Autor: FRANCISCO JAVIER POLO MARTINEZ
Actualizado 02 Marzo 2020
Arquitecto Técnico
Master Big Data

CREADO EL 14/02/2020

Vamos a proceder a explicar el código en r con la inserción de gráficos de relevancia para la tarea sobre la base de datos eleccion.xlsx proporcionada.
Primero instalamos los paquetes a utilizar así como sus librerias
Cargamos la base de datos en una variable que inicialmente llamamos elección y lo convertimos en un data frame

eleccion <- read_excel("C:/Users/polo/OneDrive/MASTER/00 COMPLUTENSE/ APUNTES MASTER/MODULO 8 MINERIA DE DATOS/TAREA 05-03-2020/eleccion.xlsx")

eleccion<-data.frame(eleccion) y lo visualizamos previamente convirtiendo las variables CCAA y ActividadPpal como factor y también eliminamos la variable Name que realmente se comporta como un ID

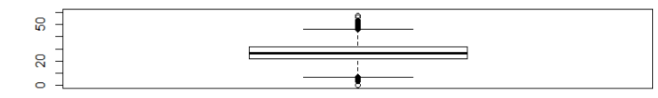
```
> str(eleccion)
'data.frame':   7716 obs. of  35 variables:
 $ CodigoProvincia            : num  10 10 10 10 10 10 10 10 10 10 ...
 $ CCAA                       : Factor w/ 19 levels "Andalucía","Aragón",..: 12 12 12 12 12 12 12 12 12 12 ...
 $ Population                 : num  336 429 569 822 787 471 774 434 650 80 ...
 $ TotalCensus                : num  282 364 569 704 835 427 695 400 648 77 ...
 $ AbstentionPtge             : num  20.2 25.3 27.2 30.1 26.2 ...
 $ Derecha                    : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 2 1 1 1 ...
 $ Age_0.4_Ptge               : num  3.869 1.632 1.23 4.258 0.508 ...
 $ Age_under19_Ptge           : num  18.16 13.05 9.14 14.96 10.29 ...
 $ Age_19_65_pct              : num  55.1 56.6 54.8 60.1 54 ...
 $ Age_over65_pct             : num  26.8 30.3 36 24.9 35.7 ...
 $ WomanPopulationPtge        : num  44 50.1 49 51.1 46.6 ...
 $ ForeignersPtge             : num  0.89 1.63 0.7 0.12 0.64 0.21 1.29 0.23 0.62 3.75 ...
 $ SameComAutonPtge           : num  79.8 90.9 78.9 93.9 92 ...
 $ SameComAutonDiffProvPtge   : num  0.298 2.797 0.703 0.487 2.922 ...
 $ DifComAutonPtge            : num  19.34 7.23 18.1 5.11 5.97 ...
 $ UnemployLess25_Ptge        : num  2.38 16.22 8.2 7.41 13.43 ...
 $ Unemploy25_40_Ptge         : num  54.8 32.4 36.1 61.1 50.7 ...
 $ UnemployMore40_Ptge        : num  42.9 51.4 55.7 31.5 35.8 ...
 $ AgricultureUnemploymentPtge : num  4.76 8.11 22.95 16.67 19.4 ...
 $ IndustryUnemploymentPtge   : num  9.52 8.11 9.84 5.56 2.98 ...
 $ ConstructionUnemploymentPtge: num  11.9 10.8 13.1 16.7 19.4 ...
 $ ServicesUnemploymentPtge   : num  73.8 67.6 49.2 59.3 53.7 ...
 $ totalEmpresas              : num  15 11 49 50 37 17 56 18 31 0 ...
 $ Industria                  : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Construccion               : num  0 0 0 0 0 0 0 0 0 0 ...
 $ ComercTTEHosteleria        : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Servicios                  : num  0 0 0 0 0 0 0 0 0 0 ...
 $ ActividadPpal              : Factor w/ 5 levels "ComercTTEHosteleria",..: 4 4 4 4 4 4 4 4 4 4 ...
 $ inmuebles                  : num  216 382 918 599 983 451 742 415 664 117 ...
 $ Pob2010                    : num  326 459 674 842 978 493 751 473 665 92 ...
 $ SUPERFICIE                 : num  4508 6271 5702 9106 11551 ...
 $ Densidad                   : chr  "MuyBaja" "MuyBaja" "MuyBaja" "MuyBaja" ...
 $ PobChange_pct              : num  3.07 -6.54 -15.58 -2.38 -19.53 ...
 $ PersonasInmueble           : num  1.56 1.12 0.62 1.37 0.8 1.04 1.04 1.05 0.98 0.68 ...
 $ Explotaciones              : num  28 67 74 66 232 118 30 128 58 28 ...
```

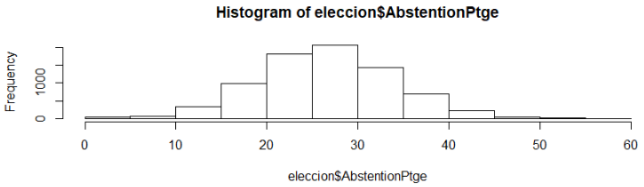
```
# entre otras cosas podemos visualizar los MISSING:
## 22 totalEmpresas                4
## 23 Industria                  176
## 24 Construccion               131
## 25 ComercTTEHosteleria          8
## 26 Servicios                   56
## 27 inmuebles                  129
## 28 Pob2010                      6
## 29 SUPERFICIE                   7
## 30 PobChange_pct                6
## 31 PersonasInmueble           129
# la variable Age_0-4-Ptg >- tiene sentido, o se considera incluida en la siguiente variable
# la variable Age_under19_Ptge >- DEBERIAN SUMAR 100%
# la variable ForeignerPtge -> TIENE VALORES NEGATIVOS
```

Contamos el número de valores diferentes para las numéricas (línea de codigo 109)

```
> sapply(Filter(is.numeric, eleccion),function(x) length(unique(x)))
          CodigoProvincia                Population               TotalCensus              AbstentionPtge              Age_0.4_Ptge
                       52                      3495                      3224                      5440                      3663
            Age_under19_Ptge             Age_19_65_pct            Age_over65_pct        WomanPopulationPtge             ForeignersPtge
                     5664                      5965                      6482                      4385                      2278
          SameComAutonPtge   SameComAutonDiffProvPtge            DifComAutonPtge        UnemployLess25_Ptge         Unemploy25_40_Ptge
                     5911                      4039                      5368                      2264                      2591
       UnemployMore40_Ptge AgricultureUnemploymentPtge IndustryUnemploymentPtge ConstructionUnemploymentPtge   ServicesUnemploymentPtge
                     2651                      2437                      2461                      2419                      2804
             totalEmpresas                 Industria                Construccion        ComercTTEHosteleria                  Servicios
                     1193                       304                       445                       787                       741
                 inmuebles                  Pob2010                SUPERFICIE                PobChange_pct           PersonasInmueble
                     3012                      3522                      7709                      2983                       282
              Explotaciones
                      741
```

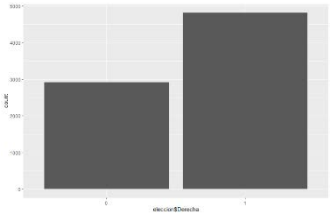
Realizamos un bloxplot e histograma de la **variable continua** y vemos su distribución





Contamos la **variable discreta** y las representamos

```
> #CONTAR DISCRETAS
> eleccion %>%
+   count(Derecha)
# A tibble: 2 x 2
  Derecha     n
  <fct>   <int>
1 0        2912
2 1        4804
```



Autor: FRANCISCO JAVIER POLO MARTINEZ
Actualizado 02 Marzo 2020
Arquitecto Técnico
Master Big Data

Creamos una nueva variable calculada sobre la variable Pob2010 denominada densidadCalculada y eliminamos la antigua:

```
eleccion <- mutate (eleccion, densidadCalculado=(Pob2010/SUPERFICIE)*1000)
eleccion<-select(eleccion, -Densidad)
```

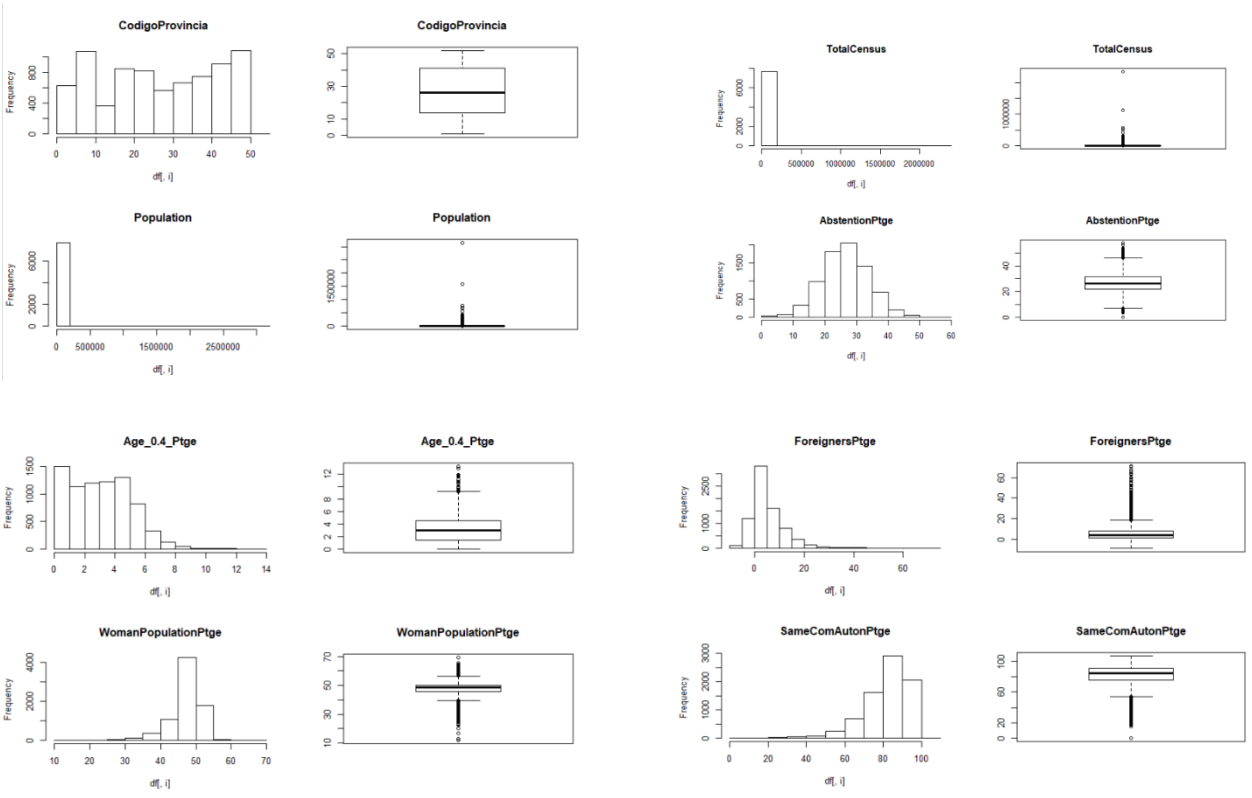
Ahora vamos con las variables de edades: Age_under19_Ptge, Age_19_65_pct, Age_over65_pct, las hacemos proporcional según sus pesos para que la suma sea el 100%

```
eleccion <- mutate (eleccion, sumaEdades = Age_under19_Ptge+Age_19_65_pct+Age_over65_pct)
view(eleccion
eleccion<-mutate(eleccion, edadmenor19 = Age_under19_Ptge+(repartir)*100)
eleccion<-mutate(eleccion, edadentre1965 = Age_19_65_pct+repartir)
eleccion<-mutate(eleccion, edadmayor65 = Age_over65_pct+repartir)
eleccion<-select(eleccion, -Age_under19_Ptge, -Age_19_65_pct, -Age_over65_pct, -repartir, -sumaEdades)
```

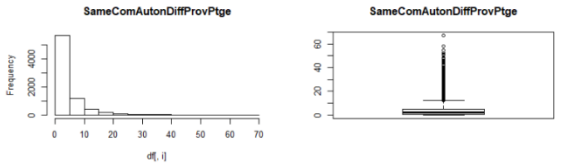
Recategorizar por cuantiles. La tramificacion por árbol está indicada como comentario en el archivo de código, aunque al final no se realizo

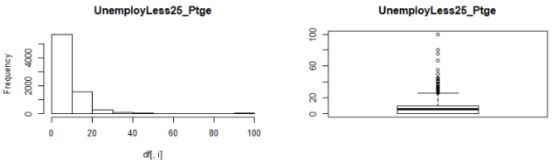
```
Industria1<- questionr::quant.cut(eleccion$Industria, 4)
Construccion1<- questionr::quant.cut(eleccion$Construccion, 4)
Servicios1<- questionr::quant.cut(eleccion$Servicios, 4)
Pob20101<- questionr::quant.cut(eleccion$Pob2010, 4)
SUPERFICIE1<-questionr::quant.cut(eleccion$SUPERFICIE, 4)

eleccion<-mutate(eleccion, Industria1 = Industria)
eleccion<-mutate(eleccion, Construccion1 = Construccion)
eleccion<-mutate(eleccion, Servicios1 = Servicios)
eleccion<-mutate(eleccion, Pob20101 = Pob2010)
eleccion<-mutate(eleccion, SUPERFICIE1 = SUPERFICIE)

eleccion<-select(eleccion, -Industria, -Construccion, -Servicios, -Pob2010, -SUPERFICIE)
```

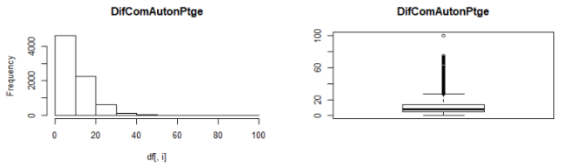
Una ver realizados estos cambios en data set hacemos una inspección grafica inicial de las variables (líneas 180-182)

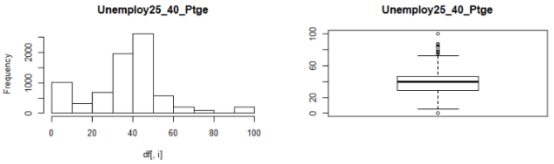


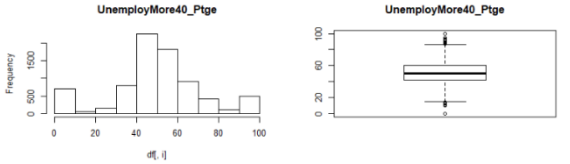
Autor: FRANCISCO JAVIER POLO MARTINEZ
Actualizado 02 Marzo 2020
Arquitecto Técnico
Master Big Data

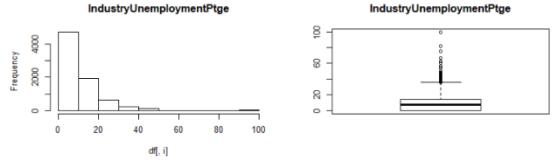


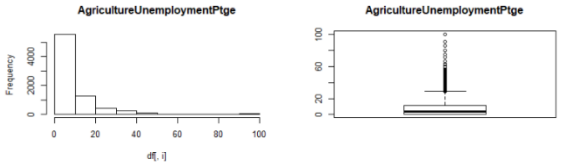


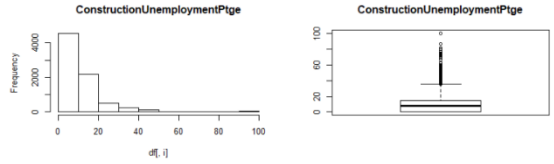


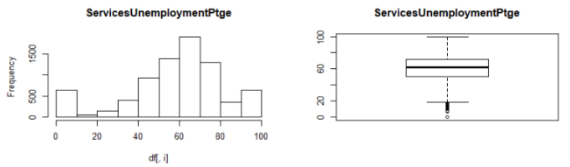


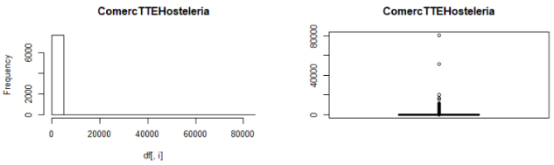


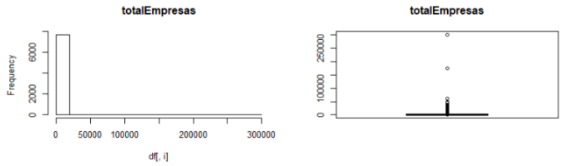


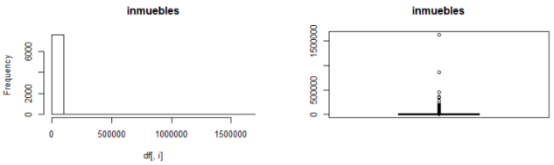


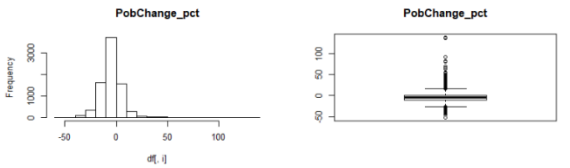


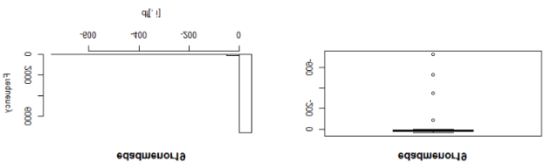


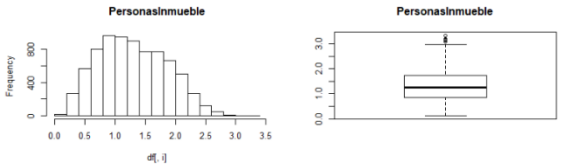


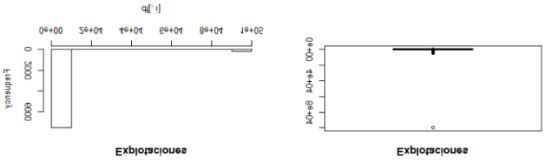




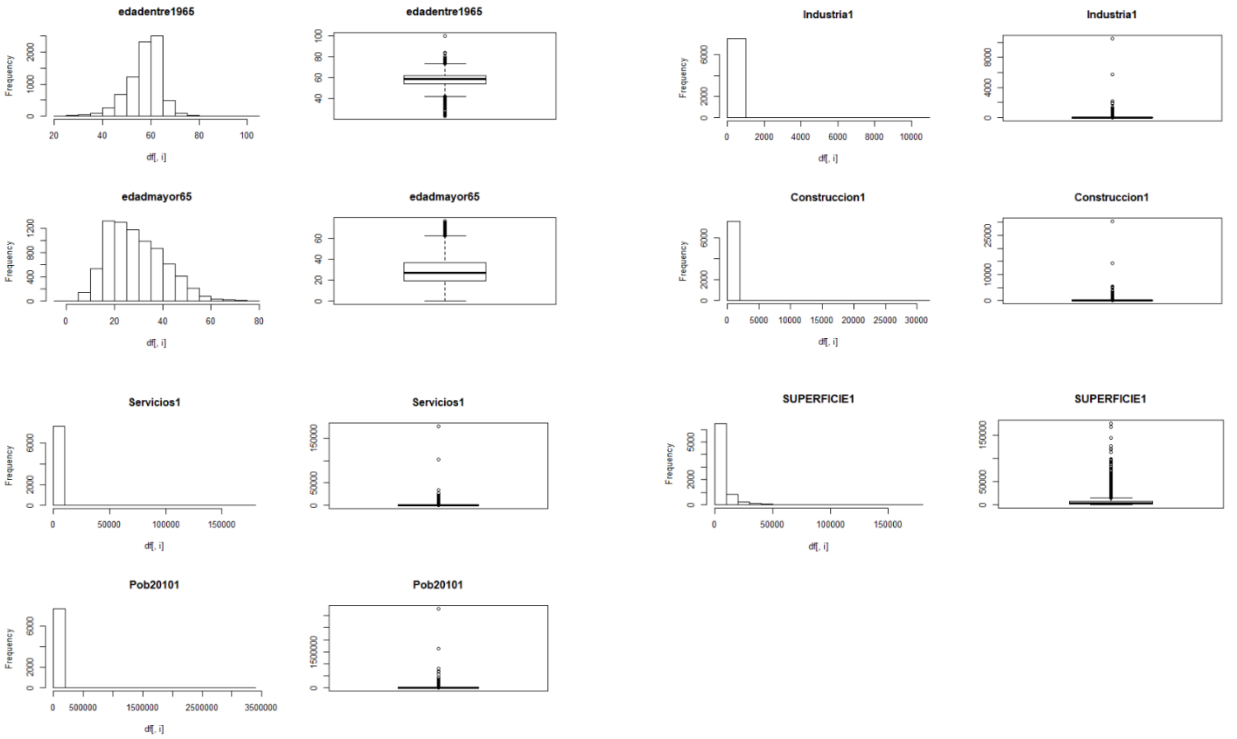








Autor: FRANCISCO JAVIER POLO MARTINEZ
Actualizado 02 Marzo 2020
Arquitecto Técnico
Master Big Data



Ahora procedemos al tratamiento de los missing no declarados y al de los valores fuera de rango:

Missings no declarados variables cualitativas (NSNC, ?)

```
eleccion$densidadCalculado<-recode.na(eleccion$densidadCalculado,"?")
```

Missings no declarados variables cuantitativas (-1, 99999)

```
eleccion$Explotaciones<-
replace(eleccion$Explotaciones,which(eleccion$Explotaciones==99999),NA)
```

Valores fuera de rango. el primero con valores mayores a 100 y el segundo con negativos

```
eleccion$SameComAutonPtge<-replace(eleccion$SameComAutonDiffProvPtge,
which((eleccion$SameComAutonDiffProvPtge < 0)|(eleccion$SameComAutonDiffProvPtge>100)),
NA)
eleccion$ForeignersPtge<-replace(eleccion$ForeignersPtge, which((eleccion$ForeignersPtge
< 0)|(eleccion$ForeignersPtge>100)), NA)
```

Ahora indicamos la variable objeto, el id y las input

```
varObjCont<-eleccion$AbstentionPtge
varObjBin<-eleccion$Derecha
input<-as.data.frame(eleccion[,-(5:6)])
row.names(input)<-eleccion$ID
```

Cuento el porcentaje de atipicos de cada variable. Si son muchos, elimino esas variables en la siguiente linea de codigo

```
sapply(Filter(is.numeric, input),function(x) atipicosAmissing(x)[[2]])/nrow(input)
```

Modifico los atapicos como missings

```
input[,as.vector(which(sapply(input, class)=="numeric"))]<-sapply(
  Filter(is.numeric, input),function(x) atipicosAmissing(x)[[1]])
sum(is.na(input))
```

Busco si existe algun patron en los missings, que me pueda ayudar a entenderlos

```
corrplot(cor(is.na(input[colnames(input)[colSums(is.na(input))>0]])),method =
"ellipse",type = "upper") # variables de servicios
```

```
#Proporcion de missings por variable y observacion
input$prop_missings<-apply(is.na(input),1,mean) # Por observacion
summary(input$prop_missings)
(prop_missingsVars<-apply(is.na(input),2,mean)) # Por variable
```

Autor: FRANCISCO JAVIER POLO MARTINEZ
Actualizado 02 Marzo 2020
Arquitecto Técnico
Master Big Data

```
> (prop_missingsVars<-apply(is.na(input),2,mean)) # Por variable
          CodigoProvincia                    CCAA              Population             TotalCensus            Age_0.4_Ptge 
                0.0000000000          0.0000000000          0.0992742354          0.0961638154          0.0000000000 
       WomanPopulationPtge          ForeignersPtge          SameComAutonPtge    SameComAutonDiffProvPtge          DifComAutonPtge 
                0.0023328149          0.0857957491          0.0187921203          0.0187921203          0.0045360290 
       UnemployLess25_Ptge          Unemploy25_40_Ptge      UnemployMore40_Ptge AgricultureUnemploymentPtge IndustryUnemploymentPtge 
                0.0032400207          0.0000000000          0.0000000000          0.0202177294          0.0058320373 
ConstructionUnemploymentPtge ServicesUnemploymentPtge             totalEmpresas       ComercTTEHosteleria          ActividadPpal 
                0.0062208398          0.0000000000          0.1049766719          0.0982374287          0.0000000000 
                   inmuebles             PobChange_pct       PersonasInmueble          Explotaciones      densidadCalculado 
                0.1047174702          0.0020736133          0.0167185070          0.0751684811          0.0012960083 
                 edadmenor19          edadentre1965             edadmayor65             Industria1          Construccion1 
                0.0005184033          0.0025920166          0.0000000000          0.1108087092          0.1065318818 
                  Servicios1                Pob20101            SUPERFICIE1          prop_missings 
                0.1246759979          0.0982374287          0.0281233800          0.0000000000 
```

Imputo todas las cuantitativas, seleccionar el tipo de imputacion: media, mediana o aleatorio

```
input[,as.vector(which(sapply(input, class)=="numeric"))]<-sapply(
  Filter(is.numeric, input),function(x) ImputacionCuant(x,"aleatorio"))
```

Si solo se quiere imputar una, variable<-ImputacionCuali(variable,"moda")

```
input[,as.vector(which(sapply(input, class)=="factor"))]<-sapply(
  Filter(is.factor, input),function(x) ImputacionCuali(x,"aleatorio"))
```

A veces se cambia el tipo de factor a character al imputar, asi que hay que indicarle que es factor

```
input[,as.vector(which(sapply(input, class)=="character"))] <- lapply(
  input[,as.vector(which(sapply(input, class)=="character"))] , factor)
```

```
>   summary(input)
 CodigoProvincia            CCAA       Population      TotalCensus      Age_0.4_Ptge    WomanPopulationPtge ForeignersPtge  SameComAutonPtge
 Min.   : 1.0    CastillaLeón  :2167   Min.   :   5.0   Min.   :   5.0   Min.   : 0.000   Min.   :26.47       Min.   : 0.00   Min.   : 0.0000  
 1st Qu.:14.0    Cataluña      : 909   1st Qu.: 153.0   1st Qu.: 132.0   1st Qu.: 1.408   1st Qu.:45.75       1st Qu.: 1.64   1st Qu.: 0.6325  
 Median :26.0    CastillaMancha: 829   Median : 452.4   Median : 370.5   Median : 3.008   Median :48.51       Median : 4.15   Median : 2.0925  
 Mean   :26.9    Andalucía     : 725   Mean   :1293.3   Mean   :1012.5   Mean   : 3.032   Mean   :47.37       Mean   : 6.09   Mean   : 3.7031  
 3rd Qu.:41.0    Aragón        : 703   3rd Qu.:1553.7   3rd Qu.:1226.2   3rd Qu.: 4.545   3rd Qu.:50.00       3rd Qu.: 8.71   3rd Qu.: 5.0000  
 Max.   :52.0    ComValenciana : 506   Max.   :9372.0   Max.   :7061.0   Max.   :13.245   Max.   :69.23       Max.   :39.86   Max.   :25.9770  
                 (Other)       :1877                                                                                                           
 SameComAutonDiffProvPtge DifComAutonPtge  UnemployLess25_Ptge Unemploy25_40_Ptge UnemployMore40_Ptge AgricultureUnemploymentPtge IndustryUnemploymentPtge
 Min.   : 0.000           Min.   : 0.000   Min.   : 0.000      Min.   :  0.00     Min.   :  0.00      Min.   : 0.000               Min.   : 0.000          
 1st Qu.: 0.629           1st Qu.: 4.922   1st Qu.: 0.000      1st Qu.: 28.83     1st Qu.: 41.76      1st Qu.: 0.000               1st Qu.: 0.000          
 Median : 2.099           Median : 8.273   Median : 5.875      Median : 40.00     Median : 50.00      Median : 3.226               Median : 7.143          
 Mean   : 3.701           Mean   :10.474   Mean   : 7.041      Mean   : 37.06     Mean   : 50.25      Mean   : 7.149               Mean   : 9.559          
 3rd Qu.: 4.982           3rd Qu.:13.843   3rd Qu.:10.345      3rd Qu.: 46.67     3rd Qu.: 60.00      3rd Qu.:10.959               3rd Qu.:14.286          
 Max.   :25.977           Max.   :56.091   Max.   :75.000      Max.   :100.00     Max.   :100.00      Max.   :45.833               Max.   :82.000          

 ConstructionUnemploymentPtge ServicesUnemploymentPtge totalEmpresas    ComercTTEHosteleria            ActividadPpal    inmuebles       PobChange_pct      
 Min.   : 0.000               Min.   :  0.00           Min.   :  0.00   Min.   :  0.00      ComercTTEHosteleria:2428   Min.   :   6.0   Min.   :-52.2700  
 1st Qu.: 0.000               1st Qu.: 50.00           1st Qu.:  7.00   1st Qu.:  0.00      Construccion       :  14   1st Qu.: 169.0   1st Qu.:-10.3700  
 Median : 8.333               Median : 62.20           Median : 24.00   Median :  0.00      Industria          :  13   Median : 414.1   Median : -4.9200  
 Mean   :10.214               Mean   : 58.83           Mean   : 76.27   Mean   : 31.15      Otro               :4666   Mean   : 904.4   Mean   : -4.9874  
 3rd Qu.:14.114               3rd Qu.: 72.22           3rd Qu.: 88.00   3rd Qu.: 42.00      Servicios          : 595   3rd Qu.:1127.0   3rd Qu.:  0.1125  
 Max.   :86.486               Max.   :100.00           Max.   :575.00   Max.   :268.00                                 Max.   :5883.0   Max.   : 54.0500  

 PersonasInmueble Explotaciones          densidadCalculado  edadmenor19      edadentre1965    edadmayor65         Industria1     Construccion1    
 Min.   :0.110    Min.   :  1.00   1173.39578944655:   2   Min.   :-0.1667   Min.   :30.36   Min.   :-0.00067   Min.   : 0.00   Min.   :  0.00  
 1st Qu.:0.860    1st Qu.: 21.00   1179.96054748022:   2   1st Qu.: 8.3661   1st Qu.:53.97   1st Qu.:19.76800   1st Qu.: 0.00   1st Qu.:  0.00  
 Median :1.260    Median : 46.00   15.644830574909 :   2   Median :13.9650   Median :58.72   Median :27.46000   Median : 0.00   Median :  0.00  
 Mean   :1.303    Mean   : 81.61   160.107063060423:   2   Mean   :13.6018   Mean   :57.48   Mean   :28.98015   Mean   : 6.69   Mean   : 12.04  
 3rd Qu.:1.730    3rd Qu.:107.00   27.4547829020192:   2   3rd Qu.:19.0851   3rd Qu.:61.84   3rd Qu.:36.78150   3rd Qu.: 9.00   3rd Qu.: 17.00  
 Max.   :3.330    Max.   :451.00   411.413986344302:   2   Max.   :33.7627   Max.   :83.78   Max.   :76.47133   Max.   :56.00   Max.   :100.00  
                                   (Other)         :7704                                                                                        
   Servicios1       Pob20101       SUPERFICIE1        prop_missings      
 Min.   :  0.0   Min.   :   5.0   Min.   :    2.578   Min.   :0.00000  
 1st Qu.:  0.0   1st Qu.: 165.0   1st Qu.: 1796.388   1st Qu.:0.00000  
 Median :  0.0   Median : 474.5   Median : 3365.050   Median :0.00000  
 Mean   : 17.9   Mean   :1330.7   Mean   : 5070.950   Mean   :0.03745  
 3rd Qu.: 21.0   3rd Qu.:1595.7   3rd Qu.: 6433.190   3rd Qu.:0.03030  
 Max.   :164.0   Max.   :9468.0   Max.   :27681.201   Max.   :0.36364  
```

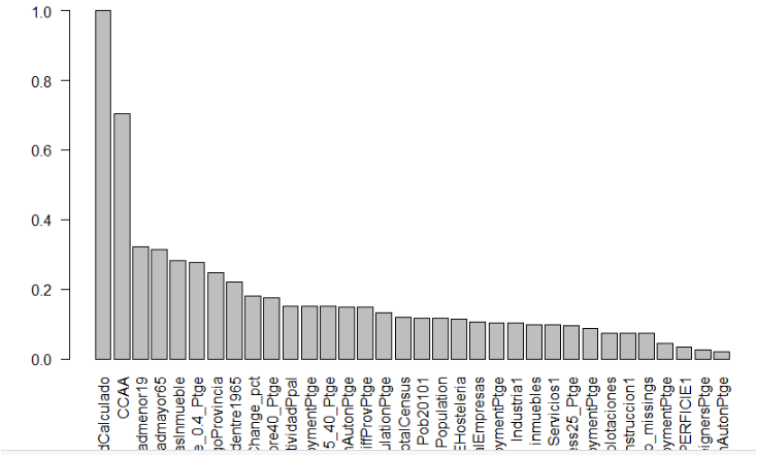
Vemos que ya no hay missing ni valores atípicos por lo que lo salvamos en un fichero de nombre datosLimpios

```
saveRDS(cbind(varObjBin,varObjCont,input),"datosLimpios")
```

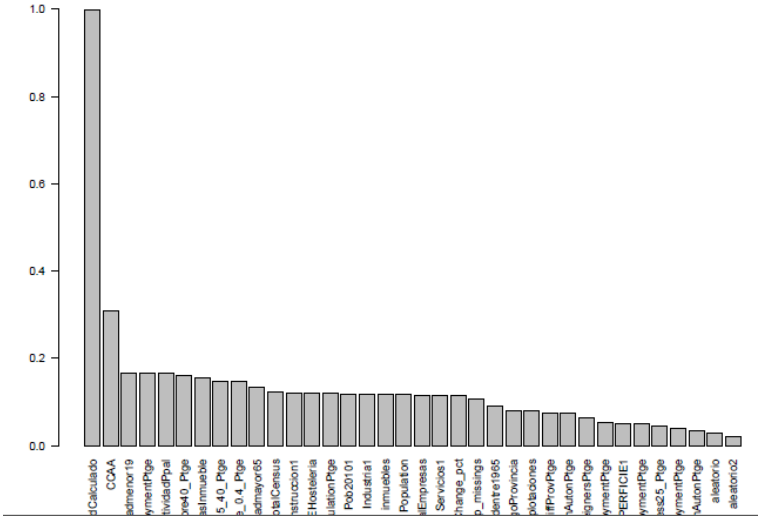
Ya podemos empezar con la regresión lineal. Para ello abrimos el archivo anteriormente creado, declaramos las variables objetos, creamos un data frame llamado dlimpios y creamos dos variables aleatorias las cuales no van a influir en los modelos que hagamos.

Autor: FRANCISCO JAVIER POLO MARTINEZ
Actualizado 02 Marzo 2020
Arquitecto Técnico
Master Big Data

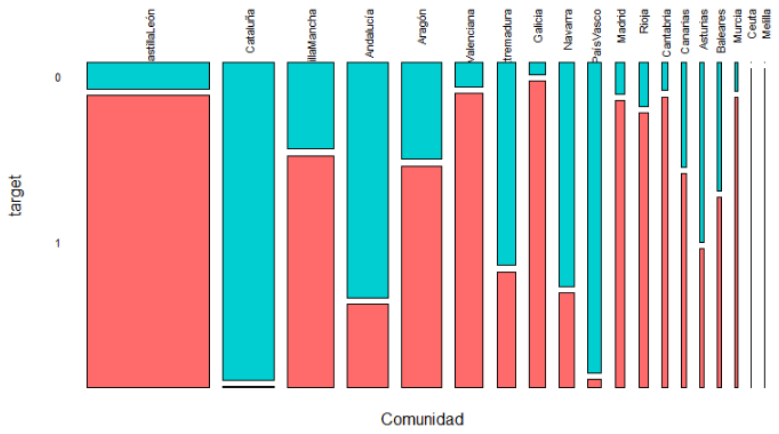
```
#Obtengo la importancia de las variables. Falla si hay alguna variable cuantitativa con menos de 6 valores diferentes
graficoVcramer(input,varObjBin) #densidadCalculado, CCAA, edadmenor19, edadmayor65
```




```
graficoVcramer(input,varObjCont) #densidadCalculado, CCAA, edadmenor19, ymenPtge, actividadPpal
```

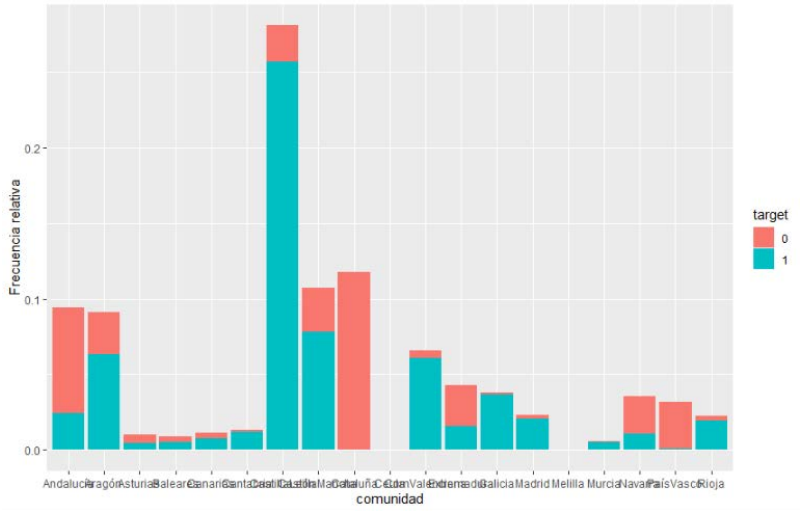



```
mosaico_targetbinaria(input$CCAA,varObjBin,"Comunidad") #esta influye
```



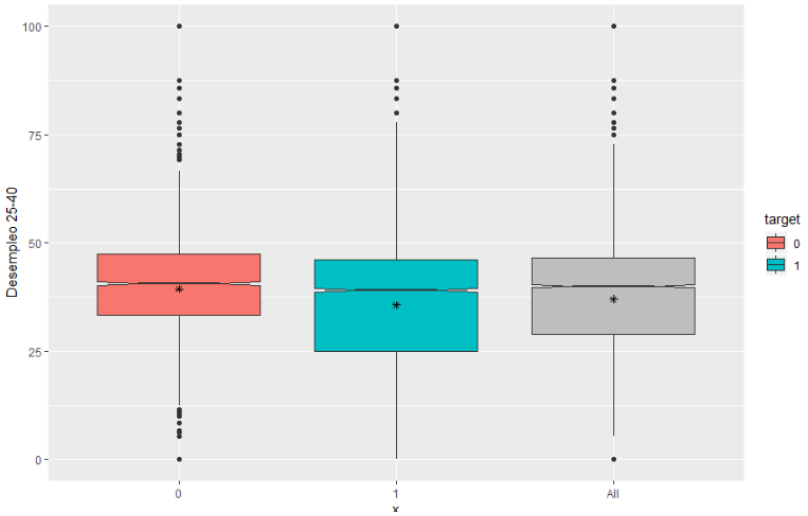
Autor: FRANCISCO JAVIER POLO MARTINEZ
Actualizado 02 Marzo 2020
Arquitecto Técnico
Master Big Data

```
barras_targetbinaria(input$CCAA,varObjBin,"comunidad")
```

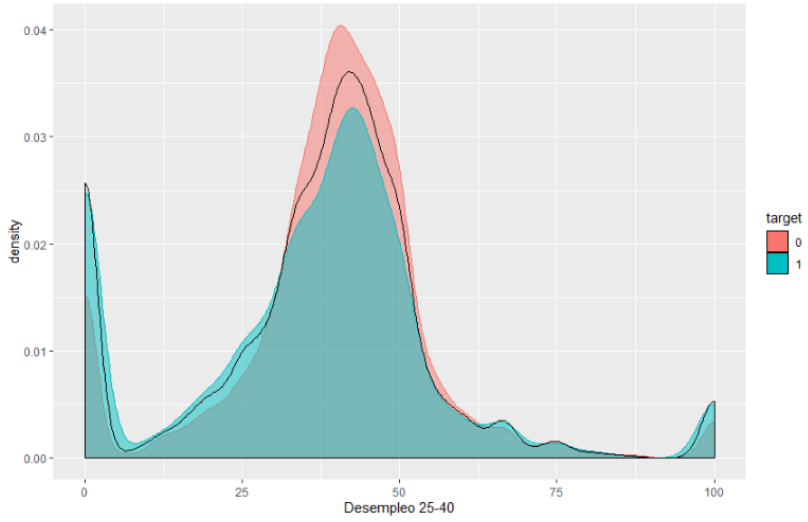


#Veo graficamente el efecto de dos variables cuantitativas sobre la binaria

```
boxplot_targetbinaria(input$Unemploy25_40_Ptge,varObjBin,"Desempleo 25-40")
```




```
hist_targetbinaria(input$Unemploy25_40_Ptge,varObjBin,"Densidad")
```



En principio no parece que influya mucho ya que el 0, 1 y la total son muy parecidas en su rango

Autor: FRANCISCO JAVIER POLO MARTINEZ
Actualizado 02 Marzo 2020
Arquitecto Técnico
Master Big Data

#Busco las mejores transformaciones para las variables numéricas con respecto a los dos tipos de variables y creamos dos ficheros: todo_bin y todo_cont

```
input_cont<-cbind(input,Transf_Auto(Filter(is.numeric, input),varObjCont))
input_bin<-cbind(input,Transf_Auto(Filter(is.numeric, input),varObjBin))

saveRDS(data.frame(input_bin,varObjBin),"todo_bin")
saveRDS(data.frame(input_cont,varObjCont),"todo_cont")
```

A partir de ahora empezamos realmente con la regresión lineal, para ello el esquema básico es el de train-test reservando el 80% para el entrenamiento del modelo y el 20% para ver si el modelo puede hacer predicciones. Gracias a la librería createDataPartition

Primero creamos un modelo 1, con todo, que tenga todas las variables, con la función lm:

```
> modelo1<-lm(varObjCont~.,data=data_train)
> summary(modelo1)

Call:
lm(formula = varObjCont ~ ., data = data_train)

Residuals:
    Min      1Q  Median      3Q     Max
-25.792  -4.483  -0.214   4.223  33.117

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.604e+01  5.549e+00   2.891 0.003848 **
CodigoProvincia           -1.087e-02  5.859e-03  -1.856 0.063498 .
Population                 5.298e-06  1.145e-04   0.046 0.963086
TotalCensus                3.696e-04  1.354e-04   2.731 0.006338 **
Age_0.4_Ptge               3.193e-01  8.665e-02   3.685 0.000230 ***
WomanPopulationPtge       -6.783e-02  2.486e-02  -2.728 0.006387 **
ForeignersPtge             2.046e-02  1.500e-02   1.364 0.172585
SameComAutonPtge           1.756e-01  1.042e-01   1.686 0.091849 .
SameComAutonDiffProvPtge   7.770e-03  1.044e-01   0.074 0.940658
DifComAutonPtge           -1.504e-02  1.206e-02  -1.247 0.212575
UnemployLess25_Ptge       -1.432e-02  1.276e-02  -1.122 0.261889
Unemploy25_40_Ptge        -2.096e-03  7.174e-03  -0.292 0.770144
UnemployMore40_Ptge        4.382e-03  6.905e-03   0.635 0.525667
AgricultureUnemploymentPtge -1.341e-02  1.138e-02  -1.179 0.238625
IndustryUnemploymentPtge   5.713e-02  1.010e-02   5.658 1.60e-08 ***
ConstructionUnemploymentPtge 7.379e-02  9.654e-03   7.644 2.43e-14 ***
ServicesUnemploymentPtge   2.770e-02  6.086e-03   4.551 5.44e-06 ***
totalEmpresas              3.677e-03  1.400e-03   2.627 0.008647 **
ComercTTEHosteleria       -3.845e-03  2.905e-03  -1.324 0.185643
inmuebles                 -8.165e-05  1.268e-04  -0.644 0.519516
PobChange_pct              2.190e-03  1.079e-02   0.203 0.839183
PersonasInmueble           1.870e+00  2.496e-01   7.491 7.82e-14 ***
Explotaciones              2.420e-03  1.127e-03   2.148 0.031754 *
edadmenor19                1.406e-01  6.215e-02   2.262 0.023754 *
edadentre1965              1.587e-02  5.531e-02   0.287 0.774119
edadmayor65                9.393e-02  5.385e-02   1.744 0.081148 .
Industria1                 2.011e-02  1.050e-02   1.915 0.055577 .
Construccion1             -7.298e-03  6.629e-03  -1.101 0.270963
Servicios1                 4.088e-03  3.992e-03   1.024 0.305916
Pob20101                   1.135e-04  1.072e-04   1.059 0.289681
SUPERFICIE1                6.202e-05  1.823e-05   3.401 0.000675 ***
prop_missings              1.231e+01  1.281e+00   9.613  < 2e-16 ***
aleatorio                  1.902e-01  2.977e-01   0.639 0.522950
aleatorio2                -5.651e-01  2.983e-01  -1.895 0.058173 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.775 on 6141 degrees of freedom
Multiple R-squared:  0.1919,    Adjusted R-squared:  0.1876
F-statistic: 44.19 on 33 and 6141 DF,  p-value: < 2.2e-16
```

Se puede ver que muchas de ellas no son significativas

Mientras que estas si lo son

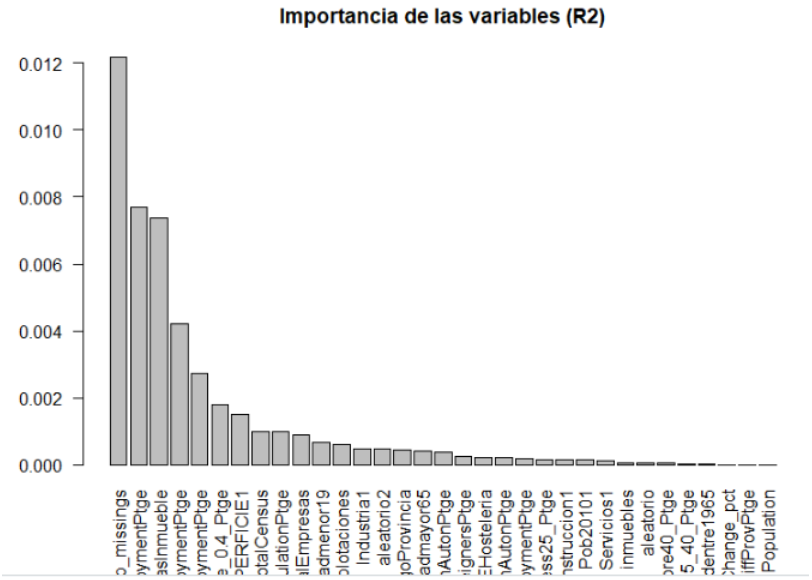
```
> Rsq(modelo1,"varObjCont",data_test) #En test hay bastante diferencia, seguramente sobren variables
$r2
[1] 0.1526128

$r2_adj
[1] 0.1334819
```

MODELO 1

Autor: FRANCISCO JAVIER POLO MARTINEZ
Actualizado 02 Marzo 2020
Arquitecto Técnico
Master Big Data

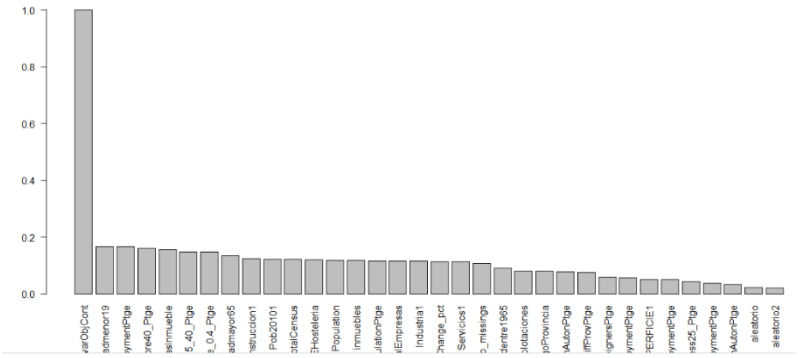
Nos fijamos en la importancia de las variables. Podemos sacar un grafico que muestra lo que se pierde en R2 en train al quitarlas del modelo

```
modelEffectSizes(modelo1)
barplot(sort(modelEffectSizes(modelo1)$Effects[-1,4],decreasing
=T),las=2,main="Importancia de las variables (R2)")
```



Las que mas importan son missing,
ConstructionUnemploymentPtge
PersonasInmueble
IndustryUnemploymentPtge

Empezamos el modelo 2, probamos el todo con la variable continua con un gráfico V de Cramer, y cogemos las variables mas significativas



```
> modelo2<-lm(varObjCont~edadmenor19+AgricultureUnemploymentPtge+UnemployMore40_Ptge+PersonasInmueble,data=data_train)
> summary(modelo2)

Call:
lm(formula = varObjCont ~ edadmenor19 + AgricultureUnemploymentPtge +
    UnemployMore40_Ptge + PersonasInmueble, data = data_train)

Residuals:
    Min      1Q  Median      3Q     Max
-26.376  -4.578  -0.157   4.323  32.238

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                19.096292   0.309699  61.661  < 2e-16 ***
edadmenor19                 0.234659   0.018608  12.611  < 2e-16 ***
AgricultureUnemploymentPtge -0.028754   0.009496  -3.028  0.00247 **
UnemployMore40_Ptge          0.024647   0.003956   6.231 4.94e-10 ***
PersonasInmueble             2.457391   0.223788  10.981  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.994 on 6170 degrees of freedom
Multiple R-squared:  0.1349,    Adjusted R-squared:  0.1344
F-statistic: 240.6 on 4 and 6170 DF,  p-value: < 2.2e-16

> Rsq(modelo2,"varObjCont",data_train)
$r2
[1] 0.1349282

$r2_adj
[1] 0.134227

> Rsq(modelo2,"varObjCont",data_test)
$r2
[1] 0.1177833

$r2_adj
[1] 0.1149096
```

MODELO 2

Vemos que ha bajado un poco. Aunque lógicamente tiene muchas menos variables

Autor: FRANCISCO JAVIER POLO MARTINEZ
Actualizado 02 Marzo 2020
Arquitecto Técnico
Master Big Data

Pasamos al modelo 3

```
> modelo3<-lm(varObjCont~edadmenor19+UnemployMore40_Ptge+PersonasInmueble,data=data_train)
> summary(modelo3)

Call:
lm(formula = varObjCont ~ edadmenor19 + UnemployMore40_Ptge +
    PersonasInmueble, data = data_train)

Residuals:
    Min      1Q  Median      3Q     Max
-26.189  -4.620  -0.188   4.317  32.419

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         18.972747   0.307203  61.760  < 2e-16 ***
edadmenor19          0.233152   0.018613  12.526  < 2e-16 ***
UnemployMore40_Ptge  0.024524   0.003958   6.196 6.16e-10 ***
PersonasInmueble     2.412023   0.223433  10.795  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.998 on 6171 degrees of freedom
Multiple R-squared:  0.1336,    Adjusted R-squared:  0.1332
F-statistic: 317.3 on 3 and 6171 DF,  p-value: < 2.2e-16

> Rsq(modelo3,"varObjCont",data_train)
$r2
[1] 0.1336425

$r2_adj
[1] 0.1330809

> Rsq(modelo3,"varObjCont",data_test)
$r2
[1] 0.1173116

$r2_adj
[1] 0.115013
```

MODELO 3

Muy similar al modelo 2

Probamos el modelo 4 con una interacción

```
Call:
lm(formula = varObjCont ~ edadmenor19 + UnemployMore40_Ptge +
    PersonasInmueble + edadmenor19:PersonasInmueble, data = data_train)

Residuals:
    Min      1Q  Median      3Q     Max
-26.597  -4.658  -0.184   4.288  32.308

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 18.05295    0.47594  37.931  < 2e-16 ***
edadmenor19                  0.30331    0.03340   9.082  < 2e-16 ***
UnemployMore40_Ptge          0.02368    0.00397   5.966 2.57e-09 ***
PersonasInmueble             3.34374    0.43076   7.762 9.69e-15 ***
edadmenor19:PersonasInmueble -0.05914    0.02338  -2.530   0.0114 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.995 on 6170 degrees of freedom
Multiple R-squared:  0.1345,    Adjusted R-squared:  0.134
F-statistic: 239.8 on 4 and 6170 DF,  p-value: < 2.2e-16

> Rsq(modelo4,"varObjCont",data_train)
$r2
[1] 0.1345401

$r2_adj
[1] 0.1338386

> Rsq(modelo4,"varObjCont",data_test)
$r2
[1] 0.1171166

$r2_adj
[1] 0.1142407
```

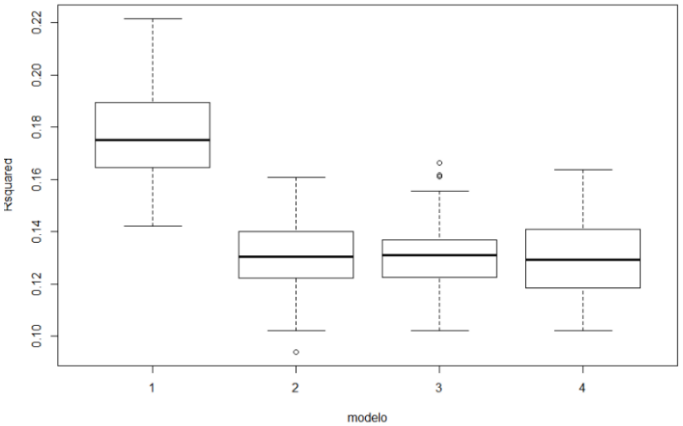
MODELO 4

Muy similar al modelo 2

Autor: FRANCISCO JAVIER POLO MARTINEZ
Actualizado 02 Marzo 2020
Arquitecto Técnico
Master Big Data

Ahora hacemos validación cruzada repetida. Esto consiste, en dividir el conjunto en un numero de veces, por ejemplo cinco y con esas cinco particiones considerar que una de ellas es test y las otras cuatro train y hacer las cinco permutaciones y luego vamos a repetir ese mismo proceso veinte veces haciéndolo de manera diferente. Esto nos va a eliminar la posible influencia de esa semilla, entonces en este caso vamos con todos los modelos

Y juntamos todos los resultados en un data set, para hacer un blxplot y valorar
Para en R2, el modelo 1 supera a los demás, pero tenemos que ver si esta compensada con el número de parámetros o no. Para ello




```
> aggregate(Rsquared~modelo, data = results, mean) #el 1 tiene mayor R2 medio
  modelo  Rsquared
1      1 0.1768676
2      2 0.1309325
3      3 0.1301797
4      4 0.1304467
```

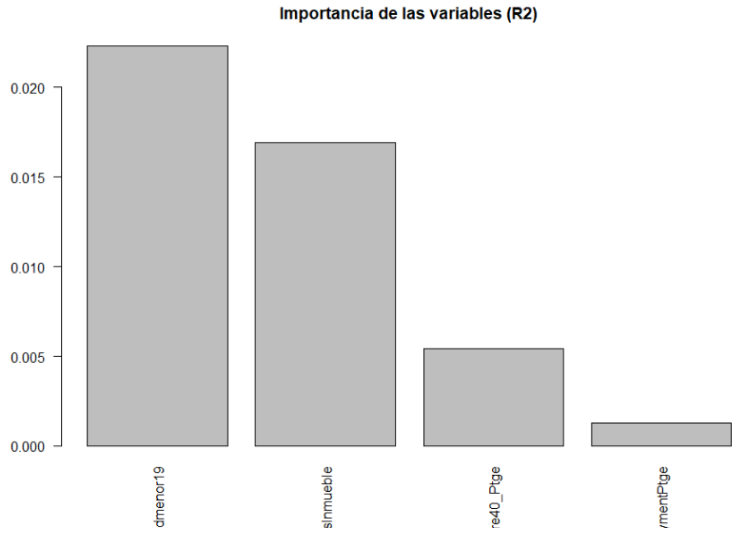
```
> aggregate(Rsquared~modelo, data = results, sd) #tb tiene mayor variabilidad
  modelo   Rsquared
1      1 0.01609291
2      2 0.01279964
3      3 0.01231722
4      4 0.01422531
```

```
[1] 5
```

Vemos que el modelo 1 tiene mejor media y mejor varianza, pero está super parametrizado, por lo que en principio nos quedamos con el modelo 2

```
> # Vemos los coeficientes del modelo ganador
> coef(modelo2)
                (Intercept)                edadmenor19 AgricultureUnemploymentPtge          UnemployMore40_Ptge               PersonasInmueble
                19.09629154                 0.23465898                -0.02875435                 0.02464704                 2.45739083
>
```

```
> #Evaluamos la estabilidad del modelo a partir de las diferencias en train y test:
> Rsq(modelo2,"varObjCont",data_train)
$r2
[1] 0.1349282

$r2_adj
[1] 0.134227

> Rsq(modelo2,"varObjCont",data_test)
$r2
[1] 0.1177833

$r2_adj
[1] 0.1149096

> # Vemos las variables mÁ¡s importantes del modelo ganador
> modelEffectSizes(modelo2)
lm(formula = varObjCont ~ edadmenor19 + AgricultureUnemploymentPtge +
    UnemployMore40_Ptge + PersonasInmueble, data = data_train)

Coefficients
                                SSR df pEta-sqr dR-sqr
(Intercept)                185960.234  1   0.3813     NA
edadmenor19                  7778.537  1   0.0251 0.0223
AgricultureUnemploymentPtge   448.493  1   0.0015 0.0013
UnemployMore40_Ptge          1898.912  1   0.0063 0.0054
PersonasInmueble             5897.640  1   0.0192 0.0169

Sum of squared errors (SSE): 301777.6
Sum of squared total  (SST): 348846.9
```



Autor: FRANCISCO JAVIER POLO MARTINEZ
Actualizado 02 Marzo 2020
Arquitecto Técnico
Master Big Data

Empezamos ahora con la Regresión Logística

Para ello, igual que siempre cargamos las funciones, las librerías que vamos a utilizar así como el archivo todo_bin creado anteriormente dentro de la variable todo.

Vemos el reparto original de la variable objeto binaria

```
> #veo el reparto original. Compruebo que la variable objetivo tome valor 1 para el evento y 0 para el no evento
> freq(todo$varObjBin) #ese ha de ser el error de referencia
     n    % val%
0 2912 37.7 37.7
1 4804 62.3 62.3
```

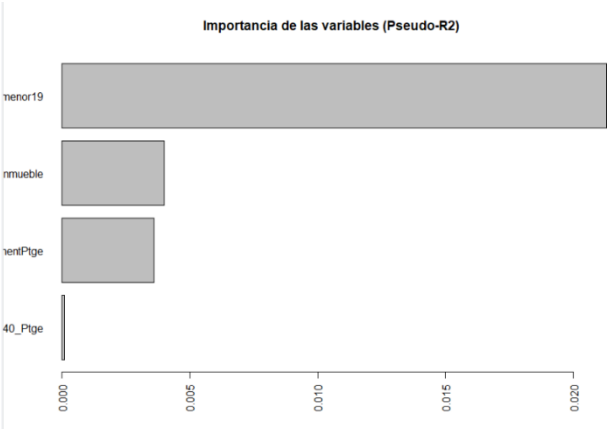
Hacemos la partición 80 - 20 con la correspondiente semilla

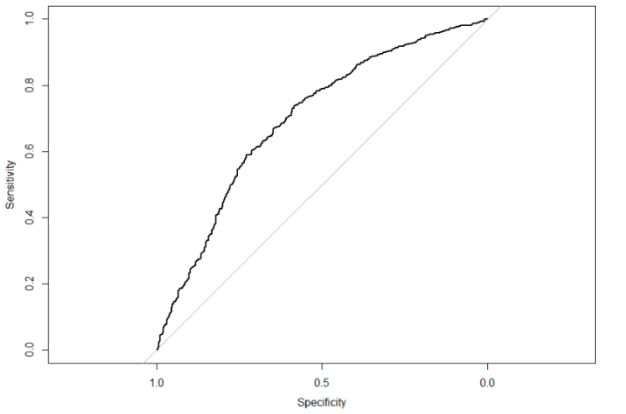
Empezamos con los modelos: Para ello usamos la función glm, que es mas general que la lm anterior, con el siguiente Código tipo para los modelos

```
modelo<-glm(varObjBin~CCAA+AgricultureUnemploymentPtge,data=data_train,family=binomial)
summary(modeloC)
pseudoR2(modeloC,data_train,"varObjBin")
pseudoR2(modeloC,data_test,"varObjBin")
modeloC$rank
```

MODELO A

```
Call:
glm(formula = varObjBin ~ edadmenor19 + AgricultureUnemploymentPtge +
    UnemployMore40_Ptge + PersonasInmueble, family = binomial,
    data = data_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1554  -1.1426   0.6640   0.9407   1.9520

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                 2.290523   0.112034  20.445  < 2e-16 ***
edadmenor19                -0.076801   0.005918 -12.977  < 2e-16 ***
AgricultureUnemploymentPtge -0.016134   0.002964  -5.444 5.22e-08 ***
UnemployMore40_Ptge         -0.001155   0.001389  -0.831    0.406
PersonasInmueble            -0.386511   0.067894  -5.693 1.25e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8183.9  on 6173  degrees of freedom
Residual deviance: 7520.2  on 6169  degrees of freedom
AIC: 7530.2

Number of Fisher Scoring iterations: 4

> pseudoR2(modeloA,data_train,"varObjBin")
[1] 0.08110349
> pseudoR2(modeloA,data_test,"varObjBin") #En test se obtienen mejor resu
[1] 0.08357881
> modeloA$rank
[1] 5
```





Autor: FRANCISCO JAVIER POLO MARTINEZ
Actualizado 02 Marzo 2020
Arquitecto Técnico
Master Big Data

Para el modelo B con una interaccion

```
> modeloB<-glm(varObjBin~-CCAA+edadmenor19+AgricultureUnemploymentPtge+edadmenor19:CodigoProvincia,data=data_train,family=binomial)
> summary(modeloB)

Call:
glm(formula = varObjBin ~ -CCAA + edadmenor19 + AgricultureUnemploymentPtge +
    edadmenor19:CodigoProvincia, family = binomial, data = data_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0745  -1.1237   0.6517   0.9234   2.0816

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  2.0200211  0.0741290  27.250  < 2e-16 ***
edadmenor19                 -0.1341025  0.0054386 -24.658  < 2e-16 ***
AgricultureUnemploymentPtge -0.0158389  0.0029841  -5.308 1.11e-07 ***
edadmenor19:CodigoProvincia  0.0013703  0.0001197  11.452  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8183.9  on 6173  degrees of freedom
Residual deviance: 7418.4  on 6170  degrees of freedom
AIC: 7426.4

Number of Fisher Scoring iterations: 4

> pseudoR2(modeloB,data_train,"varObjBin")#No parece muy buena idea
[1] 0.09353165
> pseudoR2(modeloB,data_test,"varObjBin")
[1] 0.08276563
> modeloB$rank
[1] 4
```

MODELO B

Para el modelo C

```
Call:
glm(formula = varObjBin ~ CCAA + AgricultureUnemploymentPtge +
    edadmenor19, family = binomial, data = data_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5578  -0.2044   0.3977   0.4592   3.4197

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                -0.931470   0.158387  -5.881 4.08e-09 ***
CCAAAragón                  1.825890   0.142201  12.840  < 2e-16 ***
CCAAAsturias                0.751975   0.269168   2.794 0.005211 **
CCAABaleares                1.302046   0.295469   4.407 1.05e-05 ***
CCAACanarias                1.780966   0.274936   6.478 9.31e-11 ***
CCAACantabria               3.262608   0.409871   7.960 1.72e-15 ***
CCAACastillaLeón            3.395643   0.145849  23.282  < 2e-16 ***
CCAACastillaMancha          1.939179   0.134312  14.438  < 2e-16 ***
CCAACataluña               -4.955789   0.715462  -6.927 4.31e-12 ***
CCAACeuta                  13.417129 324.743721   0.041 0.967044
CCAAComValenciana           3.542142   0.214048  16.548  < 2e-16 ***
CCAAExtremadura             0.457723   0.162483   2.817 0.004847 **
CCAAGalicia                 4.132862   0.340651  12.132  < 2e-16 ***
CCAAMadrid                  3.336617   0.320485  10.411  < 2e-16 ***
CCAAMelilla                13.415636 324.743725   0.041 0.967048
CCAAMurcia                  3.206571   0.537283   5.968 2.40e-09 ***
CCAANavarra                 0.069835   0.182410   0.383 0.701832
CCAAPaísVasco              -2.931193   0.515990  -5.681 1.34e-08 ***
CCAARioja                   2.750837   0.261904  10.503  < 2e-16 ***
AgricultureUnemploymentPtge -0.014041   0.003961  -3.544 0.000394 ***
edadmenor19                 0.003215   0.006689   0.481 0.630750
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8183.9  on 6173  degrees of freedom
Residual deviance: 4587.7  on 6153  degrees of freedom
AIC: 4629.7

Number of Fisher Scoring iterations: 11

> pseudoR2(modeloC,data_train,"varObjBin")
[1] 0.4394224
> pseudoR2(modeloC,data_test,"varObjBin")
[1] 0.4114298
> modeloC$rank
[1] 21
```

MODELO C

Mas parametrizado pero con mejores valores.

Autor: FRANCISCO JAVIER POLO MARTINEZ
Actualizado 02 Marzo 2020
Arquitecto Técnico
Master Big Data

Para el modelo D

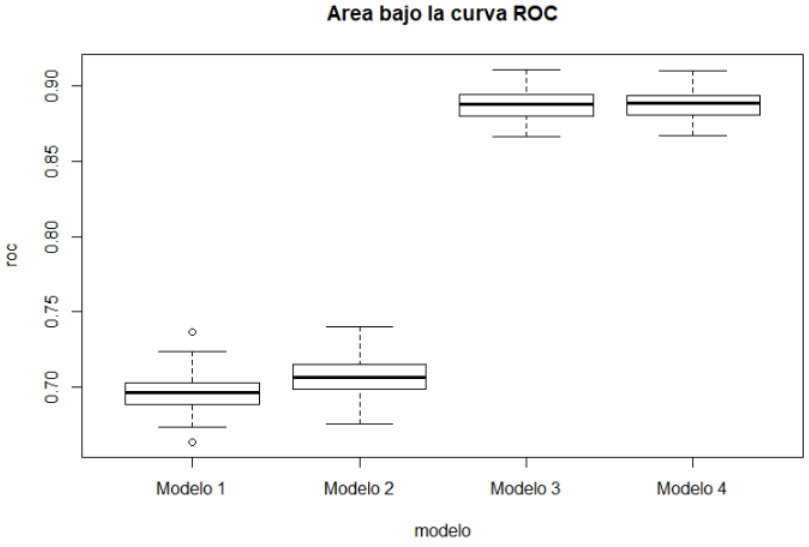
```
> modeloD<-glm(varObjBin~CCAA+AgricultureUnemploymentPtge+edadmenor19+AgricultureUnemploymentPtge:edadmenor19,data=data_train,family=binomial)
> summary(modeloD)

Call:
glm(formula = varObjBin ~ CCAA + AgricultureUnemploymentPtge + 
    edadmenor19 + AgricultureUnemploymentPtge:edadmenor19, family = binomial, 
    data = data_train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.5613  -0.2013   0.3942   0.4402   3.4281  

Coefficients:
                                            Estimate Std. Error z value Pr(>|z|)    
(Intercept)                               -1.014e+00  1.584e-01  -6.399 1.56e-10 ***
CCAAAragón                                 1.737e+00  1.437e-01  12.087  < 2e-16 ***
CCAAAsturias                               6.410e-01  2.702e-01   2.372 0.017682 *  
CCAABaleares                               1.143e+00  2.983e-01   3.832 0.000127 ***
CCAACanarias                               1.649e+00  2.770e-01   5.952 2.64e-09 ***
CCAACantabria                              3.146e+00  4.108e-01   7.657 1.91e-14 ***
CCAACastillaLeón                           3.314e+00  1.472e-01  22.517  < 2e-16 ***
CCAACastillaMancha                         1.877e+00  1.351e-01  13.887  < 2e-16 ***
CCAACataluña                              -5.090e+00  7.164e-01  -7.105 1.20e-12 ***
CCAACeuta                                  1.313e+01  3.247e+02   0.040 0.967737    
CCAAComValenciana                          3.449e+00  2.152e-01  16.028  < 2e-16 ***
CCAAExtremadura                            4.033e-01  1.637e-01   2.463 0.013761 *  
CCAAGalicia                                4.013e+00  3.418e-01  11.739  < 2e-16 ***
CCAAMadrid                                 3.149e+00  3.240e-01   9.719  < 2e-16 ***
CCAAMelilla                                1.312e+01  3.247e+02   0.040 0.967775    
CCAAMurcia                                 3.244e+00  5.387e-01   6.021 1.74e-09 ***
CCAANavarra                               -4.261e-02  1.848e-01  -0.231 0.817617    
CCAAPaisVasco                             -3.077e+00  5.176e-01  -5.944 2.78e-09 ***
CCAARioja                                  2.698e+00  2.623e-01  10.287  < 2e-16 ***
AgricultureUnemploymentPtge                1.714e-02  9.602e-03   1.786 0.074176 .  
edadmenor19                                1.726e-02  7.674e-03   2.249 0.024503 *  
AgricultureUnemploymentPtge:edadmenor19 -2.383e-03  6.568e-04  -3.629 0.000285 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8183.9  on 6173  degrees of freedom
Residual deviance: 4573.8  on 6152  degrees of freedom
AIC: 4617.8

Number of Fisher Scoring iterations: 11

> pseudoR2(modeloD,data_train,"varObjBin")#No parece muy buena idea
[1] 0.4411163
> pseudoR2(modeloD,data_test,"varObjBin")
[1] 0.411055
> modeloD$rank
[1] 22
```

MODELO D

Un poquito mejor que el anterior

Los representamos gráficamente con un boxplot conjunto

```
boxplot(roc~modelo,data=total,main="Area bajo la curva ROC")
aggregate(roc~modelo, data = total, mean)
aggregate(roc~modelo, data = total, sd)
```



Vemos el número de parámetros de cada modelo:

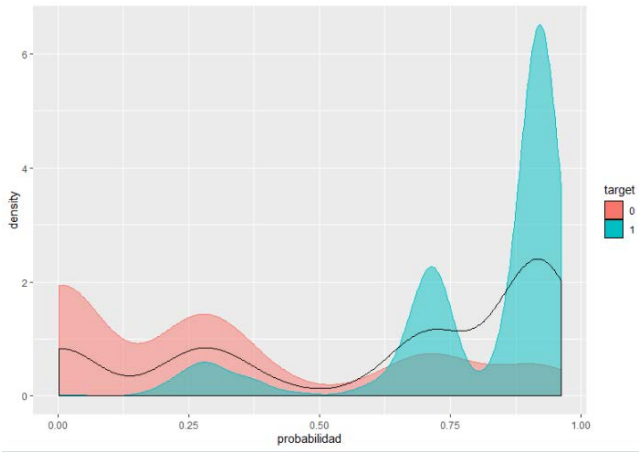
```
> #miro el numero de parametros
> modeloA$rank
[1] 5
> modeloB$rank
[1] 4
> modeloC$rank
[1] 21
> modeloD$rank
[1] 22
```

Autor: FRANCISCO JAVIER POLO MARTINEZ
Actualizado 02 Marzo 2020
Arquitecto Técnico
Master Big Data

Buscamos el mejor punto de corte:

gráfico de las probabilidades obtenidas

hist_targetbinaria(predict(modeloC, newdata=data_test,type="response"),data_test$varObjBin,"probabilidad")



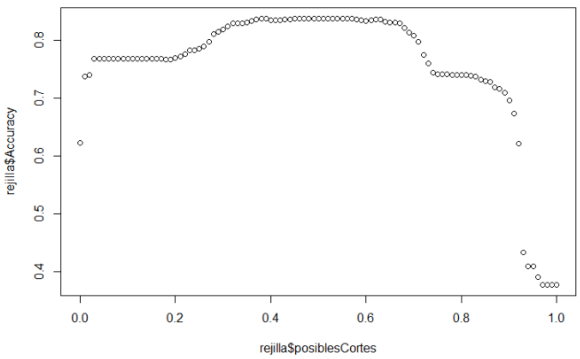
Y probamos 2

```
sensEspCorte(modeloC,data_test,"varObjBin",0.5,"1")
sensEspCorte(modeloC,data_test,"varObjBin",0.75,"1")
```

```
> #probamos dos
> sensEspCorte(modeloC,data_test,"varObjBin",0.5,"1")
      Accuracy    Sensitivity    Specificity Pos Pred Value Neg Pred Value
     0.8365759      0.9000000      0.7319588      0.8470588      0.8160920
> sensEspCorte(modeloC,data_test,"varObjBin",0.75,"1")
      Accuracy    Sensitivity    Specificity Pos Pred Value Neg Pred Value
     0.7412451      0.6500000      0.8917526      0.9082969      0.6070175
>
```

En el primero estamos reconociendo muy bien a los 1 (Sensitivity 0.90) y peor a los 0. En el segundo caso es peor. Esta mejor balanceado el primero.

Ahora aplicamos la rejilla a fin de encontrar el modelo mas balanceado



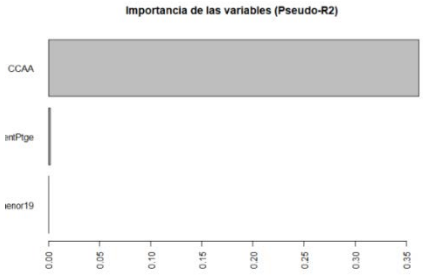

```
> rejilla$Youden<-rejilla$Sensitivity+rejilla$Specificity-1
> plot(rejilla$posiblesCortes,rejilla$Youden)
> plot(rejilla$posiblesCortes,rejilla$Accuracy)
> rejilla$posiblesCortes[which.max(rejilla$Youden)]
[1] 0.63
> rejilla$posiblesCortes[which.max(rejilla$Accuracy)]
[1] 0.46
```

Vemos que el máximo esta en 0.46 y a partir de 0.63 baja drásticamente.

```
> #Los comparamos
> sensEspCorte(modeloC,data_test,"varObjBin",0.40,"1")
      Accuracy    Sensitivity    Specificity Pos Pred Value Neg Pred Value
     0.8346304      0.9041667      0.7199313      0.8419011      0.8199609
> sensEspCorte(modeloC,data_test,"varObjBin",0.60,"1")
      Accuracy    Sensitivity    Specificity Pos Pred Value Neg Pred Value
     0.8326848      0.8916667      0.7353952      0.8475248      0.8045113
>
```

Esta mejor balanceado en este caso el segundo caso entre 0.60 y 1

Autor: FRANCISCO JAVIER POLO MARTINEZ
Actualizado 02 Marzo 2020
Arquitecto Técnico
Master Big Data



La variable mas importante de modelo C es CCAA

Con los siguiente coeficientes:

```
> # Vemos los coeficientes del modelo ganador
> coef(modeloC)
                  (Intercept)                    CCAAAragón                  CCAAAsturias                CCAABaleares                CCAACanarias 
                -0.931470147                  1.825889704                  0.751974744                  1.302045946                  1.780965679 
               CCAACantabria            CCAACastillaLeón            CCAACastillaMancha               CCAACataluña                 CCAACeuta 
                3.262608489                  3.395643287                  1.939178982                 -4.955788570                 13.417128959 
        CCAAComValenciana               CCAAExtremadura                  CCAAGalicia                  CCAAMadrid                  CCAAMelilla 
                3.542142276                  0.457722851                  4.132862004                  3.336616559                 13.415635810 
             CCAAMurcia                  CCAANavarra                  CCAAPaísVasco                   CCAARioja AgricultureUnemploymentPtge 
                3.206571009                  0.069835332                 -2.931192960                  2.750836684                 -0.014040672 
             edadmenor19 
                0.003215343 
> |
```

Evaluamos la estabilidad del modelo a partir de las diferencias en train y test:

```
pseudoR2(modeloC,data_train,"varObjBin")
pseudoR2(modeloC,data_test,"varObjBin")
roc(data_train$varObjBin, predict(modeloC,data_train,type = "response"), direction="<")
roc(data_test$varObjBin, predict(modeloC,data_test,type = "response"), direction="<")
sensEspCorte(modeloC,data_train,"varObjBin",0.60,"1")
sensEspCorte(modeloC,data_test,"varObjBin",0.60,"1")
```

```
> #Evaluamos la estabilidad del modelo a partir de las diferencias en train y test:
> pseudoR2(modeloC,data_train,"varObjBin")
[1] 0.4394224
> pseudoR2(modeloC,data_test,"varObjBin")
[1] 0.4114298
> roc(data_train$varObjBin, predict(modeloC,data_train,type = "response"), direction="<")
Setting levels: control = 0, case = 1

Call:
roc.default(response = data_train$varObjBin, predictor = predict(modeloC,     data_train, type = "response"), direction = "<")

Data: predict(modeloC, data_train, type = "response") in 2330 controls (data_train$varObjBin 0) < 3844 cases (data_train$varObjBin 1).
Area under the curve: 0.8928
> roc(data_test$varObjBin, predict(modeloC,data_test,type = "response"), direction="<")
Setting levels: control = 0, case = 1

Call:
roc.default(response = data_test$varObjBin, predictor = predict(modeloC,     data_test, type = "response"), direction = "<")

Data: predict(modeloC, data_test, type = "response") in 582 controls (data_test$varObjBin 0) < 960 cases (data_test$varObjBin 1).
Area under the curve: 0.8801
> sensEspCorte(modeloC,data_train,"varObjBin",0.60,"1")
      Accuracy    Sensitivity    Specificity Pos Pred Value Neg Pred Value 
     0.8419177      0.9071280      0.7343348      0.8492450      0.8273694 
> sensEspCorte(modeloC,data_test,"varObjBin",0.60,"1")
      Accuracy    Sensitivity    Specificity Pos Pred Value Neg Pred Value 
     0.8326848      0.8916667      0.7353952      0.8475248      0.8045113 
> |
```

```
> # Odds ratios
> epiDisplay::logistic.display(modeloC)

Logistic regression predicting varObjBin : 1 vs 0

                                  crude OR(95%CI)                                 adj. OR(95%CI)                  P(Wald's test) P(LR-test)
CCAA: ref.=Andalucía                                                                                              < 0.001
   Aragón                         6.71 (5.16,8.71)                                6.21 (4.7,8.2)                  < 0.001
   Asturias                       2.36 (1.41,3.96)                                2.12 (1.25,3.59)                0.005
   Baleares                       4.37 (2.47,7.73)                                3.68 (2.06,6.56)                < 0.001
   Canarias                       6.75 (3.96,11.51)                               5.94 (3.46,10.17)               < 0.001
   Cantabria                      30.36 (13.67,67.42)                             26.12 (11.7,58.32)              < 0.001
   CastillaLeón                   32.37 (25.08,41.78)                             29.83 (22.42,39.71)             < 0.001
   CastillaMancha                 7.32 (5.69,9.41)                                6.95 (5.34,9.05)                < 0.001
   Cataluña                       0.01 (0,0.03)                                   0.01 (0,0.03)                   < 0.001
   Ceuta                          834464.56 (0,2.20676050681942e+282)             671405.77 (0,1.77558617931349e+282) 0.967
   ComValenciana                  37.84 (24.96,57.38)                             34.54 (22.71,52.54)             < 0.001
   Extremadura                    1.58 (1.16,2.17)                                1.58 (1.15,2.17)                0.005
   Galicia                        67.53 (34.89,130.71)                            62.36 (31.98,121.57)            < 0.001
   Madrid                         33.72 (18.16,62.62)                             28.12 (15.01,52.71)             < 0.001
   Melilla                        834464.56 (0,2.20676051553397e+282)             670404.01 (0,1.77294948507694e+282) 0.967
   Murcia                         25.47 (8.9,72.89)                               24.69 (8.62,70.78)              < 0.001
   Navarra                        1.2 (0.85,1.71)                                 1.07 (0.75,1.53)                0.702
   PaísVasco                      0.06 (0.02,0.17)                                0.05 (0.02,0.15)                < 0.001
   Rioja                          16.22 (9.83,26.77)                              15.66 (9.37,26.16)              < 0.001

AgricultureUnemploymentPtge (cont. var.) 0.98 (0.97,0.98)                       0.99 (0.98,0.99)                < 0.001        < 0.001

edadmenor19 (cont. var.)          0.9043 (0.8966,0.9121)                        1.0032 (0.9902,1.0165)          0.631          0.631

Log-likelihood = -2293.8559
No. of observations = 6174
AIC value = 4629.7119
```

CONCLUSION:

Nos fijamos en el OR ajustados, teniendo como referencia Andalucía por lo que la probabilidad de tener el evento DERECHA con valor 1 en Madrid es 28.12 veces superior a la probabilidad de ser 1 en Andalucía