

# Regresión Logística para Clasificar Spam en E-mails

*Javier Adanaqué*

*24 de Abril, 2017*

## Resumen

Se elabora un modelo de Regresión Logística que nos permita clasificar emails como spam.

## 1. Introducción

El objetivo del presente documento es elaborar un primer modelo de Regresión Logística Simple que nos permita clasificar los emails como spam o no-spam usando lo aprendido en clase (introductoria). Para ello se está usando una base de datos de emails con diferentes características sobre las mismas.

Primero se realizará una exploración rápida de algunas variables y luego se contruirá el modelo, mostrando la mayoría de las veces el código usado para llegar a ellos.

Al ser un modelo de Regresión Logística **Simple** nos concentraremos en conocer si una sola variable nos ayuda a predecir si un mail es spam o no. Para ello se usará, en este primero modelo, nuestra intuición y conocimiento sobre emails para determinar qué variable nos podría predecir mejor si es spam o no.

## 2. Data

Para el análisis se está usando la data `SPAM.txt`, ubicada entre los datasets distribuidos para la clase. Esta es una base de 4601 e-mails (correos electrónicos) y sus características:

```
spam_dat <- read.table("data/SPAM.txt", header = TRUE,
                      stringsAsFactors = FALSE)
spam_dat[1:6, c(1:6, 53:58)]
```

make	address	all	num3d	our	over	charDollar	charHash	capitalAve	capitalLong	capitalW
0.00	0.64	0.64	0	0.32	0.00	0.000	0.000	3.756	61	0.000
0.21	0.28	0.50	0	0.14	0.28	0.180	0.048	5.114	101	0.000
0.06	0.00	0.71	0	1.23	0.19	0.184	0.010	9.821	485	0.000
0.00	0.00	0.00	0	0.63	0.00	0.000	0.000	3.537	40	0.000
0.00	0.00	0.00	0	0.63	0.00	0.000	0.000	3.537	40	0.000
0.00	0.00	0.00	0	1.85	0.00	0.000	0.000	3.000	15	0.000

Como se puede observar en las primeras seis observaciones (y algunas columnas), nos muestra diferentes características sobre los e-mails, incluido el tipo de email: “spam” o “nonspam”.

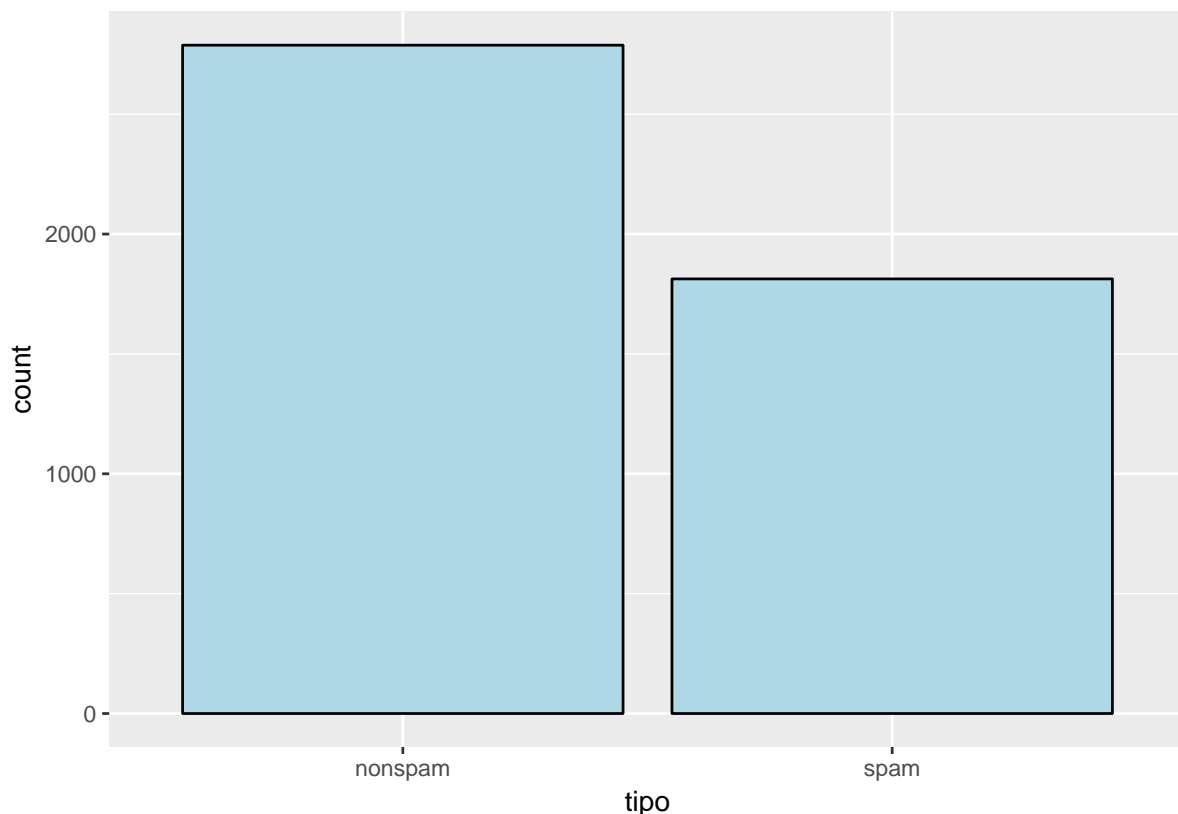
Descripción de variables:

- Primeras 48 variables: Frecuencia (relativa) del nombre de la variable en el correo electrónico. Si el nombre de la variable empieza con 'num (e.g., num857), entonces indica la frecuencia del número correspondiente (e.g., 857).
- Las variables 49-54 indican la frecuencia de los caracteres ';', '(', '[', '!', '\$', and '#'.
- Las variables 55-57 contienen características relacionadas a las letras mayúsculas en el mail.
- tipo: Indica si el e-mail es “spam” o “nonspam”.

### 3. Análisis exploratorio

A continuación exploraremos algunas variables que parecen relevantes para determinar si un e-mail es spam o no. Empezaremos explorando nuestra variable que queremos explicar, tipo:

```
ggplot(spam_dat, aes(tipo)) + geom_bar(colour = "black", fill = "lightblue")
```



Vemos que hay un buen número de e-mails clasificados como “spam” y “nonspam”, lo cual ayuda para la elaboración del modelo (más difícil sería si tuviéramos mucho de uno y muy poco del otro).

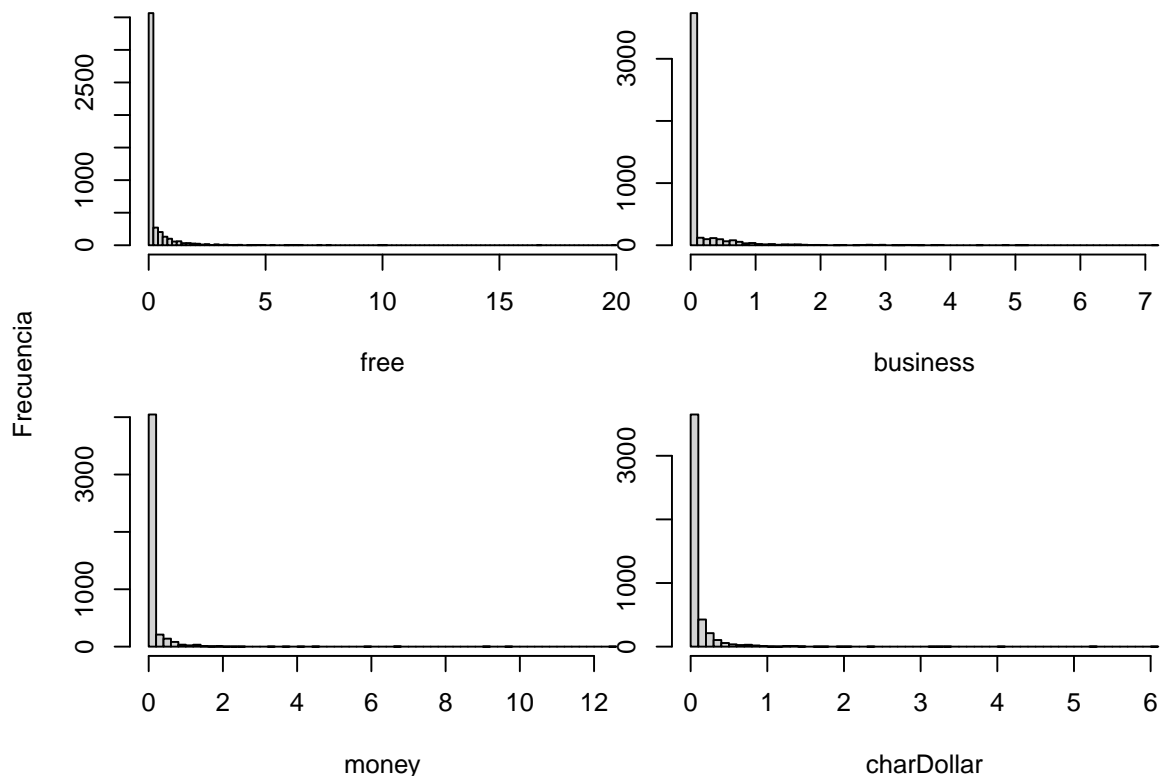
Ahora exploramos las variables que nos pueden ayudar a predecir si un e-mail es spam. Nuestra intuición nos dice que una gran mayoría de spams buscan ofrecerte algo tentador, de manera que lo leas y abras el link que se encuentre adentro (de haber alguno). Entonces, nos concentraremos en alguna variable que nos indique la frecuencia de términos/caracteres como estos: “dinero”, “gratis”, “\$”, “negocio” y similares;

```
names(spam_dat)
```

```
## [1] "make"          "address"       "all"
## [4] "num3d"         "our"           "over"
## [7] "remove"        "internet"      "order"
## [10] "mail"          "receive"       "will"
## [13] "people"        "report"        "addresses"
## [16] "free"          "business"      "email"
## [19] "you"           "credit"        "your"
## [22] "font"          "num000"        "money"
## [25] "hp"            "hpl"           "george"
## [28] "num650"        "lab"           "labs"
## [31] "telnet"        "num857"        "data"
## [34] "num415"        "num85"         "technology"
## [37] "num1999"       "parts"         "pm"
## [40] "direct"        "cs"            "meeting"
## [43] "original"      "project"       "re"
## [46] "edu"           "table"         "conference"
## [49] "charSemicolon" "charRoundbracket" "charSquarebracket"
## [52] "charExclamation" "charDollar"    "charHash"
## [55] "capitalAve"    "capitalLong"   "capitalTotal"
## [58] "tipo"
```

Exploremos las variables `free`, `business`, `money` y `charDollar` (carácter: \$), que parecen relevantes:

```
par(mfrow = c(2, 2),
    mar = c(3.9, 0, 1, 1.2),
    oma = c(1, 3.9, 0, 0))
for (i in c("free", "business", "money", "charDollar")) {
  hist(spam_dat[, i], breaks = 80, col = "lightgray",
       xlab = i, main = NULL)
}
title(ylab = "Frecuencia", outer = TRUE, line = 3)
```



```
par(mfrow = c(1, 1),
    mar = c(5.1, 4.1, 4.1, 2.1),
    oma = rep(0, 4))
```

Observamos que las 4 variables se encuentran altamente sesgadas. Esto es entendible, dado que en un correo no se suele repetir una palabra/caracter demasiadas veces; por eso es que, en las 4 variables seleccionadas, la mayoría de observaciones se encuentran entre 0 y 2 (por ciento), salvo unos cuantos outliers.

```
summary(spam_dat[, c("free", "business", "money", "charDollar")])
```

##	free	business	money	charDollar
## Min.	: 0.0000	Min. :0.0000	Min. : 0.00000	Min. :0.00000
## 1st Qu.:	0.0000	1st Qu.:0.0000	1st Qu.: 0.00000	1st Qu.:0.00000
## Median :	0.0000	Median :0.0000	Median : 0.00000	Median :0.00000
## Mean :	0.2488	Mean :0.1426	Mean : 0.09427	Mean :0.07581
## 3rd Qu.:	0.1000	3rd Qu.:0.0000	3rd Qu.: 0.00000	3rd Qu.:0.05200
## Max. :	20.0000	Max. :7.1400	Max. :12.50000	Max. :6.00300

De ahora en adelante, continuaremos el análisis usando `charDollar` como variable explicativa. Esto, porque es una variable que definitivamente se repite bastante en los correos spam (ver siguiente gráfico) y porque es una variable cuyo rango no muy extenso (entre 0 y 6, como se puede observar en la tabla anterior).

```
spam_dat %>%
  ggplot(aes(tipo, charDollar+0.01)) + geom_boxplot() +
  scale_y_continuous(trans = "log") + geom_hline(yintercept = 6, colour = "red")
```

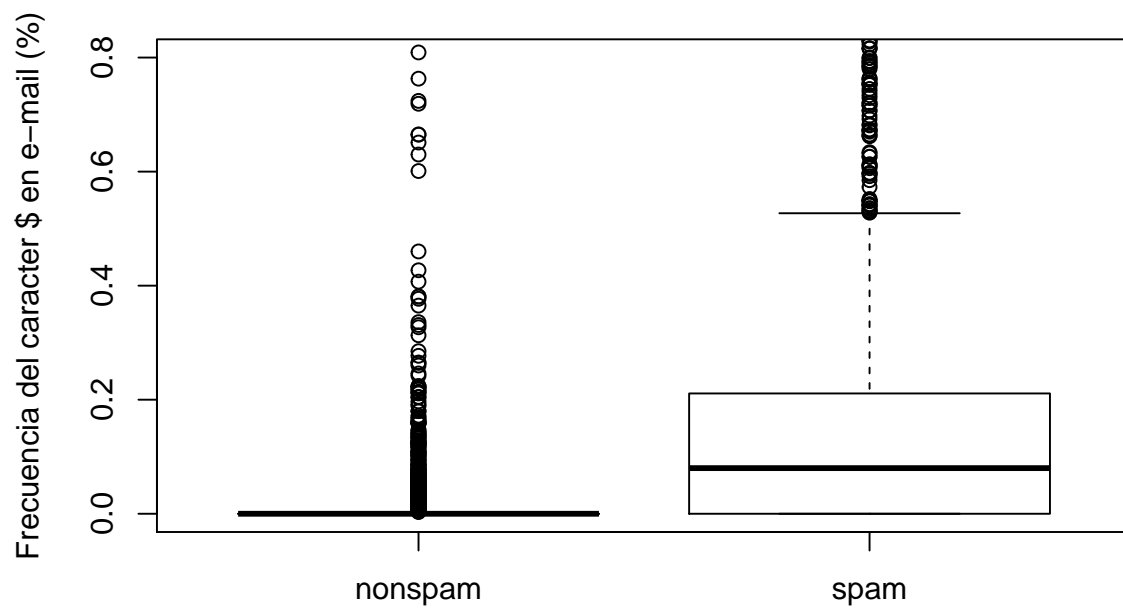
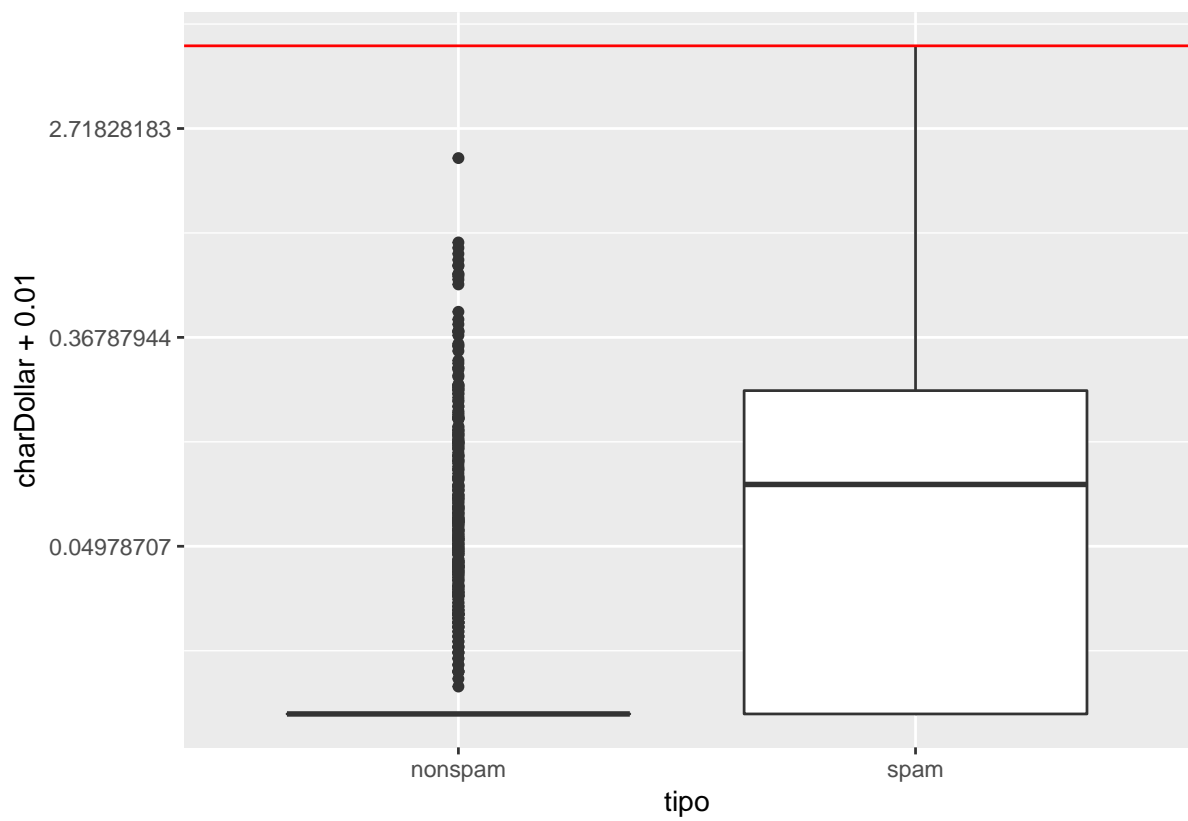


Figura 1: Diagrama de Cajas de 'charDollar' según 'tipo' de e-mail. Se limitó el eje vertical (zoom) para mayor claridad; de lo contrario, los outliers llegarían hasta 6.003.



```
spam_dat %>%  
  ggplot(aes(tipo, charDollar+0.01)) + geom_violin(trim = FALSE) +  
  scale_y_continuous(trans = "log") +  
  geom_hline(yintercept = 6, colour = "red")
```

