Bayer Execise 1

Borja

16/3/2022

```
df <- read.csv("data/real_case/data_usecase.csv", sep = ';')
head(df)
```

Exercise 1

```
##   ï..HCP_id news_id Message_id Message_type Message_creation_date
## 1      3569      22         71      Biotech            12/10/2019
## 2      3569      22         70      Biotech            31/07/2018
## 3      5941      22         70      Biotech            31/07/2018
## 4      5941      22         71      Biotech            12/10/2019
## 5      8262       5         19      Biotech            12/10/2018
## 6      3569      24         75      Biotech            08/06/2018
##        Message_TA  news_date office_or_hospital_based gender is_cardiologist
## 1 Cardio Vascular 09/08/2020                Hospital  False            True
## 2          Global 09/08/2020                Hospital  False            True
## 3          Global 09/08/2020                Hospital   True            True
## 4 Cardio Vascular 09/08/2020                Hospital   True            True
## 5 Cardio Vascular 22/07/2020                  Office   True            True
## 6          Global 23/05/2020                Hospital  False            True
##   is_gp years_since_graduation Message_read Message_click
## 1 False                     10            1             0
## 2 False                     10            0             0
## 3 False                     10            1             1
## 4 False                     10            1             1
## 5 False                     10            0             0
## 6 False                     10            0             0
```

```
summary(df)
```

```
##    ï..HCP_id       news_id        Message_id     Message_type
##  Min.   :   1   Min.   : 1.00   Min.   :  1.00   Length:55543
##  1st Qu.:2092   1st Qu.: 8.00   1st Qu.: 26.00   Class :character
##  Median :4182   Median :17.00   Median : 51.00   Mode  :character
##  Mean   :4170   Mean   :15.43   Mean   : 50.67
##  3rd Qu.:6256   3rd Qu.:24.00   3rd Qu.: 76.00
##  Max.   :8349   Max.   :30.00   Max.   :100.00
##
##  Message_creation_date  Message_TA         news_date
##  Length:55543          Length:55543       Length:55543
##  Class :character      Class :character   Class :character
##  Mode  :character      Mode  :character   Mode  :character
##
##
```

```
##
##
##  office_or_hospital_based    gender          is_cardiologist
##  Length:55543             Length:55543       Length:55543
##  Class :character         Class :character   Class :character
##  Mode  :character         Mode  :character   Mode  :character
##
##
##
##
##     is_gp           years_since_graduation  Message_read    Message_click
##  Length:55543       Min.   : 3.00          Min.   :0.0000   Min.   :0.0000
##  Class :character   1st Qu.:17.00          1st Qu.:0.0000   1st Qu.:0.0000
##  Mode  :character   Median :28.00          Median :1.0000   Median :0.0000
##                     Mean   :27.14          Mean   :0.5049   Mean   :0.2601
##                     3rd Qu.:37.00          3rd Qu.:1.0000   3rd Qu.:1.0000
##                     Max.   :68.00          Max.   :1.0000   Max.   :1.0000
##                     NA's   :6
```

Clean NAs

```
df_nona <- na.omit(df)
dim(df)
```

```
## [1] 55543    14
```

```
dim(df_nona)
```

```
## [1] 55503    14
```

Little visualization:

```
interest_features <- c("Message_type","Message_TA","office_or_hospital_based","gender","is_cardiologist
feature_read <- 'Message_read'
feature_clicked <- 'Message_click'
for (feature in interest_features)
{
  if (!((feature == feature_read) || (feature == feature_clicked))){
      print(paste('Feature: ',feature))
      print(data.frame(table(df_nona[feature])) %>% kbl() %>%  kable_paper("hover", full_width = F))
  }

}
```

```
## [1] "Feature:  Message_type"
## \begin{table}
## \centering
## \begin{tabular}[t]{l|r}
## \hline
## Var1 & Freq\\
## \hline
## Biotech & 8829\\
## \hline
## Clinical trial update & 11437\\
## \hline
## Medical study & 8282\\
## \hline
```

```
## Product launch & 9510\\
## \hline
## Service / Applications & 8996\\
## \hline
## Webinar & 8449\\
## \hline
## \end{tabular}
## \end{table}
## [1] "Feature:  Message_TA"
## \begin{table}
## \centering
## \begin{tabular}[t]{l|r}
## \hline
## Var1 & Freq\\
## \hline
## Cardio Vascular & 9638\\
## \hline
## Else & 9361\\
## \hline
## Global & 17702\\
## \hline
## Oncology & 9402\\
## \hline
## Ophtalmology & 9400\\
## \hline
## \end{tabular}
## \end{table}
## [1] "Feature:  office_or_hospital_based"
## \begin{table}
## \centering
## \begin{tabular}[t]{l|r}
## \hline
## Var1 & Freq\\
## \hline
## Hospital & 6103\\
## \hline
## Office & 49400\\
## \hline
## \end{tabular}
## \end{table}
## [1] "Feature:  gender"
## \begin{table}
## \centering
## \begin{tabular}[t]{l|r}
## \hline
## Var1 & Freq\\
## \hline
## False & 36568\\
## \hline
## True & 18935\\
## \hline
## \end{tabular}
## \end{table}
## [1] "Feature:  is_cardiologist"
```

```
## \begin{table}
## \centering
## \begin{tabular}[t]{l|r}
## \hline
## Var1 & Freq\\
## \hline
## False & 48779\\
## \hline
## True & 6724\\
## \hline
## \end{tabular}
## \end{table}
## [1] "Feature:  is_gp"
## \begin{table}
## \centering
## \begin{tabular}[t]{l|r}
## \hline
## Var1 & Freq\\
## \hline
## False & 6724\\
## \hline
## True & 48779\\
## \hline
## \end{tabular}
## \end{table}
## [1] "Feature:  years_since_graduation"
## \begin{table}
## \centering
## \begin{tabular}[t]{l|r}
## \hline
## Var1 & Freq\\
## \hline
## 3 & 7\\
## \hline
## 4 & 267\\
## \hline
## 5 & 558\\
## \hline
## 6 & 916\\
## \hline
## 7 & 1030\\
## \hline
## 8 & 1262\\
## \hline
## 9 & 1308\\
## \hline
## 10 & 1217\\
## \hline
## 11 & 1449\\
## \hline
## 12 & 1190\\
## \hline
## 13 & 1124\\
## \hline
```

```
## 14 & 1091\\
## \hline
## 15 & 1098\\
## \hline
## 16 & 1025\\
## \hline
## 17 & 1097\\
## \hline
## 18 & 838\\
## \hline
## 19 & 1132\\
## \hline
## 20 & 1187\\
## \hline
## 21 & 1338\\
## \hline
## 22 & 1396\\
## \hline
## 23 & 1317\\
## \hline
## 24 & 1233\\
## \hline
## 25 & 1418\\
## \hline
## 26 & 1469\\
## \hline
## 27 & 1373\\
## \hline
## 28 & 1382\\
## \hline
## 29 & 1442\\
## \hline
## 30 & 1585\\
## \hline
## 31 & 1657\\
## \hline
## 32 & 1666\\
## \hline
## 33 & 1615\\
## \hline
## 34 & 1395\\
## \hline
## 35 & 1625\\
## \hline
## 36 & 1528\\
## \hline
## 37 & 1302\\
## \hline
## 38 & 1507\\
## \hline
## 39 & 1494\\
## \hline
## 40 & 1309\\
## \hline
```

```
## 41 & 1394\\
## \hline
## 42 & 1022\\
## \hline
## 43 & 1130\\
## \hline
## 44 & 939\\
## \hline
## 45 & 749\\
## \hline
## 46 & 714\\
## \hline
## 47 & 456\\
## \hline
## 48 & 496\\
## \hline
## 49 & 410\\
## \hline
## 50 & 277\\
## \hline
## 51 & 234\\
## \hline
## 52 & 215\\
## \hline
## 53 & 167\\
## \hline
## 54 & 140\\
## \hline
## 55 & 122\\
## \hline
## 56 & 60\\
## \hline
## 57 & 24\\
## \hline
## 58 & 35\\
## \hline
## 59 & 32\\
## \hline
## 60 & 17\\
## \hline
## 63 & 15\\
## \hline
## 68 & 8\\
## \hline
## \end{tabular}
## \end{table}
```

Option A: Decision tree

Option B: Logistic Regression