Control 2 - Computación Estadística con R

Joshua Kunst

22 de julio de 2019

## Instrucciones

* Esto se puede resolver con todo el material visto en clases. Pero no es restrictivo en el sentido de que puedes usar otros comandos que no hemos visto.
* Se puede realizar esta evaluación en grupo de hasta tres (3) personas.
* Usted(es) debe(n) crear un script en donde se especifique los pasos necesarios (código) y comentarios (#). En el caso ideal el script debiese correr sin problemas al ejecutar los comandos en orden.
* La fecha límite para enviar la tarea es el jueves 31 de Julio de 2019 hasta las 23:59 hrs. El método de envío es via correo electrónico a [jbkunst@gmail.com](mailto:jbkunst@gmail.com) con el asunto “Control 2 - Computación Estadística con R” (copiar y pegar por favor).
* Referencias que complementan las presentaciones y scripts vistos en clases:
  + Tutorial de Riva Quiroga <https://github.com/rivaquiroga/RLadies-Santiago/blob/master/primeros_pasos_en_R.md>
  + Traducción parcial al español de “R for data science” <https://es.r4ds.hadley.nz/>

## Los datos

Usando la función read\_csv lea un archivo desde la URL https://raw.githubusercontent.com/jbkunst/puc-introduccion-a-R/master/data/clientes.csv, y cree con ello una tabla llamada clientes.

hint:

library(tidyverse)  
  
clientes <- read\_csv("https://raw.githubusercontent.com/jbkunst/puc-introduccion-a-R/master/data/clientes.csv")

Estos datos corresponden a ventas en cierto banco. Los datos de los clientes corresponden a la información al momento de la entrega del producto financiero (digamos crédito de consumo), y la variable malo\_en\_los\_12m fue un registro a posterior indicado si el cliente en el producto vendido alcanzó una mora mayor a 90, es decir: malo :(

## Explorando:

* ¿Cuántas filas y cuantas columnas contienes los datos?
* Indique como puede conocer los tipos de cada columna

## Caracterizando a los clientes

* ¿Cuantos clientes tienen 0 tarjetas de credito al momento del otorgamiento/venta?
* ¿Cual fue la proporción (número entre 0 y 1, porcentaje) de clientes que en los 12 meses posteriores a la venta que fueron *malos*?
* Genere un histograma de la variable ingreso y comente
* Estudie (pregunta abierta) si existe diferencia del ingreso según género y comente.

## Agregando más datos

Nos dicen que la edad puede ser relevante para conocer mejor a nuestros clientes. La información está en otra tabla la cual se puede leer de la siguiente manera:

clientes\_edad <- read\_csv("https://raw.githubusercontent.com/jbkunst/puc-introduccion-a-R/master/data/clientes-edad.csv")

## Parsed with column specification:  
## cols(  
## rut\_cliente = col\_double(),  
## edad = col\_double()  
## )

clientes\_edad

## # A tibble: 39,514 x 2  
## rut\_cliente edad  
## <dbl> <dbl>  
## 1 18946000 42  
## 2 25824500 25  
## 3 1849500 36  
## 4 3741500 29  
## 5 9663500 37  
## 6 3792000 35  
## 7 14183500 19  
## 8 8288500 45  
## 9 16519500 30  
## 10 6860000 53  
## # ... with 39,504 more rows

* El siguiente código obtiene la distribución de las ventas segun ciertos cortes de edad:

clientes\_edad <- clientes\_edad %>%   
 # CREAR (o modificar) variable  
 mutate(  
 edad\_categorizada = cut(edad, breaks = c(0, 40, 50, 60, 100))  
 )  
  
clientes\_edad

## # A tibble: 39,514 x 3  
## rut\_cliente edad edad\_categorizada  
## <dbl> <dbl> <fct>   
## 1 18946000 42 (40,50]   
## 2 25824500 25 (0,40]   
## 3 1849500 36 (0,40]   
## 4 3741500 29 (0,40]   
## 5 9663500 37 (0,40]   
## 6 3792000 35 (0,40]   
## 7 14183500 19 (0,40]   
## 8 8288500 45 (40,50]   
## 9 16519500 30 (0,40]   
## 10 6860000 53 (50,60]   
## # ... with 39,504 more rows

clientes\_edad %>%   
 count(edad\_categorizada)

## # A tibble: 3 x 2  
## edad\_categorizada n  
## <fct> <int>  
## 1 (0,40] 30152  
## 2 (40,50] 6636  
## 3 (50,60] 2726

¿Qué puede comentar/criticar respecto a los grupos generados?

* Ahora, nos mencionan que se requiere observar las edades usando una categorización de 20 en 20. Use el código anterior para obtener lo requerido.
* ¿Cuantos clientes son/fueron menores de edad (21 años)?

## Juntando los datos

* Consolide la información en una sola tabla.
* Con los grupos de la pregunta del punto anterior calcule el ingreso promedio para cada grupo. Comente si hace sentido o no el resultado obtenido.

## Árbol de decisión

* Con los datos consolidados obtenga un árbol de decisión y dibújelo (plot). Para esto debe primero cargar el paquete parykiy y remitirse al archivo 08-script-arbol.R. Suponiendo que los datos consolidados (con edad) se llama clientes use lo siguiente:

library(partykit)  
  
clientes <- clientes %>%   
 mutate(  
 malo\_en\_los\_12m = factor(malo\_en\_los\_12m)  
 )  
  
arbol <- ctree(malo\_en\_los\_12m ~ ., data = clientes, control = ctree\_control(maxdepth = 3))

* Comente acerca de que variables fueron las más importantes/relevantes para determinar el *riesgo* de una venta.
* Comente que perfiles de clientes son más propensos a ser *malos* en los 12 meses siguientes a la obtención de un crédito de consumo, de la misma forma los clientes menos propensos.
* Recuerde que el árbol se obtuvo (se entrenó) con datos del pasado ahora supongo que en el día de hoy se entregaron 5 créditos de consumo los cuales se pueden obtener con el siguiente código:

clientes\_nuevos <- read\_csv("https://raw.githubusercontent.com/jbkunst/puc-introduccion-a-R/master/data/clientes-nuevos.csv")

clientes\_nuevos

## # A tibble: 5 x 7  
## aux rut\_cliente malo\_en\_los\_12m cantidad\_tarjet~ ingreso edad  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 [19,~ 18984000 1 0 418500 22  
## 2 (22,~ 25836500 0 0 525000 24  
## 3 (28,~ 2542500 0 0 463500 30  
## 4 (34,~ 6973000 0 0 1320000 36  
## 5 (42,~ 8015500 0 0 324000 45  
## # ... with 1 more variable: meses\_en\_su\_actual\_trabajo <dbl>

Ahora usando el siguiente código, que entrega la probabilidad de que sea o no malo en los próximos siguientes 12 meses, indique cual es el cliente más propenso a serlo y por comenté la razón de eso:

predict(arbol, clientes\_nuevos, type = "node")