Curating best marketing strategy for term-deposit renewals

Group 5:Archit Agarwal, Sarasija Gaddameedi, Sourabh Koul, Jingwen (Echo) Pei

5/7/2022

Problem Statement

As data scientist with the marketing division at a multinational bank we wanted to help the business by identifying the optimum marketing channel (including telephone and cellphone) and communication frequency to improve loan renewal rates. The marketing team is pursuing customers to up-sell term-loans and wants to identify the best channels and frequency to approach the customer for the same. The team has data from both channels, i.e., The Telephone and The cellphone. Further the team wanted to see the impact of changing the frequency of contact by either channels on the customer term-deposit renewal.

Executive Summary

We conducted statistical experimentation to determine the impact of change in marketing channels from telephone to a cellphone. Our analysis showed that if we reach a customer via cellphone, the chances of term deposit renewal are 24% more than the traditional telephone method. Further, we identified that if we get the customer more than four times, the chances of term-deposit renewals start decreasing.

Based on our analysis, we would recommend the marketing teams design future marketing campaigns using a cellphone as a mode of communication and limit the frequency of the call to 4 during the campaign period, maximizing the chances of term-deposit renewals.

Detailed Analysis and Methodology:

Dataset/Measures

We extracted the data set from UCI Repository which contains information about the marketing efforts invested by the bank in last campaign.

```
data <- read.csv('C:/Users/archi/Downloads/bank/bank-additional (1)/bank-additional/bank-additional-full
```

It consists of customer demographics such as age, marital and employment status. Also, it contains information about past marketing features such as marketing channels and last engagement time. Moreover, it has data about the general economic indicators, such as employment rate and consumer price index.

For our analysis, the unit of observation is a single customer for whom we are considering the campaign period. We don't have the start or end date of the campaign. Our variable of interest is a renewal of term deposits. Next, we explored the distribution of the same.

Exploratory Data Analysis

```
### Libraries to use
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)
```

```
## Warning: package 'ggplot2' was built under R version 4.1.2
```

```
library(splitstackshape)
```

```
## Warning: package 'splitstackshape' was built under R version 4.1.3
```

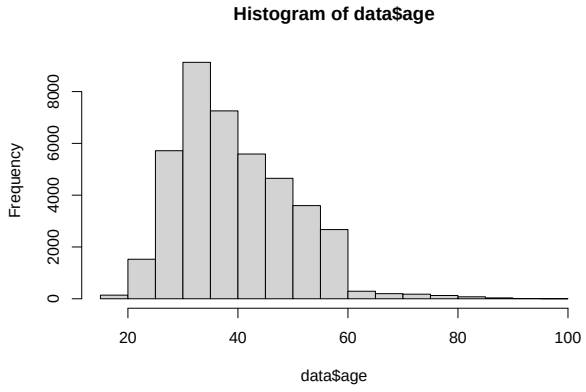
```
library(MatchIt)
```

```
## Warning: package 'MatchIt' was built under R version 4.1.2
```

```
set.seed(1)
```

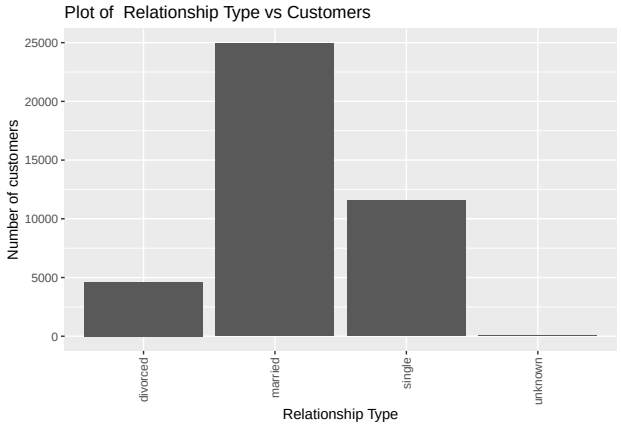
The first feature of interest is "Age" as it is directly affects customer's decision to renew a term loan or not. A aged person is expected to invest more as compared to young person.

```
hist(data$age)
```



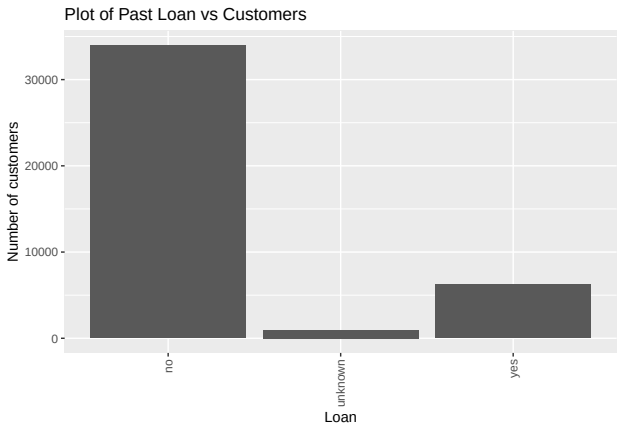
Next, we analyzed the family status of the customer. We aggregated the customer family status by combining the divorced, single and unknown into a single bucket of single, the married status remained same.

```
ggplot(data, aes(x=marital)) + geom_bar(stat='count') +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
  ggtitle("Plot of  Relationship Type vs Customers") +
  xlab("Relationship Type") + ylab("Number of customers")
```



The customer who have defaulted have less chances of renewal, and they might not be an ideal customer for the bank as well. Similarly, a customer who has taken a loan with the bank is more likely to make a deposit as compared to the customer who haven't.

```
ggplot(data, aes(x=loan)) + geom_bar(stat='count') +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
  ggtitle("Plot of Past Loan vs Customers") +
  xlab("Loan") + ylab("Number of customers")
```



Data Transformations:

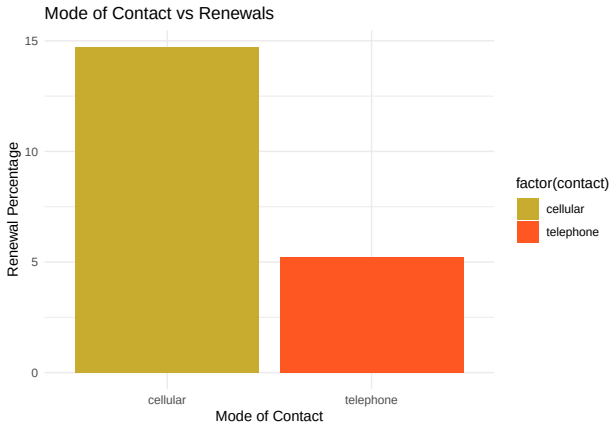
Based on our exploration, we reduced the imbalance in the dataset by combining various fields.

```
data <- data %>% mutate(employed = ifelse(job == 'unemployed',0,1),
                        relationship = ifelse(marital== 'married',1,0),
                        past_default = ifelse(default == 'yes', 1,0),
                        house_owners = ifelse(default == "yes",1,0),
                        past_loan = ifelse(previous == 0, 0, 1),
                        treatment = ifelse(contact == 'cellular', 1, 0),
                        result = ifelse(y == 'yes', 1, 0))
```

Post one-variable data distribution analysis, we performed the two variable analysis. Following visualizations will help identifying the impact of the each mode of communication on renewals, which is a major factor in improving the marketing strategy.

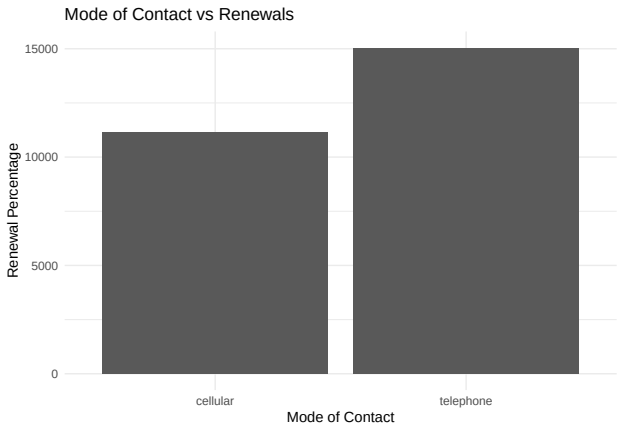
```
temp_2 <- data %>% group_by(contact) %>% summarise(count=n(),renewals = sum(result))
temp_2 <- temp_2 %>% mutate(new = 100*renewals/count)
```

```
ggplot(temp_2, aes(x=contact, y=new, fill=factor(contact))) + geom_bar(stat='identity') +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
  scale_fill_manual(values=c("#c7ac30","#ff5722")) +
  ggtitle("Mode of Contact vs Renewals") +
  xlab("Mode of Contact") + ylab("Renewal Percentage") +
  theme_minimal()
```



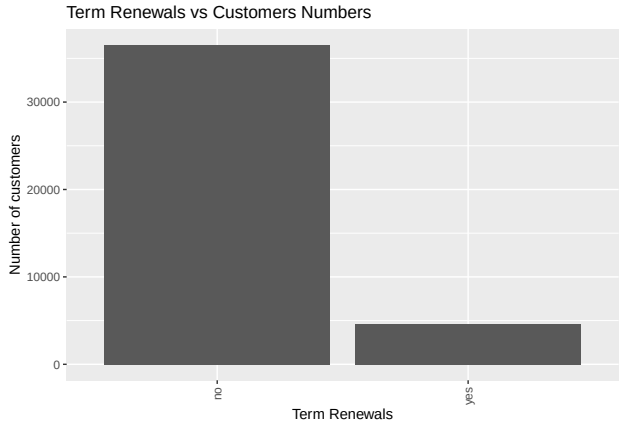
We can see that the customers on the cellular are more likely to renew a term deposit but counterintiutively the bank is using the telephone as major channel of communication. Following graph illustrates the count of customers targeted via each channel.

```
temp3 <- data %>% group_by(contact) %>% summarise(new = n())
## Removing Null values
temp3 <- temp3 %>% mutate(new = ifelse(contact=="cellular",new-15000,new))
ggplot(temp3, aes(x=contact, y=new )) + geom_bar(stat='identity') +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
  scale_fill_manual(values=c("#c7ac30","#ff5722")) +
  ggtitle("Mode of Contact vs Renewals") +
  xlab("Mode of Contact") + ylab("Renewal Percentage") +
  theme_minimal()
```



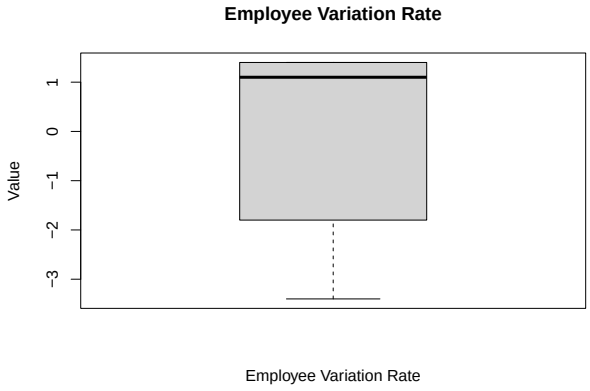
We also observed that a small fraction of users renew the term deposits.

```
ggplot(data, aes(x=y)) + geom_bar(stat='count') +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
  ggtitle("Term Renewals vs Customers Numbers") +
  xlab("Term Renewals") + ylab("Number of customers")
```



Other possible confounds that impact customer decision to term-deposit renewals are federal economic factors such as employment variation rate. During the course of the campaign, the employee variation rate changed which might have impacted customers' employment too.

```
boxplot(data$emp.var.rate, main="Employee Variation Rate",
   xlab="Employee Variation Rate", ylab="Value")
```



Next, based on our business intuition we identified the relevant features that are potential confounders to our analysis.

```
#### Preparing the A/B Testing data set
data1 <- data %>% select(age, employed, relationship, past_default,
                         house_owners, past_loan, duration, treatment,
                         campaign, pdays, emp.var.rate, cons.price.idx,
                         nr.employed, result)
```

Threats to Causal Inference

To conduct a causal study, we undertook the following assumptions:

1. **Omitted Variable:** Based on the business knowledge, we identified most factors that can influence customer choice to renew a term loan, such as demographics, marketing information, and economic indicators.
2. **Selection Bias:** The dataset comes from a bank's direct marketing campaign, which comprises all current term-Deposit holders. Hence we are selecting the entire population concerning our bank.
3. **Simultaneity Bias:** If a person renews a term deposit account, it doesn't change demographics, age,

or country's economic performance.

4. **Measurement Error:** Demographic and finance-related information is collected by the bank's automated process. So, we can safely assume the data has not been tampered with.
5. **Interference:** We assume that the marketing team uses a single communication channel to reach its existing customers. This assumption allows us to create cell phone users as a "group of interest" or treatment in statistical language.

Which channel is better: Telephone or Cellphone?

Next, we have to randomize the variables here to detect the isolated effect of marketing channel on treatment group. We conducted overall T-test to check for the randomization.

```
data_summary = data1 %>% group_by(treatment, result) %>%
  summarise_all(mean) %>% ungroup()

### Conducting T-test for all Age
t.test(data1$treatment, data1$age)
```

```
##
##  Welch Two Sample t-test
##
## data:  data1$treatment and data1$age
## t = -766.27, df = 41363, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -39.49007 -39.28856
## sample estimates:
## mean of x mean of y
##  0.634748 40.024060
```

The above test signifies a nearly <1% chance that the two groups i.e., Telephone users and Cellphone users, are similar. Hence, the randomization check fails, demanding the groups to be nearly identical.

```
#### Conducting the T-test for Duration
t.test(data1$employed, data1$treatment)
```

```
##
##  Welch Two Sample t-test
##
## data:  data1$employed and data1$treatment
## t = 136.67, df = 49628, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.3357481 0.3455183
## sample estimates:
## mean of x mean of y
## 0.9753812 0.6347480
```

Again, the group differs on the duration feature which again violate the requirement to conduct statistical analysis.

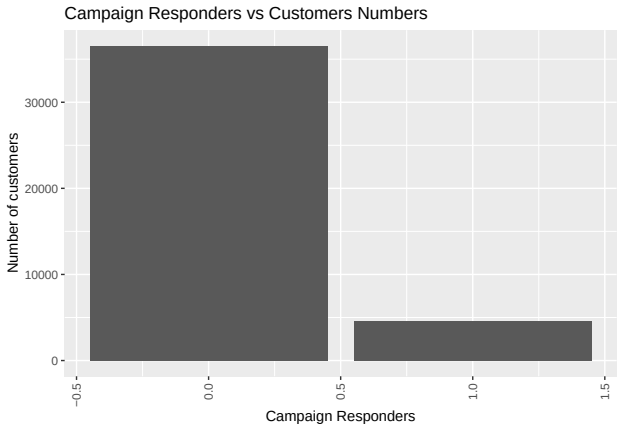
```
#### Conducting the T-test for past_default
chisq.test(data1$treatment, data1$past_default)
```

```
## Warning in chisq.test(data1$treatment, data1$past_default): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  data1$treatment and data1$past_default
## X-squared = 0.51033, df = 1, p-value = 0.475
```

Although, the groups are similar on past loan. That means the both groups has similar distribution for past loan hoalders.

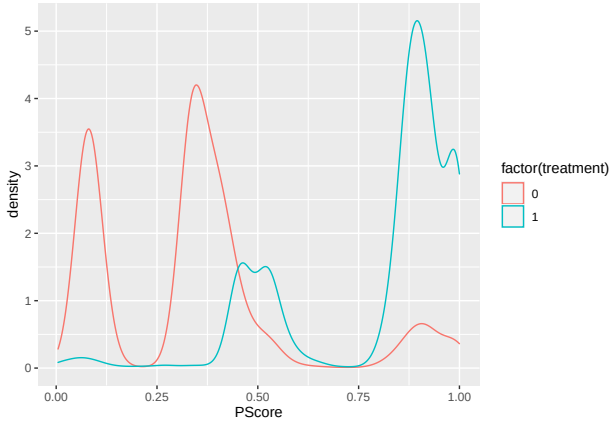
```
ggplot(data, aes(x=result)) + geom_bar(stat='count') +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
  ggtitle("Campaign Responders vs Customers Numbers") +
  xlab("Campaign Responders") + ylab("Number of customers")
```



As the variables are not randomized, we apply the Propensity Score Matching(PSM) technique to randomize them.PSM resolves the confounded differences between treatment and control(aside from treatment). Assumptions: PSM assumes that observed variables determine treatment and since we have already considered that there is no other significant confounder, we can conduct PSM.

First, we obtained predicted probabilities of treatment against all the confounders.

```
### Generating the PScores for the treatment and confounders
PScore = glm(treatment ~ age + duration + employed +
               relationship + past_default + past_loan
             + campaign + pdays + emp.var.rate + cons.price.idx ,
             data = data1,
             family = "binomial")$fitted.values
data1$PScore = PScore
ggplot(data, aes(x = PScore, color = factor(treatment))) +
  geom_density()
```



The above graph shows the probability distribution for the two groups where 1 is Treatment(Cellphone) and 0 is Control(Telephone). For each treated observation, we will find the control observation with the closest propensity score (up to a specified threshold, aka caliper) and match it.

```
match_output <- matchit(treatment ~ age + duration + employed +
                relationship + past_default + past_loan
              + campaign + pdays + emp.var.rate + cons.price.idx,
              data= data1,
              method = 'nearest',
                        distance = "logit", caliper = 0.00001,
                        replace = FALSE, ratio = 1)
```

```
## Warning: Fewer control units than treated units; not all treated units will get
## a match.
```

```
summary(match_output)
```

```
##
## Call:
## matchit(formula = treatment ~ age + duration + employed + relationship +
##     past_default + past_loan + campaign + pdays + emp.var.rate +
##     cons.price.idx, data = data1, method = "nearest", distance = "logit",
##     replace = FALSE, caliper = 1e-05, ratio = 1)
##
## Summary of Balance for All Data:
##                 Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## distance               0.8010        0.3459          2.1289     0.7827    0.4049
## age                   39.9686       40.1205         -0.0139     1.3504    0.0155
## duration             263.5278      249.1738          0.0558     0.9641    0.0103
## employed               0.9763        0.9738          0.0163          .    0.0025
## relationship           0.5816        0.6462         -0.1310          .    0.0646
## past_default           0.0001        0.0000          0.0107          .    0.0001
## past_loan              0.1997        0.0268          0.4326          .    0.1730
## campaign               2.4050        2.8501         -0.1828     0.5601    0.0106
## pdays                945.7492      991.5429         -0.2047     6.8104    0.0342
## emp.var.rate          -0.3871        0.8970         -0.7752     2.9131    0.2123
## cons.price.idx        93.3160       94.0270         -1.4494     1.3507    0.2605
##                 eCDF Max
## distance          0.7744
## age               0.0618
## duration          0.0503
## employed          0.0025
```

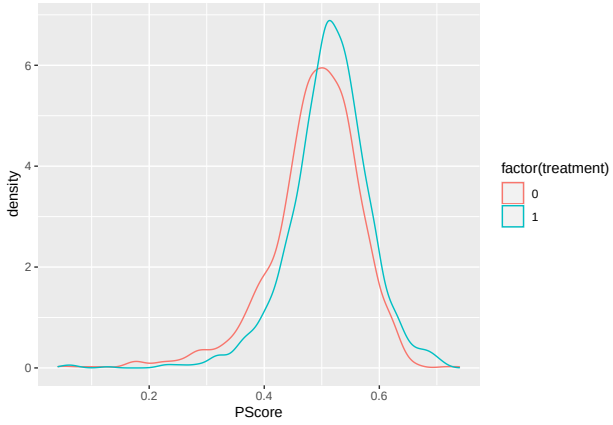
```
## relationship     0.0646
## past_default     0.0001
## past_loan        0.1730
## campaign         0.0587
## pdays            0.0461
## emp.var.rate     0.4721
## cons.price.idx   0.7827
##
##
## Summary of Balance for Matched Data:
##                Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## distance              0.7437        0.7437          0.0000     1.0000    0.0000
## age                  38.7329       39.5652         -0.0760     0.9382    0.0117
## duration            236.3164      242.5109         -0.0241     0.6691    0.0097
## employed              0.9725        0.9782         -0.0371          .    0.0056
## relationship          0.5969        0.5659          0.0629          .    0.0310
## past_default          0.0000        0.0000          0.0000          .    0.0000
## past_loan             0.2199        0.1727          0.1181          .    0.0472
## campaign              2.3221        2.9542         -0.2596     0.5984    0.0153
## pdays               952.1254      958.4038         -0.0281     1.1465    0.0042
## emp.var.rate         -0.3834       -0.3286         -0.0331     0.9588    0.0315
## cons.price.idx       93.3790       93.3611          0.0365     0.8810    0.0199
##                eCDF Max Std. Pair Dist.
## distance         0.0014          0.0000
## age              0.0430          0.8858
## duration         0.0606          0.7461
## employed         0.0056          0.3057
## relationship     0.0310          0.6029
## past_default     0.0000          0.0000
## past_loan        0.0472          0.2732
## campaign         0.1198          0.8467
## pdays            0.0078          0.1534
## emp.var.rate     0.0627          0.4381
## cons.price.idx   0.0578          0.3029
##
## Percent Balance Improvement:
##                Std. Mean Diff. Var. Ratio eCDF Mean eCDF Max
## distance                100.0      100.0     100.0     99.8
## age                    -447.7       78.8      24.4     30.4
## duration                56.8     -998.9       5.0    -20.4
## employed              -127.8          .    -127.8   -127.8
## relationship            52.0          .      52.0     52.0
## past_default           100.0          .     100.0    100.0
## past_loan               72.7          .      72.7     72.7
## campaign               -42.0       11.4     -44.8   -104.0
## pdays                   86.3       92.9      87.8     83.2
## emp.var.rate            95.7       96.1      85.2     86.7
## cons.price.idx          97.5       57.8      92.4     92.6
##
## Sample Sizes:
##           Control Treated
## All         15044   26144
## Matched      1419    1419
## Unmatched   13625   24725
```

```
## Discarded       0       0
```

```
data_match = match.data(match_output)
```

The matched units are nearly similar to the confounders and hence our sample will be randomized. Finally, we get predicted probabilities of treatment from the model after matching to check if the scores are similar or not.

```
## For matched data
PScore = glm(treatment ~ age + duration + employed +
               relationship + past_default + past_loan
             + campaign + pdays + emp.var.rate + cons.price.idx,
             data=data_match, family = "binomial")$fitted.values
data_match$PScore = PScore
ggplot(data_match, aes(x = PScore, color = factor(treatment))) +
  geom_density()
```



The above graph shows the probability distribution for the matched samples. The P-Scores are nearly the same for both groups now. Next, we again conduct some T-tests to identify if our model is random.

```
#### Stratify the sample [Control]
control <- data_match %>% filter(treatment == 0)
treatment <- data_match %>% filter(treatment ==1)
control1 <- control %>% group_by(result) %>% sample_frac(size=0.20, replace=TRUE)
treatment1 <- treatment %>% group_by(result) %>% sample_frac(size=0.20, replace=TRUE)
```

```
#### Randomization Checks
treat_age <- treatment1$age
control_age <- control1$age
t.test(treat_age, control_age, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  treat_age and control_age
## t = -0.50643, df = 563.83, p-value = 0.6128
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.958260  1.155443
## sample estimates:
## mean of x mean of y
##  38.05634  38.45775
```

The above test signifies a nearly 62% chance that the two groups, i.e., Telephone users and Cellphone users, are similar. Hence, we can conclude that the samples are pretty similar and it holds our randomization check.

```
### cons.price.idx
treat_cons.price.idx <- treatment1$cons.price.idx
control_cons.price.idx <- control1$cons.price.idx
t.test(treat_cons.price.idx, control_cons.price.idx, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  treat_cons.price.idx and control_cons.price.idx
## t = 1.5894, df = 565.98, p-value = 0.1125
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.01720381  0.16311931
## sample estimates:
## mean of x mean of y
##  93.44684  93.37388
```

Similarly, the CPI value are randomized too. Next to infer causality we identified the maximum value of change in the term-deposit renewals that we can observe:

```
###Power Test
n = nrow(treatment1)
n1 = nrow(control1)
power.t.test(n=n,type=c("two.sample"),power=0.9,sig.level=0.05,delta=NULL)
```

```
##
##      Two-sample t test power calculation
##
##               n = 284
##           delta = 0.2724569
##              sd = 1
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Based on the Power test we can infer 27% of change. Now, we conduct a logistic regression analysis to observe the effect.

A/B Testing: Applying logistic regression on result against treatment. This will give us isolated effect of treatment on result.

```
m1 = glm(result ~ treatment, data = data_match, family="binomial")
summary(m1)
```

```
##
## Call:
## glm(formula = result ~ treatment, family = "binomial", data = data_match)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5209  -0.5209  -0.4660  -0.4660   2.1326
##
## Coefficients:
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.16552    0.08737 -24.785   <2e-16 ***
## treatment    0.23642    0.11831   1.998   0.0457 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2023.9  on 2837  degrees of freedom
## Residual deviance: 2019.9  on 2836  degrees of freedom
## AIC: 2023.9
##
## Number of Fisher Scoring iterations: 4
```

The analysis shows a 24% increase in Term Deposit renewals on average if the mode of conversation switches from Telephone to Cellphone since the observed effect is less than the maximum identifiable effect of 27%.

Sensitivity Analysis for PSM:

Next, we want to test the robustness of the solution. So we changed the caliper settings of the PSM model and re-ran the above regression. Ideally, the inferences should be nearly similar to the above:

1. Run 1 (Caliper: 0.0001)

```
match_output1 <- matchit(treatment ~ age + duration + employed +
               relationship + past_default + past_loan
             + campaign + pdays + emp.var.rate + cons.price.idx,
             data= data1,
             method = 'nearest',
                        distance = "logit", caliper = 0.0001,
                        replace = FALSE, ratio = 1)
```

```
## Warning: Fewer control units than treated units; not all treated units will get
## a match.
```

```
data_match1 = match.data(match_output1)
summary(glm(result ~ treatment, data = data_match1, family="binomial"))
```

```
##
## Call:
## glm(formula = result ~ treatment, family = "binomial", data = data_match1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5567  -0.5567  -0.5122  -0.5122   2.0474
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.96473    0.05507 -35.679   <2e-16 ***
## treatment    0.17868    0.07545   2.368   0.0179 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4801.4  on 6115  degrees of freedom
## Residual deviance: 4795.8  on 6114  degrees of freedom
```

```
## AIC: 4799.8
##
## Number of Fisher Scoring iterations: 4
```

2. Run 2 (Distance : Probit)

```
match_output1 <- matchit(treatment ~ age + duration + employed +
               relationship + past_default + past_loan
             + campaign + pdays + emp.var.rate + cons.price.idx,
             data= data1,
             method = 'nearest',
                        distance = "glm", caliper = 0.0003,
                        replace = FALSE, ratio = 1)
```

```
## Warning: Fewer control units than treated units; not all treated units will get
## a match.
```

```
data_match1 = match.data(match_output1)
summary(glm(result ~ treatment, data = data_match1, family="binomial"))
```

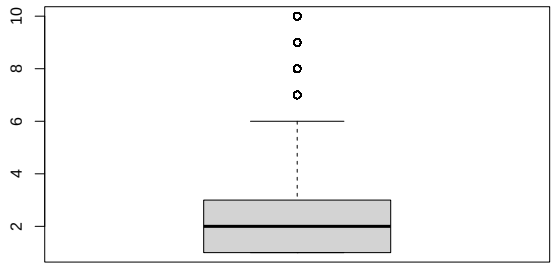
```
##
## Call:
## glm(formula = result ~ treatment, family = "binomial", data = data_match1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5781  -0.5781  -0.5237  -0.5237   2.0271
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.91747    0.05177  -37.04  < 2e-16 ***
## treatment    0.21308    0.07056    3.02  0.00253 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5434.9  on 6679  degrees of freedom
## Residual deviance: 5425.8  on 6678  degrees of freedom
## AIC: 5429.8
##
## Number of Fisher Scoring iterations: 4
```

As we can see that the last two PSM runs with the different caliper and distance settings yielded ~18% and ~22% results. Hence, our analysis is even confident that the causal effect is within the 17%-22% limit.

How many times should the bank contact the customer?

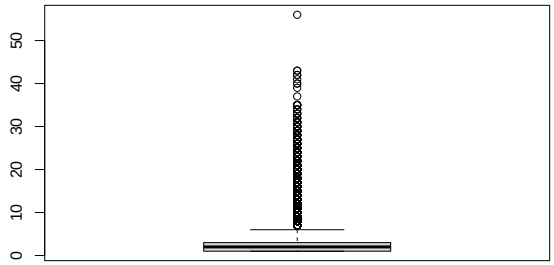
Since the frequency of communication is also an essential aspect for the marketing teams, the following section of the report explores the impact of frequency on the customer term-deposit renewals. Our team quantified frequency as the sum r of times the customer has been reached out in the recently concluded campaign between the last and recently completed campaign.

```
data$freq <- data$previous + data$campaign
temp3 = data %>% filter(freq <= 10)
boxplot(temp3$freq)
```



We can see that most customers are contacted only once or twice. If we identify the median of frequency of contact, we can take that number as a baseline threshold to see if we should increase our marketing efforts or not.

```
boxplot(data$freq)
```



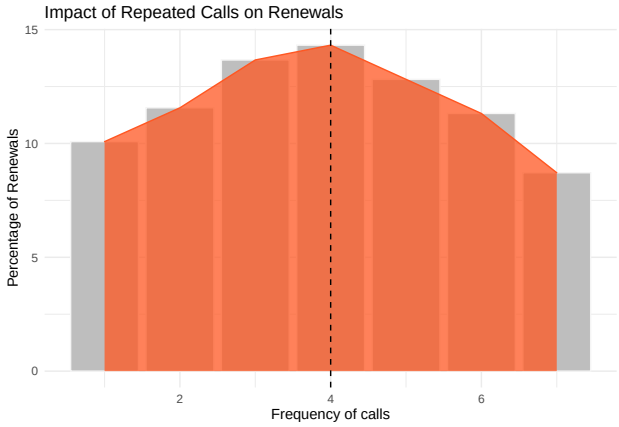
The graph visualizes the impact of changing the frequency. The graph visualizes the effect of changing the frequency. We can identify that increasing the number of calls beyond four will decline Term Deposit renewals. Hence, we split the data into two sections, i.e., Frequency < 4 and Frequency $>=4$.

```
data_freq <- data %>% filter(freq > 0)
data_freq1 <- data %>% group_by(freq) %>% summarise(result = sum(result),
                                                    count=n()) %>%
  mutate(pct = 100*result/count)

ggplot(data_freq1 %>% filter(freq<8), aes(x=freq, y=pct)) +
  geom_histogram(aes(x=freq, y=pct), stat="identity", color="#f5f5f5", fill="grey")+
  geom_density(stat="identity",color="#ff5722", fill="#ff5722", alpha=0.75) +
  geom_vline(data=data_freq1, aes(xintercept=4),
             linetype="dashed") +
  scale_color_brewer(palette="red") +
  ggtitle("Impact of Repeated Calls on Renewals") + xlab("Frequency of calls")+
  ylab("Percentage of Renewals") + theme_minimal()
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
## Warning in pal_name(palette, type): Unknown palette red
```



Using the two subsections of data we analyzed the impact of change in the frequency on the renewals:

```
m2 = glm(result ~ freq,
         data = data %>% filter(freq <4) %>% mutate(freq_2 = freq**2)
         , family="binomial")
summary(m2)
```

```
##
## Call:
## glm(formula = result ~ freq, family = "binomial", data = data %>%
##     filter(freq < 4) %>% mutate(freq_2 = freq^2))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5399  -0.4985  -0.4599  -0.4599   2.1443
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.36397    0.04446 -53.169  < 2e-16 ***
## freq         0.17065    0.02243   7.609 2.77e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 22958  on 32528  degrees of freedom
## Residual deviance: 22901  on 32527  degrees of freedom
## AIC: 22905
##
## Number of Fisher Scoring iterations: 4
```

Using the regression results, we can see that increasing 1 unit of frequency will result in nearly 17% higher chances of term-deposit renewal. We also ran our regression with the squared term of the frequency, which resulted in non-significant p-values i.e. there is no quadratic trend in the percentage of renewals and the frequency of contact.

```
m3 = glm(result ~ freq + freq_2,
         data = data %>% filter(freq <4) %>% mutate(freq_2 = freq**2)
         , family="binomial")
summary(m3)
```

```
##
```

```
## Call:
## glm(formula = result ~ freq + freq_2, family = "binomial", data = data %>%
##     filter(freq < 4) %>% mutate(freq_2 = freq^2))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5422  -0.4958  -0.4610  -0.4610   2.1422
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.30550    0.12577 -18.331   <2e-16 ***
## freq         0.09872    0.14653   0.674    0.500
## freq_2       0.01846    0.03716   0.497    0.619
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 22958  on 32528  degrees of freedom
## Residual deviance: 22900  on 32526  degrees of freedom
## AIC: 22906
##
## Number of Fisher Scoring iterations: 4
```

Similarly, we performed the regression for the frequency greater than 4:

```
m4 = glm(result ~ freq,
         data = data %>% filter(freq >=4) %>% mutate(freq_2 = freq**2)
         , family="binomial")
summary(m4)
```

```
##
## Call:
## glm(formula = result ~ freq, family = "binomial", data = data %>%
##     filter(freq >= 4) %>% mutate(freq_2 = freq^2))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5586  -0.5586  -0.4765  -0.3735   3.1656
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.09986    0.09771 -11.257   <2e-16 ***
## freq        -0.16974    0.01720  -9.869   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6040.6  on 8658  degrees of freedom
## Residual deviance: 5888.2  on 8657  degrees of freedom
## AIC: 5892.2
##
## Number of Fisher Scoring iterations: 6
```

Our analysis showed that increasing the frequency beyond 4 results in a decline of ~17% chances of renewal

at a high statistical significance level.

Results

The above statistical analysis allows us to infer that there is a higher chance of term deposit renewals if we change the marketing channel from Telephone to Cellphone. The study has shown that we can expect an increase in the possibilities of renewal by 24%.

```
5000*0.24
```

```
## [1] 1200
```

That means we can additionally add 1000+ renewals based on our analysis which is quite a significant number based on business expected numbers. Further, we identified that we connect with customer more than 4 for the renewals if start hampering our chances of renewals. Our analysis showed that increasing the value till 4 help in increasing chances by 17% for each additional increase in frequency. This can further help us in securing more renewals on and above the change in marketing channel.

Limitations

For our analysis, we have assumed that no other factors might impact the customers' judgment of renewal. But some factors such as the time of the day the marketing team is connecting with the customers, the amount of term deposit, and banks' recent reputation will impact customers' decision to renew. These features are not accounted for in the analysis. Therefore, we advise the bank marketing team to capture such data points.

Moreover, we have considered that the marketing team reaches a customer via a single channel, i.e. a customer was contacted via telephone and not via Cellphone. Hence, we would further like to ascertain the case. Since the results obtained from the above analysis are for the observational data, we wanted to design future tests that can help us establish real-time chances for the same.
