

Proyecto

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

Curso 2021-22. Última actualización: 2021-09-30

Contents

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

Introducción

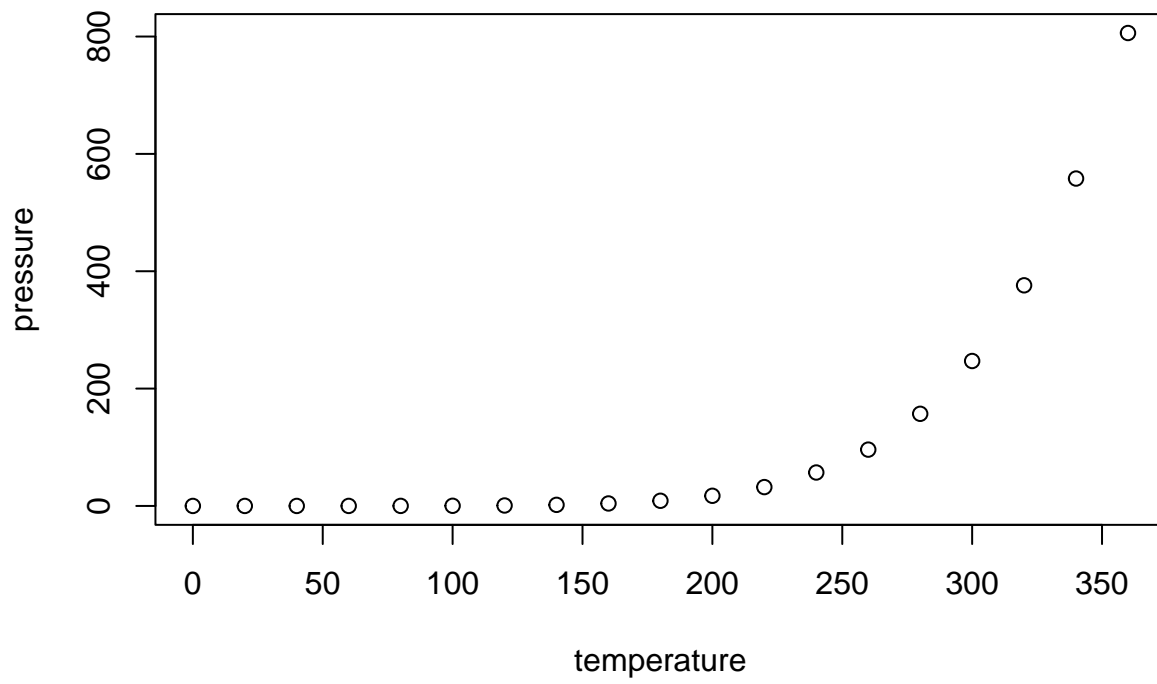
Cargamos las librerías

Leemos los datos

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

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.

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