternatively, first name, last name, phone and email could potentially be used in combination to match customers.

**Code:**

# Match based on the "Email" field in both datasets

match\_email <- abd$Email[complete.cases(abd$Email)] %in% rs$Email[complete.cases(rs$Email)]

table(match\_email)

# Matching based on 'Incoming\_Phone'

match\_incoming\_phone <- abd$Incoming\_Phone[complete.cases(abd$Incoming\_Phone)] %in% rs$Incoming\_Phone[complete.cases(rs$Incoming\_Phone)]

table(match\_incoming\_phone)

# Matching based on 'Contact\_Phone'

match\_contact\_phone <- abd$Contact\_Phone[complete.cases(abd$Contact\_Phone)] %in% rs$Contact\_Phone[complete.cases(rs$Contact\_Phone)]

table(match\_contact\_phone)

Result:

> # Match based on the "Email" field in both datasets

> match\_email <- abd$Email[complete.cases(abd$Email)] %in% rs$Email[complete.cases(rs$Email)]

> table(match\_email)

match\_email

FALSE TRUE

955 75

>

> # Matching based on 'Incoming\_Phone'

> match\_incoming\_phone <- abd$Incoming\_Phone[complete.cases(abd$Incoming\_Phone)] %in% rs$Incoming\_Phone[complete.cases(rs$Incoming\_Phone)]

> table(match\_incoming\_phone)

match\_incoming\_phone

FALSE TRUE

6935 327

>

> # Matching based on 'Contact\_Phone'

> match\_contact\_phone <- abd$Contact\_Phone[complete.cases(abd$Contact\_Phone)] %in% rs$Contact\_Phone[complete.cases(rs$Contact\_Phone)]

> table(match\_contact\_phone)

match\_contact\_phone

FALSE TRUE

8218 185

# Merging datasets based on Email

merged\_data <- merge(abd, rs, by = "Email", all.x = TRUE, all.y = TRUE)

A screenshot of a computer

Description automatically generated

5. Detail your procedure to identify customers in:

• Treatment group who purchased.

* **The "test" group is exposed to the new treatment or feature, while the "control" group remains unchanged and serves as a baseline for comparison.**
* **The "test" group is typically the one where you would assess the impact of changes or interventions like a retargeting campaign.**
* **The below code shows the individuals are part of the "test" group and have provided complete and non-empty information for their address and contact details.**

**Code:**

**merged\_data <- merge(abd, rs, by = "Email", all.x = TRUE, all.y = TRUE)**

**test\_purchased <- subset(merged\_data, Test\_Control.x == "test" & !is.na(Email))**

**head(test\_purchased)**

**Result:**

**A screenshot of a computer

Description automatically generated**

• Treatment group who didn’t purchase.

**Code**:  
  
**test\_not\_purchased <- subset(merged\_data, Test\_Control.x == "test" & is.na(Email))**

**head(test\_not\_purchased)**

**Result:**

A screenshot of a computer

Description automatically generated

• Control group who purchased.

**Code:**

**control\_purchased <- subset(merged\_data, Test\_Control.x == "control" & !is.na(Email))**

**head(control\_purchased)**

**Result:**

**A screenshot of a computer

Description automatically generated**

• Control group who didn’t purchase.

**Code:**

**control\_not\_purchased <- subset(merged\_data, Test\_Control.x == "control" & is.na(Email))**

**# Displaying the result**

**head(control\_not\_purchased)**

**Result:**

**A screenshot of a computer

Description automatically generated**

6. Are there unmatchable records? If yes, provide examples and exclude them from the analysis.

Unmatchable customers are those individuals who cannot be effectively targeted for retargeting campaigns based on the established criteria. These customers may fall into one of the following categories:

**Unmatched Criteria:**

Unmatched customers could be those who have no record of interaction with your brand. In this case, you would exclude customers with **missing or empty values** in the **"Email," "Incoming\_Phone," and "Contact\_Phone"** columns. These columns are typically used for communication and interaction, so customers without this information might be considered unmatched.

**Code:**

**library(dplyr)**

**# Create a new column "Flag" based on "Email," "Incoming\_Phone," and "Contact\_Phone" values**

**merged\_data <- merged\_data %>%**

**mutate(Flag = ifelse(!is.na(Email) & !is.na(Incoming\_Phone.x) & !is.na(Contact\_Phone.x), "Purchased", "Not\_Purchased"))**

**# Print the data frame**

**print( merged\_data)**

**Result:**

A screenshot of a computer

Description automatically generated

7. Provide a cross-tabulation of outcomes for treatment and control groups.

**Code:**

**#Cross tabulation for the people in abd dataset**

**cross\_tab <- table(merged\_data$Test\_Control.x, merged\_data$Flag)**

**# Printing the cross-tabulation**

**print(cross\_tab)**

**#Cross tabulation for the people in rs dataset**

**cross\_tab <- table(merged\_data$Test\_Control.y, merged\_data$Flag)**

**# Printing the cross-tabulation**

**print(cross\_tab)This command generates the cross-tabulation. It counts the occurrences of each unique value in the "Test\_Control" column and creates a table displaying the counts and show how many records belong to each group ("test" and "control") in the "Test\_Control" column of filtered\_combined\_data dataset.**

**Result:**

**A screenshot of a computer

Description automatically generated**

**8.** Replicate the cross-tabulation for five randomly chosen states, detailing your selections

**Code:**

**set.seed(12) # Set a seed for reproducibility**

**selected\_states <- sample(unique(merged\_data$Address.x), 5)**

**# Create cross-tabulation for each selected state**

**for (Address in selected\_states) {**

**cat("State:", Address, "\n")**

**cross\_tab <- merged\_data %>%**

**filter(Address.x == Address | Address.y == Address) %>%**

**count(Flag)**

**print(cross\_tab)**

**cat("\n")**

**}**

**Output:**

State: VA

Flag n

1 Not\_Purchased 436

2 Purchased 15

State: TX

Flag n

1 Not\_Purchased 496

2 Purchased 9

State: VT

Flag n

1 Not\_Purchased 476

2 Purchased 16

State: NC

Flag n

1 Not\_Purchased 430

2 Purchased 19

State: ME

Flag n

1 Not\_Purchased 446

2 Purchased 12

9. Generate a cleaned dataset with columns: Customer ID — Test Group — Outcome — State Available— Email Available. Each row should correspond to a matched customer from the datasets. (Ensure you attach this cleaned dataset upon submission.)

Code:

cleaned <- data.frame(

Customer\_ID = ifelse(is.na(merged\_data$Email), as.character(merged\_data$Caller\_ID.x), as.character(merged\_data$Email)),

Test\_Group = as.character(merged\_data$Test\_Control.x),

Outcome = ifelse(is.na(merged\_data$Email), "No\_Purchase", "Purchased"),

State\_Available = ifelse(is.na(merged\_data$Address.x), "No", "Yes"),

Email\_Available = ifelse(is.na(merged\_data$Email), "No", "Yes")

)

write.csv(cleaned, "P:\\cleaned\_dataset.csv", row.names = FALSE)

10. Execute a linear regression for the formula: Outcome = α + β \* Test Group + error. Share the results.

Code:

**library(stargazer)**

**library(dplyr)**

**# Assuming Outcome is determined by whether Email is NA or not**

**merged\_data$Outcome <- ifelse(is.na(merged\_data$Email), 0, 1)**

**# Assuming Test Group is determined by the value in Test\_Control.x column ("control" or "test")**

**merged\_data$Test\_Group\_binary <- ifelse(merged\_data$Test\_Control.x == "control", 0, 1)**

**linear\_model <- lm(Outcome ~ Test\_Group\_binary, data = merged\_data)**

**# Displaying the summary of the regression model**

**summary(linear\_model)**

**Output:**

> summary(linear\_model)

Call:

lm(formula = Outcome ~ Test\_Group\_binary, data = merged\_data)

Residuals:

Min 1Q Median 3Q Max

-0.1338 -0.1338 -0.1131 -0.1131 0.8869

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.113104 0.005087 22.234 < 2e-16 \*\*\*

Test\_Group\_binary 0.020697 0.007155 2.893 0.00383 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.329 on 8455 degrees of freedom

(20726 observations deleted due to missingness)

Multiple R-squared: 0.0009888, Adjusted R-squared: 0.0008706

F-statistic: 8.368 on 1 and 8455 DF, p-value: 0.003828

* **According to the model, there is a statistically significant relationship between the "Test\_Group\_binary" variable and the "Outcome." The R-squared values, however, are quite near to zero, suggesting that this variable only contributes a very minor amount to the explanation of the outcome's variation. This implies that the association between the variables may be weak or that there could be other factors at work.**

**11.** Justify that this regression is statistically comparable to an ANOVA/t-test.

**Code:**

**# One-way ANOVA**

**anova\_result <- anova(lm(Outcome ~ Test\_Group\_binary, data = merged\_data))**

**# Display the results**

**print(anova\_result)**

**Result:**

**A screenshot of a computer

Description automatically generated**

**ANOVA:**

**The grouping factor (0 for control, 1 for test) is represented by the "Test\_Group\_binary" variable, which significantly affects the "Outcome."**

**With a p-value of 0.003828 and an F-statistic of 8.3684, the results are less significant than the usual threshold of 0.05.**

**Based on sample estimates, group 0 has a mean of 0.1131038, but group 1 has a mean of 0.1338012.**

**In conclusion, there is proof of a statistically significant difference in the means of "Outcome" between the control and test groups according to the findings of the ANOVA.**

**12.** Debate the appropriateness of the regression model in making causal claims about the retargeting campaign’s efficacy.

* **These statistics might suggest a very strong statistical relationship, they don't imply a causal relationship. Since causation is a complicated topic, proper research design is needed. Randomized controlled trials or other techniques are frequently used to demonstrate cause and effect. Finding a significant statistical association in a real-world dataset does not indicate that you may draw conclusions about the retargeting campaign's causality.**
* **The relatively low adjusted R-squared reveals that the "Test\_Group\_binary" variable explains very little variance in the "Outcome," despite the regression model suggesting a statistically significant association between the test group and "Outcome." This implies that strong causal assertions regarding the effectiveness of the retargeting campaign may not be supported by the model.**
* **A deeper investigation that takes into consideration the experimental design, data over time, and other relevant factors is required in order to draw conclusions regarding the campaign's efficacy's causality. In order to confirm the causal association, more studies and tests must be carried out.**

**13.** Integrate State and Email dummies into the regression. Also consider interactions with the treatment group. Compare these results to the previous regression and provide insights.

**Code:**

**model\_x <- lm(Outcome ~ Test\_Control.x \* Address.x + Email, data = merged\_data)**

**# Print the regression summary**

**summary(model\_x)**

**OutPut:**

**A screenshot of a computer

Description automatically generated**

**Interpretation:**

**Overfitting may be suggested by the exceptionally low residual standard error, while a poor fit between the model and the data is indicated by zero or negative R-squared values. This might be the result of problems with variable selection or the existence of missing data. To solve these problems, further research and even an alternative modeling strategy are required.**

* **5 Reflections**

14. Reflect on the project:

• Would you modify the experiment design if given a chance?

* The experiment could be improved by having a larger and more representative sample size. I can think about including a few extra columns that offer more context and data in order to improve the analysis and model construction utilizing the provided dataset. These are some ideas for new demographic-related columns that may be added. Like the state, the size of the city, the region, and the time-related elements like the day of the week, the year, and the customer behavior. This indicates a chance to expand the campaign's target audience.

• Could alternative paths be taken with better-quality data?

With higher-quality data, other options could be considered to improve the retargeting campaign analysis.

* **Better Data Collection:** Gather more thorough and precise information about the interactions, preferences, and behavior of your customers. Monitoring website traffic, click-through rates, product views, and cart abandonment rates are a few examples of this.
* **Improved client Profiling:** Compile extra demographic, psychographic, and prior purchase data to create more thorough client profiles. Segments for more specialized marketing may be created as a result.
* **Longitudinal Data:** Compile information before and after the campaign over an extended period of time. This offers a more thorough understanding of how retargeting attempts affect consumer behavior.

* **Data Quality Assurance:** Reduce missing or incorrect data by putting data quality assurance procedures into place. This includes confirming contact details.

• Are there actionable business implications from this analysis?

* **The examination of the data from the retargeting campaign has practical commercial ramifications. These ramifications have the potential to influence strategic choices and spur advancements in consumer interaction and marketing tactics. The following are some possible takeaways.**
* **Targeting according to Segment:**

**It is possible from the study that some consumer categories react better to retargeting campaigns. Marketing resources may be allocated more wisely to the categories that have the best conversion rates based on this data.**

* **Optimizing Creative and Messaging:**

**Through performance analysis of various creatives, ad formats, and message, companies may improve the effectiveness of their advertising material and make it more appealing to consumers. This might entail customizing the message in accordance with the tastes and actions of the audience.**

* **Customer Retention: The analysis may indicate that certain customers who did not make an immediate purchase are more likely to become repeat buyers with a tailored follow-up strategy. This knowledge can inform customer retention efforts.**

**Mapping the Customer Journey:**

* **Businesses may discover critical decision points by having a thorough understanding of the client path and touchpoints prior to conversion. Having this knowledge can help with more accurate targeting throughout crucial parts of the trip.**

**Suggested Products:**

* **Businesses might suggest goods or services based on the analysis of what customers have viewed or left in their shopping carts. The experience of purchasing may be improved with personalized recommendations.**

15. Self-assessment: Rate your effort (0-100) and anticipated performance. Elaborate if needed, mentioning any collaborations.

**I would rate myself 100/100. I strongly believe that my work aligned with project goals.**