Causal Inference via Econometrics and Experimentation:

Final Project Report

Team 05

Team members: Samira Arondekar, Leo (Chuchen) Xiong, Shobhit Mishra, Mona Kan, Maya Carnie

Table of Contents

Executive Summary

A Portuguese banking institution wants to understand the impact of making over two calls to prospective customers on them subscribing to term deposit. Using matching technique, we found a randomized treatment and control group and measured the treatment effect. Doing robustness and sensitivity analysis, we have identified randomized treatment and control groups. Here treatment effect is the effect of making over two calls to prospective customers compared to making only one or two calls to them.

We found that making over two calls to prospective customers negatively affects their probability to subscribe to term deposit by 1.3%. So, we recommend making only one or two calls and not annoying prospective customers with more calls.

Further, we also found that the prospective customers who have loans should not be contacted, instead we should contact those without loans as having loans negatively impacts probability of subscribing to term deposit. Also, prospective customers should be contacted on telephones instead of cellphones as this increases probability of subscribing to term deposit. Lastly, prospective customers should not be contacted on Wednesdays as this reduces probability to subscribe to term deposit.

Business Problem

A Portuguese banking institution conducts a direct marketing campaign using phone calls to promote subscription to a term deposit at their institution. Often reminding people about availability of such services and its providers is important to drive sales. As each incremental phone call costs them time and money, while not placing the promotional phone call to prospective clients might cause them to lose out on that customer. Hence, they need to identify the right amount of phone calls per prospective client to drive subscription to term deposit.

To aid with the above effort, we will run an analysis to detect the effect (positive/negative) of making over 2 phone calls compared to making just one or two phone calls.

Causal Question: Does contacting a prospective customer more than twice via calls have an effect on the client subscribing to term deposit compared to subscribing once or twice?

Assumptions

- The audience has not been contacted by other forms of advertisement or campaigns for term deposits.
- All the prospective customers can be contacted
- There is no interference between the two groups, i.e. they do not talk to each other about this campaign
- The audience is not aware that they are part of an experiment and there is also no social desirability bias
- Making calls to advertise for term deposit has significantly positive effect on subscriptions compared to no advertisement

Threats to Causal Inference

1. Selection Bias

   Selection bias would exist if the sample was not representative of the population. However, since we have no reason to believe that the dataset we have is a subset of the population, it must be representative of the population and hence there is no reason to believe that there is a selection bias.

2. Simultaneity

   Since according to current setup whether call is placed or not is not determined by the probability of the prospective client to subscribe to term deposit, we do not have reason to believe that simultaneity bias exists.

3. Omitted Variables

   Making calls is currently not based on any other input like effect of previous contact. If there was a factor that affected both the placing of calls and the subscription to term deposit, then it would lead to omitted variable bias.

4. Measurement Error

   Since a prospective customer can either subscribe to term deposit or not subscribe, there is no reason to believe that the outcome variable does not reflect subscription to a term deposit.

Data used

The dataset used for this analysis is obtained from UCI, Machine learning repository. It contains data related to a direct phone marketing campaign from May 2008 to November 2010. The outcome variable is captured by field “y” indicating ‘yes’ for subscription and ‘no’ for no subscription.

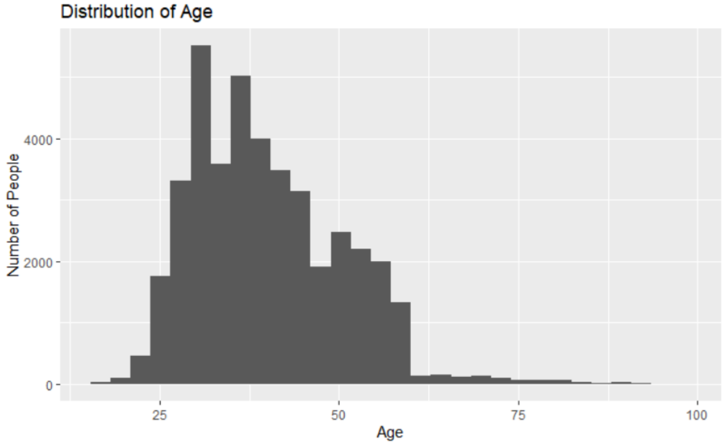
The important independent variables are:

Variables	Description
Age	Age of prospective customer (numeric)
Job	Type of job of prospective customer (categorical)
Marital	Marital status (categorical)
Education	Education level (categorical)
Default	Does the prospect have credit in default? (categorical)
Housing	Does the prospect have housing loan? (categorical)
Loan	Does the prospect have personal loan? (categorical)
Contact	Contact communication type (categorical)
Month	Last contact month of year (categorical)
Day_of_week	Last contact day of the week (categorical)
Duration	Last contact duration, in seconds (numeric)
Treatment	Indicated by “campaigns” field. Indicates number of contacts performed during this campaign and for this client (numeric, includes last contact).

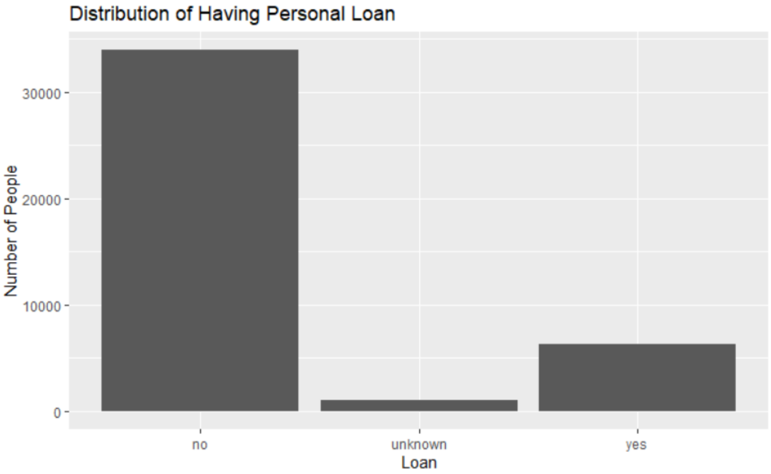
We have done data quality check to ensure that the variables we include in the analysis don't have any missing values or NA values.

Let's further look into the distribution of the variables to understand this dataset better:

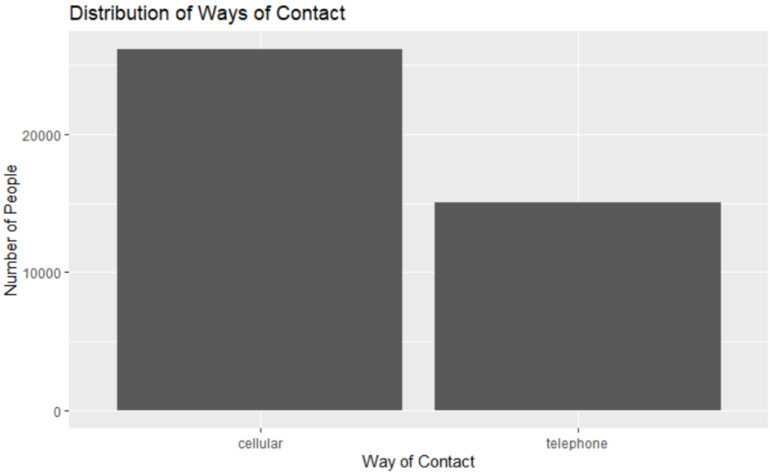
1. Distribution of age: Almost 98% of the recipients are in the age range of 20-60



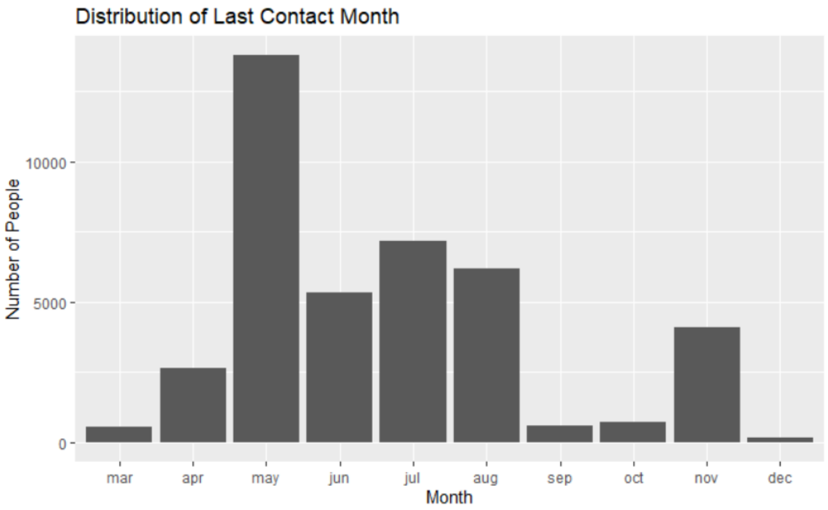
2. Distribution of having personal loan or not: 82% of the recipients in the dataset don't have personal loan.



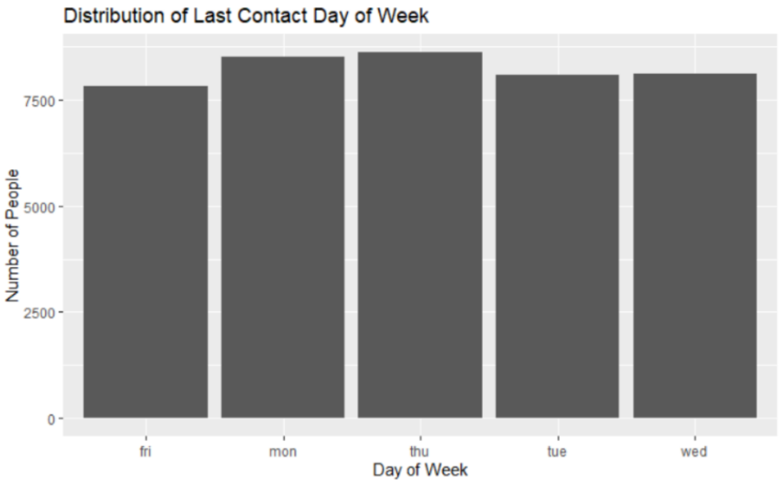
3. Distribution of ways of contact: 64% of the recipients were contacted via cellular, the rest were via telephone.



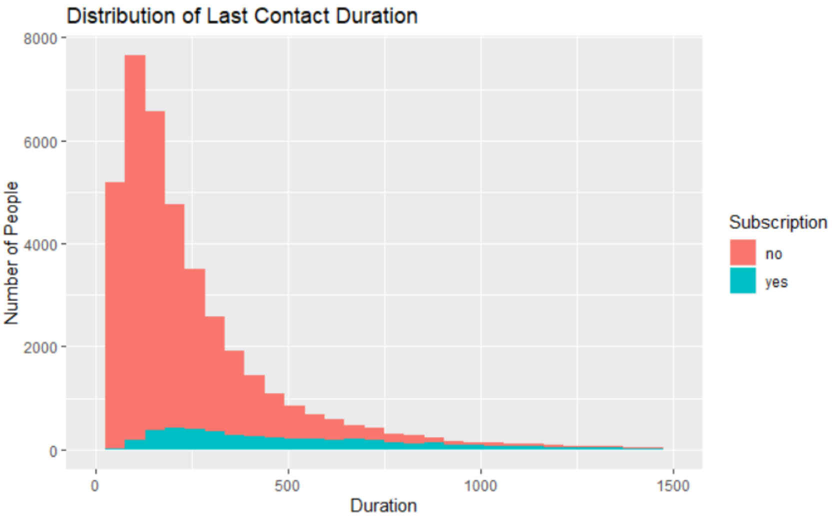
4. Distribution of last contact month: Most people were contacted between May and August and May is especially the high point of being contacted.



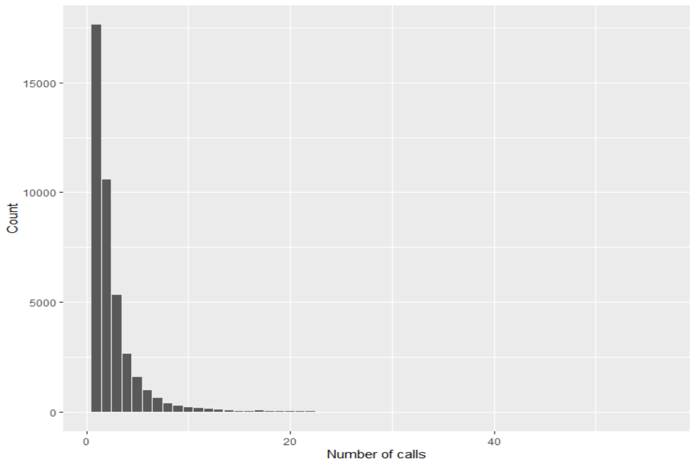
5. Distribution of last contact day of week: People’s last contact day of week is almost equally distributed across the weekdays.



6. Distribution of last contact duration: Even though some clients had very long conversation duration but still didn’t subscript the term deposit at the end. The last contact durations of those clients who eventually subscribed are more evenly distributed than who didn’t.



Analysis



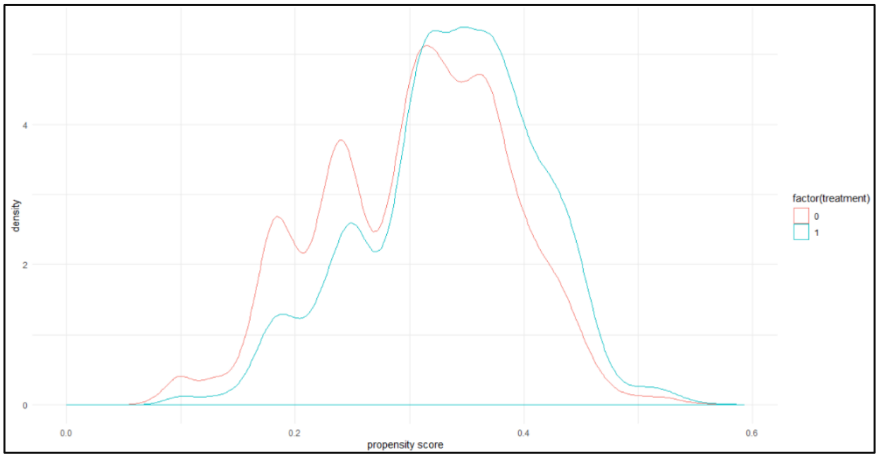
Over half of the prospective customers get one or two campaign calls, while others get above two calls. So we take divide the audience into treatment group as people receiving more than two calls and control group as people receiving one or two calls.

To measure treatment effect using the created treatment and control groups, we need to do randomization checks to ensure that there are no statistically significant differences between the two groups. Using t-tests and chi-sq tests to compare the two groups based on independent variables, we get the below results for randomization check:

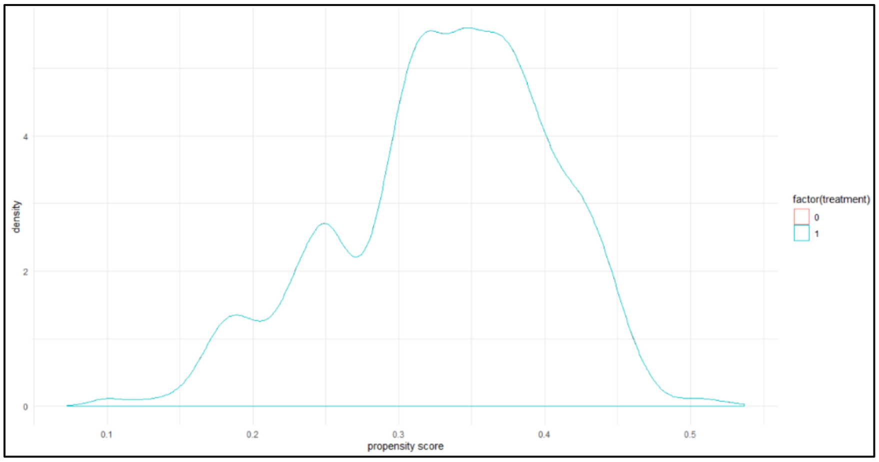
Independent variable	Randomized?
Age	Yes
Marital Status	Yes
Education	Yes
Job	No
Loan	Yes
Contact	No
Month	No
Day of Week	No

Since, there are statistically significant differences between the two groups, we will have to find groups that do not have differences using other methods. We will use propensity score matching to find the two groups.

In propensity score matching, we selected the control variables for matching based on whether treatment varied with them. Treatment varies with call duration, day of week, month, contact, loan, default (yes status gets no calls), job, little based on education, marital status, age. Calculating probability of treatment based on these variables, we get the following propensity score distribution for treatment and control groups.

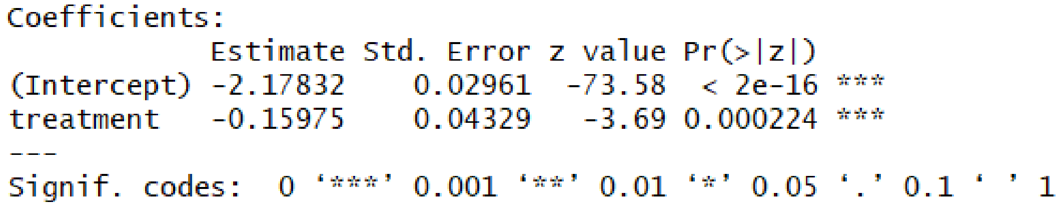


Where blue represents treatment group and red represents control group. Our goal is to obtain groups whose propensity score distribution overlaps for the most part. After matching based on propensity scores, we get the following propensity score distribution for matched treatment and control groups.



After matching we get propensity score distributions that overlap well. We will use these matched groups for detecting treatment effect. We get 12,486 observations in treatment and control group each.

We use logistic regression to estimate the treatment effect.



```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -2.17832    0.02961  -73.58  < 2e-16 ***
treatment   -0.15975    0.04329   -3.69 0.000224 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Finding: Greater than 2 calls result in decline in clients subscribing to term deposit.

Since, logistic regression gives us change in log odds, we will convert it to corresponding probability value.

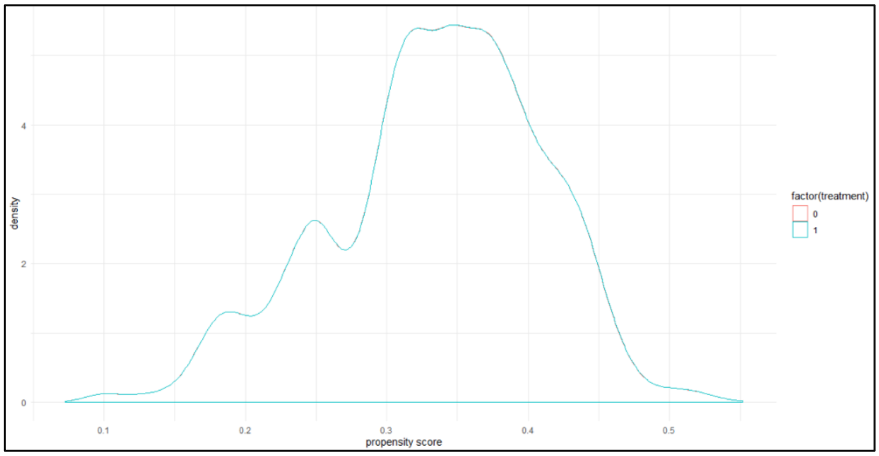
odds = exp(beta)

probability = odds / (1 + odds)

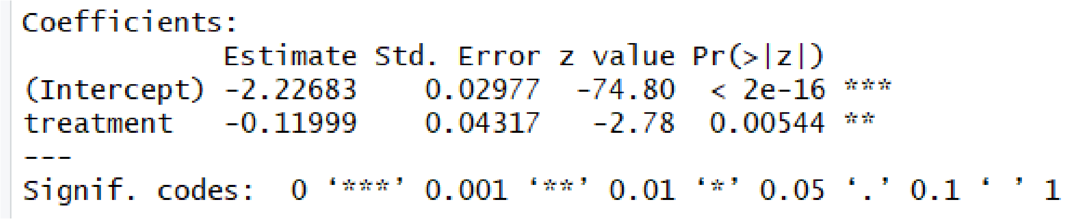
Treatment effect = probability of treatment group – probability of control group

Hence, probability reduces by 0.013 (1.3%) when more than 2 calls are made compared to when only 1 or 2 calls are made.

We will perform sensitivity check by changing the caliper used for matching from 0.001 to 0.005. We get 12,838 observations in treatment and control group each.



The matching still looks pretty good, but it has slightly less overlap than with caliper 0.001. Using logistic regression to estimate the treatment effect, we observe beta coefficient of -0.12.

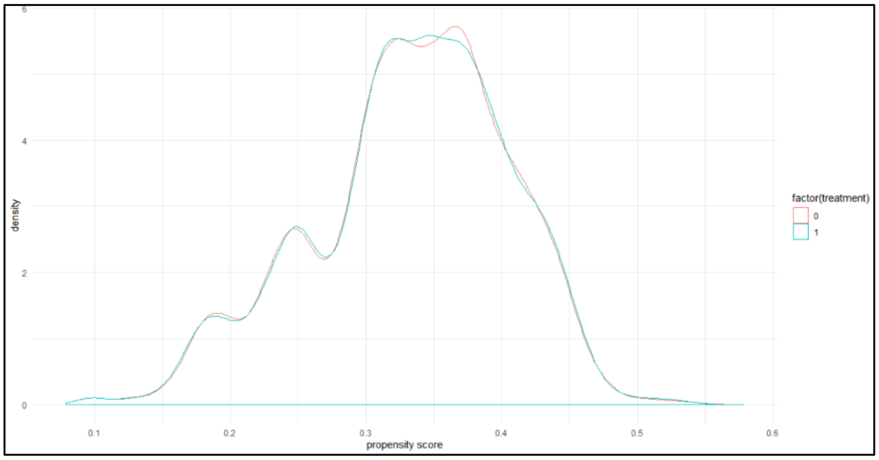


```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -2.22683    0.02977  -74.80  < 2e-16 ***
treatment   -0.11999    0.04317   -2.78  0.00544 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

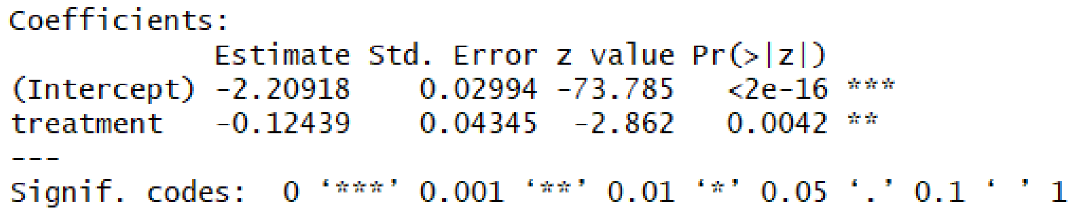
There is a negative change in probability due to treatment. It is -1%.

Although we observe a significant effect, it is at a higher significance level i.e. it allows higher type 1 error.

Further, we will check robustness by matching on different subset of variables. Let's eliminate parameters that showed little variation w.r.t. treatment. So we match on job, default, loan, contact, month, day_of_week, duration. We get 12,514 observations in treatment and control group each.



The matching is not very good as we see that the two distributions do not overlap well.

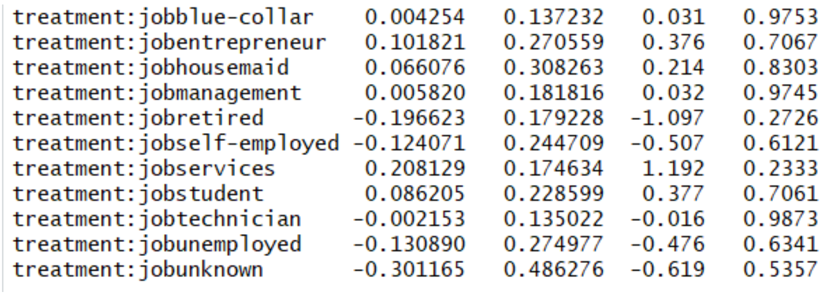


```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -2.20918    0.02994 -73.785   <2e-16 ***
treatment   -0.12439    0.04345  -2.862   0.0042 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Again, although we observe a significant effect, it is at a higher significance level i.e. it allows higher type 1 error.

Check for heterogenous treatment effect:

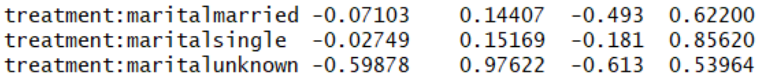
1. Checking for job field



```
treatment:jobblue-collar     0.004254   0.137232   0.031   0.9753
treatment:jobentrepreneur    0.101821   0.270559   0.376   0.7067
treatment:jobhousemaid       0.066076   0.308263   0.214   0.8303
treatment:jobmanagement      0.005820   0.181816   0.032   0.9745
treatment:jobretired        -0.196623   0.179228  -1.097   0.2726
treatment:jobself-employed -0.124071   0.244709  -0.507   0.6121
treatment:jobservices        0.208129   0.174634   1.192   0.2333
treatment:jobstudent         0.086205   0.228599   0.377   0.7061
treatment:jobtechnician     -0.002153   0.135022  -0.016   0.9873
treatment:jobunemployed     -0.130890   0.274977  -0.476   0.6341
treatment:jobunknown        -0.301165   0.486276  -0.619   0.5357
```

Finding: We see no significant effect of having a different job type compared to admin job type when checking for effect of treatment over control.

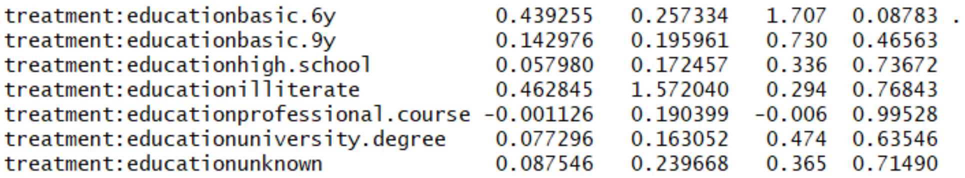
2. Checking for marital field



```
treatment:maritalmarried -0.07103    0.14407  -0.493  0.62200
treatment:maritalsingle  -0.02749    0.15169  -0.181  0.85620
treatment:maritalunknown -0.59878    0.97622  -0.613  0.53964
```

Finding: We see no significant effect of having a different marital status compared to divorced marital status when checking for effect of treatment over control.

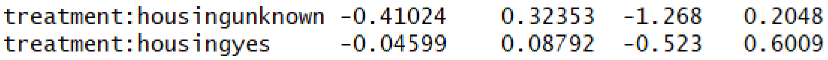
3. Checking for education field



```
treatment:educationbasic.6y              0.439255   0.257334   1.707  0.08783 .
treatment:educationbasic.9y              0.142976   0.195961   0.730  0.46563
treatment:educationhigh.school           0.057980   0.172457   0.336  0.73672
treatment:educationilliterate            0.462845   1.572040   0.294  0.76843
treatment:educationprofessional.course -0.001126   0.190399  -0.006  0.99528
treatment:educationuniversity.degree     0.077296   0.163052   0.474  0.63546
treatment:educationunknown               0.087546   0.239668   0.365  0.71490
```

Finding: We see no significant effect of having a different education level compared to basic 4 year education level when checking for effect of treatment over control, except for 6 year education level compared to basic 4 year education level when checking for effect of treatment over control (but with a significance level of 10%).

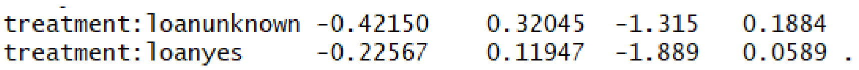
4. Checking for housing field



```
treatment:housingunknown -0.41024    0.32353  -1.268   0.2048
treatment:housingyes     -0.04599    0.08792  -0.523   0.6009
```

Finding: We see no significant effect of having a house compared to having no house when checking for effect of treatment over control.

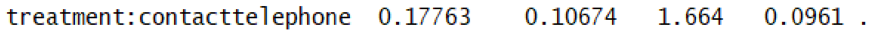
5. Checking for loan field



```
treatment:loanunknown -0.42150    0.32045  -1.315   0.1884
treatment:loanyes     -0.22567    0.11947  -1.889   0.0589 .
```

Finding: We do see a significant effect of having a loan compared to having no loan when checking for effect of treatment over control but with a significance level of 10%.

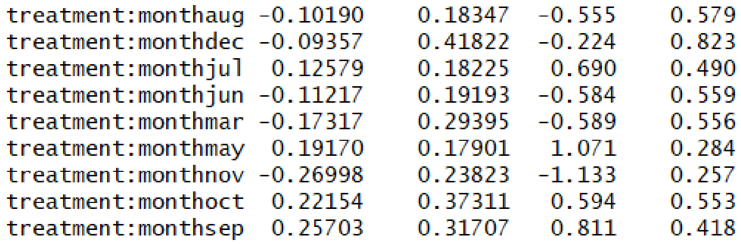
6. Checking for contact field



```
treatment:contacttelephone  0.17763    0.10674   1.664   0.0961 .
```

Finding: We do see a significant effect of contacting by telephone compared to contacting by cellphone when checking for effect of treatment over control with a significance level of 10%.

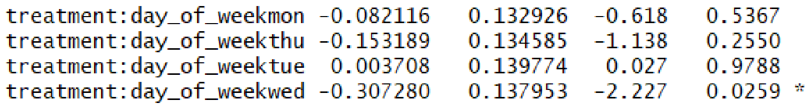
7. Checking for month field



```
treatment:monthaug -0.10190    0.18347  -0.555    0.579
treatment:monthdec -0.09357    0.41822  -0.224    0.823
treatment:monthjul  0.12579    0.18225   0.690    0.490
treatment:monthjun -0.11217    0.19193  -0.584    0.559
treatment:monthmar -0.17317    0.29395  -0.589    0.556
treatment:monthmay  0.19170    0.17901   1.071    0.284
treatment:monthnov -0.26998    0.23823  -1.133    0.257
treatment:monthoct  0.22154    0.37311   0.594    0.553
treatment:monthsep  0.25703    0.31707   0.811    0.418
```

Finding: We see no significant effect of contacting in a different month compared to contacting in month of April when checking for effect of treatment over control.

8. Checking for day_of_week field



```
treatment:day_of_weekmon -0.082116   0.132926  -0.618   0.5367
treatment:day_of_weekthu -0.153189   0.134585  -1.138   0.2550
treatment:day_of_weektue  0.003708   0.139774   0.027   0.9788
treatment:day_of_weekwed -0.307280   0.137953  -2.227   0.0259 *
```

Finding: We see a significant effect of contacting on Wednesday compared to contacting on Friday when checking for effect of treatment over control.

Conclusions and Recommendations

From above analysis we conclude that making over two calls to prospective customers negatively affects their probability to subscribe to term deposit by 1.3%. So, we recommend making only one or two calls and not annoying prospective customers with more calls.

Further, we also found that the prospective customers who have loans should not be contacted, instead we should contact those without loans as having loans negatively impacts probability of subscribing to term deposit. Also, prospective customers should be contacted on telephones instead of cellphones as this increases probability of subscribing to term deposit. Lastly, prospective customers should not be contacted on Wednesdays as this reduces probability to subscribe to term deposit.

Limitations and Future Steps

We don’t have data about what happens if no phone call advertisement is done, we are not sure whether campaigns increase the probability of subscribing to term deposit.

Currently the analysis has been done to compare effect of one or two calls compared to higher than two calls. We can do further analysis to check with a different cutoff point (example 3 calls, 4 calls etc.)

Ideal numbers of calls for prospective customers with different characteristics can be found through further analysis. This can be done as an extension to the heterogeneity analysis.

Appendix

Code for the analysis:

```
# Load libraries
library(dplyr)
library(MatchIt)
library(ggplot2)

# read data
bank_mrktng <- read.csv("<path> /bank-additional-full.csv", sep=";", header=TRUE)
# Data source: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

# Visualization
## Distribution of Age
ggplot(bank_mrktng, aes(age)) + geom_histogram() +
  labs(title = 'Distribution of Age', y="Number of People",x='Age')

## Distribution of Having Credit Card
ggplot(bank_mrktng, aes(default)) + geom_bar(stat = 'count',position=position_dodge()) +
  labs(title = 'Distribution of Having Credit Card', x='Credit Card', y="Number of People")

## Distribution of Having Personal Loan
ggplot(bank_mrktng, aes(loan)) + geom_bar(stat = 'count',position=position_dodge())+
  labs(title = 'Distribution of Having Personal Loan', y='Number of People', x='Loan')

## Distribution of Ways of Contact
ggplot(bank_mrktng, aes(contact)) + geom_bar(stat = 'count',position=position_dodge()) +
  labs(title = 'Distribution of Ways of Contact', x='Way of Contact', y='Number of People')

## Distribution of Last Contact Month
bank_mrktng %>%
  mutate(month = factor(month, levels = c("mar","apr","may","jun","jul","aug","sep","oct","nov","dec")))
  %>% ggplot(aes(month)) + geom_bar(stat = 'count',position=position_dodge()) +
  labs(title = 'Distribution of Last Contact Month', x='Month', y='Number of People')

## Distribution of Last Contact Day of Week
ggplot(bank_mrktng, aes(day_of_week)) +
  geom_bar(stat = 'count',position=position_dodge()) +
  labs(title = 'Distribution of Last Contact Day of Week', x='Day of Week', y='Number of People')

## Distribution of Last Contact Duration
ggplot(bank_mrktng, aes(duration,fill=y)) +
  geom_histogram() +
  xlim(0,1500) +
  labs(title = 'Distribution of Last Contact Duration', x='Duration', y='Number of People') +
  scale_fill_discrete(name = "Subscription")
```

```
# Distribution of campaign calls
ggplot(bank_mrktng %>% group_by(campaign) %>% count() , aes(x=campaign, y=n)) +
  geom_bar(stat = "identity") +
  xlab("Number of calls") +
  ylab("Count")

# Add treatment flag
bank_mrktng <- bank_mrktng %>% mutate(treatment = ifelse(campaign <=2, 0, 1))
bank_mrktng %>% group_by(treatment) %>% count()

# Perform randomization checks on treatment vs control
t.test(age ~ treatment, data = bank_mrktng)
chisq.test(bank_mrktng$treatment, bank_mrktng$job)
chisq.test(bank_mrktng$treatment, bank_mrktng$marital)
chisq.test(bank_mrktng$treatment, bank_mrktng$education)
chisq.test(bank_mrktng$treatment, bank_mrktng$loan)
chisq.test(bank_mrktng$treatment, bank_mrktng$contact)
chisq.test(bank_mrktng$treatment, bank_mrktng$month)
chisq.test(bank_mrktng$treatment, bank_mrktng$day_of_week)

# Divide based on contact frequency
mktg_lessThan3 <- bank_mrktng %>% filter(campaign <= 2)
mktg_moreThan2 <- bank_mrktng %>% filter(campaign > 2)
# Check distribution of propensity scores
PScore = glm(treatment ~ age + job + marital + education + default + loan + contact + month +
day_of_week + duration , data = bank_mrktng, family = "binomial")$fitted.values
bank_mrktng$PScore = PScore
ggplot(bank_mrktng, aes(x = PScore, color = factor(treatment))) +
  geom_density() +
  xlab("propensity score") +
  theme_minimal()
# Control matching to get similar treatment and control groups
matched_IDs <- matchit(treatment ~ age + job + marital + education + default + loan + contact + month +
day_of_week + duration, data = bank_mrktng, method = 'nearest', distance = "logit", caliper = 0.001,
replace = FALSE, ratio = 1)
summary(matched_IDs)
matched_data = match.data(matched_IDs)

# Evaluatepropensity score distribution, after matching
ggplot(matched_data, aes(x = PScore, color = factor(treatment))) +
  geom_density() + xlab("propensity score") +  theme_minimal()

# Check if greater than 2 calls have an effect on client subscribing to term deposit
matched_data$y_num <- ifelse(matched_data$y == "yes", 1, 0)
summary(glm(y ~ treatment, data = matched_data, family="binomial"))

# Sensitivity analysis by changing caliper
matched_IDs_2 <- matchit(treatment ~ age + job + marital + education + default + loan + contact +
month + day_of_week + duration, data = bank_mrktng, method = 'nearest', distance = "logit", caliper =
0.005, replace = FALSE, ratio = 1)
summary(matched_IDs_2)
```

```
matched_data_2 = match.data(matched_IDs_2)

# Plot propensity score distribution
ggplot(matched_data_2, aes(x = PScore, color = factor(treatment))) +
  geom_density() + xlab("propensity score") + theme_minimal()

# Estimate effect
summary(glm(y ~ treatment, data = matched_data_2, family="binomial"))

# Stress test by changing parameters
# Let's change parameters that showed little variation wrt treatment
matched_IDs_3 <- matchit(treatment ~ job + default + loan + contact + month + day_of_week + duration,
data = bank_mrktng, method = 'nearest', distance = "logit", caliper = 0.001, replace = FALSE, ratio = 1)
summary(matched_IDs_3)
matched_data_3 = match.data(matched_IDs_3)

# Plot propensity score distribution
ggplot(matched_data_3, aes(x = PScore, color = factor(treatment))) +
  geom_density() + xlab("propensity score") + theme_minimal()

# Estimate effect
summary(glm(y ~ treatment, data = matched_data_3, family="binomial"))

# Checks for Heterogeneous treatment effect
print(summary(glm(y ~ treatment*job , data = matched_data, family="binomial")))
print(summary(glm(y ~ treatment*marital , data = matched_data, family="binomial")))
print(summary(glm(y ~ treatment*education , data = matched_data, family="binomial")))
print(summary(glm(y ~ treatment*housing , data = matched_data, family="binomial")))
print(summary(glm(y ~ treatment*loan , data = matched_data, family="binomial")))
print(summary(glm(y ~ treatment*contact , data = matched_data, family="binomial")))
print(summary(glm(y ~ treatment*month , data = matched_data, family="binomial")))
print(summary(glm(y ~ treatment*day_of_week , data = matched_data, family="binomial")))
```