

Proyecto FMAD

ICAI. Máster en Big Data. Fundamentos Matemáticos del Análisis de Datos (FMAD).

Curso 2021-22. Última actualización: 2021-10-01



Índice

Introducción	3
Definición de las variables	4
Including Plots	5
TODO Resumen de datos	6
TODO Preprocesamiento de los datos	7
TODO Visualización de los datos	8
TODO Buscar la relación posible entre distintas variable	8
TODO Realizar algún modelo predictivo sobre variables target como el ‘complain’	8
TODO Convertir a booleana la variable target	8

Introducción

Cargamos las librerías

Leemos los datos

```
datos <- read.csv("marketing_campaign.csv", header = TRUE, sep = "")
head(datos)
```

##	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer
## 1	5524	1957	Graduation	Single	58138	0	0	04-09-2012
## 2	2174	1954	Graduation	Single	46344	1	1	08-03-2014
## 3	4141	1965	Graduation	Together	71613	0	0	21-08-2013
## 4	6182	1984	Graduation	Together	26646	1	0	10-02-2014
## 5	5324	1981	PhD	Married	58293	1	0	19-01-2014
## 6	7446	1967	Master	Together	62513	0	1	09-09-2013

##	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts
## 1	58	635	88	546	172	88
## 2	38	11	1	6	2	1
## 3	26	426	49	127	111	21
## 4	26	11	4	20	10	3
## 5	94	173	43	118	46	27
## 6	16	520	42	98	0	42

##	MntGoldProds	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases
## 1	88	3	8	10
## 2	6	2	1	1
## 3	42	1	8	2
## 4	5	2	2	0
## 5	15	5	5	3
## 6	14	2	6	4

##	NumStorePurchases	NumWebVisitsMonth	AcceptedCmp3	AcceptedCmp4	AcceptedCmp5
## 1	4	7	0	0	0
## 2	2	5	0	0	0
## 3	10	4	0	0	0
## 4	4	6	0	0	0
## 5	6	5	0	0	0
## 6	10	6	0	0	0

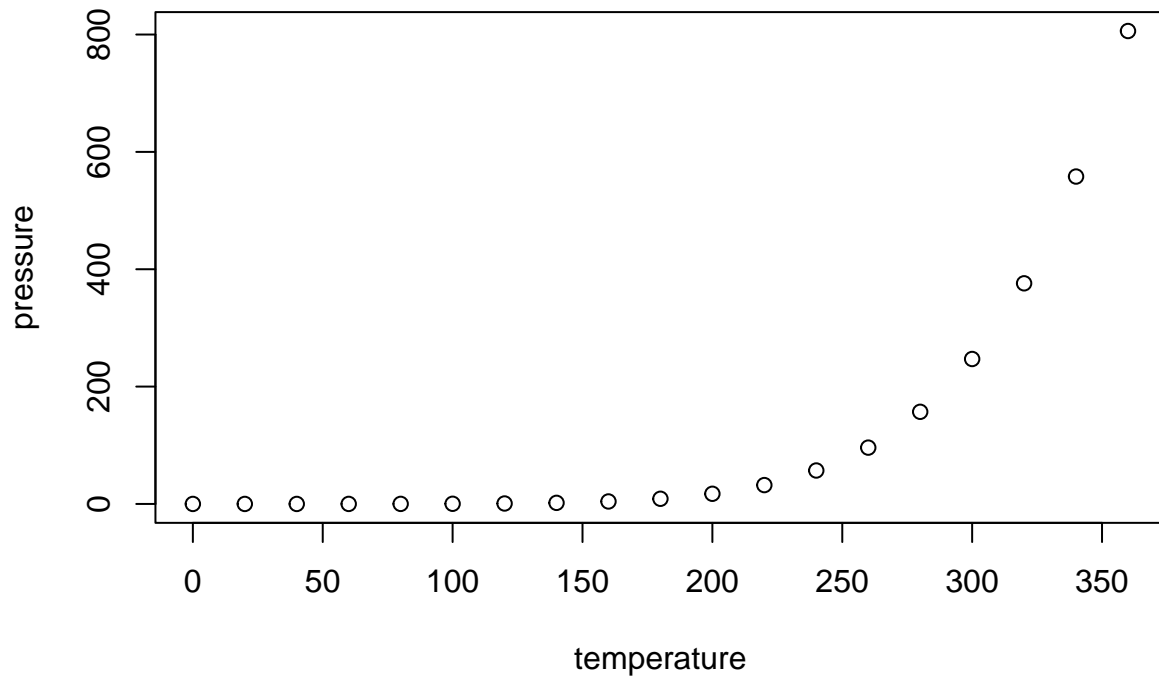
##	AcceptedCmp1	AcceptedCmp2	Complain	Z_CostContact	Z_Revenue	Response
## 1	0	0	0	3	11	1
## 2	0	0	0	3	11	0
## 3	0	0	0	3	11	0
## 4	0	0	0	3	11	0
## 5	0	0	0	3	11	0
## 6	0	0	0	3	11	0

Definición de las variables

- **ID:** El ID del cliente.
- **Year__Birth:** Año de nacimiento.
- **Education:** Nivel de educación del cliente.
- **Marital__Status:** Estado civil del cliente.
- **Income:** Ingreso familiar anual del cliente.
- **Kidhome:** Número de niños pequeños en casa del cliente.
- **Teenhome:** Número de adolescentes en el hogar del cliente.
- **Dt__Customer:** Fecha de inscripción del cliente en la empresa.
- **Recency:** Número de días desde la última compra.
- **MntWines:** Gasto en productos vitivinícolas en los últimos 2 años.
- **MntGoldProds:** Gasto en productos “premium”?? en los últimos 2 años.
- **NumDealsPurchases:** Número de compras con uso de descuento.
- **NumWebPurchases:** Número de compras a través de la web.
- **NumCatalogPurchases:** Número de compras usando catalogo.
- **NumWebVisitsMonth:** ¿¿¿¿Dicen que es lo mismo que las compras a traves de la web???
Pero no seria visitas por mesa la web
- **AcceptedCmp1:** 1 si el cliente acepta la oferta en la 1ra campaña, 0 si no lo acepta.
- **AcceptedCmp2:** 1 si el cliente acepta la oferta en la 2nd campaña, 0 si no lo acepta.
- **Complain:** 1 si el cliente se ha quejado en los dos últimos años.
- **Z__CostContact:** Coste de contactar con cliente.
- **Z__Revenue:** Ingresos/Beneficios después de que el cliente acepte la campaña
- **Response:** 1 si el cliente acepta la oferta en la última campaña y 0 si no la acepta.

Including Plots

You can also embed plots, for example:



TODO Resumen de datos

```
cat(cat(cat(cat("El conjunto de datos tiene", nrow(datos)), "filas y"), ncol(datos)), "columnas")
```

```
## El conjunto de datos tiene 2440 filas y 29 columnas
```

```
head(datos)
```

```
##      ID Year_Birth Education Marital_Status Income Kidhome Teenhome Dt_Customer
## 1 5524      1957 Graduation      Single  58138      0      0 04-09-2012
## 2 2174      1954 Graduation      Single  46344      1      1 08-03-2014
## 3 4141      1965 Graduation Together  71613      0      0 21-08-2013
## 4 6182      1984 Graduation Together  26646      1      0 10-02-2014
## 5 5324      1981      PhD      Married  58293      1      0 19-01-2014
## 6 7446      1967      Master Together  62513      0      1 09-09-2013
##      Recency MntWines MntFruits MntMeatProducts MntFishProducts MntSweetProducts
## 1      58      635      88      546      172      88
## 2      38      11      1      6      2      1
## 3      26     426     49     127     111     21
## 4      26      11      4     20     10      3
## 5      94     173     43     118     46     27
## 6      16     520     42     98      0     42
##      MntGoldProds NumDealsPurchases NumWebPurchases NumCatalogPurchases
## 1      88      3      8      10
## 2      6      2      1      1
## 3     42      1      8      2
## 4      5      2      2      0
## 5     15      5      5      3
## 6     14      2      6      4
##      NumStorePurchases NumWebVisitsMonth AcceptedCmp3 AcceptedCmp4 AcceptedCmp5
## 1      4      7      0      0      0
## 2      2      5      0      0      0
## 3     10      4      0      0      0
## 4      4      6      0      0      0
## 5      6      5      0      0      0
## 6     10      6      0      0      0
##      AcceptedCmp1 AcceptedCmp2 Complain Z_CostContact Z_Revenue Response
## 1      0      0      0      3      11      1
## 2      0      0      0      3      11      0
## 3      0      0      0      3      11      0
## 4      0      0      0      3      11      0
## 5      0      0      0      3      11      0
## 6      0      0      0      3      11      0
```

Hay 2 columnas al final: “Z_CostContact” y “Z_Revenue” que no sé qué son.

TODO Preprocesamiento de los datos

En esta sección haría los siguientes cambios para dejar un datasets más simple. Por un lado cogería las columnas “NumWebPurchases”, “NumCatalogPurchases” y “NumStorePurchases” que indican el lugar por donde se han hecho las ofertas a cada cliente y las sumaría todas en una única columna que indique el número de ofertas totales que ha recibido el cliente.

```
datos %>%
  replace(is.na(.), 0) %>%
  rowwise(ID) %>%
  mutate(suma_ofertas = sum(c(NumWebPurchases, NumStorePurchases)))

## # A tibble: 2,440 x 30
## # Rowwise: ID
##       ID Year_Birth Education Marital_Status Income Kidhome Teenhome Dt_Customer
##   <int>   <dbl> <chr>      <chr>          <chr>   <dbl> <chr>   <chr>
## 1  5524    1957 Graduation Single         58138     0 0     04-09-2012
## 2  2174    1954 Graduation Single         46344     1 1     08-03-2014
## 3  4141    1965 Graduation Together        71613     0 0     21-08-2013
## 4  6182    1984 Graduation Together        26646     1 0     10-02-2014
## 5  5324    1981 PhD          Married         58293     1 0     19-01-2014
## 6  7446    1967 Master       Together        62513     0 1     09-09-2013
## 7   965    1971 Graduation Divorced         55635     0 1     13-11-2012
## 8  6177    1985 PhD          Married         33454     1 0     08-05-2013
## 9  4855    1974 PhD          Together        30351     1 0     06-06-2013
## 10 5899    1950 PhD          Together         5648     1 1     13-03-2014
## # ... with 2,430 more rows, and 22 more variables: Recency <chr>,
## #   MntWines <dbl>, MntFruits <dbl>, MntMeatProducts <dbl>,
## #   MntFishProducts <dbl>, MntSweetProducts <dbl>, MntGoldProds <dbl>,
## #   NumDealsPurchases <dbl>, NumWebPurchases <dbl>, NumCatalogPurchases <dbl>,
## #   NumStorePurchases <dbl>, NumWebVisitsMonth <dbl>, AcceptedCmp3 <dbl>,
## #   AcceptedCmp4 <dbl>, AcceptedCmp5 <dbl>, AcceptedCmp1 <dbl>,
## #   AcceptedCmp2 <dbl>, Complain <dbl>, Z_CostContact <dbl>, ...
```

Por otro lado trataría las columnas: “AcceptedCmp1”, “AcceptedCmp2”, “AcceptedCmp3”, “AcceptedCmp4” y “AcceptedCmp5”. que indican si el cliente aceptó la oferta i-ésima. La columna “Response” indica si el cliente aceptó la última oferta. Cogería todas esas columnas y las sumaría y dejaría una única que indique el número de ofertas que ha aceptado el cliente de las últimas 6. Con todas estas columnas habría que pivotar y dejar una única columna que sea en el primer grupo: lugar, y en el segundo una columna que sea la oferta que se aceptó. Si no se pone la oferta i indica que no se aceptó.

Aquí tengo mis dudas. Con las siguiente columnas: “MntWines”, “MntFruits”, “MntMeatProducts”, “MntFishProducts”, “MntSweetProducts”, “MntGoldProds” que marcan la cantidad de dinero que se ha gastado cada cliente en un tipo de producto.

TODO Visualización de los datos

TODO Buscar la relación posible entre distintas variable

TODO Realizar algún modelo predictivo sobre variables target como el ‘complain’

TODO Convertir a booleana la variable target

Preprocesamiento de los datos