Máster en Big Data y Business Analytics

Universidad Complutense de Madrid

Curso 2020-2021

Guillermo Sánchez-Mariscal

- Introducción al objetivo del problema y las variables implicadas.
- Importación del conjunto de datos y asignación correcta de los tipos de variables.
  - Importacion de datos
- Análisis descriptivo de datos en el conjunto de training. Numero de observaciones, numero y naturaleza de variables, datos erróneos etc.
- Corrección de los errores detectados.
- Análisis de valores atípicos. Decisiones.
  - Missings
- Análisis de valores perdidos. Imputaciones.
- Transformaciones de variables y relaciones con las variables objetivo.
- Detección de las relaciones entre las variables input y objetivo.
- Regresion lineal
  - Seleccion de variables clasica
    - Generacion de iteraciones
    - Transformaciones y las variables originales
    - Trans e interacciones
    - Validacion cruzada repetida
  - Seleccion aleatoria
  - Seleccion modelo ganador
  - Interpretación de los coeficientes de dos variables incluidas en el modelo, una binaria y otra continua
  - Justificación de porque es el mejor modelo y medir la calidad del mismo
- Regresion logistica
  - Seleccion clasica
    - Generacion de interacciones
    - Transformaciones y las variables originales
    - Transformaciones e interacciones
    - Validacion cruzada repetida
  - Seleccion aleatoria
  - Punto de corte
  - Interpretación de los coeficientes de dos variables incluidas en el modelo, una binaria y otra continua
  - Justificación del mejor modelo y medir la calidad del mismo

1. Introducción al objetivo del problema y las variables implicadas.

En nuestro caso hemos decidido seleccionar como variable objetivo continua el porcentaje de votos a partidos de izquierda y de variable binaria Izquierda, que toma el valor 1 si la suma de los votos de izquierdas es superior a la de derechas y otros y, 0, en otro caso.

2. Importación del conjunto de datos y asignación correcta de los tipos de variables.

Cargo las librerías y funciones que necesitaré en los demás apartados

```
library(questionr)
library(psych)
library(car)
library(corrplot)
library(caret)
library(ggplot2)
library(lmSupport)
library(unmarked)
library(VGAM)
library(pROC)
library(glmnet)
source("./FuncionesRosa.R")
library(readxl)
```

3.1 Importación de datos

```
elecciones_data <- read_excel("./DatosEleccionesEspaña.xlsx")
datos <- elecciones_data
```

Comprobamos el tipo asignado a cada variable.

```
str(datos)
```

Nos damos cuenta de que no todas las variables categóricas están como factores. A las siguientes variables les cambiaremos el tipo a factor:

- Código provincia
- CCAA
- AbstencionAlta
- Izquierda
- Derecha
- Actividad Ppal
- Densidad

```
datos[,c(2,3,7,11,12,34,38)] <- lapply(datos[,c(2,3,7,11,12,34,38)], factor)
```

3.Análisis descriptivo de datos en el conjunto de training. Número de observaciones, número y naturaleza de variables, datos erróneos etc. 4. Corrección de los errores detectados.

Vamos a ver un resumen de los datos para poder así, limpiar aquellos que están fuera de los rangos o que toman valores raros.

```
summary(datos)
```

Podemos sacar las siguientes conclusiones:

* Valores fura de rango en ForeignersPtge y en SameComAutonPtge

* Posible NA en Explotaciones

* Densidad tiene una categoría '?'

Corrijo los valores fuera de rango

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

Cambio los valores 99999 de Explotaciones por NA

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

Cambio ? por Na en la variable densidad

```
datos$Densidad<-recode.na(datos$Densidad,"?")
```

Cuento el número de valores diferentes para las numéricas

```
sapply(Filter(is.numeric, datos),function(x) length(unique(x)))
```

No encuentro ningún valor raro, como por ejemplo alguna binaria que de más de dos distintos.

Ahora voy a ver el reparto de las categorías de las variables cualitativas.

En las siguientes variables no encontramos categorías poco representadas por lo que no muestro su output.

```
freq(datos$CodigoProvincia)
freq(datos$Izquierda)
freq(datos$Derecha)
freq(datos$AbstencionAlta)
freq(datos$Densidad)
```

Comunidad Autónoma: Vemos que Ceuta y Melilla no están nada representadas, habría que juntarlas con otra comunidad, en este caso elijo Andalucía por proximidad, aunque no sería mala opción juntarlas con Murcia que es la categoría menos representada.

```
freq(datos$CCAA)
datos$CCAA<-recode(datos$CCAA, "c('Andalucía','Melilla','Ceuta')='Andalucía'")
```

- Actividad Principal: En este caso construcción e industria están poco representada por lo que, he decicido juntarla con servicios que es la siguiente categoría menos representada.

```
freq(datos$ActividadPpal)
datos$ActividadPpal<-recode(datos$ActividadPpal,
"c('Construccion','Industria','Servicios')='Servicios-Construccion-Industria'")
```

5.Análisis de valores atipicos. Decisiones.

Indico la variableObj continua y binaria.

```
varObjCont<-datos$Izda_Pct
varObjBin<-datos$Izquierda
```

Declaro los inputs que vamos a utilizar (los atípicos y los missings se gestionan sólo de las input). Quito las otras variables objetivos que se podían utilizar, y utilizo de índices el nombre de la población más su provincia

```
input<-as.data.frame(datos[,-c(6:12)])
row.names(input)<- paste(datos$Name,'-',datos$CodigoProvincia)
```

Cuento el porcentaje de atípicos de cada variable. Si son muchos, elimino esas categoríias

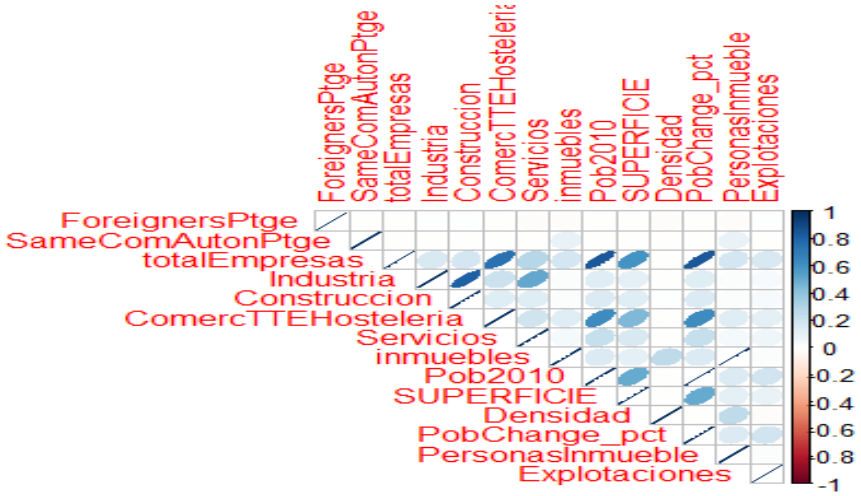
```
sapply(Filter(is.numeric, input),function(x) atipicosAmissing(x)[[2]])/nrow(input)
```

Ninguna variable tiene muchos atípicos por lo que no los convertiremos a missings

5.1 Missings

Busco si existe algún patrón 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")
```



De aquí podemos decir que los missings totalEmpresas están relacionados con ComercTTEHosteleria, Pob2010, SUPERFICIE, PobChange_pct

A continuación, miro la proporción de missings por variable y observación

```
input$prop_missings<-apply(is.na(input),1,mean)
summary(input$prop_missings)
(prop_missingsVars<-apply(is.na(input),2,mean))
```

No hay ninguna variable con más de la mitad de sus datos missings por lo que no elimino ninguna.

Tampoco recategorizo los missings de ninguna variable a otra categoría poque ninguna tiene suficientes, la densidad es la la categórica con mayor porcentaje de missings, y es muy bajo (1.13%), por lo que no hacemos nada

6.Análisis de valores perdidos. Imputaciones.

Imputo todas las variables cuantitativas, seleccionando la mediana como tipo de imputación

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

Imputo todas las variables cualitativas, seleccionar un valor aleatorio como tipo de imputación

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

Se cambia el tipo de factor a character al imputar, así 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)
```

Ya tenemos los datos depurados y los guardamos.

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

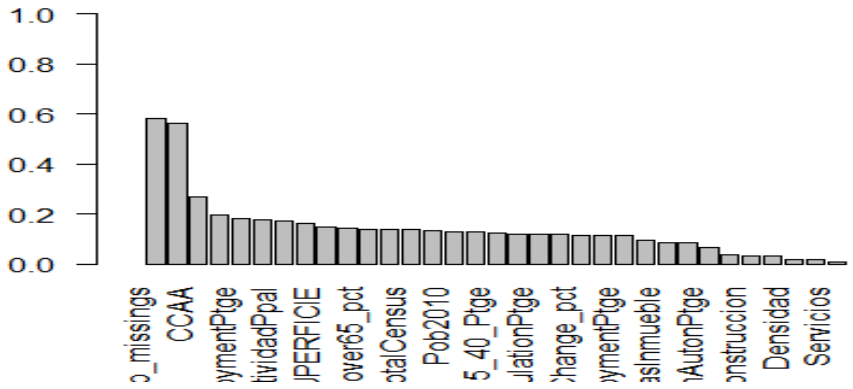
7.Transformaciones de variables y relaciones con las variables objetivo.

```
datos <- readRDS("datosElecc")
varObjCont <- datos$varObjCont
varObjBin <- datos$varObjBin
input <- datos[,-c(1:3)]
```

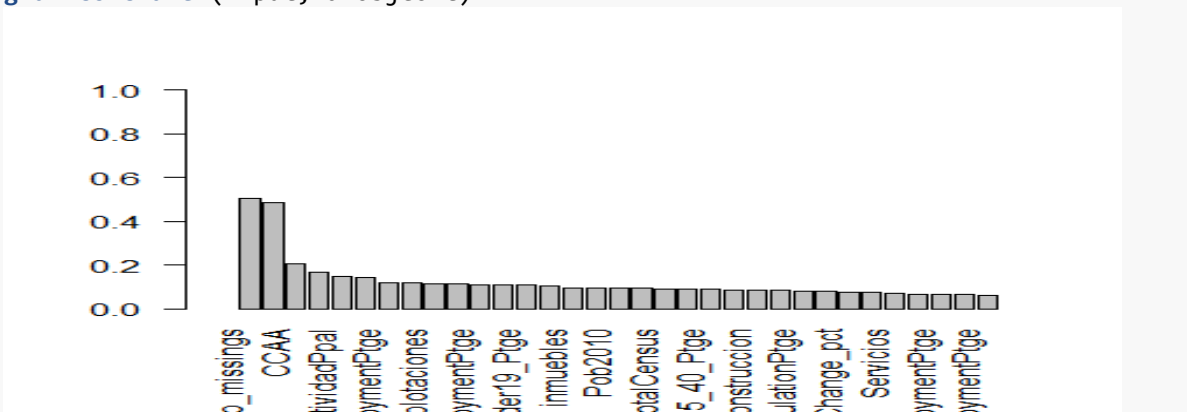
8.Detección de las relaciones entre las variables input y objetivo.

A continuación, vamos a ver la importancia de las variables tanto para la variable objetivo binaria como para la continua.

```
graficoVcramer(input,varObjBin)
```



```
graficoVcramer(input,varObjCont)
```

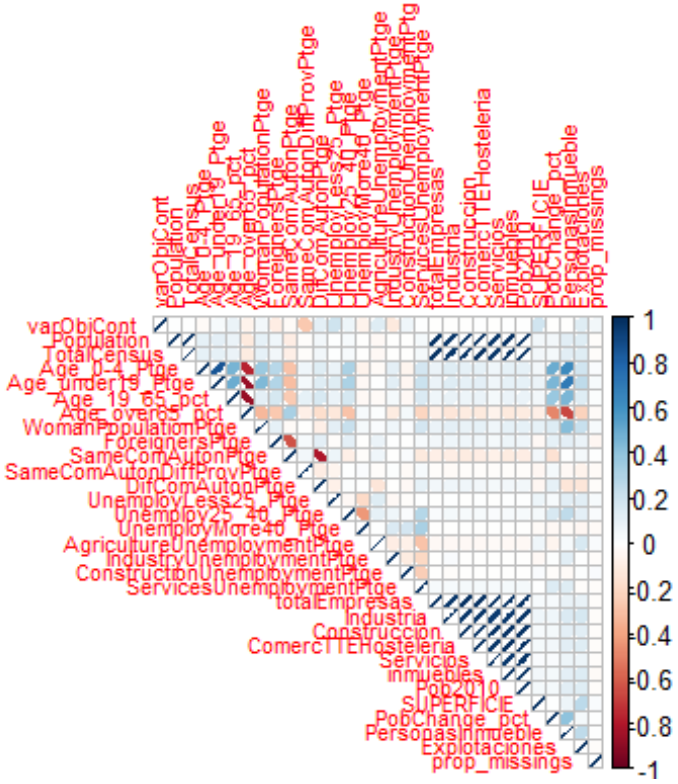


Podemos ver que tanto provincia como CCAA tienen una importancia muy parecida, por lo que seguramente representen lo mismo. Voy a eliminar provincia porque tienen muchas más categorías.

```
input <-as.data.frame(input[,-c(1)])
```

Ahora vamos a analizar el grafico de correlaciones de las distintas variables.

```
corrplot(cor(cbind(varObjCont,Filter(is.numeric, input)), use="pairwise",
method="pearson"), method = "ellipse",type = "upper" , tl.cex = 0.7)
```



Vemos como las siguientes variables están muy correlacionadas, prácticamente correlación 1.

- Population
- TotalCensus

- totalEmpresas
- Industria
- Construccion
- ComercTTEHosteleria
- Servicios
- Inmuebles

Por lo que decido quedarme únicamente con una, en este caso he elegido Population

```
input <-as.data.frame(input[,-c(3,20:24,26,27)])
```

9. Regresión lineal

Paso a buscar las mejores transformaciones para las variables numéricas con respecto a los dos tipos de variables

```
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")
```

Hago la partición train y test de los datos con transformaciones.

```
todo<-readRDS("todo_cont")
set.seed(12345678)
trainIndex <- createDataPartition(todo$varObjCont, p=0.8, list=FALSE)
data_train <- todo[trainIndex,]
data_test <- todo[-trainIndex,]
```

Selección de variables clásica

```
null <- lm(varObjCont ~ 1, data = data_train) #Modelo minimo
full <- lm(varObjCont ~ ., data = data_train[,c(1:25,48)]) #Modelo maximo, con las transformaciones

modeloStepAIC <- step(null, scope=list(lower=null, upper=full), direction="both")
summary(modeloStepAIC)
Rsq(modeloStepAIC,"varObjCont",data_test) #R^2 =0.6249


modeloBackAIC <- step(full, scope=list(lower=null, upper=full), direction="backward")
summary(modeloBackAIC)
Rsq(modeloBackAIC,"varObjCont",data_test) #R^2 = 0.6248


modeloStepBIC <- step(null, scope=list(lower=null, upper=full), direction="both",
k=log(nrow(data_train)))
summary(modeloStepBIC)
Rsq(modeloStepBIC,"varObjCont",data_test) #R^2 = 0.6239


modeloBackBIC <- step(full, scope=list(lower=null, upper=full), direction="backward",
k=log(nrow(data_train)))
summary(modeloBackBIC)
Rsq(modeloBackBIC,"varObjCont",data_test) #R^2 = 0.6238
```

Por r cuadrado entre el método Stepwise y el backward me daría igual cual elegir. Voy a ver el número de parámetros.

```
modeloStepAIC$rank #33
modeloBackAIC$rank #33
modeloStepBIC$rank #28
modeloBackBIC$rank #29
```

Si nos fijamos en el número de variables sería más interesante escoger modeloStepBIC o modeloBackBI

Generación de iteraciones

```
formInt <- formulaInteracciones(todo[,c(1:25,48)],26)
fullInt <- lm(formInt, data=data_train) #Modelo con todas las variables y todas las interacciones

modeloStepAIC_int <- step(null, scope=list(lower=null, upper=fullInt), direction="both")
summary(modeloStepAIC_int)
Rsq(modeloStepAIC_int,"varObjCont",data_test) # R^2 = 0.6588

modeloStepBIC_int<-step(null, scope=list(lower=null, upper=fullInt),
direction="both",k=log(nrow(data_train)))
summary(modeloStepBIC_int)
Rsq(modeloStepBIC_int,"varObjCont",data_test) #0.6447

modeloStepAIC_int$rank #122
modeloStepBIC_int$rank #50
```

Por el principio de parsimonia, es preferible el modeloStepBIC_int, puesto que aunque tenga un menor r cuadrado el número de variables es mucho más bajo que el modeloStepAIC_int.

Transformaciones y las variables originales

```
fullT <- lm(varObjCont~., data=data_train)

modeloStepAIC_trans <- step(null, scope=list(lower=null, upper=fullT), direction="both")
summary(modeloStepAIC_trans)
Rsq(modeloStepAIC_trans,"varObjCont",data_test) #0.6369434

modeloStepBIC_trans <- step(null, scope=list(lower=null, upper=fullT),
direction="both",k=log(nrow(data_train)))
summary(modeloStepBIC_trans)
Rsq(modeloStepBIC_trans,"varObjCont",data_test) #0.6364797

modeloStepAIC_trans$rank #42
modeloStepBIC_trans$rank #30
```

El mejor es modeloStepBIC_trans en este caso elegimos este modelo porque la diferencia de su r cuadrado es muy pequeña y en cambio, tiene 12 variables menos

Transformaciones e interacciones

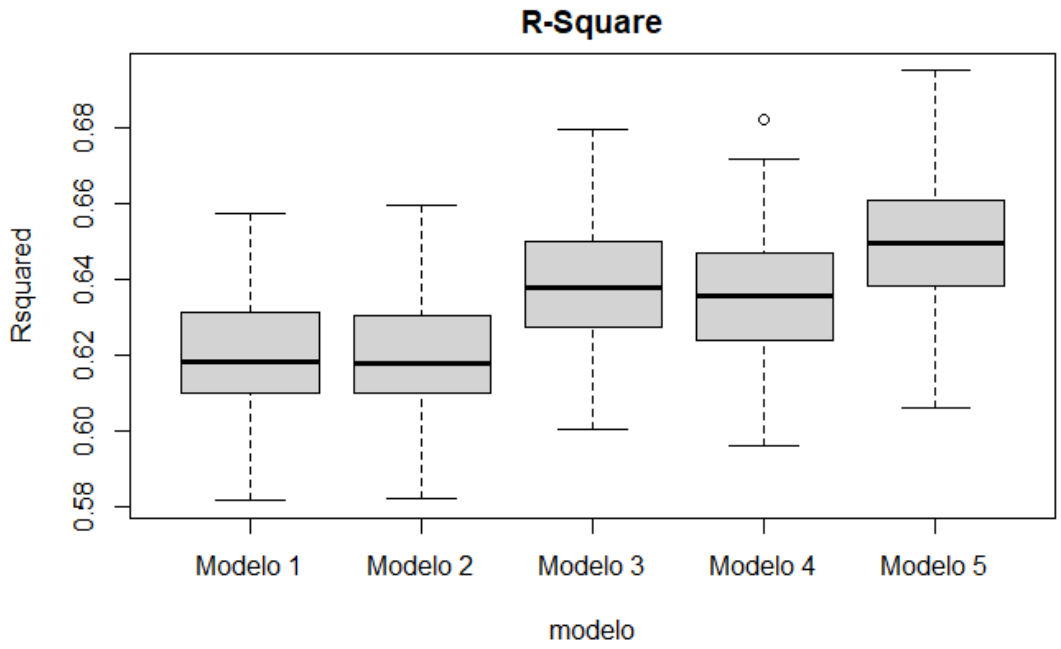
```
formIntT <- formulaInteracciones(todo,48)
fullIntT <- lm(formIntT, data=data_train)

modeloStepAIC_transInt<-step(null, scope=list(lower=null, upper=fullIntT), direction="both")
summary(modeloStepAIC_transInt)
Rsq(modeloStepAIC_transInt,"varObjCont",data_test) #0.6661591

modeloStepBIC_transInt<-step(null, scope=list(lower=null, upper=fullIntT),
direction="both",k=log(nrow(data_train)))
summary(modeloStepBIC_transInt)
Rsq(modeloStepBIC_transInt,"varObjCont",data_test) #0.6490593

modeloStepAIC_transInt$rank #140
modeloStepBIC_transInt$rank #48
```

Por el principio de parsimonia, es preferible modeloStepBIC_transInt, modeloStepAIC_transInt tiene 140 variables

Validación cruzada repetida

Recordemos que este método consiste en dividir el conjunto de datos en submuestras e iterativamente construir el modelo con todas las observaciones menos las de una submuestra y evaluarlo a continuación con las observaciones de dicha submuestra excluida. La comparación de modelos se suele llevará a cabo a partir del R2. Pruebo los mejores modelos obtenidos anteriormente.

```
total <- c()
modelos <-
sapply(list(modeloStepBIC,modeloBackBIC,modeloStepBIC_int,modeloStepBIC_trans,modeloStepBIC_transInt
),formula)
for (i in 1:length(modelos)){
  set.seed(1712)
  vcr<-train(as.formula(modelos[[i]]), data = data_train,
             method = "lm",
             trControl = trainControl(method="repeatedcv", number=5, repeats=20,
                                      returnResamp="all")
  )
  total<-rbind(total,cbind(vcr$resample[,1:2],modelo=rep(paste("Modelo",i),
                                                         nrow(vcr$resample))))
}
boxplot(Rsquared~modelo,data=total,main="R-Square")
aggregate(Rsquared~modelo, data = total, mean) #muy parecidos
aggregate(Rsquared~modelo, data = total, sd)
```



A simple vista parece que el modelo 5 es el mejor 5 en términos de "bondad media", su deviación es mayor que la de los modelos 1,2,4,5 pero sus bondades son más bajas.

```
modeloStepBIC$rank #28
modeloBackBIC$rank #29
modeloStepBIC_int$rank # 50
modeloStepBIC_trans$rank #30
modeloStepBIC_transInt$rank #48
```

[1] 28

[1] 29

[1] 50

[1] 30

[1] 48

Vemos el número de parámetros, el modelo 1,2,4 tienen menos pero su r cuadrado es mucho menor, nos quedamos el 5: modeloStepBIC_transInt

Selección aleatoria

```
rep <- 100
prop <- 0.7
modelosGenerados <- c()
for (i in 1:rep){
  set.seed(12345+i)
  subsample<-data_train[sample(1:nrow(data_train),prop*nrow(data_train),replace = T),]
  full<-lm(formIntT,data=subsample)
  null<-lm(varObjCont~1,data=subsample)
  modeloAux<-
step(null,scope=list(lower=null,upper=full),direction="both",trace=0,k=log(nrow(subsample)))
  modelosGenerados<-
c(modelosGenerados,paste(sort(unlist(strsplit(as.character(formula(modeloAux))[3]," [+]
"))),collapse = "+"))
}
frecuencias <- freq(modelosGenerados,sort="dec")
frecuencias
```

Selección modelo ganador

De las 100 repeticiones no hay ninguno que se repita, y no puedo probar con más repeticiones porque mi ordenador se satura. Por lo que, vamos a dar por hecho que modeloStepBIC_transInt (nuestro

	n	%	val%
ActividadPpal+ActividadPpal:Age_0.4_Ptge+ActividadP...	1	1	1
ActividadPpal+ActividadPpal:Age_0.4_Ptge+Age_0.4_Pt...	1	1	1
ActividadPpal+ActividadPpal:Age_over65_pct+Age_0.4_...	1	1	1
ActividadPpal+ActividadPpal:Age_under19_Ptge+Activi...	1	1	1
ActividadPpal+ActividadPpal:Age_under19_Ptge+Age_u...	1	1	1
ActividadPpal+ActividadPpal:AgricultureUnemploymen...	1	1	1
ActividadPpal+ActividadPpal:AgricultureUnemploymen...	1	1	1
ActividadPpal+ActividadPpal:AgricultureUnemploymen...	1	1	1
ActividadPpal+ActividadPpal:AgricultureUnemploymen...	1	1	1
ActividadPpal+ActividadPpal:AgricultureUnemploymen...	1	1	1
ActividadPpal+ActividadPpal:AgricultureUnemploymen...	1	1	1
ActividadPpal+ActividadPpal:ForeignersPtge+Actividad...	1	1	1
ActividadPpal+ActividadPpal:ForeignersPtge+Actividad...	1	1	1
ActividadPpal+ActividadPpal:ForeignersPtge+Age_0.4_...	1	1	1
ActividadPpal+ActividadPpal:ForeignersPtge+Age_0.4_...	1	1	1
ActividadPpal+ActividadPpal:ForeignersPtge+Agricultu...	1	1	1
ActividadPpal+ActividadPpal:logxAge_0.4_Ptge+Activid...	1	1	1

ganador de la seleccion clásica) es el modelo ganador.

```
ModeloGanador <- lm(varObjCont ~ CCAA + raiz4DifComAutonPtge + logxPopulation +
    ActividadPpal + ForeignersPtge + Age_19_65_pct + raiz4AgricultureUnemploymentPtge +
    raiz4Explotaciones + Age_0.4_Ptge + logxServicesUnemploymentPtge +
    logxConstructionUnemploymentPtge + logxPopulation:ActividadPpal +
    CCAA:raiz4DifComAutonPtge + ActividadPpal:raiz4AgricultureUnemploymentPtge,
    data = data_train)


summary(ModeloGanador)
```

```
##
## Call:
## lm(formula = varObjCont ~ CCAA + raiz4DifComAutonPtge + logxPopulation +
##     ActividadPpal + ForeignersPtge + Age_19_65_pct + raiz4AgricultureUnemploymentPtge +
##     raiz4Explotaciones + Age_0.4_Ptge + logxServicesUnemploymentPtge +
##     logxConstructionUnemploymentPtge + logxPopulation:ActividadPpal +
##     CCAA:raiz4DifComAutonPtge + ActividadPpal:raiz4AgricultureUnemploymentPtge,
##     data = data_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.580  -5.604  -0.449   5.128  56.626
##
## Coefficients:
##                                                                          Estimate
## (Intercept)                                                             64.02232
## CCAAAragón                                                             -33.18301
## CCAAAsturias                                                           -27.85506
## CCAABaleares                                                           -38.16461
## CCAACanarias                                                           -49.15606
## CCAACantabria                                                          -48.17221
## CCAACastillaLeón                                                       -43.63250
## CCAACastillaMancha                                                     -26.56313
## CCAACataluña                                                           -75.00955
## CCAAComValenciana                                                      -44.43849
## CCAAExtremadura                                                         -4.76602
## CCAAGalicia                                                            -51.18520
## CCAAMadrid                                                              -9.26296
## CCAAMurcia                                                             -19.45833
## CCAANavarra                                                            -30.12096
## CCAAPaísVasco                                                          -81.09773
## CCAARioja                                                              -35.44956
## raiz4DifComAutonPtge                                                   -15.14962
## logxPopulation                                                           0.09280
## ActividadPpalOtro                                                       12.48497
## ActividadPpalServicios-Construccion-Industria                           -2.63564
## ForeignersPtge                                                          -0.14866
## Age_19_65_pct                                                            0.11900
## raiz4AgricultureUnemploymentPtge                                         6.65099
## raiz4Explotaciones                                                      -6.20151
## Age_0.4_Ptge                                                            -0.47668
## logxServicesUnemploymentPtge                                             0.20890
## logxConstructionUnemploymentPtge                                         0.13259
## logxPopulation:ActividadPpalOtro                                         2.41445
## logxPopulation:ActividadPpalServicios-Construccion-Industria             0.27357
## CCAAAragón:raiz4DifComAutonPtge                                         26.35480
## CCAAAsturias:raiz4DifComAutonPtge                                       26.49828
## CCAABaleares:raiz4DifComAutonPtge                                       31.90736
## CCAACanarias:raiz4DifComAutonPtge                                       40.17987
## CCAACantabria:raiz4DifComAutonPtge                                      33.37062
## CCAACastillaLeón:raiz4DifComAutonPtge                                   27.30337
## CCAACastillaMancha:raiz4DifComAutonPtge                                 19.09320
## CCAACataluña:raiz4DifComAutonPtge                                       34.18141
## CCAAComValenciana:raiz4DifComAutonPtge                                  16.69837
## CCAAExtremadura:raiz4DifComAutonPtge                                     3.61228
## CCAAGalicia:raiz4DifComAutonPtge                                        22.49038
## CCAAMadrid:raiz4DifComAutonPtge                                          0.48339
## CCAAMurcia:raiz4DifComAutonPtge                                          5.55293
## CCAANavarra:raiz4DifComAutonPtge                                        20.37329
## CCAAPaísVasco:raiz4DifComAutonPtge                                      58.99959
## CCAARioja:raiz4DifComAutonPtge                                          23.28565
## ActividadPpalOtro:raiz4AgricultureUnemploymentPtge                      -5.42655
## ActividadPpalServicios-Construccion-Industria:raiz4AgricultureUnemploymentPtge   1.29771
##                                                                        Std. Error
## (Intercept)                                                              3.83261
## CCAAAragón                                                               3.78251
## CCAAAsturias                                                            10.79273
## CCAABaleares                                                            10.53164
## CCAACanarias                                                             6.85808
## CCAACantabria                                                            7.44189
## CCAACastillaLeón                                                         3.37850
## CCAACastillaMancha                                                       3.74190
## CCAACataluña                                                             4.03856
## CCAAComValenciana                                                        4.36393
## CCAAExtremadura                                                          5.42571
## CCAAGalicia                                                              5.49308
## CCAAMadrid                                                              10.19950
## CCAAMurcia                                                              15.07096
## CCAANavarra                                                              5.10788
## CCAAPaísVasco                                                            4.97609
## CCAARioja                                                                5.54212
## raiz4DifComAutonPtge                                                     3.70042
## logxPopulation                                                           0.23585
## ActividadPpalOtro                                                        1.53957
## ActividadPpalServicios-Construccion-Industria                            1.71762
## ForeignersPtge                                                           0.01990
## Age_19_65_pct                                                            0.02123
## raiz4AgricultureUnemploymentPtge                                         1.04910
## raiz4Explotaciones                                                       0.81144
## Age_0.4_Ptge                                                             0.08867
## logxServicesUnemploymentPtge                                             0.05624
## logxConstructionUnemploymentPtge                                         0.03810
## logxPopulation:ActividadPpalOtro                                         0.30622
## logxPopulation:ActividadPpalServicios-Construccion-Industria             0.35339
## CCAAAragón:raiz4DifComAutonPtge                                          4.14991
## CCAAAsturias:raiz4DifComAutonPtge                                       11.21979
## CCAABaleares:raiz4DifComAutonPtge                                        9.53405
## CCAACanarias:raiz4DifComAutonPtge                                        7.77047
## CCAACantabria:raiz4DifComAutonPtge                                       7.21520
## CCAACastillaLeón:raiz4DifComAutonPtge                                    3.84184
## CCAACastillaMancha:raiz4DifComAutonPtge                                  4.04055
## CCAACataluña:raiz4DifComAutonPtge                                        4.39317
## CCAAComValenciana:raiz4DifComAutonPtge                                   4.68885
## CCAAExtremadura:raiz4DifComAutonPtge                                     5.68912
## CCAAGalicia:raiz4DifComAutonPtge                                         6.62505
```

```
## CCAAMadrid:raiz4DifComAutonPtge                                                 9.08062
## CCAAMurcia:raiz4DifComAutonPtge                                                15.41355
## CCAANavarra:raiz4DifComAutonPtge                                                5.20727
## CCAAPaísVasco:raiz4DifComAutonPtge                                              5.11695
## CCAARioja:raiz4DifComAutonPtge                                                  5.44423
## ActividadPpalOtro:raiz4AgricultureUnemploymentPtge                              1.08362
## ActividadPpalServicios-Construccion-Industria:raiz4AgricultureUnemploymentPtge  2.27565
##                                                                                t value
## (Intercept)                                                                     16.705
## CCAAAragón                                                                      -8.773
## CCAAAsturias                                                                    -2.581
## CCAABaleares                                                                    -3.624
## CCAACanarias                                                                    -7.168
## CCAACantabria                                                                   -6.473
## CCAACastillaLeón                                                               -12.915
## CCAACastillaMancha                                                              -7.099
## CCAACataluña                                                                   -18.573
## CCAAComValenciana                                                              -10.183
## CCAAExtremadura                                                                 -0.878
## CCAAGalicia                                                                     -9.318
## CCAAMadrid                                                                      -0.908
## CCAAMurcia                                                                      -1.291
## CCAANavarra                                                                     -5.897
## CCAAPaísVasco                                                                  -16.297
## CCAARioja                                                                       -6.396
## raiz4DifComAutonPtge                                                            -4.094
## logxPopulation                                                                   0.393
## ActividadPpalOtro                                                                8.109
## ActividadPpalServicios-Construccion-Industria                                   -1.534
## ForeignersPtge                                                                  -7.469
## Age_19_65_pct                                                                    5.605
## raiz4AgricultureUnemploymentPtge                                                 6.340
## raiz4Explotaciones                                                              -7.643
## Age_0.4_Ptge                                                                    -5.376
## logxServicesUnemploymentPtge                                                     3.714
## logxConstructionUnemploymentPtge                                                 3.481
## logxPopulation:ActividadPpalOtro                                                 7.885
## logxPopulation:ActividadPpalServicios-Construccion-Industria                     0.774
## CCAAAragón:raiz4DifComAutonPtge                                                  6.351
## CCAAAsturias:raiz4DifComAutonPtge                                                2.362
## CCAABaleares:raiz4DifComAutonPtge                                                3.347
## CCAACanarias:raiz4DifComAutonPtge                                                5.171
## CCAACantabria:raiz4DifComAutonPtge                                               4.625
## CCAACastillaLeón:raiz4DifComAutonPtge                                            7.107
## CCAACastillaMancha:raiz4DifComAutonPtge                                          4.725
## CCAACataluña:raiz4DifComAutonPtge                                                7.781
## CCAAComValenciana:raiz4DifComAutonPtge                                           3.561
## CCAAExtremadura:raiz4DifComAutonPtge                                             0.635
## CCAAGalicia:raiz4DifComAutonPtge                                                 3.395
## CCAAMadrid:raiz4DifComAutonPtge                                                  0.053
## CCAAMurcia:raiz4DifComAutonPtge                                                  0.360
## CCAANavarra:raiz4DifComAutonPtge                                                 3.912
## CCAAPaísVasco:raiz4DifComAutonPtge                                              11.530
## CCAARioja:raiz4DifComAutonPtge                                                   4.277
## ActividadPpalOtro:raiz4AgricultureUnemploymentPtge                              -5.008
## ActividadPpalServicios-Construccion-Industria:raiz4AgricultureUnemploymentPtge   0.570
##                                                                                Pr(>|t|)
## (Intercept)                                                                     < 2e-16
## CCAAAragón                                                                      < 2e-16
## CCAAAsturias                                                                   0.009876
## CCAABaleares                                                                   0.000293
## CCAACanarias                                                                   8.48e-13
## CCAACantabria                                                                  1.03e-10
## CCAACastillaLeón                                                                < 2e-16
## CCAACastillaMancha                                                             1.39e-12
## CCAACataluña                                                                    < 2e-16
## CCAAComValenciana                                                               < 2e-16
## CCAAExtremadura                                                                0.379751
## CCAAGalicia                                                                     < 2e-16
## CCAAMadrid                                                                     0.363818
## CCAAMurcia                                                                     0.196710
## CCAANavarra                                                                    3.89e-09
## CCAAPaísVasco                                                                   < 2e-16
## CCAARioja                                                                      1.70e-10
## raiz4DifComAutonPtge                                                           4.29e-05
## logxPopulation                                                                 0.693979
## ActividadPpalOtro                                                              6.04e-16
## ActividadPpalServicios-Construccion-Industria                                  0.124963
## ForeignersPtge                                                                 9.15e-14
## Age_19_65_pct                                                                  2.17e-08
## raiz4AgricultureUnemploymentPtge                                               2.46e-10
## raiz4Explotaciones                                                             2.44e-14
## Age_0.4_Ptge                                                                   7.88e-08
## logxServicesUnemploymentPtge                                                   0.000206
## logxConstructionUnemploymentPtge                                               0.000504
## logxPopulation:ActividadPpalOtro                                               3.68e-15
## logxPopulation:ActividadPpalServicios-Construccion-Industria                   0.438881
## CCAAAragón:raiz4DifComAutonPtge                                                2.29e-10
## CCAAAsturias:raiz4DifComAutonPtge                                              0.018219
## CCAABaleares:raiz4DifComAutonPtge                                              0.000823
## CCAACanarias:raiz4DifComAutonPtge                                              2.40e-07
## CCAACantabria:raiz4DifComAutonPtge                                             3.82e-06
## CCAACastillaLeón:raiz4DifComAutonPtge                                          1.32e-12
## CCAACastillaMancha:raiz4DifComAutonPtge                                        2.35e-06
## CCAACataluña:raiz4DifComAutonPtge                                              8.35e-15
## CCAAComValenciana:raiz4DifComAutonPtge                                         0.000372
## CCAAExtremadura:raiz4DifComAutonPtge                                           0.525487
## CCAAGalicia:raiz4DifComAutonPtge                                               0.000691
## CCAAMadrid:raiz4DifComAutonPtge                                                0.957548
## CCAAMurcia:raiz4DifComAutonPtge                                                0.718663
## CCAANavarra:raiz4DifComAutonPtge                                               9.23e-05
## CCAAPaísVasco:raiz4DifComAutonPtge                                              < 2e-16
## CCAARioja:raiz4DifComAutonPtge                                                 1.92e-05
## ActividadPpalOtro:raiz4AgricultureUnemploymentPtge                             5.65e-07
## ActividadPpalServicios-Construccion-Industria:raiz4AgricultureUnemploymentPtge 0.568522
##
## (Intercept)                                                                    ***
## CCAAAragón                                                                     ***
## CCAAAsturias                                                                   **
## CCAABaleares                                                                   ***
## CCAACanarias                                                                   ***
```

```
## CCAACantabria                                                                  ***
## CCAACastillaLeón                                                               ***
## CCAACastillaMancha                                                             ***
## CCAACataluña                                                                   ***
## CCAAComValenciana                                                              ***
## CCAAExtremadura
## CCAAGalicia                                                                    ***
## CCAAMadrid
## CCAAMurcia
## CCAANavarra                                                                    ***
## CCAAPaísVasco                                                                  ***
## CCAARioja                                                                      ***
## raiz4DifComAutonPtge                                                           ***
## logxPopulation
## ActividadPpalOtro                                                              ***
## ActividadPpalServicios-Construccion-Industria
## ForeignersPtge                                                                 ***
## Age_19_65_pct                                                                  ***
## raiz4AgricultureUnemploymentPtge                                               ***
## raiz4Explotaciones                                                             ***
## Age_0.4_Ptge                                                                   ***
## logxServicesUnemploymentPtge                                                   ***
## logxConstructionUnemploymentPtge                                               ***
## logxPopulation:ActividadPpalOtro                                               ***
## logxPopulation:ActividadPpalServicios-Construccion-Industria
## CCAAAragón:raiz4DifComAutonPtge                                                ***
## CCAAAsturias:raiz4DifComAutonPtge                                              *
## CCAABaleares:raiz4DifComAutonPtge                                              ***
## CCAACanarias:raiz4DifComAutonPtge                                              ***
## CCAACantabria:raiz4DifComAutonPtge                                             ***
## CCAACastillaLeón:raiz4DifComAutonPtge                                          ***
## CCAACastillaMancha:raiz4DifComAutonPtge                                        ***
## CCAACataluña:raiz4DifComAutonPtge                                              ***
## CCAAComValenciana:raiz4DifComAutonPtge                                         ***
## CCAAExtremadura:raiz4DifComAutonPtge
## CCAAGalicia:raiz4DifComAutonPtge                                               ***
## CCAAMadrid:raiz4DifComAutonPtge
## CCAAMurcia:raiz4DifComAutonPtge
## CCAANavarra:raiz4DifComAutonPtge                                               ***
## CCAAPaísVasco:raiz4DifComAutonPtge                                             ***
## CCAARioja:raiz4DifComAutonPtge                                                 ***
## ActividadPpalOtro:raiz4AgricultureUnemploymentPtge                             ***
## ActividadPpalServicios-Construccion-Industria:raiz4AgricultureUnemploymentPtge
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.664 on 6449 degrees of freedom
## Multiple R-squared:  0.6546, Adjusted R-squared:  0.6521
## F-statistic:   260 on 47 and 6449 DF,  p-value: < 2.2e-16
```

Interpretación de los coeficientes de dos variables incluidas en el modelo, una binaria y otra continua

```
coef(ModeloGanador)
```

```
(Intercept)
64.02232484

CCAAAragón

-33.18300920

CCAAAsturias

-27.85505824

CCAABaleares

-38.16460990

CCAACanarias

-49.15605638

CCAACantabria

-48.17221350

CCAACastillaLeón

-43.63250236

CCAACastillaMancha

-26.56312669

CCAACataluña

-75.00955284

CCAAComValenciana

-44.43848623

CCAAExtremadura

Age_0.4_Ptge

-0.47667638

logxServicesUnemploymentPtge

0.20889669

logxConstructionUnemploymentPtge
0.13259304

logxPopulation:ActividadPpalOtro

2.41445308

                    logxPopulation:ActividadPpalServicios-
Construccion-Industria

0.27356844

CCAAAragón:raiz4DifComAutonPtge

26.35479669

CCAAAsturias:raiz4DifComAutonPtge

26.49827796

CCAABaleares:raiz4DifComAutonPtge

31.90735686

CCAACanarias:raiz4DifComAutonPtge

40.17987242

CCAACantabria:raiz4DifComAutonPtge

33.37061516
```

```
-4.76602093
CCAAGalicia
-51.18519548
CCAAMadrid
-9.26295636
CCAAMurcia
-19.45832860
CCAANavarra
-30.12096230
CCAAPaísVasco
-81.09772501
CCAARioja
-35.44955701
raiz4DifComAutonPtge
-15.14962080
logxPopulation
0.09280261
ActividadPpalOtro
12.48497285
                                    ActividadPpalServicios-
Construccion-Industria
-2.63564445
ForeignersPtge
-0.14865872
Age_19_65_pct
0.11900255
raiz4AgricultureUnemploymentPtge
6.65098594
raiz4Explotaciones
-6.20150713
CCAACastillaLeón:raiz4DifComAutonPtge
27.30336938
CCAACastillaMancha:raiz4DifComAutonPtge
19.09319560
CCAACataluña:raiz4DifComAutonPtge
34.18141087
CCAAComValenciana:raiz4DifComAutonPtge
16.69836745
CCAAExtremadura:raiz4DifComAutonPtge
3.61227893
CCAAGalicia:raiz4DifComAutonPtge
22.49037965
CCAAMadrid:raiz4DifComAutonPtge
0.48339040
CCAAMurcia:raiz4DifComAutonPtge
5.55292553
CCAANavarra:raiz4DifComAutonPtge
20.37328609
CCAAPaísVasco:raiz4DifComAutonPtge
58.99958702
CCAARioja:raiz4DifComAutonPtge
23.28565224
ActividadPpalOtro:raiz4AgricultureUnemploymentPtge
                                                    -5.42655250
## ActividadPpalServicios-Construccion-Industria:raiz4AgricultureUnemploymentPtge
##                                                     1.29771115
```

• Primero voy a analizar CCAA, que es una variable categórica. Podemos observar que se han creado una variable por cada CA, quitando una comunidad, que sirve como variable referencia. Su valor lo podremos obtener a través del resto de variables de comunidades. Si nos fijamos en una comunidad en particular, como por ejemplo Aragón, vemos que la diferencia con la comunidad referencia es de -33.18300920, significa que la media de variable objetivo en Aragón es menor que la de la comunidad referencia.

• En la variable continua Age_0.4_Ptge (porcentaje de ciudadanos con menos de 5 años) podemos observar que la variable objetivo disminuye -0.476 cada vez que incrementa Age_0.4_Ptge en una unidad.

Justificación de porque es el mejor modelo y medir la calidad del mismo

Un buena manera para comprobar que nuestro modelo es estable es calculando el r-cuadrado para nuestros datos train y test, si hay poca diferencia será estable.

```
Rsq(ModeloGanador,"varObjCont",data_train)

## [1] 0.654595

Rsq(ModeloGanador,"varObjCont",data_test)

## [1] 0.6490593
```

Hay poca diferencia entre sus r-cuadrados por lo que podemos decir que es estable.

10. Regresión logística

Voy a volver a cargar los datos de nuevo y cambio la semilla, para así comparar resultados con los compañeros.

```
RNGkind(sample.kind = "Rejection") #fijamos la semilla
set.seed(123456)

trainIndex <- createDataPartition(todo$varObjBin, p=0.8, list=FALSE)
data_train <- todo[trainIndex,]
data_test <- todo[-trainIndex,]
```

Seleccion clasica

```
null <- glm(varObjBin ~ 1, data = data_train, family=binomial) #Modelo minimo
full <- glm(varObjBin ~ ., data = data_train[,c(1:25,48)], family=binomial) #Modelo maximo, con las
transformaciones

modeloStepAIC_log <- step(null, scope=list(lower=null, upper=full), direction="both")
summary(modeloStepAIC_log)
pseudoR2(modeloStepAIC_log,data_test,"varObjBin") #0.3525915

modeloBackAIC_log <- step(full, scope=list(lower=null, upper=full), direction="backward")
summary(modeloBackAIC_log)
pseudoR2(modeloBackAIC_log,data_test,"varObjBin") #0.3495761

modeloStepBIC_log <- step(null, scope=list(lower=null, upper=full), direction="both",
k=log(nrow(data_train)))
summary(modeloStepBIC_log)
pseudoR2(modeloStepBIC_log,data_test,"varObjBin") #0.3506825

modeloBackBIC_log <- step(full, scope=list(lower=null, upper=full), direction="backward",
k=log(nrow(data_train)))
summary(modeloBackBIC_log)
pseudoR2(modeloBackBIC_log,data_test,"varObjBin") #0.3443574

modeloStepAIC_log$rank #35
modeloBackAIC_log$rank #34
modeloStepBIC_log$rank #27
modeloBackBIC_log$rank #26
```

Los modelos modeloStepAIC_log y modeloStepBIC_log tienen un mayor pseudo R cuadrado que los otros dos, entre estos dos el modelo modeloStepBIC_log tien 8 variables menos por lo que legimos etse.

Generación de interacciones

```
formInt <- formulaInteracciones(todo[,c(1:25,48)],26)
fullInt <- glm(formInt, data=data_train, family=binomial)

modeloStepAIC_int_log <- step(null, scope=list(lower=null, upper=fullInt),
direction="both")
summary(modeloStepAIC_int_log)
```

```
pseudoR2(modeloStepAIC_int_log,data_test,"varObjBin") #0.3885997

modeloStepBIC_int_log<-step(null, scope=list(lower=null, upper=fullInt),
direction="both",k=log(nrow(data_train)))
summary(modeloStepBIC_int_log)
pseudoR2(modeloStepBIC_int_log,data_test,"varObjBin") #0.3585504

modeloStepAIC_int_log$rank #125
modeloStepBIC_int_log$rank #42
```

Por el principio de parsimonia es preferible modeloStepBIC_int_log, cuya cantidad de variables es mucho menos elevada.

Transformaciones y las variables originales

```
fullT <- glm(varObjBin~., data=data_train, family=binomial)

modeloStepAIC_trans_log <- step(null, scope=list(lower=null, upper=fullT), direction="both")
summary(modeloStepAIC_trans_log)
pseudoR2(modeloStepAIC_trans_log,data_test,"varObjBin") #0.3525915

modeloStepBIC_trans_log <- step(null, scope=list(lower=null, upper=fullT),
direction="both",k=log(nrow(data_train)))
summary(modeloStepBIC_trans_log)
pseudoR2(modeloStepBIC_trans_log,data_test,"varObjBin") #0.3506825

modeloStepAIC_trans_log$rank #35
modeloStepBIC_trans_log$rank #27
```

El mejor es modeloStepBIC_trans_log, la diferencia de su pseudo R-cuadrado es poca y en cambio tiene 8 variables menos.

Transformaciones e interacciones

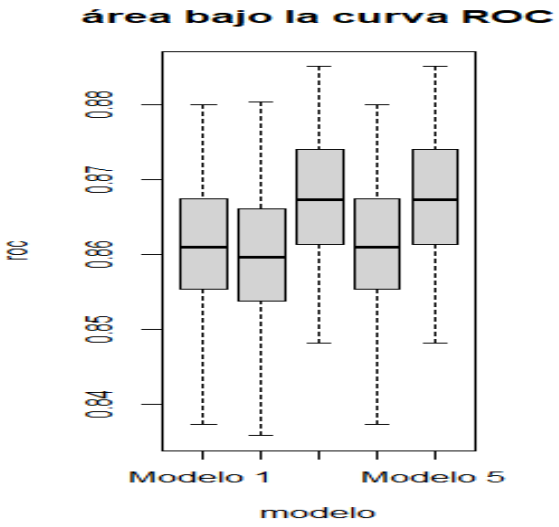
```
formIntT <- formulaInteracciones(todo,48)
fullIntT <- glm(formIntT, data=data_train, family = binomial)

modeloStepAIC_transInt_log <- step(null, scope=list(lower=null, upper=fullIntT),
direction="both")
summary(modeloStepAIC_transInt_log)
pseudoR2(modeloStepAIC_transInt_log,data_test,"varObjBin") #0.3885997

modeloStepBIC_transInt_log <- step(null, scope=list(lower=null, upper=fullIntT),
direction="both",k=log(nrow(data_train)))
summary(modeloStepBIC_transInt_log)
pseudoR2(modeloStepBIC_transInt_log,data_test,"varObjBin") #0.3585504

modeloStepAIC_transInt_log$rank #125
modeloStepBIC_transInt_log$rank #42
```

El número de variables de modeloStepAIC_transInt_log es muy elevado y el pseudo r cuadrado no es mucho mayor, por lo que escogo el modelo modeloStepBIC_transInt_log

Validacion cruzada repetida

```
total <- c()
modelos <-
sapply(list(modeloStepBIC_log,modeloBackBIC_log,modeloStepBIC_int_log,modeloStepBIC_trans
_log,modeloStepBIC_transInt_log),formula)
for (i in 1:length(modelos)){
  set.seed(1712)
```

```
  vcr<-train(as.formula(modelos[[i]]), data = todo,
             method = "glm", family="binomial",metric = "ROC",
             trControl = trainControl(method="repeatedcv", number=5, repeats=20,
                                      summaryFunction=twoClassSummary,
                                      classProbs=TRUE,returnResamp="all")
  )
  total<-rbind(total,data.frame(roc=vcr$resample[,1],modelo=rep(paste("Modelo",i),
                                                                nrow(vcr$resample))))
}
boxplot(roc~modelo,data=total,main="área bajo la curva ROC")
aggregate(roc~modelo, data = total, mean)
aggregate(roc~modelo, data = total, sd)
```



	modelo	roc
1	Modelo 1	0.8615765
2	Modelo 2	0.8600937
3	Modelo 3	0.8678952
4	Modelo 4	0.8615765
5	Modelo 5	0.8678952

	modelo	roc
1	Modelo 1	0.008531858
2	Modelo 2	0.008607594
3	Modelo 3	0.008269334
4	Modelo 4	0.008531858
5	Modelo 5	0.008269334

El modelo 3 y 5 son los dos mejores modelos, de hecho, revisando sus variables me he dado cuenta de que es el mismo modelo. Tiene las siguientes variables.

glm(formula = varObjBin ~ CCAA + ForeignersPtge + ActividadPpal + DifComAutonPtge + Explotaciones + CCAA:DifComAutonPtge + ForeignersPtge:ActividadPpal, family = binomial, data = data_train)

Selección aleatoria

```
rep <- XXX
prop <- 0.7 # se realiza con el 70% de los datos de entrenamiento por velocidad. El resultado es el
mismo.
modelosGenerados <- c()
for (i in 1:rep){
  set.seed(12345+i)
  subsample<-data_train[sample(1:nrow(data_train),prop*nrow(data_train),replace = T),]
  full<-glm(formIntT,data=subsample,family = binomial)
  null<-glm(varObjBin~1,data=subsample,family = binomial)
  modeloAux<-
step(null,scope=list(lower=null,upper=full),direction="both",trace=0,k=log(nrow(subsample)))
  modelosGenerados<-
c(modelosGenerados,paste(sort(unlist(strsplit(as.character(formula(modeloAux))[3]," [+]
"))),collapse = "+"))
```

```
}
frecuencia <- freq(modelosGenerados,sort="dec")
```

No puedo sacar modelos que se repitan, porque me tarda demasiado, y estamos de obra en casa y me cortan la luz sin avisar.

Habría que comparar los distintos modelos, que se repitan con el seleccionado anteriormente. En nuestro caso nos quedamos con el de la selección clásica.

```
ModeloGanador <- glm(varObjBin ~ CCAA + ForeignersPtge + ActividadPpal + DifComAutonPtge +
                     Explotaciones + CCAA:DifComAutonPtge + ForeignersPtge:ActividadPpal,
                     data=data_train, family = binomial)
summary(ModeloGanador)
```

```
Call: glm(formula = varObjBin ~ CCAA + ForeignersPtge + ActividadPpal + DifComAutonPtge + Explotaciones +
CCAA:DifComAutonPtge + ForeignersPtge:ActividadPpal, family = binomial, data = data_train)

Deviance Residuals: Min 1Q Median 3Q Max
-2.1705 -0.5291 -0.3327 -0.0001 3.1412

Coefficients: (2 not defined because of singularities) Estimate Std. Error (Intercept) 2.595e+00 2.603e-01

CCAAAragón -2.716e+00 2.894e-01

CCAAAsturias -1.647e+00 6.744e-01

CCAABaleares -3.094e+00 7.106e-01

CCAACanarias -3.614e+00 4.907e-01

CCAACantabria -5.500e+00 7.315e-01

CCAACastillaLeón -4.532e+00 2.764e-01

CCAACastillaMancha -2.540e+00 2.718e-01

CCAACataluña -1.992e+01 4.584e+02

CCAACeuta -1.871e+01 6.523e+03

CCAAComValenciana -6.280e+00 7.620e-01

CCAAExtremadura -4.047e-01 3.591e-01

CCAAGalicia -5.764e+00 9.421e-01

CCAAMadrid -4.646e+00 1.075e+00

CCAAMelilla -1.844e+01 6.523e+03

CCAAMurcia -2.088e+00 2.127e+00

CCAANavarra -1.621e+00 3.668e-01

CCAAPaísVasco -7.813e+00 8.962e-01

CCAARioja -4.419e+00 5.940e-01

ForeignersPtge -7.176e-02 1.057e-02

ActividadPpalConstIndServ -1.665e+00 2.903e-01

ActividadPpalOtro -9.255e-01 1.322e-01

DifComAutonPtge -1.020e-01 3.754e-02

Explotaciones -9.796e-04 1.971e-04

CCAAAragón:DifComAutonPtge 1.352e-01 3.894e-02

CCAAAsturias:DifComAutonPtge 1.361e-01 7.947e-02

CCAABaleares:DifComAutonPtge 2.135e-01 5.255e-02

CCAACanarias:DifComAutonPtge 3.275e-01 6.657e-02

CCAACantabria:DifComAutonPtge 1.765e-01 4.805e-02
```

CCAACastillaLeón:DifComAutonPtge 1.506e-01 3.858e-02

CCAACastillaMancha:DifComAutonPtge 1.007e-01 3.789e-02

CCAACataluña:DifComAutonPtge 9.130e-02 3.721e+01

CCAACeuta:DifComAutonPtge NA NA

CCAAComValenciana:DifComAutonPtge 1.617e-01 5.929e-02

CCAAExtremadura:DifComAutonPtge 1.733e-02 4.437e-02

CCAAGalicia:DifComAutonPtge 5.208e-02 2.141e-01

CCAAMadrid:DifComAutonPtge 1.603e-01 6.732e-02

CCAAMelilla:DifComAutonPtge NA NA

CCAAMurcia:DifComAutonPtge -1.500e-01 3.113e-01

CCAANavarra:DifComAutonPtge 1.058e-01 4.179e-02

CCAAPaísVasco:DifComAutonPtge 3.668e-01 5.729e-02

CCAARioja:DifComAutonPtge 1.432e-01 4.513e-02

ForeignersPtge:ActividadPpalConstIndServ 6.011e-02 2.254e-02 ForeignersPtge:ActividadPpalOtro 5.196e-02 1.344e-02 z value Pr(>|z|)
(Intercept) 9.970 < 2e-16 ***

CCAAAragón -9.386 < 2e-16 ***

CCAAAsturias -2.442 0.014621 *

CCAABaleares -4.354 1.33e-05 ***

CCAACanarias -7.365 1.77e-13 ***

CCAACantabria -7.519 5.53e-14 ***

CCAACastillaLeón -16.400 < 2e-16 ***

CCAACastillaMancha -9.347 < 2e-16 ***

CCAACataluña -0.043 0.965340

CCAACeuta -0.003 0.997711
CCAAComValenciana -8.240 < 2e-16 ***

CCAAExtremadura -1.127 0.259738

CCAAGalicia -6.118 9.49e-10 ***

CCAAMadrid -4.323 1.54e-05 ***

CCAAMelilla -0.003 0.997745

CCAAMurcia -0.982 0.326205

CCAANavarra -4.420 9.89e-06 ***

CCAAPaísVasco -8.718 < 2e-16 ***

CCAARioja -7.439 1.02e-13 ***

ForeignersPtge -6.788 1.14e-11 ***

ActividadPpalConstIndServ -5.734 9.83e-09 ***

ActividadPpalOtro -7.002 2.52e-12 ***

DifComAutonPtge -2.717 0.006579 **

Explotaciones -4.969 6.73e-07 ***

CCAAAragón:DifComAutonPtge 3.472 0.000517 ***

CCAAAsturias:DifComAutonPtge 1.713 0.086771 .

CCAABaleares:DifComAutonPtge 4.064 4.83e-05 ***

CCAACanarias:DifComAutonPtge 4.920 8.64e-07 ***

```
CCAACantabria:DifComAutonPtge 3.673 0.000240 ***

CCAACastillaLeón:DifComAutonPtge 3.904 9.45e-05 ***

CCAACastillaMancha:DifComAutonPtge 2.657 0.007887 **

CCAACataluña:DifComAutonPtge 0.002 0.998042

CCAACeuta:DifComAutonPtge NA NA

CCAAComValenciana:DifComAutonPtge 2.728 0.006375 **

CCAAExtremadura:DifComAutonPtge 0.391 0.696154

CCAAGalicia:DifComAutonPtge 0.243 0.807816

CCAAMadrid:DifComAutonPtge 2.381 0.017275 *

CCAAMelilla:DifComAutonPtge NA NA

CCAAMurcia:DifComAutonPtge -0.482 0.630031

CCAANavarra:DifComAutonPtge 2.533 0.011320 *

CCAAPaísVasco:DifComAutonPtge 6.403 1.52e-10 ***

CCAARioja:DifComAutonPtge 3.173 0.001509 **

ForeignersPtge:ActividadPpalConstIndServ 2.667 0.007651 **

ForeignersPtge:ActividadPpalOtro 3.865 0.000111 ***
```

```
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 6893.6  on 6496  degrees of freedom

Residual deviance: 4539.6 on 6455 degrees of freedom AIC: 4623.6

Number of Fisher Scoring iterations: 17
```

Punto de corte

Generamos una rejilla de puntos de corte y calculamos los puntos que maximicen el índice de Youden.

```
rejilla$posiblesCortes[which.max(rejilla$Youden)]
```

[1] 0.16

```
rejilla$posiblesCortes[which.max(rejilla$Accuracy)]
```

[1] 0.46

Ahora vamos a ver cual es mejor

```
sensEspCorte(ModeloGanador,data_test,"varObjBin",0.16,"1")
sensEspCorte(ModeloGanador,data_test,"varObjBin",0.46,"1")
```

```
  Accuracy Sensitivity Specificity Pos Pred Value Neg Pred Value
 0.7621688   0.9002770   0.7226624      0.4814815      0.9620253

  Accuracy Sensitivity Specificity Pos Pred Value Neg Pred Value
 0.8471965   0.5512465   0.9318542      0.6982456      0.8789238
```

Elegimos el punto dado por el Accuracy (0.46).

Interpretación de los coeficientes de dos variables incluidas en el modelo, una binaria y otra continua

```
coef(ModeloGanador)
```

```
                                                  (Intercept) 
                                                2.739937432 
                                                  CCAAAragón 
                                               -2.910235036 
                                                CCAAAsturias 
                                               -1.891998049 
                                                CCAABaleares 
                                               -4.106553129 
                                                CCAACanarias 
                                               -3.501315538 
                                               CCAACantabria 
                                               -5.002278989 
                                            CCAACastillaLeón 
                                               -4.660223640 
                                           CCAACastillaMancha 
                                               -2.725471876 
                                               CCAACataluña 
                                               -6.219118595 
                                                  CCAACeuta 
                                              -12.201811113 
                                           CCAAComValenciana 
                                               -5.721720939 
                                             CCAAExtremadura 
                                               -0.315044629 
                                                 CCAAGalicia 
                                               -6.021698707 
                                                  CCAAMadrid 
                                               -4.836839698 
                                                  CCAAMelilla 
                                              -12.001313430 
                                                  CCAAMurcia 
                                               -2.966851678 
                                                 CCAANavarra 
                                               -1.810458864 
                                               CCAAPaísVasco 
                                               -7.629217416 
                                                    CCAARioja 
                                               -4.805583006 
                                            ForeignersPtge 
                                               -0.076342664 
                                  ActividadPpalConstIndServ 
                                               -1.483374956 
                                          ActividadPpalOtro 
                                               -0.889763917 
                                                DifComAutonPtge 
                                                  -0.133792133 
                                                Explotaciones 
                                                  -0.001051996 
                                  CCAAAragón:DifComAutonPtge 
                                                   0.159611841 
                                 CCAAAsturias:DifComAutonPtge 
                                                   0.165791529 
                          CCAABaleares:DifComAutonPtge 
                                                   0.290683113 
                                 CCAACanarias:DifComAutonPtge 
                                                   0.358809327 
                                CCAACantabria:DifComAutonPtge 
                                                   0.186297167 
                             CCAACastillaLeón:DifComAutonPtge 
                                                   0.179916045 
                            CCAACastillaMancha:DifComAutonPtge 
                                                   0.132055115 
                                 CCAACataluña:DifComAutonPtge 
                                                  -0.037561444 
                                    CCAACeuta:DifComAutonPtge 
                                                            NA 
                           CCAAComValenciana:DifComAutonPtge 
                                                   0.166951612 
                              CCAAExtremadura:DifComAutonPtge 
                                                   0.024253106 
                                  CCAAGalicia:DifComAutonPtge 
                                                   0.102892658 
                                   CCAAMadrid:DifComAutonPtge 
                                                   0.209303732 
                                  CCAAMelilla:DifComAutonPtge 
                                                            NA 
                                   CCAAMurcia:DifComAutonPtge 
                                                   0.059149372 
                                  CCAANavarra:DifComAutonPtge 
                                                   0.136205473 
                                 CCAAPaísVasco:DifComAutonPtge 
                                                   0.391750534 
                                      CCAARioja:DifComAutonPtge 
                                                   0.198130685 
```

ForeignersPtge:ActividadPpalConstIndServ 0.046681532 ForeignersPtge:ActividadPpalOtro 0.061880408

- Primero voy a analizar CCAA, que es una variable categórica. Si nos fijamos en una comunidad en Baleares, vemos que la probabilidad de se dé la variable objetivo, es -4.106553129 veces menor si votas allí.
- En la variable continua DifComAutonPtge (porcentaje de ciudadanos que reside en la distinta CCAA de la que nacieron), podemos observar que, al aumentar este campo en una unidad, produce un cambio de -0.133792133 en el logit, por lo que si el porcentaje de residentes de distinta CA aumenta en una unidad disminuye en un 13.3792133 % que se dé la variableObjetivo.

Justificación del mejor modelo y medir la calidad del mismo

```
pseudoR2(ModeloGanador,data_train,"varObjBin") #0.3289335
pseudoR2(ModeloGanador,data_test,"varObjBin") #0.3585504

roc(data_train$varObjBin, predict(ModeloGanador,data_train,type = "response")) #0.8687
roc(data_test$varObjBin, predict(ModeloGanador,data_test,type = "response")) #0.8847
```

Ambos pseudo r cuadrados son muy parecidos por lo que podemos confirmar que nuestro modelo es robusto.