The Advertising Campaign Analysis

Venkata Satvik Reddy Tanuboddi

2023-10-25

Imports

```
library(dplyr)

library(rio)

library(rmarkdown)
```

Import data sets

```
abd <- read.csv("Abandoned.csv", header = TRUE, stringsAsFactors = FALSE)
rs <- read.csv("Reservation.csv", header = TRUE, stringsAsFactors = FALSE)

cat("Shape of data in Abandoned\t=\t", dim(abd), "\n")

## Shape of data in Abandoned   =    8442 12

cat("Shape of data in Reservation\t=\t", dim(rs), "\n")

## Shape of data in Reservation =    20814 12
```

1.Business Justification

1.1 Why retargeting customers who initially didn't buy a package makes business sense:

Prior Engagement: Retargeting customers who showed interest but didn't purchase capitalizes on previous marketing efforts. By focusing on these individuals, a business can further leverage its prior investment in marketing.

Warm Leads: These individuals have already demonstrated interest, making them warmer leads compared to entirely new prospects. Warm leads are often more responsive than cold audiences since they've shown some level of interest in the past.

Potential for Higher ROI: Retargeting can lead to higher conversion rates, which can significantly boost the return on investment (ROI). Reminding these customers of the product can provide the necessary nudge they need to finally make a purchase.

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

We use the following to asses the division: -

Balance: A well-executed test/control division should be balanced in size, meaning that the number of customers in the test group should be roughly equal to the number in the control group.

```
balance_check <- table(abd$Test_Control)
print("Split of control and test group")

## [1] "Split of control and test group"

print(100 * balance_check / sum(balance_check))

##
##  control     test
## 49.46695 50.53305
```

The control and test are approximately equally split.

```
state_distribution <- table(abd$Address, abd$Test_Control)
print(state_distribution)

##
##      control test
##         2321 2309
##   AK      32   29
##   AL      42   38
##   AR      46   38
##   AZ      44   54
##   CA      37   48
##   CO      37   40
##   CT      33   42
##   DE      46   34
##   FL      37   38
##   GA      33   47
##   HI      39   40
##   IA      34   39
##   ID      28   32
##   IL      47   37
##   IN      35   29
##   KS      41   37
##   KY      33   33
##   LA      36   39
##   MA      34   36
```

```
##   MD      40   38
##   ME      42   32
##   MI      33   43
##   MN      24   45
##   MO      31   43
##   MS      33   32
##   MT      34   35
##   NC      35   35
##   ND      36   26
##   NE      45   33
##   NH      44   28
##   NJ      36   52
##   NM      36   41
##   NV      53   45
##   NY      36   40
##   OH      39   50
##   OK      38   33
##   OR      39   39
##   PA      35   56
##   RI      28   41
##   SC      29   44
##   SD      35   38
##   TN      41   40
##   TX      33   44
##   UT      33   27
##   VA      32   49
##   VT      35   46
##   WA      37   34
##   WI      43   31
##   WV      50   47
##   WY      36   40
```

Randomness: Random assignment ensures that there's no inherent bias in the selection pr ocess. By looking at the distribution across states for both groups, we can assess the rando mness:

For most states, the division between test and control seems fairly even, indicating good ra ndomization. However, certain states like AZ, MN, NJ, PA, and VT have noticeable discrepan cies. This could either be due to the randomness inherent in the selection process or sugges t potential issues in the assignment methodology. It might be worth investigating further if these discrepancies are statistically significant.

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

The summarized count indicates the number of participants in each segment. For instance:

In Alaska (AK), 32 participants are in the control group and 29 in the test group.

In Alabama (AL), 42 participants are in the control group and 38 in the test group.

It's also notable that there are entries with an empty string as the Address, representing 2321 in control and 2309 in the test group. These might be missing data or could represent a segment of the population without specified state information. Proper investigation is required to understand the nature of these entries, and decisions should be made based on the context (e.g., filtering out, imputation, etc.).

Overall, from a business perspective, the strategy of retargeting makes sound sense. From a statistical perspective, while the division between the test and control groups is well-balanced, further investigation into the state-wise distribution can ensure there's no inherent bias or issue in the randomization process.

```
abd %>%
  group_by(Address, Test_Control) %>%
  summarise(Count = n())
```

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

```
## # A tibble: 102 × 3
## # Groups:   Address [51]
##    Address Test_Control Count
##    <chr>   <chr>        <int>
##  1 ""      control       2321
##  2 ""      test          2309
##  3 "AK"    control         32
##  4 "AK"    test            29
##  5 "AL"    control         42
##  6 "AL"    test            38
##  7 "AR"    control         46
##  8 "AR"    test            38
##  9 "AZ"    control         44
## 10 "AZ"    test            54
## # ℹ 92 more rows
```

2.Data Alignment

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

From an examination of both files, the following fields can be potential matching keys:

Incoming_Phone: This column contains the phone numbers from which calls were made. Matching phone numbers across datasets can help identify the same customer.

Contact_Phone: This column contains phone numbers that might be different from the incoming phone but serve as an alternative contact. This can also be used to match customers, especially if there is a scenario where customers might use different phones for incoming and contact purposes.

Email: Email addresses are unique identifiers for individuals in most scenarios. Matching email addresses from both datasets would provide a strong indication of the same customer.

It's essential to approach these potential keys with caution. For instance, the presence of empty strings or invalid email addresses should be accounted for to avoid false matches. The code has also taken steps to match across different phone number columns (e.g., Incoming_Phone from one dataset with Contact_Phone from another), thereby increasing the chances of identifying the same customer across datasets.

2.2 Detail your procedure to identify customers in:

The following R code is used to identify matches:

First, we identify matches for each of the potential matching keys (Incoming_Phone, Contact_Phone, and Email) across both datasets.

We then create new binary columns in the 'abd' dataset to indicate whether a match was found for each of these keys.

For categorizing the observations based on purchasing behavior and their group (test or control), we use the Test_Control and pur columns.

```
emails = intersect(rs[,c('Email')],abd[,c('Email')])
emails = emails[emails != ""]

incoming_phone = intersect(rs[,c('Incoming_Phone')],abd[,c('Incoming_Phone')]
)
incoming_phone = incoming_phone[incoming_phone != ""]
contact_phone = intersect(rs[,c('Contact_Phone')],abd[,c('Contact_Phone')])
contact_phone = contact_phone[contact_phone != ""]
incoming_Contact = intersect(rs[,c('Incoming_Phone')],abd[,c('Contact_Phone')
])
incoming_Contact = incoming_Contact[incoming_Contact != ""]
Contact_Incoming = intersect(rs[,c('Contact_Phone')],abd[,c('Incoming_Phone')
])
Contact_Incoming = Contact_Incoming[Contact_Incoming != ""]
```

```
abd$match_email <- 0
abd$match_email = 1 * (abd$Email %in% emails)

abd$match_incoming_phone <- 0
abd$match_incoming_phone = 1 * (abd$Incoming_Phone %in% incoming_phone)

abd$match_contact_phone <- 0
abd$match_contact_phone = 1 * (abd$Contact_Phone %in% contact_phone)

abd$match_incoming_contact <- 0
abd$match_incoming_contact = 1 * (abd$Contact_Phone %in% incoming_Contact)

abd$match_contact_incoming <- 0
abd$match_contact_incoming = 1 * (abd$Incoming_Phone %in% Contact_Incoming)


abd$pur = 1 * (abd$match_email | abd$match_incoming_phone | abd$match_contact
_phone | abd$match_incoming_contact | abd$match_contact_incoming)
```

• Treatment group who purchased.

```
group_condition <- abd$Test_Control == "test"
purchase_condition <- abd$pur == 1
test_purchase <- subset(abd, group_condition & purchase_condition)
dim(test_purchase)

## [1] 345  18
```

• Treatment group who didn't purchase.

```
group_condition <- abd$Test_Control == "test"
purchase_condition <- abd$pur != 1
test_no_purchase <- subset(abd, group_condition & purchase_condition)
dim(test_no_purchase)

## [1] 3921   18
```

• Control group who purchased.

```
group_condition <- abd$Test_Control == "control"
purchase_condition <- abd$pur == 1
control_purchase <- subset(abd, group_condition & purchase_condition)
dim(control_purchase)

## [1] 93 18
```

• Control group who didn't purchase.

```
group_condition <- abd$Test_Control == "control"
purchase_condition <- abd$pur != 1
```

```
control_no_purchase <- subset(abd, group_condition & purchase_condition)
dim(control_no_purchase)

## [1] 4083   18
```

- Treatment group who purchased: Filter on treat == 1 and purchased == 1.
- Treatment group who didn't purchase: Filter on treat == 1 and purchased == 0.
- Control group who purchased: Filter on treat == 0 and purchased == 1.
- Control group who didn't purchase: Filter on treat == 0 and purchased == 0.

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

Yes, there were unmatchable records. Unmatchable records are nothing but ones from the abandoned who never returned to make a pruchase/reservation.

```
unmatachable <- subset(abd, abd$pur == 0)
cat("Number of unmatched records = ", dim(unmatachable)[1], "\n")

## Number of unmatched records =  8004
```

Example of them are

```
head(unmatachable)

##          Caller_ID             Session First_Name Last_Name Street City Ad
dress
## 1 68359340ZFZYECVJ 2014.01.06 04:09:15     Bertha
## 2 83119994SISMLSRO 2014.01.06 04:15:43       Kyle
## 3 58448995NDCOTARU 2014.01.06 04:20:05     Paxton
## 4 84112006ELLKIDBK 2014.01.06 04:25:05     Thelma
## 5 10840705OUMWVGBO 2014.01.06 04:30:35      Lorna
## 6 65443966UFFCGEYN 2014.01.06 04:36:22      Leann
##   Zipcode Email Incoming_Phone  Contact_Phone Test_Control match_email
## 1               (864)-004-6354 (864)-004-6354         test           0
## 2               (703)-220-0148 (703)-220-0148      control           0
## 3               (559)-299-7745 (559)-299-7745      control           0
## 4               (636)-611-4439 (636)-611-4439         test           0
## 5               (253)-461-5118 (253)-461-5118      control           0
## 6               (407)-910-9280 (407)-910-9280         test           0
##   match_incoming_phone match_contact_phone match_incoming_contact
## 1                    0                   0                      0
## 2                    0                   0                      0
## 3                    0                   0                      0
## 4                    0                   0                      0
## 5                    0                   0                      0
## 6                    0                   0                      0
##   match_contact_incoming pur
## 1                      0   0
## 2                      0   0
## 3                      0   0
```

```
## 4                       0   0
## 5                       0   0
## 6                       0   0
```

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

```
cross_table_outcomes <- table(abd$Test_Control, abd$pur)
colnames(cross_table_outcomes) <- c("No Purchase", "Purchase")

# Print the cross-tabulation
print(cross_table_outcomes)
```

```
##
##           No Purchase Purchase
##   control        4083       93
##   test           3921      345
```

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

```
set.seed(1000)
random_states <- sample(unique(abd$Address), 5)
for (state in random_states)
{
  cat(state, "\n")
  state_subset <- subset(abd, abd$Address == state)
  cross_table_outcomes <- table(state_subset$Test_Control, state_subset$pur)
  colnames(cross_table_outcomes) <- c("No Purchase", "Purchase")
  # Print the cross-tabulation
  print(cross_table_outcomes)
}
```

```
## MS
##
##           No Purchase Purchase
##   control          31        2
##   test             29        3
## NV
##
##           No Purchase Purchase
##   control          51        2
##   test             40        5
## OK
##
##           No Purchase Purchase
##   control          38        0
##   test             32        1
## AZ
##
##           No Purchase Purchase
##   control          43        1
```

```
##   test              51        3
## MT
##
##           No Purchase Purchase
##   control          34        0
##   test             32        3
```

3.Data Refinement

To generate a refined dataset, we will be selecting a subset of columns from the 'abd' dataset and rename them as per the requirements. Our cleaned dataset will consist of columns:

Customer ID: To represent each unique customer. It seems the 'Caller_ID' from the 'abd' dataset can serve as a unique identifier for each customer.

Test Group: To represent whether a customer is in the control or test group.

Outcome: To show if a purchase was made (1 for a purchase, 0 otherwise).

State Available: A binary column indicating whether a state address is available (1 if yes, 0 otherwise).

Email Available: A binary column indicating whether an email is available (1 if yes, 0 other wise).

```
cleaned_dataset <- abd[, c("Test_Control", "pur", "Email", "Address")]
# cleaned_dataset = subset(cleaned_dataset,cleaned_dataset$Email!='' & cleane
d_dataset$Address!='')
head(cleaned_dataset)
```

```
##   Test_Control pur Email Address
## 1         test   0
## 2      control   0
## 3      control   0
## 4         test   0
## 5      control   0
## 6         test   0
```

```
write.csv(cleaned_dataset, file = './cleaned.csv', row.names = FALSE)
```

---

4. Statistical Assessment

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

We performed a linear regression with the formula: Outcome = α + β * Test Group + error.

```
model1 <- lm(formula = pur ~ Test_Control, data = cleaned_dataset)
summary(model1)

##
## Call:
## lm(formula = pur ~ Test_Control, data = cleaned_dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08087 -0.08087 -0.02227 -0.02227  0.97773
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.022270   0.003402   6.545 6.28e-11 ***
## Test_Controltest 0.058602   0.004786  12.244  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2199 on 8440 degrees of freedom
## Multiple R-squared:  0.01745,    Adjusted R-squared:  0.01733
## F-statistic: 149.9 on 1 and 8440 DF,  p-value: < 2.2e-16
```

Results:

The linear regression model attempted to predict pur (purchase outcome) based on whether a customer was part of the Test_Control group.

Intercept: The estimated intercept is 0.02227. This is the predicted mean value of pur when all predictors are set to zero. Given the nature of categorical predictors, it represents the mean outcome for the reference group (likely the control group).

Test_Controltest: The coefficient of 0.058602 represents the average difference in the pur value between the test group and the control group. It indicates that, on average, being in the test group increases the outcome by about 0.0586 units compared to being in the control group. The t-value of 12.244 and an extremely low p-value (< 2e-16) means this effect is highly statistically significant.

R-squared: Only 1.745% of the variability in the outcome (pur) is explained by the predictor Test_Control. This suggests that while the test group assignment has a statistically significant effect on the outcome, it only explains a small fraction of its variability. Other factors not included in the model might play a significant role.

F-statistic: The F-statistic of 149.9 with a very low p-value indicates that the model with Test_Control as a predictor fits the data significantly better than a model with no predictors.

The linear regression results suggest that being in the test group has a significant positive effect on the likelihood of a purchase. However, the overall effect size is relatively small, and the R-squared value suggests that other factors might also play a crucial role in determining the purchase outcome. Further analyses or inclusion of additional predictors might provide more insights.

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

```
annova <- aov(pur ~ Test_Control, cleaned_dataset)
summary(annova)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## Test_Control    1    7.2   7.247   149.9 <2e-16 ***
## Residuals    8440  408.0   0.048
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Running an ANOVA on the same data, we get results that mirror those from the linear regre ssion. The p-value for Test_Control is extremely low (< 2e-16), and the F-statistic is 149.9, mirroring the results from the regression. This suggests a statistically significant difference in purchase outcomes between the two groups when analyzed through ANOVA. ANOVA can be considered a special case of linear regression when dealing with categorical predictors. Hence, it's not surprising that their results are in agreement.

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

Regression models can be a powerful tool in identifying relationships between variables, but inferring causation requires careful consideration. While the significant p-value from the regression suggests that the retargeting campaign might be influencing purchase outcomes, other potential confounders might be at play. To truly establish causation, one would need to ensure that there's random assignment to test and control groups, no unmeasured confounding variables, and consider the possibility of reverse causation or other biases.

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.

Address and Email dummies were integrated into the regression:

```
cleaned_dataset$Address_dummy <- ifelse(cleaned_dataset$Address == "", 0, 1)
cleaned_dataset$Email_dummy <- ifelse(cleaned_dataset$Email == "", 0, 1)
model2 <- lm(pur ~ Test_Control + Address_dummy + Email_dummy, cleaned_datase
t)
summary(model2)
```

```
##
## Call:
## lm(formula = pur ~ Test_Control + Address_dummy + Email_dummy,
##     data = cleaned_dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.12161 -0.06833 -0.06399 -0.01070  0.98930
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.010703   0.004023   2.661 0.007814 **
## Test_Controltest 0.057623   0.004777  12.064  < 2e-16 ***
## Address_dummy    0.016873   0.004921   3.429 0.000609 ***
## Email_dummy      0.036416   0.007485   4.865 1.16e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2193 on 8438 degrees of freedom
## Multiple R-squared:  0.02268,    Adjusted R-squared:  0.02233
## F-statistic: 65.26 on 3 and 8438 DF,  p-value: < 2.2e-16
```

Results:

Test_Controltest has a coefficient of 0.057623, and remains statistically significant with an extremely low p-value (< 2e-16). This means, even after controlling for the other variables, being in the test group has a significant effect on the purchase outcome.

Address_dummy (indicating the presence of a state address) has a positive coefficient of 0.016873 and is statistically significant, suggesting that those with an address have a higher likelihood of making a purchase than those without.

Email_dummy (indicating the presence of an email) has a positive coefficient of 0.036416 and is highly significant, implying that having an email is strongly associated with a higher purchase outcome.

The R-squared value of this model (0.02268) is slightly higher than the previous one (0.01745), suggesting that the addition of Address and Email dummies helped in explaining more variability in the purchase outcomes.

Interpretation:

From the updated regression, it's evident that while the retargeting campaign has a statisti cally significant effect on purchase outcomes, other factors like the presence of an email or address also play an important role. In particular, having an email seems to be a strong pre dictor of purchase outcomes. This could be due to a variety of reasons, such as the possibilit y that customers with registered emails receive more personalized marketing or offers, or t hey might be more engaged with the brand. This underscores the importance of not just foc using on a single marketing effort (like the retargeting campaign) but also considering othe r customer touchpoints and data points when evaluating marketing effectiveness.

5. Reflections

5.1. Reflect on the project:

Experiment Design Modification: Given the chance, I would consider modifying the experiment design to ensure that all potential confounding variables are accounted for. The current analysis shows that having an email is a significant predictor for purchase outcomes, which indicates that customer engagement outside of the retargeting campaign might play a crucial role in influencing purchases. A more thorough experimental design could involve stratifying the test and control groups based on the presence of an email to better isolate the impact of the retargeting campaign.

Alternative Paths with Better-Quality Data: With access to better-quality data, we could explore additional variables that might influence purchase behavior, such as customer demographic information, past purchase history, or engagement with previous marketing campaigns. Incorporating these variables into the analysis could provide a more comprehensive view of what drives purchases and help to identify specific customer segments that are more responsive to retargeting campaigns.

Actionable Business Implications: The analysis provides several actionable insights for the business. The lack of a statistically significant impact of the retargeting campaign on purchase outcomes suggests that the current approach might need to be re-evaluated. The significant impact of having an email implies that customer engagement through email marketing could be a fruitful area to explore further. The business might consider investing more resources in building their email list and developing targeted email marketing campaigns to drive purchases.

5.2. Self-assessment:

Effort: I would rate my effort on this project at 90 out of 100. I diligently worked through the dataset, cleaned it, and conducted a comprehensive statistical analysis to understand the impact of the retargeting campaign. I ensured my interpretation of the results was thorough and tried to extract meaningful business implications from the data.

Anticipated Performance: I anticipate that my performance on this project would be around 90 - 100 out of 100. I believe I have conducted a robust analysis, but there is always room for improvement, particularly in exploring additional variables, providing more context, and enhancing the experiment design.

Collaborations:

completed the majority of this project independently. However, when I encountered challe nging aspects, I sought guidance from my professor to ensure clarity and accuracy in my ap proach. All interpretations and coding were primarily done by me, reflecting my understan ding of the data and results. In addition to this, I made sure to consult relevant R document ation and statistical analysis resources to further solidify the validity of my findings.