

Machine Learning Project Phase 1
Data-Driven Approach to Predict the Success of Bank Telemarketing

Names: Piyush Bhatt

Chapter 1

Introduction:

1.1 Objective

The objective of the report is a data driven approach to predict the Success of Bank Telemarketing. The data sets were sourced from the UCI Machine Learning Repository at <http://archive.ics.uci.edu/ml/datasets/Bank+Marketing#>. This project has 2 phases. Phase 1 focuses on the pre-processing and exploration of the data and in the Second Phase a model for the prediction will be build.

Section 2 describes the datasets and their attributes followed by section 3 which will have the pre-processing like checking missing values or outliers if present then replacing or removing them from the data. Further section 4 will covers important attributes and their inter relationship and the last section will present the summary.

1.2 Data Sets

From UCI machine learning repository have selected the bank.zip file and it contains 2 files bank.csv(cleaned) and bank-full.csv(unclean) files and we have selected the bank-full file for this phase for the data exploration. The training data has 45211 observations and number of attributes 17 in which 16 are the descriptive features and 1(y) is the target variable.

1.3 Target Feature

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. Y is a binary target variable.

Y = Yes (Clients has subscribed to term deposit)

No (Client has not subscribed to term deposit)

The goal is to predict whether the client subscribed to term deposit or not.

1.4 Descriptive features

The variable description is sourced from the UCI Repository.

- Age (numeric)
- Job: type of job (categorical: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
- Marital: marital status (categorical: 'divorced', 'married', 'single', 'unknown'; note: 'divorced' means divorced or widowed)
- Education (categorical: 'basic.4y', 'basic.6y', 'basic.9y', 'high. school', 'illiterate', 'professional.course', 'university.degree', 'unknown')
- Default: has credit in default? (categorical: 'no', 'yes', 'unknown')
- Housing: has housing loan? (categorical: 'no', 'yes', 'unknown')
- Loan: has personal loan? (categorical: 'no', 'yes', 'unknown')
- Contact: contact communication type (categorical: 'cellular', 'telephone')
- Month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
- Day_of_week: last contact day of the week (categorical: 'mon', 'tue', 'wed', 'thu', 'fri')
- Duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.
- Campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
- Pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
- Previous: number of contacts performed before this campaign and for this client (numeric)
- Poutcome: outcome of the previous marketing campaign (categorical: 'failure', 'non-existent', 'success')

Chapter-2

Data Pre-Processing

2.1 Preliminaries

Reading the data from the location of the file saved.

CODE:

```
#all required library
library(ggplot2)
library(car)
library(dplyr)
library(lattice)
library(tidyr)
library(caret)
library(MASS)
library(broom)
library(ROCR)

theme_set(theme_minimal())
rm(list = ls())
options(scipen = 999)
# Read the data
bank_data <- read.csv("file:///C:/Users/bhatt/Desktop/Machine Learning/Assignment-1/bank/bank-full.csv",
                      sep = ";")

# Data Pre-Processing
bank_data <- subset(bank_data, bank_data$poutcome != "other")

bank_data$education <- plyr::revalue(bank_data$education, c("unknown" = "other"))
bank_data$job <- plyr::revalue(bank_data$job, c("unknown" = "other"))
```

OUTPUT-

```
> head(bank_data)
  age  job marital education default balance housing loan contact day month duration campaign pdays previous poutcome y
1  58 management married tertiary no    2143   yes  no unknown  5  may    261         1    -1         0 unknown no
2  44 technician single  secondary no     29   yes  no unknown  5  may    151         1    -1         0 unknown no
3  33 entrepreneur married secondary no     2   yes  yes unknown  5  may     76         1    -1         0 unknown no
4  47 blue-collar married unknown no   1506   yes  no unknown  5  may     92         1    -1         0 unknown no
5  33 unknown single unknown no     1   no   no unknown  5  may    198         1    -1         0 unknown no
6  35 management married tertiary no    231   yes  no unknown  5  may    139         1    -1         0 unknown no
```

Pre-Processing the data

- Poutcome- dropped the rows with the value 'other'.
- Education- replaced the value 'unknown' with 'other'.
- Job- replaced the value 'unknown' with 'other'.
- Missing Value- Checked the missing value for the full data but didn't find any.
- Contact- removed the attribute as it has no impact on the target variable.
- Education- dropped the records with education as 'other'.
- Target- changed the name to 'Target' from 'y'.

CODE:

```
# Data Pre-Processing
bank_data <- subset(bank_data, bank_data$poutcome != "other")

bank_data$education <- plyr::revalue(bank_data$education, c("unknown" = "other"))
bank_data$job <- plyr::revalue(bank_data$job, c("unknown" = "other"))
# Check missing value in Numeric columns
num_var <- select_if(bank_data, is.numeric)
colSums(sapply(num_var, is.na))

# Check missing values in Categorical columns
cat_var <- select_if(bank_data, is.factor)
colSums(sapply(cat_var, is.na))

# Summarize the numerical variables
summary(num_var)

# Summarize the categorical variables
summary(cat_var$poutcome)

# Explore the target variable
table(bank_data$y)
```

OUTPUT-

```
colSums(sapply(num_var, is.na))
  age balance      day duration campaign      pdays previous
  0      0      0      0      0      0      0      0

colSums(sapply(cat_var, is.na))
  job marital education default housing      loan contact      month poutcome      y
  0      0      0      0      0      0      0      0      0      0

> summary(num_var)
   age      balance      day      duration      campaign      pdays
Min. :18.00 Min. : -8019 Min. : 1.00 Min. : 0.0 Min. : 1.000 Min. : -1.00
1st Qu.:33.00 1st Qu.:  70 1st Qu.: 8.00 1st Qu.:103.0 1st Qu.: 1.000 1st Qu.: -1.00
Median :39.00 Median : 443 Median:16.00 Median:180.0 Median: 2.000 Median: -1.00
Mean :40.99 Mean : 1357 Mean :15.86 Mean : 258.3 Mean : 2.777 Mean : 32.16
3rd Qu.:48.00 3rd Qu.:1417 3rd Qu.:21.00 3rd Qu.:318.0 3rd Qu.: 3.000 3rd Qu.: -1.00
Max. :95.00 Max. :102127 Max. :31.00 Max. :4918.0 Max. :63.000 Max. :871.00
 previous
Min. : 0.0000
1st Qu.: 0.0000
Median : 0.0000
Mean : 0.4349
3rd Qu.: 0.0000
Max. :55.0000

> # Summarize the categorical variables
> summary(cat_var$poutcome)
failure      other      success      unknown
 4901          0      1511      36959

> # Explore the target variable
> table(bank_data$y)

   no   yes
38389 4982
```

CODE:

```
# Visualize the balance to check the outliers and remove them if any
outliers <- boxplot(bank_data$balance, horizontal = TRUE, plot = FALSE)$out
bank_data <- bank_data[-which(bank_data$balance %in% outliers),]
boxplot(bank_data$balance, horizontal = TRUE)

# Remove the column contact as it has no impact on target variable y
bank_data$contact <- NULL
# Keep records which has call duration of more than 5 seconds
bank_data <- subset.data.frame(bank_data, bank_data$duration > 5)
# Drop the records for customer with education as other
bank_data <- subset(bank_data, bank_data$education != "other")
cat_var <- select_if(bank_data, is.factor)
summary(cat_var)

# Rename the y variable as target
names(bank_data)[length(bank_data)] <- "Target"
```

OUTPUT:

```
> summary(cat_var)
```

job	marital	education	default	housing	loan	month
blue-collar:8157	divorced: 4395	primary : 5930	no :36429	no :16170	no :30725	may :11659
management :7586	married :22304	secondary:20267	yes: 753	yes:21012	yes: 6457	jul : 6143
technician :6364	single :10483	tertiary :10985				aug : 5343
admin. :4352		other : 0				jun : 4340
services :3554						nov : 2817
retired :1746						apr : 2274
(other) :5423						(Other): 4606

outcome	y
failure: 4169	no :33113
other : 0	yes: 4069
success: 1218	
unknown:31795	

Chapter -3

Data Exploration

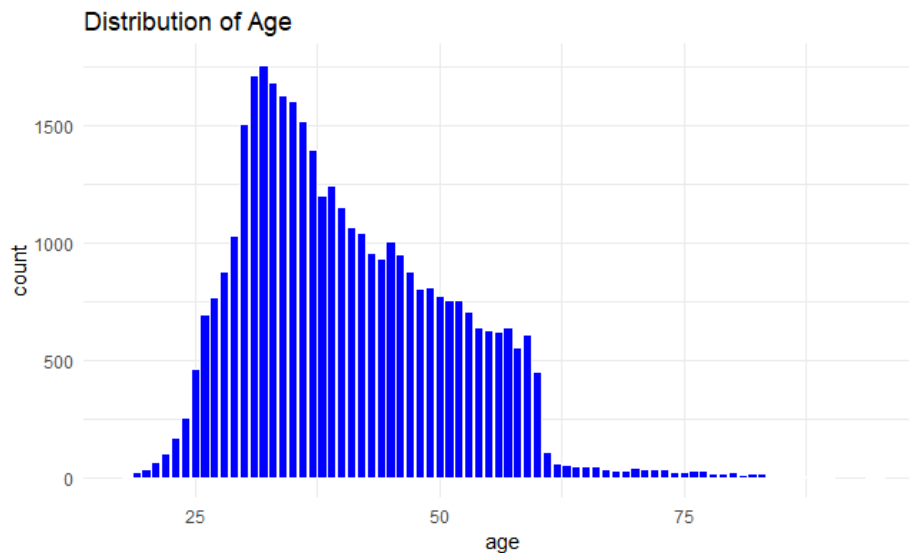
3.1 Univariate Visualisation

Defined two plots one is Bar plot for numerical variable (age) and other one is histogram for numerical variable (balance).

CODE:

```
# Distribution of age
p <- ggplot(bank_data, aes(x = age))
p + geom_bar(color = "white",
              fill = "blue") + theme_minimal() + labs(title = "Distribution of Age")
```

OUTPUT-

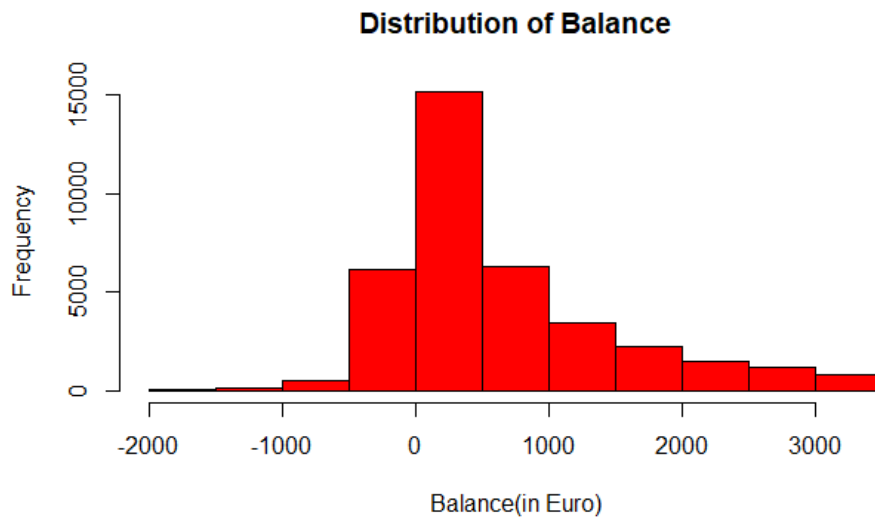


Analysing distribution of age it can interpreted that majority of the customers contacted by bank is between 25-50 years.

.CODE:

```
# Distribution of Balance
hist(bank_data$balance, fill = "red", col = "red",
     main = "Distribution of Balance",
     xlab = "Balance(in Euro)")
```

OUTPUT-



Analysing the balance histogram plot it can be said that maximum balance of the customers is between Dollars 0 to 1000.

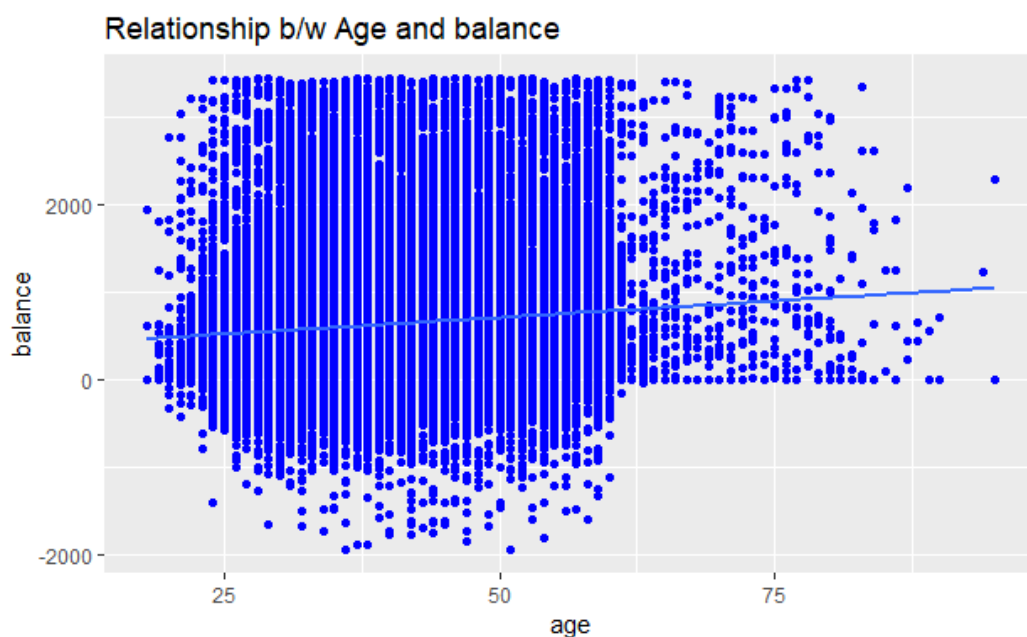
3.2 Bi-variate Visualisation

3.2.1 Relationship between Age and Balance

CODE:

```
# Relationship between age and balance
d <- ggplot(bank_data, aes(x = age, y = balance))
d + geom_point(color = "blue") + labs(title = "Relationship b/w Age and balance") + geom_smooth(method = "lm", se = F)
```

OUTPUT-



Based on the scatter plot no clear relationship can be interpreted between the clients age and their balance.

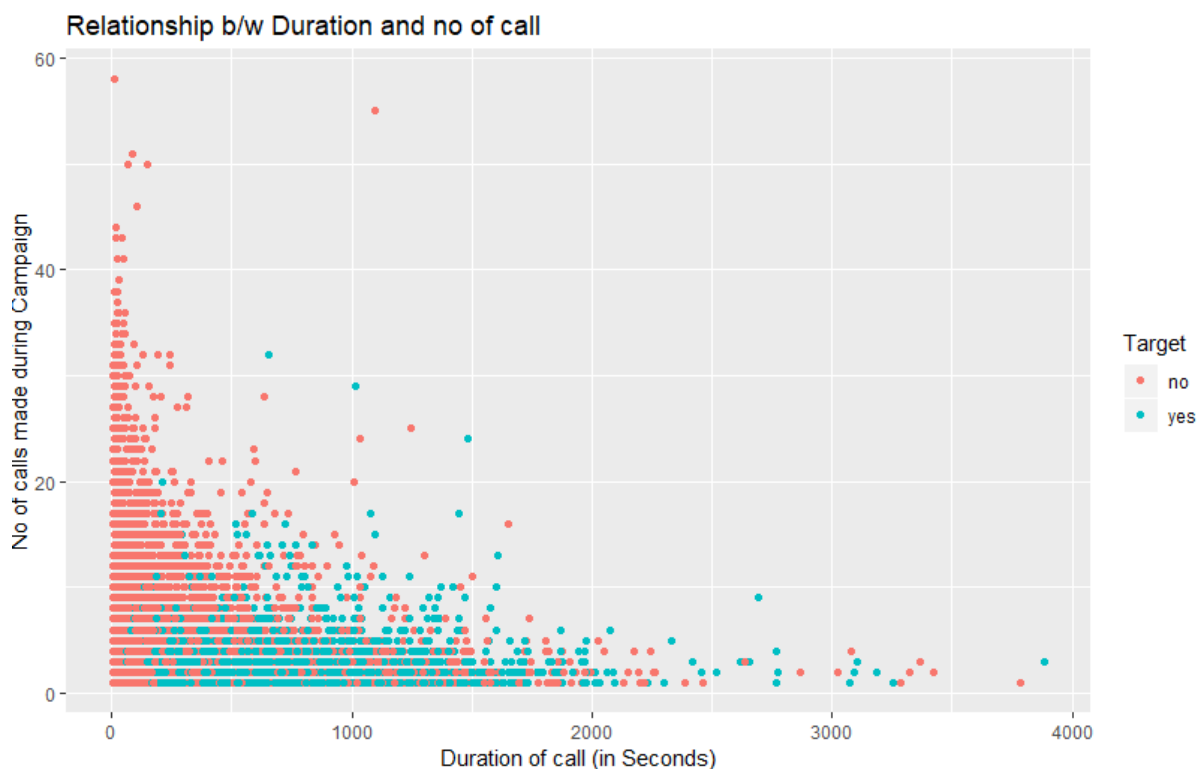
3.2.2 Relationship between duration and campaign with response rate.

CODE:

```
# Relationship b/w duration and campaign with response rate

ggplot(bank_data, aes(x = bank_data$duration, y = bank_data$campaign)) +
  geom_point(aes(col = Target)) + labs(title = "Relationship b/w Duration and no of call",
    x = "Duration of call (in Seconds)",
    y = "No of calls made during Campaign")
```

OUTPUT-



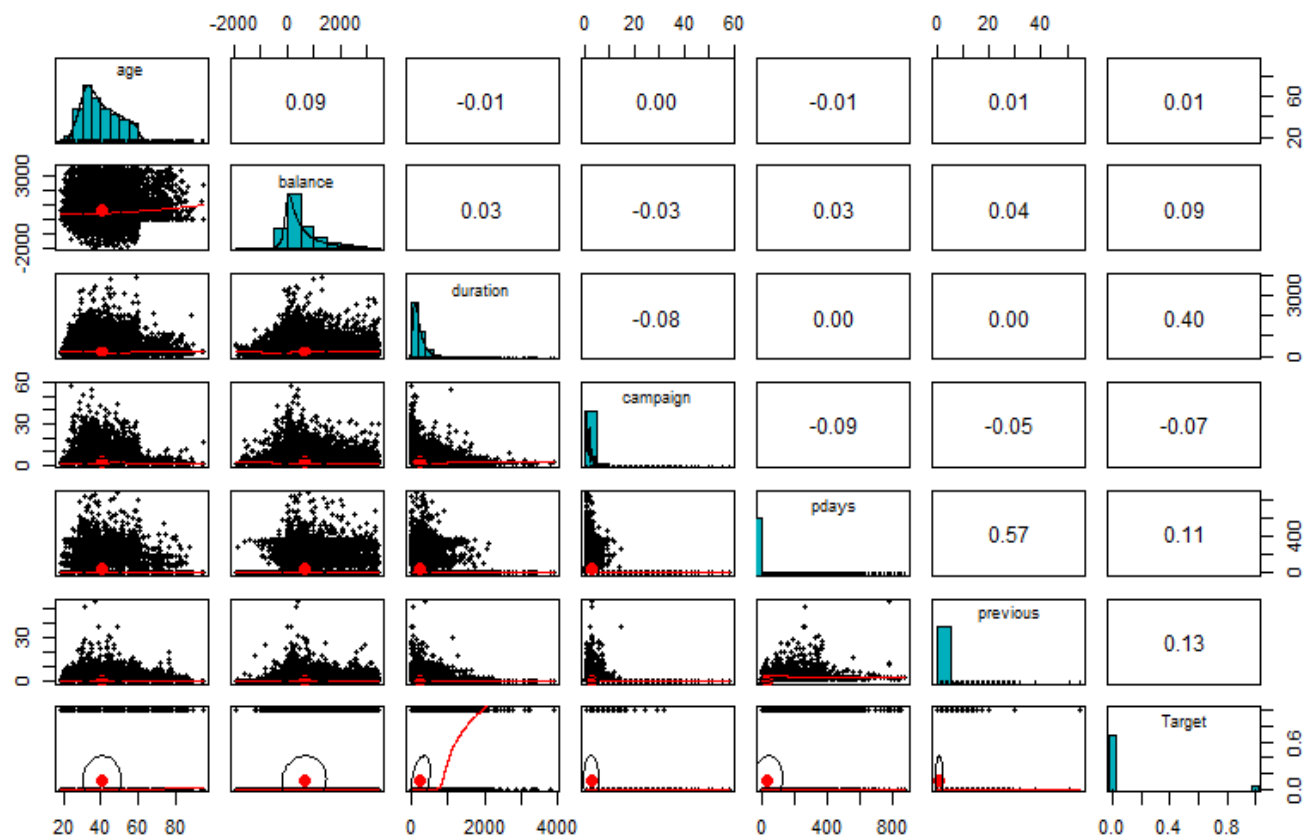
From the scatterplot it can be said the clients who subscribed were contacted few times and had longer call duration as compared to clients with 'no' response were contacted many times with shorter call duration.

3.3 Scatter matrix and correlation matrix:

CODE:

```
# Correlation between all the numerical variables with target variable
bank_data$Target <- ifelse(bank_data$Target == "yes", 1,0)
sub_data <- bank_data[,c("age", "balance", "duration", "campaign", "pdays", "previous", "Target")]
library(psych)
pairs.panels(sub_data,
  method = "pearson", # correlation method
  hist.col = "#00AFBB",
  density = TRUE, # show density plots
  ellipses = TRUE # show correlation ellipses
)
```

OUTPUT:



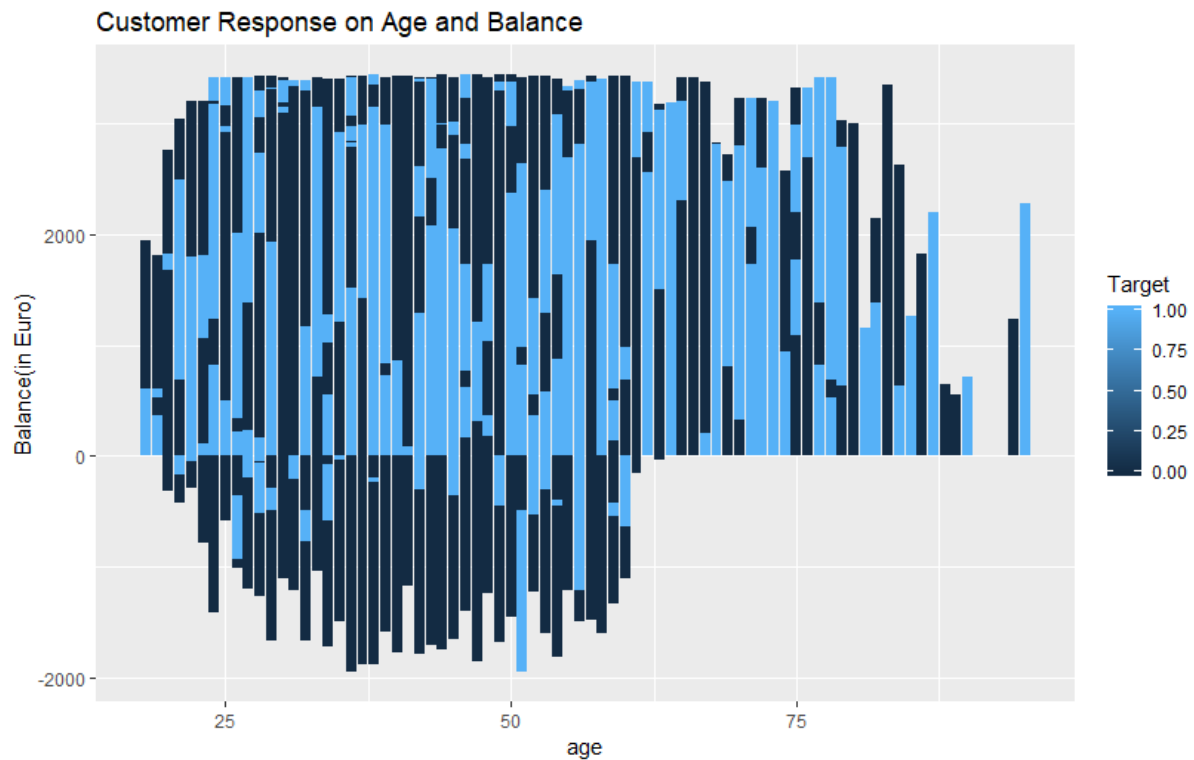
From the above scatter matrix correlation plot it can be interpreted that duration has maximum impact over the target variable as compared to other variables also 'pdays' and 'previous' has intermediate impact with campaign as less impact.

3.4 Customer response on Age and Balance with respect to Target Variable

CODE:

```
# subscribe
range(bank_data$balance)
# visualise target variable with respect to different different predictors
ggplot(data = bank_data, aes(x = age, y = bank_data$balance, fill = Target)) + geom_bar(stat = "identity", position = position_dodge()) +
  labs(title = "Customer Response on Age and Balance",
        y = "Balance(in Euro)") + scale_color_manual(values = c("#999999", "#56B4E9"))
bank_data$Target <- ifelse(bank_data$Target == 1, "Subscribed", "Not Subscribed")
```

OUTPUT-



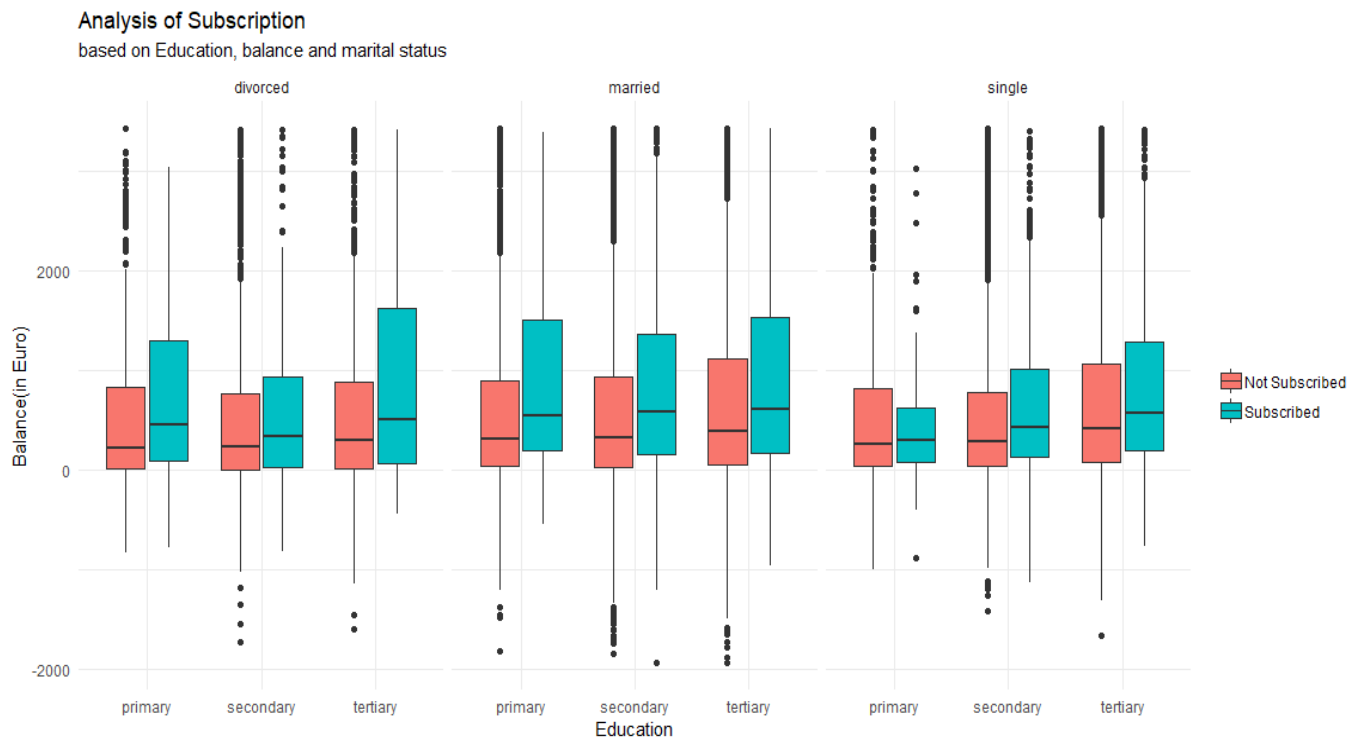
From the bar plot it can be interpreted the willingness to subscribe is higher for people aged between 25 to 60 years. and the effect of the balance on every individual can be seen. Can be said the bank should prioritise its telemarketing for the customers aged above 60 years.

3.5 Analysis of subscription based on Education, balance and marital status

CODE:

```
# Boxplot between numerical variables|
e <- ggplot(data = bank_data, aes(x = education, y = balance, fill = Target))
e + geom_boxplot() + labs(y = "Balance(in Euro)",
                          x = "Education",
                          title = "Analysis of Subscription ",
                          subtitle = "based on Education, balance and marital status") +
  theme(legend.title = element_blank()) + facet_wrap(~ marital)
```

OUTPUT:



From the above box plot it can be interpreted-

For divorced- more number of customer with tertiary education with more balance and then comes the primary.

For married- there are almost similar number of customers with primary and tertiary education and similar balance.

For single – the more the customer education level the more chances of subscribing.

3.6 Subscription rate base on job, balance and loan:

```
# Barplot between loan and balacne with respect to target variable.|  
f <- ggplot(bank_data, aes(x= loan, y = balance, fill = Target))  
f + geom_bar(stat = "identity", position = position_dodge()) +  
  labs(title = "Subscription rate",  
        subtitle = "Based on Job, Balance, and loan") + facet_wrap(~ job)
```



From the above bar plot it can be interpreted that if a customer is a student then and doesn't have any loan with a good balance is more likely to be a subscribed member. Also if the customer is a housemaid the chances of being a subscribed member is very low. Students and retired customer account for around half of the subscribed customer base.

Summary

From the data exploration and all the visualisation, further analysing the graphs it can be interpreted that the bank should target below following customer:

1. Either students or retired personnel as their probability of subscribing is more than other people.
2. Age- similarly bank should target age group of <30 or >60 years.

References

- <http://archive.ics.uci.edu/ml/datasets/Bank+Marketing#>
- https://l.facebook.com/l.php?u=http%3A%2F%2Fpubs.com%2FMentors%20Ubiquum%2Fremoving_outliers%3Ffbclid%3DIwAR1fs26SoUP2VfNdHigalSDAiyWfWlQ9eiVY1bXSux9Tua0ChxSv3o2tdC4&h=AT2vvu9I7n7IP0j3ydogGW5vd8vgYVpy-8W28CS8JrtbeMBSjt1FJe8IUisnKULfQLRxQ3L0SORqsAacTkWKeSrcFJhW8eJ9VVjrd61HPQh7idtFpvERaAPNMCJo2IEr9K9bfg
- https://l.facebook.com/l.php?u=http%3A%2F%2Fwww.sthda.com%2Fenglish%2Fwiki%2Fscatter-plot-matrices-r-base-graphs%3Ffbclid%3DIwAR316s8pn5lvDnuGDhXeSCiQ6ZaWvQo9imW1yN13Cu7R1SRQrsJX_o nvzwU&h=AT2vvu9I7n7IP0j3ydogGW5vd8vgYVpy-8W28CS8JrtbeMBSjt1FJe8IUisnKULfQLRxQ3L0SORqsAacTkWKeSrcFJhW8eJ9VVjrd61HPQh7idtFpvERaAPNMCJo2IEr9K9bfg