The Advertising Campaign Analysis

**1.Business Justification**

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

There are various reasons why retargeting customers makes business sense:

* Compared to new consumers, retargeted clients are frequently more likely to convert. They have already expressed interest in the good or service by going to the website or interacting with the business, which suggests a greater desire to buy. Retargeting can increase conversion rates, according to statistics, because the clients are already familiar with the brand.
* Compared to keeping current customers or retargeting people who have interacted with the business in the past, acquiring new customers might be considerably more expensive. To optimize return on investment (ROI), it makes sense from a business standpoint to dedicate a portion of the advertising budget to retargeting efforts.
* Retargeting enables customized advertising based on previous interactions between the customer and the company. A business can generate personalized adverts that are more relevant and enticing to the individual, improving the possibility of conversion, by evaluating client data and behavior. For instance, if a user looked at a specific package but didn't purchase, the retargeted ad can highlight that particular package or offer a special discount on it. This level of personalization can lead to better engagement and higher click-through rates.
* Retargeting can offer a sizable advantage in markets with intense competition. A business can boost its chances of acquiring these clients before they go to a rival by staying visible and pertinent to customers who have expressed interest but did not convert.

1. ***Analyze the test/control division. Does it seem well-executed?***
   * From the test/control analyses For the Abandoned dataset, the division between the test and control groups seems relatively even. There's a small difference of 90 between the two groups, which is roughly a 1% variation. This is generally acceptable and can be considered well-executed.
   * However, for the Reserved dataset, there's a significant disparity between the test and control groups. The test group has almost 9 times more entries than the control group.

Since we will only be using the Reserved dataset for matching, this does not concern us.

A blue screen with yellow and orange text

Description automatically generated

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

*Assuming by summary statistics in this context, we mean to segment the total number of test and control variables by State to see their distribution by State.*

*A blue screen with white text

Description automatically generated*

*A blue screen with orange and yellow text

Description automatically generated*

A blue screen with white text

Description automatically generated

The division for the Abandoned data set is relatively equal on the State level between test and control.

**2.Data Alignment**

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

I will be matching the files on:

* Email
* Contact\_Phone
* Incoming\_Phone
* Further, I will be matching for Contact\_Phone of Abandoned dataset with Incoming\_Phone of Reserved data set and vice versa.

1. ***Detail your procedure to identify customers in:***
   1. ***Treatment group who purchased.***
   2. ***Treatment group who didn’t purchase.***
   3. ***Control group who purchased.***
   4. ***Control group who didn’t purchase.***

* First we are mainly concerned with matching the 'Abandoned' dataset (abdn) with the 'Reservation' dataset (rser) based on different keys (Email, Incoming\_Phone, and Contact\_Phone). For each matching condition, a flag is created in the abdn dataset.

emailmatch: Matches on Email.

incom\_match: Matches on Incoming\_Phone.

contactmatch: Matches on Contact\_Phone.

incom\_contact\_match: Matches on both Incoming\_Phone from abdn and Contact\_Phone from rser.

contact\_incom\_match: Matches on Contact\_Phone from abdn and Incoming\_Phone from rser.

A screen shot of a computer code

Description automatically generatedA screen shot of a computer program

Description automatically generated

* Next we create a logical flag abdn$pur to determine if a record in the 'Abandoned' dataset (abdn) matches with a record in the 'Reservation' dataset (rser) based on any of the above keys (using OR | logic). This flag will be 1 if there's a match based on any of the conditions and 0 otherwise.
  + After that, an additional column is created :
  + 

abdn$treat - a binary column derived from the Test\_Control column where "test" is represented as 1 and "control" is represented as 0.

* + I also create dummies for State and Email for use in analyses.



* Finally , I created a Table to showcase the combination of purchase decisions (**abdn$pur**) and test/control assignment (**abdn$treat**). The results provide counts of customers in each segment.

A screen shot of a computer program

Description automatically generated

A group of numbers on a blue background

Description automatically generated

* **Treatment group who purchased:** This is represented by the "Purchased" row and "Treatment Group" column in the table, which has a count of **345**.
* **Treatment group who didn’t purchase:** This is represented by the "Not Purchased" row and "Treatment Group" column, with a count of **3921**.
* **Control group who purchased:** This is represented by the "Purchased" row and "Control Group" column, which gives a count of **93**.
* **Control group who didn’t purchase:** Represented by the "Not Purchased" row and "Control Group" column, showing a count of **4083**.

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

If we assume unmatchable to be not purchased, then below is the solution:

A blue screen with white text

Description automatically generated

A blue screen with many small letters

Description automatically generated with medium confidence

However, we cannot remove these records from the analysis as we need both “purchased” and “not purchased” records.

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

A screen shot of a computer program

Description automatically generated

A group of numbers on a blue background

Description automatically generated

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

A computer screen shot of a code

Description automatically generated

A screenshot of a computer program

Description automatically generated

**3.Data Refinement**

1. ***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.)***

A computer screen shot of a blue screen

Description automatically generated

* + First, we remove the other columns that are not needed.
  + Then we change the index of the remaining columns and change the names as per the template given above.
  + Finally, I am exporting the cleaned dataset to attach in the submission.

**4.Statistical Assessment**

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

A computer screen with white text

Description automatically generated

**Outcome = 0.222 + 0.058\*Test Group + error**

H0 : There is No effect of Retargeting(Test\_Group) on Outcome.

**Model Intrepretations:**

Test\_Group Intercept is 0.058602

This represents the difference in the expected means of Outcome between the two levels of Treatment groups. (test and control)

So, members of the test group (when Test\_Group is 1) have, on average, an Outcome value that's higher by s 0.058602 compared to the control group.

The p-values are highly significant (way below 0.05). This indicates that both coefficients are statistically significant. Based on the p-value , we can reject the Null Hypothesis.

Multiple R-squared of 0.01745. This is really low because when dealing with a binary outcome, a linear regression is not the ideal model as the outcome variable is not continuous.

Instead, logistic regression is more suitable for binary outcomes, as it models the log-odds of the probability of an event occurring.

**Conclusion:** Being in the Retargeting campaign has a statistically significant positive association with a Outcome when compared to not being Retargeted

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

A screenshot of a computer screen

Description automatically generated

* For a binary predictor, a simple linear regression is just another way of performing a t-test, and by extension, it's statistically comparable to a two-group ANOVA. The results from the linear regression and ANOVA are different ways of expressing the same underlying statistical tests and relationships. (difference in means)
* The coefficient for the Test\_Group in the linear regression gives the difference in means between the two groups. In the context of ANOVA, this is equivalent to comparing the means of two groups ( test and control)
* The p-value for the **Test\_Group** in the regression output is the same as the p-value for the ANOVA. Both are extremely small, indicating a statistically significant difference between the two groups.
* The F-statistic in the regression output (**149.9**) is the same as the F value in the ANOVA output. This is because when you run a linear regression with a binary independent variable (like the Test\_Group here), the square of the t-statistic for the binary predictor will be equal to the F-statistic from an ANOVA comparing the means of the two groups.
* In my analysis, the extremely low p-value for the Test\_Group variable in the regression output (less than 2.2e-16) and the high F-statistic in ANOVA both tell us the same thing: There's a highly significant difference in means between the control and treatment groups.
* In summary, whether you use linear regression or ANOVA in this scenario, you'll arrive at the same conclusion: Being in the treatment group significantly affects the outcome compared to the control group. This illustrates the statistical equivalence between linear regression and ANOVA

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

* Regression results alone is not enough proof of causation. A significant regression coefficient merely shows that there's an association, not a causal link.
* Regression can identify relationships, but it's more difficult to determine causality. It's not necessary to conclude that the retargeting caused the treatment group's varied results just because they were different. Confounding factors may exist that the model does not take into consideration.
* If important variables that influence the dependent variable are not included in the regression, the findings may be biased. This means that even if the independent and dependent variables have a significant relationship, it could be due to the omitted factors rather than a causal relationship.

Thus, we cannot claim that the retargeting is causing a positive outcome but only that the likelihood of a positive outcome increases if a customer is retargeted.

1. ***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.***

A screenshot of a computer program

Description automatically generated

***3 linear regression models and 1 logistic regression model created optionally.***

*It's clear that the campaign is effective. Moreover, the availability of state and email data enhances its effectiveness. The significant interaction terms further underscore the value of personalized retargeting.*

**Comparison:**

* Coefficients:
* Test\_Group: Represents the effect of being in the test (or treatment) group.
* Model (1) suggests that being in the test group increases the outcome by 0.059 units, holding other variables constant.
* Model (2) and (3) show a slightly reduced effect. This decrease may be due to controlling for state and email availability.
* In Model (2), the outcome increases by 0.017 units when state information is available.
* In Model (3), the effect is slightly reduced to 0.010.
  + In Model (2), the outcome increases by 0.036 units when an email is available.
  + However, this effect drops significantly in Model (3) to 0.008.
  + Test\_Group:Email\_Available: Represents how the effect of being in the test group changes when an email is available. This interaction has a significant effect on the Outcome.
  + In Model (3), the coefficient is 0.053, a notable positive interaction. This indicates that having both treatment and an available email has a synergistic effect, boosting the outcome more than their separate effects would suggest.

**Insights in the Context of Retargeting Campaign:**

* Significance of Test Group: Being part of the retargeting campaign consistently shows a positive effect on the outcome across all models. This suggests that the campaign is effective.
* State and Email Information: Having state and email information available enhances the effect on the outcome. This suggests that personalized retargeting (based on state or email) might be more effective.
* Interaction Importance: The positive interactions between the test group and email availability suggest that the retargeting campaign is even more effective when it can use email information for personalization.
* Model Comparisons: As we introduce more variables and interactions, the effect size of the test group decreases. This implies that some of the initial effect attributed to the test group in Model (1) can be explained by the availability of state and email information in the subsequent models.

**5. Reflections**

1. ***Reflect on the project:***
   1. ***Would you modify the experiment design if given a chance?***

Yes, I would modify the design to improve the precision and get more valuable insights:

* + - To find out which aspects of the campaign are most effective, test various variations of the advertising (alternative messaging, images, etc.) rather than just the total impact of retargeting.
    - Control the effects of geographical events, time zones, and/or seasons on user behavior. Retargeting, for instance, can be more successful during certain regions' holiday seasons.
    - Gain more information on the customers’ demographics to understand their relation on the outcome. Eg. Age can be a very important factor in deciding the success of the campaign or to know which age groups are more likely to respond positively to retargeting.
    - Finally, If we had uniquely matching customer IDs on both data sets, we could have better match quality.
  1. ***Could alternative paths be taken with better-quality data?***
* Deeper segmentation research could be performed using precise demographic, behavioral, and psychographic data to determine which specific audience categories respond best to retargeting.
* Time-Series Analysis: More granular time-series analysis would be possible if data were collected in a time-stamped manner. This could help researchers better understand patterns, seasonality, and the immediate vs delayed effects of retargeting.
* Machine Learning Models: With more data, machine learning prediction models might be constructed to forecast future campaign outcomes or to personalize retargeting methods for particular users.
* Data on user attrition or retention could provide insights into the long-term impact of retargeting advertising on consumer loyalty.
* Multi-channel Attribution: Using data from multiple channels (e.g., email, social media, direct traffic), a multi-channel attribution model could be built to understand the holistic impact of retargeting across all touchpoints.
  1. ***Are there actionable business implications from this analysis?***

Yes, there are several implications and action steps to be taken from these analyses:

* **Optimize Retargeting Efforts**: The analysis indicates that the retargeting campaign had a positive impact. The business should consider:
  + **Scaling Up**: Allocate a larger budget to retargeting campaigns given their proven efficacy.
  + **Refinement**: Continuously refine the retargeting criteria based on real-time feedback to ensure the highest impact.
* **Data Collection Focus**:
  + **State & Email Collection**: Since having information about a user's state and email appears to enhance the campaign's efficacy, encourage users to provide these details, possibly through incentives or seamless UX design.
  + **Expand Data Points**: Look for other potential data points that could be collected to further refine and personalize the retargeting efforts.
* **Tailored Campaigns**:
  + **Segmentation**: Use the state and email data to segment the audience and design tailored retargeting campaigns for each segment. For instance, different advertisements or offers might resonate differently with users from various states.
* **Personalized Messaging**: If email information is available and users have given consent, send personalized emails alongside retargeting ads to boost engagement and conversions.
* **Interaction Effect Utilization:** Given that there's a significant interaction between the test group and whether email data is available, design strategies that capitalize on this. For example, if you know the email of a user, your retargeting approach may differ from someone whose email you don’t have.
* **Integrated Marketing:** Ensure that retargeting efforts are synchronized with other marketing channels. For instance, if a user sees a retargeting ad and then receives a complementary email, the combined effect could be more potent.
* **Resource Reallocation**: Based on the success of the retargeting campaign, re-evaluate the marketing budget. Consider allocating more resources to retargeting and possibly reducing spend on less effective channels.
* Regularly revisiting the analysis and updating strategies based on fresh data is key to sustained success.

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

I would rate my effort as a 90 as I dug deep into the conducting the analyses and learned to use the statistical models in the business context to give actionable insights. I had to follow the template of the Prof for the cleaning part so that is why I did not rate myself as 100. But I did make many adjustments to that template, so I think I did really well.

**Appendix:**

library(dplyr)

library(stargazer)

# Read data

abdn = read.csv("C:/Users/mudas/OneDrive/Desktop/QMB/midterm project/Abandoned.csv", header = TRUE, na.strings = "")

rser = read.csv("C:/Users/mudas/OneDrive/Desktop/QMB/midterm project/Reservation.csv", header = TRUE, na.strings = "")

#Check for missing values

sum(is.na(abdn))

sum(is.na(rser))

#Check for Duplicates

sum(duplicated(abdn))

sum(duplicated(abdn$Caller\_ID))

sum(duplicated(rser))

sum(duplicated(rser$Caller\_ID))

#test/control division analyses

sum(abdn$Test\_Control == "test" )

sum(abdn$Test\_Control == "control")

sum(rser$Test\_Control == "test" )

sum(rser$Test\_Control == "control")

#segment by available state

if\_state <- abdn[complete.cases(abdn['Address']),]

table(if\_state$Test\_Control)

summary\_stats = abdn %>%

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

summarize(Count = n())

print(summary\_stats)

summary\_stats\_rser = rser %>%

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

summarize(Count = n())

print(summary\_stats\_rser)

# Matching based on different keys and create logical vectors for each condition

emailmatch = abdn$Email[complete.cases(abdn$Email)] %in% rser$Email[complete.cases(rser$Email)]

incom\_match = abdn$Incoming\_Phone[complete.cases(abdn$Incoming\_Phone)] %in% rser$Incoming\_Phone[complete.cases(rser$Incoming\_Phone)]

contactmatch = abdn$Contact\_Phone[complete.cases(abdn$Contact\_Phone)] %in% rser$Contact\_Phone[complete.cases(rser$Contact\_Phone)]

incom\_contact\_match = abdn$Incoming\_Phone[complete.cases(abdn$Incoming\_Phone)] %in% rser$Contact\_Phone[complete.cases(rser$Contact\_Phone)]

contact\_incom\_match = abdn$Contact\_Phone[complete.cases(abdn$Contact\_Phone)] %in% rser$Incoming\_Phone[complete.cases(rser$Incoming\_Phone)]

# Create flags for matches

abdn$emailmatch = 0

abdn$emailmatch[complete.cases(abdn$Email)] = 1 \* emailmatch

abdn$incom\_match = 0

abdn$incom\_match[complete.cases(abdn$Incoming\_Phone)] = 1 \* incom\_match

abdn$contactmatch= 0

abdn$contactmatch[complete.cases(abdn$Contact\_Phone)] = 1 \* contactmatch

abdn$incom\_contact\_match= 0

abdn$incom\_contact\_match[complete.cases(abdn$Incoming\_Phone)] = 1 \* incom\_contact\_match

abdn$contact\_incom\_match= 0

abdn$contact\_incom\_match[complete.cases(abdn$Contact\_Phone)] = 1 \* contact\_incom\_match

# Logical selection for matching records for those who purchased

abdn$pur = 1 \* (abdn$callerid\_match | abdn$emailmatch | abdn$incom\_match |abdn$contactmatch | abdn$incom\_contact\_match | abdn$contact\_incom\_match)

# Create additional columns for analyses

abdn$email = 1 \* complete.cases(abdn$Email)

abdn$state = 1 \* complete.cases(abdn$Address)

abdn$treat = ifelse(abdn$Test\_Control == "test", 1, 0)

tab = table(abdn$pur, abdn$treat)

# Adding row labels for 'Outcome'

rownames(tab) = c("Not Purchased", "Purchased")

# Adding column labels for 'Treatment'

colnames(tab) = c("Control Group", "Treatment Group")

print(tab)

unmatchable\_abdn <- abdn[abdn$pur == 0, ]

print(unmatchable\_abdn)

#Dropping unmatched records and selecting only those that matched (purchased)

abdn\_match <- abdn[abdn$pur == 1, ]

#Cross tabulations for all records(purchased and not purchased)

tab = table(abdn$pur, abdn$treat)

rownames(tab) = c("Not Purchased", "Purchased")

# Add column labels for 'Outcome'

colnames(tab) = c("Control Group", "Treatment Group")

print(tab)

all\_states = abdn$Address[!is.na(abdn$Address)]

set.seed(123) # Setting a seed for reproducibility

random\_states = sample(all\_states, 5)

cross\_tabulations = list()

for (state in random\_states) {

subset\_data = abdn[abdn$Address == state, ]

cross\_tabulation = table(subset\_data$pur, subset\_data$treat)

rownames( cross\_tabulation ) = c("Not Purchased", "Purchased")

colnames( cross\_tabulation ) = c("Control Group", "Treatment Group")

cross\_tabulations[[state]] = cross\_tabulation

}

# Print the cross-tabulations

for (state in random\_states) {

cat("Cross-tabulation for State:", state, "\n")

print(cross\_tabulations[[state]])

cat("\n")

}

#Cleaning dataset

# Remove multiple columns

abdnclean = abdn %>%

select(-(2:17))

#Changing index of columns and their column names

abdnclean = abdnclean %>%

select(1, 5, 2, 4,3:ncol(abdnclean))

colnames(abdnclean) = c("Customer\_ID", "Test\_Group","Outcome","State\_Available",

"Email\_Available")

#Exporting the clean data set as a csv

write.csv(abdnclean, file = "abdnclean.csv", row.names = FALSE)

#Statistical tests

# Run regression analyses

out1 = lm(Outcome ~ Test\_Group, data = abdnclean)

summary(out1)

out2 = aov(Outcome ~ Test\_Group, data = abdnclean)

summary(out2)

out3 = lm(Outcome ~ Test\_Group + State\_Available + Email\_Available , data = abdnclean)

summary(out3)

out4 = lm(Outcome ~ Test\_Group + State\_Available + Email\_Available + State\_Available\*Test\_Group + Email\_Available\*Test\_Group, data = abdnclean)

summary(out4)

#optional : logistic model

logmodel = glm(Outcome ~ Test\_Group + State\_Available + Email\_Available + State\_Available\*Test\_Group + Email\_Available\*Test\_Group ,family = binomial(link="logit"), data = abdnclean)

summary(logmodel)

# Generate summary table

stargazer(out1,out3,out4,logmodel, type = "text")