**Naramsetty Easwar Manikanta Srinivas – U02793943  
MIDTERM PROJECT**

**I. The Business Problem**

ABD contains data for all the customers in the dataset that were already pursued (advertised) but did not buy a vacation package.

Business Problem: Should we retarget those customers?

**Q1:** In light of your experience as a businesswoman/man, argue why this is a sensible business question.

Yes, this is a sensible business question as customer acquisition depends on various factors. Some of those factors can be the time they receive the call i.e during the month end people may run short of money to spend on vacation or they might have other unavoidable plans. Another major challenge commoners find issue with, is the customer service and their awareness of package offered by the firm based on seasonality. Advertising plays a major role, and it adds more budget to the companies’ balance sheet, so every company must be aware of their return on investment.

The more effective way would be like targeting the customers whose income is above a certain level or the customer's who showed interest during the first conversation and that can be monitored by checking the call duration.

An experiment is run where customers in the abandoned dataset are randomly placed in a treatment or a control group (see column L in both files).

Those marked as “test” are retargeted (treated); the others marked as control are part of the control group.

**Q2:** Investigate the test/control variable. Does the experiment seem to be run properly?

> table(abd$Test\_Control)

control test

4176 4266

From the above result we can say that the experiment was random and has been run properly without any bias.

**Q3:** compute the same summary statistics for this Test\_variable by stratifying on States (meaning considering only the entries with known “State”), wherever this information is available.

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

> known\_states <- abd[complete.cases(abd['Address']),]

> table(known\_states$Test\_Control)

control test

1855 1957

From the above table also, we can say that the experiment was random

**II. Data Matching**

About three months later, the experiment/retargeting campaign is over.

Customers, presented in the ABD excel file, who bought vacation packages during the time frame, are recorded in the RS excel file.

**Q5:** After observing the data in both files, argue that customers can be matched across some “data keys” (column labels). Correctly identify all these data keys (feel free to add a few clarifying examples if needed)

Yes, customers can be matched using the Email, Incoming\_Phone, Contact\_Phone and also by combining the First\_Name and Last\_Name thereby forming a single field named as First\_Last\_Name.

Example 1: Matching using First\_Name and Last\_Name fields by combining into one field.

abd<-read.csv("C:/Users/manik/Desktop/QMB\_Class/Project/Abandoned.csv")

res<-read.csv("C:/Users/manik/Desktop/QMB\_Class/Project/Reservation.csv")

abd$First\_Name[abd$First\_Name==""] <- 0

abd$Last\_Name[abd$Last\_Name==""] <- 0

abd$First\_Last\_Name <- paste(abd$First\_Name,abd$Last\_Name,sep = "\_")

res$First\_Name[res$First\_Name==""] <- 0

res$Last\_Name[res$Last\_Name==""] <- 0

res$First\_Last\_Name <- paste(res$First\_Name,res$Last\_Name,sep = "\_")

match\_name=abd$First\_Last\_Name %in% res$First\_Last\_Name

> table(match\_name)

match\_name

FALSE TRUE

7446 996

From the above table we can conclude that there are 996 matches while comparing the abandoned and reservation files by combining First\_Name and Last\_Name columns into one column namely First\_Last\_Name i.e 996 customers have purchased during the campaign.

Example 2: Matching using the field Email.

abd<-read.csv("C:/Users/manik/Desktop/QMB\_Class/Project/Abandoned.csv")

res<-read.csv("C:/Users/manik/Desktop/QMB\_Class/Project/Reservation.csv")

abd[abd==""] <- NA

res[res==""] <- NA

match\_email=abd$Email[complete.cases(abd$Email)] %in% res$Email[complete.cases(res$Email)]

> table(match\_email)

match\_email

FALSE TRUE

955 75

From the above table we can say that there are 75 matches in reservation file when matched with field Email.

Example 3: Matching using the field Incoming\_Phone

match\_incoming=abd$Incoming\_Phone[complete.cases(abd$Incoming\_Phone)] %in% res$Incoming\_Phone[complete.cases(res$Incoming\_Phone)]

> table(match\_incoming)

match\_incoming

FALSE TRUE

6935 327

From the above table we can say that there are 327 matches in the reservation table when compared using Incoming\_Phone field.

**Q6**: EXTREMELY CAREFULLY DESCRIBE YOUR DATA MATCHING PROCEDURE to IDENTIFY: (1) Customers in the TREATMENT group who bought (2) Customers in the TREATMENT group who did not buy (3) Customers in the Control group who bought, and (4) Customers in the Control group who did not buy. Be as precise as possible.

**For TREATMENT GROUP**

**test\_group<-subset(abd,abd$Test\_Control=="test")**

**match\_email=test\_group$Email[complete.cases(test\_group$Email)] %in% res$Email[complete.cases(res$Email)]**

**match\_incoming=test\_group$Incoming\_Phone[complete.cases(test\_group$Incoming\_Phone)] %in% res$Incoming\_Phone[complete.cases(res$Incoming\_Phone)]**

**match\_contact=test\_group$Contact\_Phone[complete.cases(test\_group$Contact\_Phone)] %in% res$Contact\_Phone[complete.cases(res$Contact\_Phone)]**

**match\_incoming\_contact= test\_group$Incoming\_Phone[complete.cases(test\_group$Incoming\_Phone)] %in% res$Contact\_Phone[complete.cases(res$Contact\_Phone)]**

**match\_contact\_incoming= test\_group$Contact\_Phone[complete.cases(test\_group$Contact\_Phone)] %in% res$Incoming\_Phone[complete.cases(res$Incoming\_Phone)]**

**test\_group$match\_email <-0**

**test\_group$match\_email[complete.cases(test\_group$Email)] <- 1\* match\_email**

**test\_group$match\_incoming <- 0**

**test\_group$match\_incoming[complete.cases(test\_group$Incoming\_Phone)] <- 1\* match\_incoming**

**test\_group$match\_contact <- 0**

**test\_group$match\_contact[complete.cases(test\_group$Contact\_Phone)] <- 1\* match\_contact**

**test\_group$match\_incoming\_contact <- 0**

**test\_group$match\_incoming\_contact[complete.cases(test\_group$Incoming\_Phone)] <- 1\* match\_incoming\_contact**

**test\_group$match\_contact\_incoming <- 0**

**test\_group$match\_contact\_incoming[complete.cases(test\_group$Contact\_Phone)] <- 1\* match\_contact\_incoming**

**test\_group$pur<-0**

**test\_group$pur <- 1\*(test\_group$match\_contact | test\_group$match\_contact\_incoming | test\_group$match\_email | test\_group$match\_incoming\_contact | test\_group$match\_incoming)**

**sum(test\_group$pur)**

**> sum(test\_group$pur)**

**[1] 345**

**Therefore 345 customers in the treatment group who have bought and 3921 customers who didn’t purchase.**

**FOR CONTROL GROUP**

**Upon doing similar analysis for the control group, we get the following results.**

**> sum(control\_group$pur)**

**[1] 93**

**Therefore 93 customers in the control group have purchased and 4083 customers didn’t buy.**

**Q7:** Are there problematic cases? i.e. data records not matchable? If so, provide a few examples and toss those cases out of the analysis.

There are high chances of having duplicate records i.e a person can provide different email or he can be contacted from different incoming mobile or there is also a chance of the person changing his contact number. If the person doesn’t provide his First or Last Name and does any one of the above mentioned cases there are high chances of having duplicate records.

**Q8: Complete the following cross-tabulation:**

|  |  |  |
| --- | --- | --- |
| **Group \ Outcome** | **Buy** | **No Buy** |
| **Treatment** | **345** | **3921** |
| **Control** | **93** | **4083** |

**Q9: Repeat Q8 for 5 randomly picked states. Report 5 different tables by specifying the states you “randomly picked”.**

**State 1: KS**

**state1 = subset(test\_group,test\_group$Address=="KS")**

**> nrow(state1)**

**[1] 37**

**> sum(state1$pur)**

**[1] 5**

**> state1c = subset(control\_group,control\_group$Address=="KS")**

**> nrow(state1c)**

**[1] 41**

**> sum(state1c$pur)**

**[1] 0**

|  |  |  |
| --- | --- | --- |
| **Group \ Outcome** | **Buy** | **No Buy** |
| **Treatment** | **5** | **32** |
| **Control** | **0** | **41** |

**State 2: TX**

**> state2 = subset(test\_group,test\_group$Address=="TX")**

**> nrow(state2)**

**[1] 44**

**> sum(state2$pur)**

**[1] 3**

**> state2c = subset(control\_group,control\_group$Address=="TX")**

**> nrow(state2c)**

**[1] 33**

**> sum(state2c$pur)**

**[1] 0**

|  |  |  |
| --- | --- | --- |
| **Group \ Outcome** | **Buy** | **No Buy** |
| **Treatment** | **3** | **41** |
| **Control** | **0** | **33** |

**State 3: FL**

**> state3 = subset(test\_group,test\_group$Address=="FL")**

**> nrow(state3)**

**[1] 38**

**> sum(state3$pur)**

**[1] 4**

**> state3c = subset(control\_group,control\_group$Address=="FL")**

**> nrow(state3c)**

**[1] 37**

**> sum(state3c$pur)**

**[1] 0**

|  |  |  |
| --- | --- | --- |
| **Group \ Outcome** | **Buy** | **No Buy** |
| **Treatment** | **4** | **34** |
| **Control** | **0** | **37** |

**State 4: AZ**

**> state4 = subset(test\_group,test\_group$Address=="AZ")**

**> nrow(state4)**

**[1] 54**

**> sum(state4$pur)**

**[1] 3**

**> state4c = subset(control\_group,control\_group$Address=="AZ")**

**> nrow(state4c)**

**[1] 44**

**> sum(state4c$pur)**

**[1] 1**

|  |  |  |
| --- | --- | --- |
| **Group \ Outcome** | **Buy** | **No Buy** |
| **Treatment** | **3** | **51** |
| **Control** | **1** | **43** |

**State 5: UT**

**> state5 = subset(test\_group,test\_group$Address=="UT")**

**> nrow(state5)**

**[1] 27**

**> sum(state5$pur)**

**[1] 4**

**> state5c = subset(control\_group,control\_group$Address=="UT")**

**> nrow(state5c)**

**[1] 33**

**> sum(state5c$pur)**

**[1] 3**

|  |  |  |
| --- | --- | --- |
| **Group \ Outcome** | **Buy** | **No Buy** |
| **Treatment** | **4** | **23** |
| **Control** | **3** | **30** |

**III. Data Cleaning:**

You have now identified all the relevant customers for the analysis and their outcome, and you also know if they are in a treated or in a control group.

Produce an Excel File with the following columns

Customer ID | Test Variable | Outcome | D\_State | D\_Email |

Where Test Variable indicates the treatment or the control group, the Outcome is a binary variable indicating whether a vacation package was ultimately bought. D\_State and D\_Email identify whether the information is present on file.

(Note that you should have as many rows as customers you were able to match across the two data sets. Be sure to attach this excel file to the submission for proper verification.)

**new\_df<-subset(abd,abd$pur==1)**

**new\_df$D\_Email<- ifelse(new\_df$Email != 'NA',1,0)**

**new\_df["D\_Email"][is.na(new\_df["D\_Email"])] <- 0**

**new\_df$D\_State<- ifelse(new\_df$Address != 'NA',1,0)**

**new\_df["D\_State"][is.na(new\_df["D\_State"])] <- 0**

**new\_file<- data.frame(**

**Customer\_ID=new\_df$Caller\_ID,**

**Test\_Variable =new\_df$Test\_Control,**

**Outcome = new\_df$pur,**

**D\_Email = new\_df$D\_Email,**

**D\_State = new\_df$D\_State**

**)**

**write\_xlsx(new\_file,"C:/Users/manik/Desktop/QMB\_Class/Project/Matched.xlsx")**

**IV. Statistical Analysis**

We are finally in a condition to try to answer the relevant business question.

**Q10:** Run a Linear regression model for

Outcome = alpha + beta \* Test\_Variable + error

And Report the output.

> linear\_model<-lm(abd$pur~abd$Test\_Control, data = abd)

> summary(linear\_model)

Call:

lm(formula = abd$pur ~ abd$Test\_Control, data = abd)

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 \*\*\*

abd$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

The p-value for the Test\_Control variable is less than 0.05, so we reject null hypothesis. We have enough evidence that the beta coefficients are not zero.

Purchase = 0.022270+0.058602\* Test\_Control

**Q11:** Argue this is statistically equivalent to an ANOVA/t-test.

> anova\_out=aov(abd$pur ~ abd$Test\_Control, data = abd)

> summary(anova\_out)

Df Sum Sq Mean Sq F value Pr(>F)

abd$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

The p-value associated with F-statistic is less than 0.05 so we reject the null hypothesis, which means that purchases made by treatment group are different from the purchases made by control group.

**Q12:** Argue whether this is a properly specified linear regression model, if so if we can draw any causal statement about the effectiveness of the retargeting campaign. Is this statistically significant?

Besides the t-statistic and p-value, this is our most important metric for measuring regression model fit. R² measures the linear relationship between our predictor variable (test control) and our response / target variable (Purchase). It always lies between 0 and 1. A number near 0 represents a regression that does not explain the variance in the response variable well and a number close to 1 does explain the observed variance in the response variable. In our experiment, the adjusted R² (which adjusts for degrees of freedom) is 0.01733% of an increase in purchase can be explained. If we perform a multiple regression, we will find that the R² will increase with an increase in the number of response variables. So, in the next question, the extra response variables are added to observe the change.

**Q13:** Now add the dummies for State and Emails to the regression model. Also consider including interactions with the treatment. Report the outcome and comment on the results. (You can compare with Q9)

> model<- lm(abd$pur ~ abd$Test\_Control + abd$email+ abd$state, data = abd)

> summary(model)

Call:

lm(formula = abd$pur ~ abd$Test\_Control + abd$email + abd$state,

data = abd)

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 \*\*

abd$Test\_Controltest 0.057623 0.004777 12.064 < 2e-16 \*\*\*

abd$email 0.036416 0.007485 4.865 1.16e-06 \*\*\*

abd$state 0.016873 0.004921 3.429 0.000609 \*\*\*

---

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

From the above results the p-value is less than 0.05 so we reject null hypothesis. We have enough evidence that the beta coefficients are not zero.

> multi\_linear\_model <- lm(pur~Test\_Control \* email +Test\_Control \* state, data=abd)

> summary(multi\_linear\_model)

Call:

lm(formula = pur ~ Test\_Control \* email + Test\_Control \* state,

data = abd)

Residuals:

Min 1Q Median 3Q Max

-0.14608 -0.06218 -0.03474 -0.01684 0.98316

Coefficients:

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

(Intercept) 0.016844 0.004583 3.675 0.000239 \*\*\*

Test\_Controltest 0.045338 0.006494 6.981 3.15e-12 \*\*\*

email 0.007599 0.011016 0.690 0.490343

state 0.010301 0.006987 1.474 0.140443

Test\_Controltest:email 0.052981 0.015004 3.531 0.000416 \*\*\*

Test\_Controltest:state 0.013011 0.009833 1.323 0.185810

---

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

Residual standard error: 0.2191 on 8436 degrees of freedom

Multiple R-squared: 0.02466, Adjusted R-squared: 0.02408

F-statistic: 42.65 on 5 and 8436 DF, p-value: < 2.2e-16

library(stargazer)

stargazer(linear\_model,model,multi\_linear\_model,type = "html",out="midterm.htm")

**VI: Conclusion**

**Q14: Lesson Learned. What would you have done differently in designing the experiment? Any other directions you could have taken with better data? Are there any prescriptive managerial implications of this study? Please answer briefly**

As mentioned earlier, fields like Call duration would be effective to retarget using which we can predict the interest of the customer to purchase. This will help us in shortening the treatment group and thereby reducing the money involved in campaign. The prescriptive implication which I find is the creation of treatment group. Creating it randomly is not the best way to achieve the desired outcome.

**Q15: Self-evaluation. Please score your effort on a scale 0-100. Please score your expected performance on the same scale. Add comments if necessary, including whether you collaborate with your peers.**

Professor mentioned that this case study would take around 20 hours or so to complete which is approximately true in my case. I have spent most of the time to analyze and match the data across the files. I would say my efforts are more than 100%.

There was no need of collaboration with any of my peers because the professor already gave a gist of the assignment along with how to analyze the result. With this, I have learned how to influence the power of R and statistics to deal with real business problems.